ELSEVIER

HANDBOOK OF HEALTH ECONOMICS

Volume 1A

Anthony J. Culyer &
Joseph P. Newhouse

HANDBOOK OF HEALTH ECONOMICS
VOLUME 1A

This Page Intentionally Left Blank

HANDBOOK OF HEALTH ECONOMICS

VOLUME 1A

Edited by

ANTHONY J. CULYER
University of York

and

JOSEPH P. NEWHOUSE
Harvard University Medical School

2000
ELSEVIER
AMSTERDAM • LAUSANNE • NEW YORK • OXFORD • SHANNON • SINGAPORE • TOKYO

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2000

Library of Congress Cataloging-in-Publication Data

A catalog record from the Library of Congress has been applied for.

ISBN: 0-444-82290-9 (set, comprising vols. 1A + 1B)
ISBN: 0-444-50470-2 (vol. 1A)
ISBN: 0-444-50471-0 (vol. 1B)

⊗ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Printed in The Netherlands.

INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR

PUBLISHER'S NOTE

For a complete overview of the Handbooks in Economics Series, please refer to the listing on the last two pages of this volume.

This Page Intentionally Left Blank

CONTENTS OF THE HANDBOOK

VOLUME 1A

PART 1 – OVERVIEWS AND PARADIGMS

PART 2 – DEMAND AND REIMBURSEMENT FOR MEDICAL SERVICES

PART 3 – INSURANCE MARKETS, MANAGED CARE, AND CONTRACTING

VOLUME 1B

PART 4 – SPECIFIC POPULATIONS

PART 5 – THE MEDICAL CARE MARKET

PART 6 – LAW AND REGULATION

PART 7 – HEALTH HABITS

ACKNOWLEDGMENTS

Several of the chapters in this Handbook were discussed at a conference at the University of Chicago sponsored by Pfizer and the NIA Center for Aging at the University of Chicago. We are grateful for their having made this opportunity available.

Anthony J. Culyer
Joseph P. Newhouse

CONTENTS OF VOLUME 1A

PART 1 – OVERVIEWS AND PARADIGMS

PART 2 – DEMAND AND REIMBURSEMENT FOR MEDICAL SERVICES

PART 3 – INSURANCE MARKETS, MANAGED CARE, AND CONTRACTING

INTRODUCTION: THE STATE AND SCOPE OF HEALTH ECONOMICS

ANTHONY J. CULYER and JOSEPH P. NEWHOUSE

The health of health economics

Health economics is commonly regarded as an applied field of economics. "It draws its theoretical inspiration principally from four traditional areas of economics: finance and insurance, industrial organisation, labour and public finance. Some of the most useful work employs only elementary economic concepts but requires detailed knowledge of health technology and institutions. Policy-oriented research plays a major role and many important policy-relevant articles are published in journals read by physicians and other with direct involvement in health" [Fuchs (1987)]. It might also be reasonably claimed, and has been by Blaug (1998), that health economics has contributed more than merely the application of the standard economic and econometric toolkits of economics. These volumes provide ample opportunity for readers to evaluate these claims for themselves.

By almost any criterion, health economics has been a remarkably successful subdiscipline. It has substantively contributed to the mainstream discipline (the theory of human capital, outcome measurement and valuation, the methodology of cost-effectiveness analysis, econometric method, the foundations of welfare economics, the economics of insurance, principal-agent theory, asymmetric information, the theory of incomplete markets, supplier-induced demand, to name but a few). It has generated several comprehensive bibliographies [e.g., Jolly (1977), Griffiths et al. (1980), Blades et al. (1986)]. It has generated several specialist electronic literature (systematic review) databases (e.g., Database of Abstracts of Reviews of Effectiveness, NHS Economic Evaluation Database, Health Technology Assessment Database, each of which may be accessed at http://www.york.ac.uk/inst/crd/), Health Economic Evaluations Database (available on CD-rom from the Office of Health Economics, London) and comprehensive access to the world's electronically available resources may be gained via http://www.york.ac.uk/res/herc. There are a large number of specialised texts, each covering most of the field [e.g., Newhouse (1978), Cullis and West (1979), Evans (1984), Mooney (1986), McGuire et al. (1988), Phelps (1992), Donaldson and Gerard (1993), Santerre and Neun (1996), Jacobs (1997), Folland, Goodman and Stano (1997), Getzen (1997), Zweifel and Breyer (1997), Feldstein (1999)], and innumerable conference proceedings. There are several "readings" in health economics [e.g., Cooper and Culyer

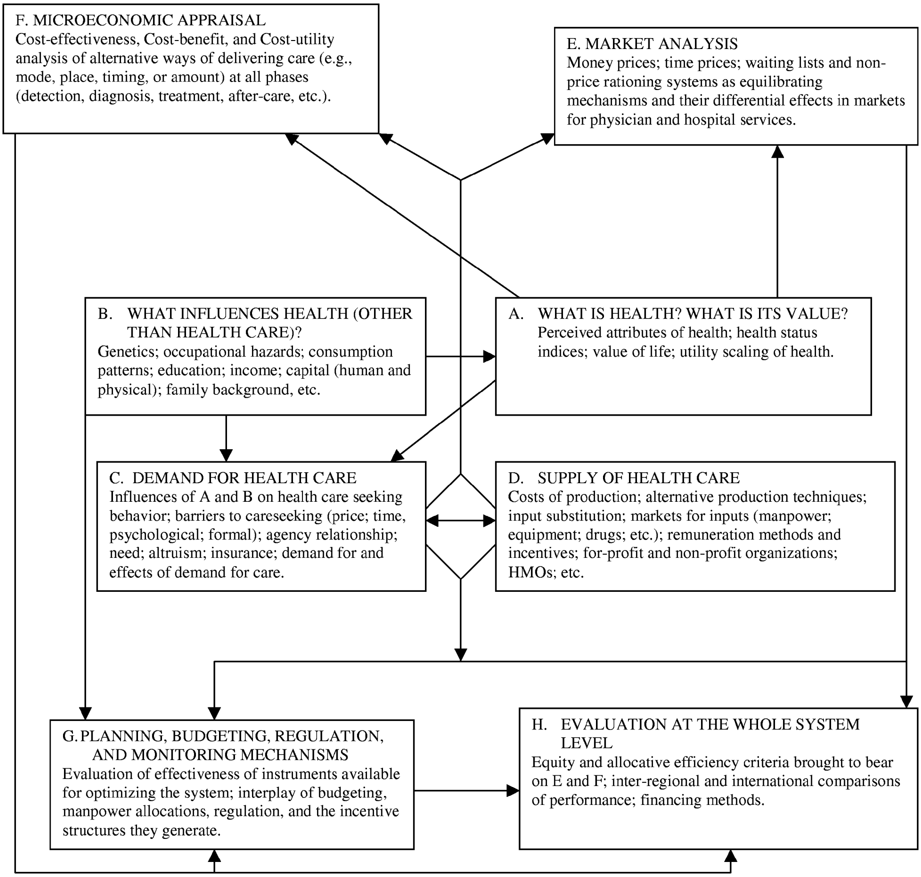
(1973), Culyer (1991)]. Health economists mounted the largest formal economic experiment in the history of economics [Newhouse and the Insurance Experiment Group (1993)]. Health economics has two international journals exclusively devoted to its subject matter (Journal of Health Economics and Health Economics), which are amongst the most frequently cited of all economics journals, and there are many others, especially multi-disciplinary journals, in which health economics features prominently. Most developed countries now have specialist professional health economics associations [for the history of one, see Croxson (1998)] and there is also an international organisation (the International Health Economics Association). There are several thriving schools of graduate study in health economics, each of which has no shortage of demand, and health economics is a common undergraduate special subject in universities. There is an ample supply of research funding, both public and private, which has led to the creation of many specialist research centres around the world. Health economists (as distinct from health economics) have even been treated as objects of study by sociologists [Ashmore et al. (1989)]! All this is powerful evidence that health economics as an academic pursuit has more than merely established itself. It is thriving.

The impact of health economics outside the economics profession has been immense. It has introduced the common currency of economists (opportunity cost, elasticity, the margin, production functions) into medical parlance (indeed, established health economists are as likely to be as heavily cited in the scientific literatures as in economics). Some major areas of research are essentially multi-disciplinary (cost-effectiveness studies and determinants of population health are two ready examples) and have led to fully integrated teams of researchers with health economists at their heart. Its policy impact has also been immense [see, e.g., Hurst (1998)]. As has been the case with other health-related professions, the language of health economics has permeated the thinking of policy makers and health service managers at all levels. Alongside academic health economics, and often in close association with it, has grown an immense cadre of health economics consultancies, servicing the demands of health care agencies, regional and national governments, and international organisations. Alongside "big issue" questions which health economists have helped public decision makers to solve, is a myriad of smaller scale research outcomes for specific clients within a great many countries' health care systems, concerning investment decisions, pricing, regulation, location, R&D, and a host of other practical issues. Policy impact is not easily measured, not least because an important class of impact has the important outcome "no change". Nonetheless, the qualitative indicators are that over the past three decades health economics has had an impact that is at least as great in its sphere of policy as that of any other branch of economics in its. The policy impact of health economics has also been heightened by the policy impact that individual policy-orientated health economists have had, where personal skills in political networking, chairing important committees, and so on, supplement the usefulness of the economics.

If one dates the real beginning of health economics as we now know it with the classic article of Arrow (1963), its start date roughly coincides with that of a related economics sub-discipline, the economics of education. From a starting point where at least one observer [Blaug (1998)] thought the prospects for education economics brighter than those for health economics, both the intellectual history and the practical relevance of the two subjects have diverged remarkably. Blaug's first commentary as an outsider on health economics appeared in an appendix to his 1970 book on the economics of education [Blaug (1970)]. His comments at that time focussed on an apparent emphasis in health economics on institutional delivery (rather than public health), health as a capital stock with rates of return, the contribution of health (or expenditures on it) to economic growth, forecasting manpower "requirements", and the special welfare characteristics of health care as a consumption good. He did not notice Arrow's (1963) article, nor Feldstein's pioneering econometrics [Feldstein (1967)] (which was certainly more than merely an application of extant methods) nor the early work on outcome measurement, cost-effectiveness analysis, or the behavioural analysis of hospitals. His main references were to Klarman (1965), Mushkin (1962), Fein (1967) and Lees (1961) (the latter being the only non-American contribution). Despite these oversights, however, the relatively primitive state of health economics in the mid-1960s was broadly as Blaug describes it. Whereas the economics of education seems to have atrophied, however, health economics has flourished and provided practical answers to practical questions as well as developing its own distinctive theoretical modes. Education economists have largely failed to resolve their own research agenda (the determination of earnings differentials, the contribution of education to economic growth, the social rate of return to training and education, the optimal size of schools and classes, the use of primitive outcome measures ...). Blaug (1998, p. S66) comments that "virtually all of the 100 articles in the 1985 International Encyclopaedia of Education devoted to the economics of education could just as well have been written in 1970 or even 1960". Blaug offers no explanation for this difference between the development patterns of these two twin subjects. For some reason, one seems to have succeeded and the other failed in capturing the creative imaginations of sufficient numbers of economists of sufficient creative ability, whether in theoretical, applied or policy-oriented (or all three) research. One factor helping to account for the success of health economics must have been the ample availability of research funding from both public and private sources (though this scarcely explains why the funding became available in the first place). Sociologists' explanations may also hold part of the truth. Ashmore et al. (1989) attribute the success of health economists (in the UK) to their assiduity in "colonising" the minds of policy makers, civil servants and health service professionals, through direct interactions with decision-makers via consultancies and the like, and through engaging in public debate.

The scope of health economics

A useful schematic structure of health economics was first drawn up by Williams (1987) and is reproduced (with some editing) as Figure 1. Although we have not used this schematic structure to organise the content of this book, it may provide some readers



F. MICROECONOMIC APPRAISAL
Cost-effectiveness, Cost-benefit, and Cost-utility analysis of alternative ways of delivering care (e.g., mode, place, timing, or amount) at all phases (detection, diagnosis, treatment, after-care, etc.).

E. MARKET ANALYSIS
Money prices; time prices; waiting lists and non-price rationing systems as equilibrating mechanisms and their differential effects in markets for physician and hospital services.

B. WHAT INFLUENCES HEALTH (OTHER THAN HEALTH CARE)?
Genetics; occupational hazards; consumption patterns; education; income; capital (human and physical); family background, etc.

A. WHAT IS HEALTH? WHAT IS ITS VALUE?
Perceived attributes of health; health status indices; value of life; utility scaling of health.

C. DEMAND FOR HEALTH CARE
Influences of A and B on health care seeking behavior; barriers to careseeking (price; time, psychological; formal); agency relationship; need; altruism; insurance; demand for and effects of demand for care.

D. SUPPLY OF HEALTH CARE
Costs of production; alternative production techniques; input substitution; markets for inputs (manpower; equipment; drugs; etc.); remuneration methods and incentives; for-profit and non-profit organizations; HMOs; etc.

G. PLANNING, BUDGETING, REGULATION, AND MONITORING MECHANISMS
Evaluation of effectiveness of instruments available for optimizing the system; interplay of budgeting, manpower allocations, regulation, and the incentive structures they generate.

H. EVALUATION AT THE WHOLE SYSTEM LEVEL
Equity and allocative efficiency criteria brought to bear on E and F; inter-regional and international comparisons of performance; financing methods.

Figure 1. A schematic of Health Economics.

with a helpful general overview of the subject and the material covered. The figure shows the principal topics in the field (with sometimes slightly arbitrary boundaries drawn between them) and the intellectual links between them. The arrows indicate the direction of logical flow between boxes, with material that is for the most part logically prior in boxes from which the flow is indicated. It is the inter-linkages that make it possible to create research programmes, and a sub-discipline, that are more than merely a collections of topics. The four central boxes, A, B, C and D, are the disciplinary "engine room" of health economics, while the four peripheral boxes E, F, G and H are the main empirical fields of application for whose sake the "engine room" exists. This is not, of course, to deny that the four central boxes contain material that is of substantive interest in its own right, and they also contain empirical work, but the purpose of the central four is mainly instrumental, needed not so much for their own sakes (or to impress fellow economists) as for the empirical leverage they enable one to bring to bear on the issues in the peripheral boxes.

Box A contains the conceptual foundation – health. It contains a multi-disciplinary literature in which one finds economists, epidemiologists, operational researchers, psychologists and sociologists all working – and sometimes even working together! The central issues in this box relate to the meaning of "health", its relationship with "welfare", and the development of valid and reliable measures of it for a variety of purposes, specific and general. It is impossible for these matters to be addressed without careful attention to the value assumptions that are to be made (and where they should come from). Chapters 2, 4 and 32 survey the content of this box.

Box B is concerned with the determinants of health, broadly genetic and environmental, as human capital, not just in the sense of a stream of discounted benefits over an expected lifetime but as a distinctive way of treating health itself – a capital stock that can be invested in, which depreciates, for the demand influences and is influenced by the demand for other human investments. It concerns the interaction between a health production function and a health demand function and has been a highly distinctive research area within health economics. Chapters 7, 29, 30, 31 and 33 develop these themes further.

Box C concerns the demand for health care. This demand is a derived demand (from the demand for health) and comes logically after boxes A and B. This is also where utility interdependencies come in (externalities), where the tensions between "need" and "demand" (and the advocates of each) are addressed, and where important questions related to the normative significance of revealed demand have been extensively discussed. Like box A, the material of box C requires the careful handling of value judgements. Chapters 2, 8, 9, 10 and 11 cover the material of box C.

Box D contains the material to be expected in supply-side economics: hospital production functions, input substitutions, behavioural relations, labour markets, the responses of institutions and health industry workers to changes in their environments and modes of payment, industrial regulation. The health care "industry" encompasses not only the more obvious health care organisations like hospitals, HMOs, and general practices, and the again obvious medical supplies sector (pharmaceuticals, equipment, etc.), but also other public and private caring agencies, often dealing with specific client groups like the elderly, the mentally infirm and the disabled, and often doing so on a community basis (for example, caring for them – and their informal carers – in their own homes). Chapters 10, 13, 21, 22, 24 and 25 covers a large segment of this vast territory.

Box E deals with the ways in which markets in all these sectors operate and is a major chunk of applied health economics, especially in countries where there is substantial dependence on market institutions for the provision of health care insurance and the delivery of health care. Even where there are no formal markers, the health care system operates as a kind of quasi-market, with, for example, contracts between non-profit public sector agencies, and pseudo-prices (including time prices) being paid. Queuing and waiting lists/times for admission to hospital are thus considered in this box. The material of this box is "positive" (i.e. concerned with "what happens", "what happened" and "what is predicted to happen" if...) as well as "normative": evaluating the performance of markets using the tools of welfare economics. The extensive material of box E is covered in Chapters 2, 3, 5, 11, 12, 20 and 23.

Box F is more specifically evaluative and normative. It is the home of applied cost-effectiveness and cost-utility analysis. The literature in this genre is now vast and a book such as this cannot do justice to the immense variety of topics, technologies and mechanisms which have been evaluated, let alone to the secondary literature of systematic reviews and meta-analyses that have developed over the past ten years. Chapters that deal wholly or largely with these topics are Chapters 2 and 4.

Box G is primarily American in its content, doubtless largely because of the great variety of health care delivery institution, insurance and reimbursement mechanisms, and the various roles played by federal and state agencies. The evolution of new forms of organisation, financing and monitoring/control has flourished apace in the US and many of these developments are reviewed in Chapters 3, 14, 15, 26, 27 and 28.

Box H is concerned with the highest level of evaluation and appraisal across systems and countries. The internationally observed differences between the mechanisms, expenditure rates, objectives and outcomes are phenomena needing explanation but they also raise difficult questions of how best to make comparisons (and for what purpose) and how best to infer "lessons" from one system for another. Chapters 1, 34 and 35 review much of this material.

Most chapters spread their wings across more than one box. Those dealing with specific client groups (e.g., Chapters 16–19) range across many. So does Chapter 6 on econometric methods.

As a "scientific research programme" [Lakatos (1978)], health economics seems to be in good shape, showing both substantial theoretical growth and immense application. Moreover, its "hard core" of neoclassical economics (especially welfare economics) is itself a part of the ongoing developmental agenda of the subject. There is, thus, something in health economics for almost every conceivable kind of economist: powerful defenders of conventional methods and aggressive challengers; pure theorists and applied economists, those who undertake academic research for its own sake and those who see it as an instrument for the improvement of societies, those who love to engage in the cut and thrust of debate on important topics and those who prefer to observe and comment on it, those whose main objective is to do research as well as those who want an exciting subject to teach and those who want to be active participants in policy formation processes. One thing is clear: the agenda is sufficiently broad and contains sufficient unanswered (and doubtless some unasked) questions to keep many health economists creatively and usefully busy for the foreseeable future.

The scope of the Handbook

We have sought, as editors, to ensure both that the practical scope of application of health economics is well illustrated in what follows and that the alternative paradigms that are in common use are represented. The latter are discussed explicitly in Chapter 2 and also in the chapters that are explicitly in applied welfare economics (with emphases that vary from author to author). As to the former, we eschewed the idea of trying to cover *every* possible field of application (for example, the economics of each of the main types of labour employed in health care) while including those that have developed a substantial literature and those that are plainly core to the sub-discipline. We hope, as a result, that there is something for everyone here though our major target readership (and one we asked all authors to bear in mind) was a typical UK masters student embarking on (or in the course of) a master's degree in health economics or US first year graduate student in a doctoral program. The book is, therefore, primarily for economists but we hope nonetheless that it may also be instructive for others who want to find out more about what is going on in this field.

We hope that all readers will find this Handbook a useful overview of the field as it currently stands. Most of the chapters in this book finish with some indication of the authors' perceptions of what the next steps in research in their subfield might be. We hope that both the main texts and these further suggestions will prove to be useful especially for readers seeking topics for masters or PhD thesis topics.

References

Arrow, K.J. (1963), "The welfare economics of medical care", American Economic Review 53:941–973.

Ashmore, M., M. Mulcahy and T. Pinch (1989), Health and Efficiency: a Sociology of Health Economics (Open University Press, Milton Keynes).

Blades, C.A., A.J. Culyer, J. Wiseman and A. Walker, eds. (1986), The International Bibliography of Health Economics (Wheatsheaf, Brighton).

Blaug, M. (1970), An Introduction to the Economics of Education (Allen Lane, London).

Blaug, M. (1998), "Where are we now in British health economics?", Health Economics 7:S63–S78.

Croxson, B. (1998), "From private club to professional network: an economic history of the Health Economists' Study Group, 1972–1997", Health Economics 7:S9–S45.

Cullis, J.G., and P.A. West (1979), The Economics of Health: An Introduction (Martin Robertson, Oxford).

Culyer, A.J., ed. (1991), The Economics of Health, The International Library of Critical Writings in Economics 12 (Edward Elgar, Aldershot).

Culyer, A.J., and M.H. Cooper (1973), Health Economics: Selected Readings (Penguin, London).

Donaldson, C., and K. Gerard (1993), The Economics of Health Care Financing (Macmillan, London).

Evans, R.G. (1984), Strained Mercy: The Economics of Canadian Health Care (Butterworths, Toronto).

Fein, R. (1967), The Doctor Shortage: An Economic Diagnosis (Brookings, Washington, DC).

Feldstein, P.J. (1999), Health Care Economics, 5th edn. (Delmar, Albany, NY).

Feldstein, M.S. (1967), Economic Analysis for Health Service Efficiency (North-Holland, Amsterdam).

Folland, S., A.C. Goodman and M. Stano (1997), The Economics of Health and Health Care, 2nd edn. (Macmillan, London).

Fuchs, V.R. (1987) "Health economics", in: J. Eatwell, M. Milgate and P. Newman, eds., The New Palgrave. A Dictionary of Economics (Macmillan, London) 614–618.

Getzen, T.E. (1997), Health Economics: Fundamentals and Flow of Funds (Wiley, New York).

Griffiths, D.A.T., R. Rigoni, P. Tacier and N.M. Prescott, eds. (1980), An Annotated Bibliography of Health Economics: Western European Sources (Martin Robertson, Oxford).

Hurst, J. (1998), "The impact of health economics on health policy in England, and the impact of health policy on health economics, 1972–1997", Health Economics 7:S47–S61.

Jolly, D. (1977), Economie de la Sante: Bibliographie Choisie et Annotee (Editions Bordas, Paris).

Jacobs, P. (1997), The Economics of Health and Medical Care, 4th edn. (Aspen, Gaithersburg).

Klarman, H.E. (1965), The Economics of Health (Columbia University Press, New York).

Lakatos, I. (1978), "The methodology of scientific research programmes", in: J. Warrall and G. Currie, eds., Philosophical Papers (Cambridge University Press, Cambridge).

Lees, D.S. (1961), Health Through Choice (Institute of Economic Affairs, London).

McGuire, A., J. Henderson and G. Mooney (1988), The Economics of Health Care (Routledge & Kegan Paul, London).

Mooney, G.H. (1986), Economics, Medicine and Health Care (Wheatsheaf, Brighton).

Mushkin, S.J. (1962) "Health as investment", Journal of Political Economy 70:129–157.

Newhouse, J.P. (1978), The Economics of Medical Care: A Policy Perspective (Addison-Wesley, Reading, MA).

Newhouse, J.P., and the Insurance Experiment Group (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge, MA).

Phelps, C.E. (1992), Health Economics (Harper-Collins, New York).

Santerre, R.E., and S.P. Neun (1996), Health Economics (Irwin, Chicago).

Williams, A. (1987), "Health economics: the cheerful face of the dismal science?", in: A. Williams, ed., Health and Economics (Macmillan, Houndmills) 1–11.

Zweifel, P., and F. Breyer (1997), Health Economics (Oxford University Press, Oxford).

PART 1

OVERVIEWS AND PARADIGMS

This Page Intentionally Left Blank

Chapter 1

INTERNATIONAL COMPARISONS OF HEALTH EXPENDITURE: THEORY, DATA AND ECONOMETRIC ANALYSIS*

ULF-G. GERDTHAM and BENGT JÖNSSON

Centre for Health Economics, Stockholm School of Economics, Stockholm, Sweden

Contents

*We are grateful to Martin Chalkley, Tony Culyer, Eddy van Doorslaer, Joe Newhouse, Albert Okunade and Tom Philipson for helpful comments.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

Comparisons of aggregate health expenditure across different countries have become popular over the last three decades as they permit a systematic investigation of the impact of different institutional regimes and other explanatory variables. Over the years, several regression analyses based on cross-section and panel data have been used to explain the international differences in health expenditure. A common result of these studies is that aggregate income appears to be the most important factor explaining health expenditure variation between countries and that the size of the estimated income elasticity is high and even higher than unity which in that case indicates that health care is a "luxury" good. Additional results indicates, for example, that the use of primary care "gatekeepers" lowers health expenditure and also that the way of remunerating physicians in the ambulatory care sector appears to influence health expenditure; capitation systems tend to lead to lower expenditure than fee-for-service systems. Finally, we also list some issues for the future. We demand more efforts on theory of the macroeconomic analysis of health expenditure, which is underdeveloped at least relative to the macroeconometrics of health expenditure. We also demand more replications based on updated data and methods that seeks to unify the many differing results of previous studies.

Keywords

international health expenditure, health system, government policy, econometric analysis

JEL classification: C20, H51, I10, I18

1. Background and overview

The growth of health expenditure and of its share in Gross Domestic Product (GDP) is a phenomenon which is constantly the subject of comments and discussions among politicians, administrators and academics in many countries. One approach to this issue has been international comparisons of health expenditure. There are substantial differences in health expenditure across countries, irrespective of how they are measured. This is even true among the relatively homogeneous industrialized market economies, e.g., the Organization for Economic Cooperation and Development (OECD) countries. An illustration of these differences is given by health expenditures per capita measured in purchasing power parities (PPPs); in 1997 these ranged from less than $1,000 (Mexico $391, Korea $587, Greece $974) to more than $2,500 (Switzerland $2,547 and the United States $4,090), with an unweighted arithmetical OECD average of $1,725 (see Table 1). In low-income countries outside the OECD area, the amounts are much smaller, both absolutely – $10 per capita in many African countries, and less than $100 per capita in most of Asia and Latin America – and as a share of income.

The global interest in health expenditure can be explained by the fact that all countries put similar emphasis on cost-containment and the cost-effectiveness of health expenditure; in addition, the great bulk of health expenditure is publicly financed, i.e. financed by taxes or compulsory social insurance contributions. This may raise health expenditure as a result of additional demand resulting from a decrease in the net price of care [see Leu (1986)], though there is disagreement on this point. Buchanan (1965) and Bird (1970) suggested the exact opposite: that the public financing of health expenditure serves as a restraining factor, while Newhouse (1977) suggested that per capita income is the only relevant variable explaining health expenditure [see Culyer (1988, 1989) for reviews]. The high fraction of public finance in health expenditure creates a problem because virtually all OECD countries have deficits in the public sector which have been increasing over time [OECD (1996)]. This increases public debt and interest payments on the debt. These macroeconomic pressures on public budgets may spill over to contain health budgets. One approach to reducing the fraction of public financing is to substitute out-of-pocket payments or private insurance. There are major problems with this substitution. First, there is a limit to how much out-of-pocket payments can be increased if the goal of equity is to be fulfilled. Most expenditure in any one year is concentrated on small segments of the population. For example in the USA, the ten percent of the population who spend most on health care accounted for 72 percent of all expenditure [Berk (1992)]. Similar results have been found in other countries [OECD (1987)]. Second, private insurance as a means of financing poses a problem, because those with the highest potential expenditure also have the lowest incomes. Since most expenditure occurs late in life, after the end of the economically active period, a significant increase in private financing must imply a very long-term perspective. Therefore, relying on competition between insurers, or allowing opting out of public systems, is not likely to work without a sound mechanism for allocating public funding to compensate for these differences.

Table 1
Public and per capita health spending, OECD countries, 1997

Countries	Public financing as percentage of total	Total health spending per capita [a]	GDP per capita [a]
Australia	68.7	1805	21671
Austria	72.0	1793	22789
Belgium	87.6	1747	22902
Canada	68.7	2095	22606
Czech Republic	92.0		
Denmark	65.0	1848	23874
Finland	77.0	1447	19821
France	78.4	2103	21290
Germany	77.4	2339	22385
Greece	74.8	974	13805
Hungary	69.1		
Iceland	83.5	2005	24937
Ireland	75.0	1324	18875
Italy	69.9	1589	20914
Japan	77.4	1741	23765
Korea	56.7	587	14578
Luxembourg	91.8	2340	33089
Mexico	60.0	391	8312
Netherlands	72.0	1825	21450
New Zealand	77.4	1352	17903
Norway	82.2	1814	24423
Poland	93.0		
Portugal	60.0	1125	13672
Spain	78.7	1168	15800
Sweden	83.3	1728	20150
Switzerland	69.9	2547	25088
Turkey			6531
United Kingdom	84.5	1347	20139
United States	46.7	4090	29195
Average	74.7	1725	20383.2

[a] Note: Data are for 1998. Expressed in purchasing power parity dollars.
Source: OECD Health Database.

Total and public financing of health care in different regions are shown in Table 2. The public fraction of health expenditure is at least 50 percent in every region except Asia. The fraction is highest in rich countries, which also have the highest total expenditure. Private financing dominates in low-income countries, where direct out-of-pocket payments are more important than private insurance as a source of revenue. A similar, though less clear, picture emerges for OECD countries (Table 1). The countries with the lowest incomes also have the lowest fraction of public finance. The USA is an exception, having the lowest fraction of public finance of all OECD countries, about 50

Table 2
Global health expenditure by region, 1990

	Share of world population (%)	Total health expenditure (billions of US dollars)	Health expenditure as percentage of world total	Public health expenditure as percentage of regional total	Share of GNP spent on health (%)	Per capita health expenditure (US dollars)
OECD countries	15	1,483	87	60	9.2	1,860
Transition economies of Europe	7	49	3	71	3.6	142
Developing countries	78	170	10	50	4.7	41
Latin America	8	47	3	60	4.0	105
Middle East	10	39	2	58	4.1	77
Other Asia and islands	13	42	2	39	4.5	61
India	16	18	1	22	6.0	21
China	22	13	1	59	3.5	11
Sub-Saharan	10	12	1	55	4.5	24
World	100	1,702	100	60	8.0	329

Source: World Bank (1993).

percent. By contrast, the government pays for nearly all health care resources in Iceland, Norway, Sweden, and the UK. However, taking into account the high health expenditure in the USA, the fraction of public finance for health care relative to the GDP is similar to that in other industrialized countries.

1.1. Why international comparisons?

Since health expenditure typically reaches $1,500–2,500 per capita in rich countries, intriguing and challenging issues of how to organize and pay for care are highly relevant. One attraction of using international comparisons in this context is that they can be used to ask positive question such as:

- Does the overall organization of the health care system have any effect on health expenditure? It is common to distinguish three types of relations between funders and providers of health care according to whether countries have reimbursement, contract, or integrated systems [OECD (1995a, 1995b)]. Under the reimbursement system, providers receive retroactive payments for services supplied. These payments may be billed directly to insurers or to patients, who may be partly or entirely reimbursed by insurers. The reimbursement system, often coupled with fee-for-service payment arrangements, can be found in systems with multiple private and public insurers and multiple (usually private) suppliers, as in the USA. In low and middle-income countries it is rare for the reimbursement model to be combined with public financing. Chile is an exception, with part of the government financing used to reimburse private providers retrospectively. The contract system involves an agreement between third-party payers (insurers) and health care providers, which aims at greater control over total funding and its distribution. This approach tends to be found in social insurance systems with predominantly private (non-profit) providers. Prospective budgets are combined with per diem, case mix (diagnostic related group DRG) and fee-for-service payments. A variant of this system is used in Brazil, where budgets are set by the state or municipality and providers are paid under a DRG tariff [Lewis (1994)]. Preferred provider organizations in the USA also use the contractual approach. In integrated systems the same agency controls both the funding and the provision of health services. Medical personnel are generally paid salaries, and budgets are the main instrument for allocating resources. Integrated public systems are found in the Nordic countries and the UK, and they are the common organizational form for ministries of health in developing countries. In many such countries the integrated approach is also used for social security systems, which have their own hospitals and clinics, although there are often also contractual relations with private providers. Health maintenance organizations (HMOs) in the USA are examples of integrated private systems. There may be a trend toward two types of relation between funders and providers [Jönsson (1996), van de Ven et al. (1994)]. The first type involves a (near) public monopoly in health care funding, through taxes or compulsory social insurance contributions, and competitive contracts with private and public providers. Thus financing and provision may be separated, in what is sometimes referred to as a purchaser-provider split. The second type is an integrated model with competition between different integrated regimes (HMOs). In accordance with what might be expected, Hurst [OECD (1992)] found for seven OECD countries that the success in controlling costs was weakest for the reimbursement approach and greatest for the integrated system, with the contract approach falling in between.[1]

- Have countries with prospective budget ceilings or fee-for-service or payment per bed-days in hospital care either lower or higher expenditure?
- Does the use of a general practitioner *gatekeeper* result in lower health expenditure? OECD (1995b) suggests that gatekeeper arrangements can provide for better continuity in health care while also acting as a barrier to moral hazard. The use of a gatekeeper reduces the risk of multiple visits for the same sickness episode, particularly where there is an over-supply of physicians, strong competition among the physicians for market shares and remuneration on a fee-for-service basis.
- Do the ways of remunerating doctors in the ambulatory sector make a difference in health expenditure, i.e. is the expenditure higher in countries which remunerate their doctors by fee-for-service and lower in countries which remunerate their doctors by means of capitation or a fixed wage per period?
- Do increases in the supply of doctors result in increases in health expenditure? This links up with the controversial supplier-inducement hypothesis as an explanation for the increase in health expenditure [see Evans (1974), Rice (1983), Cromwell and Mitchell (1986), McGuire et al. (1988), Newhouse (1992)]. Supplier-induced demand can arise for several reasons, although the form it could take and its extent depend on institutional arrangements. Under a fee-for-service system, doctors may adjust their work load in response to changes in the environment so that their target income can be maintained [Evans (1974)]. When the stock of doctors increases and the work load decreases doctors may induce the patients to use more services at higher prices, i.e. conditional on the target income hypothesis we have supplier-induced demand. Empirical relationships have been reported between the stock of physicians, the remuneration system and the number of surgical operations, and between the number of hospital beds, hospitalization rates and average length of stay in hospital, and between the physician stock and total outpatient expenditure. However, greater competition among doctors may encourage them to be more willing to comply with patient demands for referrals, prescriptions or other health services, particularly where the cost of these services is covered by insurance. Furthermore, the prediction of greater spending with additional physicians with no induced demand may also be consistent with classical microeconomics, because a positive association between the number

[1] If one compares the unweighted average health expenditure per capita between reimbursement, contract and integrated systems, respectively [see OECD (1995b) for a classification] and applies this to the expenditure figures for 1997 [OECD (1998)], it appears, as expected, that the expenditure is highest in reimbursement systems ($2,336; six countries) and lowest in integrated systems ($1,502; thirteen countries). The average figures for contract system is $2,086 (three countries).

of doctors' and doctors visits may reflect true demand factors; for example, a larger number of doctors may increase the availability of health care supply since there is less distance to travel and less time to wait, and unit price does not fall because of administratively set prices [see Carlsen and Grytten (1998)].

- Do increases in insurance or health system coverage result in higher health expenditure? There are two dimensions here: the population covered and the fraction of individual medical bills covered. Cost sharing is one way to control moral hazard. Moral hazard can manifest itself in two ways, one static and the other dynamic. People with health insurance tend to see the doctors more often and to use costly treatments even if the benefits are small [Pauly (1968), Zeckhauser (1970)]. Doctors also may change their behavior, particularly in fee-for-service systems. Since costs are not borne by the patients, it is easier for doctors to suggest more expensive treatments. The dynamic effect of moral hazard is the incentives it creates to introduce new medical technology, for which there would be no market in the absence of insurance [Weisbrod (1991)]. Both problems derive from the inability of the insurer to monitor service providers and the insured. The conclusion of all this is that increases in health insurance can influence health expenditure, both through the demand and supply of health care and through the dynamic effects that may be involved.
- Does the level of high-cost procedures (transplants, dialyses, etc.) have any impact on health expenditure? Advances in medical technology, though sometimes making existing procedures cheaper, generally increase the range of what is possible and thus lead to increasing demand and supply [Weisbrod (1991)]. This may explain differences in health expenditure across countries because of the great spread of new expensive medical technologies among countries.
- Do countries with a higher degree of public supply of health services have higher health expenditure than those where private sector supply plays a greater role?
- Do countries with a larger fraction of (usually more expensive) in-patient care have higher health expenditure?

It should be recognized that international comparisons of health expenditure are exclusively concerned with positive questions, as above, and should *not be* mixed up with normative questions such as whether health expenditure in different countries is *too* low/high or if one health system is *better/worse* than another health system.

During the last three decades several regression analyses based on international data have been used to investigate the differences in per capita health expenditure. A common result of these studies is that aggregate income appears to be the most important factor explaining health expenditure variation between countries and that the size of the estimate of the income elasticity is around or even higher than one [see Kleiman (1974), Newhouse (1977, 1987), Maxwell (1981), Leu (1986), OECD (1987), Culyer (1988, 1989), Pfaff (1990), Gerdtham et al. (1988, 1992a, 1992b), Parkin et al. (1987), Gbesemete and Gerdtham (1992), Gerdtham (1992), Hitiris and Posnett (1992), Sahn (1992), Viscusi (1994a), Gerdtham et al. (1998), Barros (1998), Roberts (1998a)]. Furthermore, as indicated above, per capita income may not be the sole determinant, and Leu (1986), Gerdtham et al. (1992a, 1992b, 1998), Gerdtham (1992), Hitiris and Posnett (1992), and Roberts (1998a) have further demonstrated that demographic and institutional factors also exhibit a measurable influence on health expenditure in the OECD countries.

1.2. Methodological problems

International comparisons are fraught with many problems. A first apparent problem is the weak theoretical base for the determinants of aggregate health expenditure, which provide little guidance as to the possible explanatory variables and the causal mechanisms involved. Parkin et al. (1987), Culyer (1988, 1989), McGuire et al. (1993), Roberts (1998a), and others, have stressed the lack of theory and the "atheoretical basis" of macroeconomic analysis of health expenditure. Culyer has called the search for missing determinants of health expenditure "A quest without a compass" (Culyer, pp. 29–34). Among the few exceptions which have attempted to provide a "theoretical compass" of public choice, e.g., Buchanan (1965) and Leu (1986), Culyer concludes that these compasses are faulty (p. 34): "In Buchanan's study, the events to be explained were highly stylized and the theory highly selective", and in Leu's study, "... the variables were (unavoidably) crude and theory again (avoidably) highly selective" (p. 45). In modeling aggregate health expenditure, it may also be important to note that the usual separation of demand from supply influences in market analysis is difficult in the case of the health care market, for a number of reasons. These include: the role of the physician both as the patient's agent in advising on health care needs, and as the key supplier of health services; the fact that health services are usually provided on the basis of "need" rather than "willingness to pay"; and the public provision of most health services coupled with various forms of non-price rationing (such as waiting times).

A second problem is that rigorous assessment of the quality (accuracy and reliability) of the cross-national data is difficult. Poullier (1989) describes the prevailing data compiling approach as: "An analyst's attempts to "massage" data from various countries, using as closely comparable units as can be obtained from the readily accessible information." (p. 111). There is ample scope for imperfect reliability with respect to international comparisons due to differential classification, especially on the borderline of health services such as care for the aged. For example, the care of the mentally retarded is not included in the expenditure for Denmark nor for Sweden after 1985, but it is included in the expenditure for Finland, Iceland and Norway. Another difference is that local nursing homes are not included in the Danish statistics, whereas they were included in Finland, Iceland, Norway and Sweden before 1992 [Gerdtham and Jönsson (1991a, 1994)].[2] Thus heterogeneous definitions are present even if one selects apparently similar countries such as the Nordic countries, which have similar GDP per capita and sim-

[2] Earlier, the latter difference in accounting was often put forward (by Danes) as an explanation for the comparatively low share of GDP used for health care in Denmark (see Enggard (1986), at that time Denmark's Home Secretary).

ilar social systems, and also that there are heterogeneous health expenditure definitions for a single country over time. A related issue is that explanatory variables of possible relevance have to be omitted from the estimation due to lack of data. Health system variables are often omitted since it is difficult to characterize countries' health systems in ways that are tractable to regression analysis. This is because these systems often combine many differing forms of provision and finance, e.g., no country fits perfectly into just one of the categories representing public reimbursement, contract and integrated health systems; indeed, many countries have elements of all three. Barr (1992, p. 782) expressed this by saying that health systems "... merge into each other like the colors of the rainbow". (For surveys and assessments of health systems in different countries, see Ham et al. (1990), Besley and Gouveia (1994), OECD (1994, 1995a, 1995b).

A third problem of international comparisons is the small sample size (commonly the OECD countries), which forces restrictions on model size and statistical inference. One may of course add to the number of countries, but only at the cost of omitted regressors and aggravated problems of heterogeneous definitions.

A fourth problem is that cross-sectional comparisons implicitly imposes the assumption of homogeneous relationships across countries which may appear unrealistic for many reasons, i.e. heterogeneous preferences, production functions, etc. [Roberts (1998a)].

A fifth problem is that cross-sectional comparisons are static, while the observed differences in health expenditure and income are the result of both real (permanent) differences and transistory differences when countries are in different stages of some adjustment process. The actual process of expenditure adjustment is not well understood and depends on many factors including organizational dynamics, accumulated surpluses and deficits, technological change and expectations [Getzen and Poullier (1992), see also Kendix and Getzen (1994)]. An alternative to cross-section studies is a separate time-series analyses of each country but then the researcher cannot consider the determinants of variations across countries.

Some studies combine cross-section with time-series data in panel analyses to overcome some of the above-mentioned problems [e.g., Gerdtham (1992), Hitiris and Posnett (1992), Viscusi (1994a)]. One advantage to using panel data is the larger sample size and hence more powerful significance tests. Another is the possibility of analyzing dynamic properties of the relationships. A third is that panel data allow us to relax the assumption of homogeneous relationships across countries. Moreover these data enable investigators to include country and time-specific effects, which help to control for the presence of mismeasured and/or unobserved variables that are correlated with the explanatory variables included in the model. Nonetheless, many difficulties remain, and Culyer (1988) concluded in his summary that: "We have had crude data, misspecified equations, contentious theory, and cavalier history" (p. 45). Taken together, these problems indicate that results obtained with international comparisons should be treated with considerable caution.

1.3. Organization of the chapter

This chapter reviews the literature on international comparisons of health expenditure. The chapter is organized into five sections. Sections 2 and 3 review the first-generation studies and second-generation studies, respectively. The first-generation studies use international cross-section data for a single year (or selected years) to analyze the cross country differences in health expenditure. One particular methodological issue of these studies concerns the choice between different conversion factors such as exchange rates or purchasing power parities (PPPs) and whether this choice affect the empirical results. The second-generation studies use panels of countries, each with a relatively long time series of annual data, which enable one to test a more extensive range of hypotheses, because of the larger sample size, and to control for country and time-invariant variables whose omission might otherwise result in inconsistent estimates of the regression coefficients. Methodological issues in the second-generation studies concerns relationships involving non-stationary variables, cointegrating and dynamic relationships, and heterogeneous relationships across countries. Section 4 summarizes and concludes the chapter, and lists some issues for the future.

2. First-generation studies

2.1. Cross-section bivariate regressions

The analysis of international health expenditure has to some extent been based on *standard* demand theory, typically focusing on the income elasticity of health expenditure estimated in functions linking per capita health expenditure (henceforth *HE*) to per capita GDP (henceforth *GDP*).

2.1.1. Newhouse (1977)

The seminal article by Newhouse attempted to identify factors determining the quantity of health care services in 13 developed countries using 1971 data. He regressed *HE* on *GDP* working in US dollars ($) at annual average exchange rates and obtained the following results (t-value in parenthesis):

$$HE_i = -60 + \underset{(11.47)}{0.079}\, GDP_i, \quad R^2 = 0.92, \tag{2.1}$$

The two principal results were that aggregate income explains almost all, about 92 percent, of the variance in the level of *HE* between countries; and the income elasticity of health care exceeds one.

On the basis of these results, Newhouse made two strong inferences:

(1) Factors other than income, for example the price paid by the consumer and the method of reimbursing the physician, are of marginal significance.

(2) Health care is technically a *luxury good*, possibly arising from the fact that, at the margin, the demand for health care may relate more to *caring* (or subjective components of health) than to *curing* (or physiological health).

The latter result is "...consistent with the view that in the developed countries, medical care services at the margin have less to do with common measures of health status such as mortality and morbidity and more to do with services that are less easily measured such as relief of anxiety, somewhat more accurate diagnosis and heroic measures near the end of life" [Newhouse (1977, p. 123)].

Parkin et al. (1987) criticized Newhouse's conclusions, saying that they were based on microeconomic concepts but employed macrodata which gave rise to the well-known – daunting – problem of aggregation, and misspecification arising from omitted variables or inadequate functional form, and the conversion factor problem (see below).[3] Most empirical research has confirmed Newhouse's *empirical* results concerning the income elasticity and the high explanatory power of the relationship, irrespective of whether it is calculated at the mean from linear regressions or estimated directly as a constant in log-linear regressions. This holds both for rather heterogeneous samples such as those in Kleiman (1974) and for more homogeneous samples such as in the OECD countries [see Leu (1986), OECD (1987), Culyer (1988, 1989), Pfaff (1990), Gerdtham et al. (1988, 1992a, 1992b)], see also Leviatan (1964), Abel-Smith (1967)]. Parkin et al. (1987) and Gbesemete and Gerdtham (1992) represent two exceptions from the regular finding that the estimated health care income elasticity exceeds unity in cross-section studies. Parkin et al. replicated Newhouse (1977) with research based on 18 OECD countries and 1980 data using different functional forms (linear, semi-log, double-log, exponential) and using different conversion factors (exchange rates and PPPs). Their results indicated that certain functional forms imply specific magnitudes of Engel income elasticities of medical care (p. 119) and that income elasticities are around unity in cross-sections when PPP conversion factors, rather than exchange rates, are used. Gbsemete and Gerdtham investigated health expenditure in 30 African countries and reported an income elasticity not significantly higher than one.[4]

2.1.2. *Is health care a "luxury" good*

However, the high income elasticity and the high explanatory power of the relationship contrast with the evidence obtained from national micro data (for example, household

[3] It appears that Newhouse later modified his position regarding the size of the income elasticity. Newhouse (1992, p. 8, footnote 7) argued that since income elasticities from time series data within countries are around unity, and that time series income elasticity would be expected to exceed the cross section elasticity (because technology is not held constant over time and new technology is likely to increase health expenditure), the cross-section income elasticity should be lower than unity.

[4] The following explanatory variables were included in the regression analysis: percentage of births attended by health staff (hospital deliveries), Gross National Product (GNP) per capita in US$, population under 15 years of age as percentage of total population, crude birth rates and foreign aid received per capita in US$. In their preferred model, three explanatory variables were positive and significant, i.e. percentage of births attended by health staff, GNP per capita and foreign aid received per capita.

surveys), where numerous studies have revealed a low income elasticity for the utilization of health care across households [Andersen and Benham (1970), Grossman (1972), Newhouse and Phelps (1974), Muurinen (1982), Okunade (1985), Wagstaff (1986), Manning et al. (1987)].

Several hypotheses have been put forward to explain this difference:

- Since insured individuals or households pay only a minor fraction of the health care costs as direct out-of-pocket payments, income may be less of a budget constraint on individual *HE*. By contrast, the nation as a whole faces the full costs of health care consumption and, where health care is largely financed by the state, the income constraint may be more binding at the aggregate level, particularly where there is non-price rationing [Newhouse (1977)]. If non-price rationing is relaxed with increasing income, then the income effect at the aggregate level will be greater than at the individual level [Culyer (1988)]. However, Blomqvist and Carter (1997, p. 208) noted further that insurance *per se* does not explain the discrepancy between micro and macro income elasticity estimates, and they observed that individuals in a private insurance system are restricted by the provisions of the insurance plan. If the higher income of rich families enables them to buy more generous insurance plans than the poor, then "there is no reason why spending patterns across rich and poor families would look any different from spending patterns across rich and poor countries" (footnote 1).
- Cross-section estimates may have been misspecified: the high income elasticity at the aggregate level may reflect omitted variables, for example differences in degrees of supplier-induced demand, so that the income coefficient may not be a measure of the pure income elasticity in an Engel curve sense [Parkin et al. (1987, 1989), McGuire et al. (1993)].
- There may be an inadequate distinction between prices and quantities. Newhouse used market exchange rates to convert expenditure and income data to a common currency unit. However, exchange rates reflect at best the relative prices of internationally traded commodities only, and income and expenditure data converted at exchange rates therefore still in nominal values. The availability of PPPs in recent years brought the issue of conversion factor into the analysis of health expenditure: the choice between exchange rates and PPPs, and if PPPs are used then the choice between PPPs for all expenditure (PPPs for GDP) which erase differences in overall price levels between countries, or specific PPPs for health care which erase differences in prices for health care. Parkin et al. (1987, 1989) argued in favor of using PPPs for health care to convert health expenditure in national currencies, since this conversion method provides a measure much closer to Newhouse's "quantity of resources a country devotes to medical care". They argued that the effect of income changes on health expenditure and quantity of health care is identical only if prices of health care do not vary with income, and they noted that this may not be the case. The production of health care is relatively labor intensive, and if labor is more scarce in rich countries than in poor countries it is compensated better; thus the relative prices of health care may increase with income across countries.

2.1.3. Empirical results on the issue of conversion factor instability

The empirical evidence in this matter seems ambiguous. Parkin et al. (1987, 1989) found that the (simple) income elasticity of health expenditure dropped when they used PPPs for health care instead of PPPs for GDP, and the elasticity was not significantly different from unity. They used cross-section data from 1980 and the PPPs estimated that year. They claimed, therefore, that countries spend resources for health care in proportion to their income, but richer countries pay more for the services. Using 1985 cross-section data and 1985 PPPs for health care, Gerdtham and Jönsson (1991b) could not replicate this result, and reported that the income elasticity is the same, and above unity, both when PPPs for health care and PPPs for GDP are used [see also Murthy (1992) and Gerdtham and Jönsson (1992)]. In sum: Parkin et al. (1987) and Gerdtham and Jönsson (1991b) focused on the sensitivity of the estimated income elasticity to the choice of different conversion methods, and therefore prices were only introduced as deflators and not as unrestricted explanatory variables. In a later study, Gerdtham and Jönsson (1991c) investigated the price/quantity issue using the same data as in Gerdtham and Jönsson (1991b) but also included the relative price of health care as an additional explanatory variable on the quantity of health care. The results showed that the income elasticity and price elasticity were 1.43 and −0.84, respectively, which indicated that the relative price of health care has a strong rationing effect on quantity, i.e. decision-makers will adjust the quantity of health care according to price changes. Moreover, the null hypothesis of a unit price elasticity with respect to health care could not be rejected, which indicated that price inflation above general price inflation is compensated fully by decreases (increases) in real resources. Milne and Molana (1991) reached about the same empirical results as Gerdtham and Jönsson (1991c) when they pooled 1980 and 1985 cross-section data for 11 EC countries (see their Table 2). Their results indicated in accordance with Gerdtham and Jönsson (1991c) that the income elasticity on health expenditure was higher than unity even when the relative price of health care were included as an explanatory variable.[5] These divergent findings of Parkin et al. and Gerdtham and Jönsson (1991b, 1991c) are probably partly attributable to differences between the 1980 and 1985 PPPs for health care.[6] Gerdtham and Jönsson (1991c) noted also that the choice between PPPs for GDP and PPPs for health care depends on what one wants to measure, whether it is

[5] Milne and Molana (1991) interpreted their results differently to Gerdtham and Jönsson (1991c) in that they stated: "Our empirical results, based on a conventional model and cross national data set for the EC, show that whereas health care may be labeled as a luxury good, the large income effect can be interpreted as merely offsetting the price effect" (p. 1221). However, it appears that they have no foundation for this conclusion since they did not estimate the income elasticity on health care relative prices. If this elasticity is zero as Gerdtham and Jönsson (1991c) found in their study and as was argued by Newhouse (1977, 1987), then the results of Milne and Molana are consistent with Gerdtham and Jönsson (1991c).

[6] There are two main technical problems with PPP adjustments of health expenditure. The first is that the number of products in the health "basket" was limited. In a sector with complex outputs and significant variations across countries in the types of outputs produced, this can be a serious problem. The second concerns

the financing burden on the country (PPPs for GDP) or the quantity of resources spent by a country on health care (PPPs for health care). Newhouse (1987) pointed out further that the finding that the income elasticity exceeds unity is only a secondary issue; what is more interesting is the finding that the income elasticity found in international cross sections substantially exceeds zero and substantially exceeds the corresponding estimates from within-county cross sections; and he asserted that it is this difference which "... is interesting and suggestive of what the marginal resources are buying" (pp. 161–162).

2.2. Cross-section multivariate regressions

2.2.1. Leu (1986) – a public-choice approach

Because of possible omitted variable bias in the income coefficient, some researchers have asked whether other variables have any significant independent impact on national *HE*. Leu (1986), using national data for 1974 data for 19 OECD countries (excluding Luxembourg, Iceland, Japan, Portugal and Turkey), included the following regressors:

- A set of relevant exogenous variables. These included the fraction of persons under 15 and over 65 (these groups tend to use more health care than others); and urbanization (the risk of contagion is higher (Kleiman, 1974), and time and travel costs are lower in cities).
- A variable to reflect the extent of public sector provision of health services. On the basis of "some well-known results in the public choice literature" (p. 42), Leu argued that an increase in the size of the public share would increase total spending. This could occur via two channels: bureaucrats in public or private non-profit hospitals would maximize budgets to increase their own utility (status, better pay, promotion possibilities etc.); and unit costs at each level of activity would be higher due to less intensive competition in the public sector. Leu also suggested that *HE* should increase with an increased fraction of public finance, assuming implicitly that this fraction reduces the price to the consumer.
- Dummies for the National Health Service (the UK and New Zealand), where centralized budgetary control might have a restraining effect; and for direct democracy (Switzerland), on the grounds that controlling *HE* would be easier if voters had greater direct control over government choice and tax levels.

the weighting method. The deflators should in principle reflect the health expenditure item in the general government consumption and private consumption components in the national accounts. There were only 6 countries which provided the expenditure breakdown (i.e. the weights) and the prices for that part of health expenditure included in general government consumption. In practice, statisticians at Eurostat used only the private sector weights in calculating PPPs for health expenditure, even in those countries where a breakdown for the general government component of health expenditure was available. For a country like the UK, the PPPs are therefore based on weights and prices for only 15 percent of the total. It may be possible in future to use better PPPs, in part by using general government weights for the countries where they are available and more appropriate weights for countries where the private sector weights do not appear to be an appropriate representation of the structure of health expenditure.

Leu confirmed the predominant effect of the income variable. He also found that a number of additional variables were significant and with the expected signs, albeit with mostly small coefficients. The stronger effects were: a 10 percent increase in the public to total bed ratio was expected to increase *HE* by 8–9 percent, and the NHS dummy suggested that this system lowered *HE* on average by 20 to 25 percent, *ceteris paribus*. An increase in the fraction of public financing by 10 percent was associated with 2–3 percent higher *HE*, *ceteris paribus*.

These conclusions have remained controversial, particularly as regards the institutional variables. Despite the reference to public choice theory, the *a priori* signs of the variables proposed by Leu remain in doubt and Gerdtham et al. (1988, 1992a, 1992b) in subsequent tests on more recent data were not able to reproduce these results.

Culyer (1988, 1989) noted that private sector bureaucrats are not necessarily better controlled than their colleagues in the public sector, that costs in the private sector may be larger due to advertising and selling costs and that market pressures may be less reliable than professional ethics and regulation (p. 28). He also quoted the conclusion from a review of empirical comparisons [Stoddart and Labelle (1985)] that privately owned for-profit hospitals do not operate at lower production costs than non-profit hospitals. Barr (1992) added that much of the private/public argument "is clouded by ideology" and "In many respects, however, managers, administrators, and bureaucrats all do broadly the same job and face similar problems" (p. 784). Culyer (1989) suggests that both of Leu's hypotheses, i.e. that both public finance and public provision increase expenditure, depend on a passive response from the financing agent, who adjusts the supply of finance to the quantities and prices of health care services. He suggests further that the financing mechanism, in particular the degree of "open-endedness" of finance, i.e. the lack of budget restriction, would be more relevant than the distribution of finance and provision between public and private institutions. Open-ended financing systems are characterized by multiple finance sources (insurance companies) and by fee-for-service remuneration. Conversely, closed systems are characterized by one or a few finance agents, prospective payments such as capitation for out-patient services, and global budgets for hospitals. Open-ended systems provide little incentive for providers and little opportunity for financiers to contain expenditure; the converse is true for closed systems. The conclusion of all this appears to be that the impact of the fraction of public finance and/or provision on health care expenditure cannot be determined *a priori*. However, countries with more closed health care financing systems are anticipated to have lower expenditure.

2.2.2. Gerdtham et al. (1992a, 1992b) – impact of open-ended finance

Gerdtham et al. (1992a, 1992b) used cross-sectional and pooled cross-sectional (over three selected years) data sets. They attempted to measure the effect of open-endedness of finance on health expenditure. For both data sets, Gerdtham et al. specified the following log-linear model (the continuous variables are transformed in natural logarithms), implying that the coefficients of the variables are to be interpreted as constant elasticities:

$$\begin{aligned} HE_i &= b_0 + b_1 GDP_i + b_2 RP_i + b_3 DOCT_i + b_4 TEXMC_i + b_5 PF_i \\ &\quad + b_6 FEE_i + b_7 GLOBAL_i + b_8 FP_i + b_9 AGE_i + b_{10} URB_i + e_i, \end{aligned} \tag{2.2}$$

$i = 1, 2, \dots, 19$; $T = 1987$ for Gerdtham et al. (1992a) and $T = 1974, 1980, 1987$ for Gerdtham et al. (1992b), *RP* is relative prices,[7] *DOCT* is the number of doctors,[8] *TEXMC* is the ratio of in-patient to total spending, *PF* is the ratio of government to total *HE*, *FP* is the female participation ratio,[9] *AGE* is the ratio between population 65 years of age and over and population aged 15 to 64, *URB* is urbanization, i.e. the fraction of population living in towns with over 500,000 inhabitants (1980), *FEE* is a dummy variable for the fee-for-service payment of doctors and *GLOBAL* is a dummy variable for global budgeting caps. The latter two dummy variables are measures of the open-endedness of finance in out-patient service and hospital care, respectively.

Strict cross-section estimates based on 1987 data for 19 OECD countries were reported in Gerdtham et al. (1992a). The preferred model had five variables: per capita *GDP*, urbanization, fraction of public financing, fraction of in-patient care expenditure and the dummy variable for countries with fee-for-service payment; this accounted for about 95 percent of the variance and nearly all variables had expected sign (Column GTH1 in Table 3). These authors also tested the most appropriate functional form and found that a logarithmic transformation was superior to linear and exponential specifications. *GDP* continued to be the most important variable in explaining *HE*, with an elasticity of 1.33 (significantly different from one). In contrast to Leu, an increase in the fraction of public financing by 10 percent was associated with 5 percent *lower HE*, while a 10 percent increase in the fraction of in-patient care had a *positive* impact on expenditure of around 2 percent. The fee-for-service dummy variable indicated that *HE* was about 11 percent higher in countries where that arrangement dominated. None of the demographic variables except urbanization was significant and this had an unexpected (negative) sign.

Gerdtham et al. (1992b) showed that two variables in addition to the five of the previous cross-section study were statistically significant (Column GHT2 in Table 3): a 10 percent increase in the fraction of those aged above 64[10] increased *HE* by about 2 percent; an increase in the number of physicians per capita by 10 percent reduced *HE* by

[7] PPPs for health relative to PPPs for *GDP*.

[8] The term provides a measure of supplier-induced demand – as the number of doctors increases, with work load held constant, doctors may try to induce patients to use more services [Evans (1974)]. It may also proxy for unmeasured demand factors. These cannot be distinguished.

[9] The participation term represents the possible replacement of informal care in the home by more formal institutional care as women increasingly go out to work [Fuchs (1972), Ståhl (1986)].

[10] Relative to those in the 15 to 64 age group.

Table 3
Results from selected works on comparisons of health expenditure across the OECD countries

Study	Newhouse	Leu 1	Leu 2	Leu 3	GTH 1	GTH 2	GTH 3	GTH 4
Design	Cross-section	Cross-section			Cross-section	Pooled 3 year	Pooled 16 year	
Sample	13 countries	19 countries			19 countries	19 countries	22 countries	
Year	1971	1974			1987	1974, 1980, 1987	1972–1987	
Estimation	OLS	OLS			OLS	OLS	WLS	
Regressor variable								
$GDPpc(i,t)$	0.078^{a} (1.31)^{2}	1.18^{a}	1.36^{a}	1.21^{a}	1.33^{a}	1.27^{a}	0.74	–
$GDPpc(i,t)/GDPpc(i,t-1)$	–	–	–	–	–	–	–	0.17
$HEXTpc(i,t-1)/GDPpc(i,t-1)$	–	–	–	–	–	–	–	−0.22^{b}
Inflation(i,t)	–	–	–	–	–	–	−0.16	−0.17^{b}
Inflation$(i,t-1)$	–	–	–	–	–	–	–	−0.00
Population $<$ 15 year(i,t)	–	0.56*	1.10^{a}	0.69^{a}	–	–	–	–
$65+/15$–64 years(i,t)	–	–	–	–	–	–	−0.11	0.21
$65+/15$–64 years$(i,t-1)$	–	–	–	–	–	–	–	−0.16
Urbanization(i,t)	–	0.11	0.28^{a}	–	−0.17^{b}	−0.23^{b}	–	–
Public financing(i,t)	–	–	0.34^{b}	0.16	−0.52^{a}	−0.48^{a}	−0.12	−0.21^{b}
Public financing$(i,t-1)$	–	–	–	–	–	–	–	0.24^{b}
Public beds(i,t)	–	0.90^{a}	–	0.85^{a}	–	–	–	–
INP%(i,t)	–	–	– –	–	0.22^{c}	0.31^{a}	–	–
Physicians/pop(i,t)	–	–	–	–	–	−0.17^{c}	–	–
NHS(i,t)	–	−0.21^{a}	−0.24^{b}	−0.23^{a}	–	–	–	–
Direct democracy(i,t)	–	−0.31^{a}	−0.20	−0.29^{a}	–	–	–	–
Fee/service(i,t)	–	–	– –	–	1.12^{b}	1.13	–	–
Constant	–	−12.41^{a}	−9.65^{a}	−10.06^{a}	25.10^{a}	−4.35^{a}	−0.03	−0.67^{b}
Country dummies	–	–	–	–	–	No	Yes	Yes
Time dummies	–	–	–	–	–	Yes	Yes	Yes
R^2		0.97	0.96	0.97	0.94	0.92	0.97	0.31

[a,b,c] Represent 1%, 5% and 10% levels of significance, respectively.

* Linear regression; elasticity estimated at the mean.

10 percent. The remaining variables had broadly the same orders of magnitude as in the strict cross-section estimates.

A further objective of this study was to shed light on econometric aspects such as temporal stability of the estimated relationships and their functional form.

Temporal stability was analyzed within a model where all coefficients and regression variances were allowed to vary over the three cross-section years. The actual tests were carried out sequentially starting with the equality restriction on the regression variances followed by the equal slope vector hypotheses. It turned out that the specified *HE* model was stable over time, both in slopes and in variance, but that the *HE* function shifted upwards, i.e. the average *HE* increased autonomously by 2.5 percent.

The functional form issue was analyzed within the framework of Box–Cox transformation analysis [Box and Cox (1964), Zarembka (1974), Spitzer (1982)]. The idea in the Box–Cox analysis is to parameterize the functional form by estimable transformation parameters. In the study two transformation parameters were specified $(\lambda_y; \lambda_x)$: one for the dependent variable and one for a relevant subset of the regression variables. Given estimates of the transformation parameters, hypotheses about particular values of these (and hence the functional form) can be tested. The results indicated that the quadratic (square-root power for response, linear for regressors) functional form fitted the data best on the likelihood criterion applied to a two-parameter Box–Cox regression model. Thus the entire regression analysis based on the double log model was replicated, with square roots of *HE* in order to examine sensitivity in the results to the choice of functional form. The square root functional form did not alter the general results. The signs of the coefficients were unchanged, and the outcomes of the model tests and misspecification analysis were similar to the results based on the double log functional form.

3. Second-generation studies

3.1. Panel data analyses

3.1.1. Methods

Panel data [Greene (1993)] enable one to test for country and time-invariant effects and carry out appropriate estimation in their presence. The error term in a typical panel data model is of the form: $\varepsilon_{it} + \mu_i + \theta_t$, where μ is the country-specific term and θ is the time-specific term. Different ways of modeling these country and time-specific terms give rise to different panel data models. Running a simple least squares (OLS) regression assumes that $\mu_i = 0$ and $\theta_t = 0$. A fixed-effects model assumes that μ_i and/or θ_t are fixed constants for each country and time period respectively, in which case an appropriate panel estimation model is OLS with country-specific and/or time-specific dummy variables. If $\mu_i \neq 0$ is the correct specification, but a strict OLS regression is estimated, the coefficient vector will be biased if μ_i is correlated with other regressors. The third possibility is that μ (or θ) is itself a random variable. In this case, there is an error components model – referred to as a random-effects model – that can be estimated using generalized least squares (GLS). The random-effects model is estimated by two-stage GLS in the following manner: the variance components are estimated using the residuals from OLS regressions and GLS estimates are calculated using these estimated variances. Conventional F-type tests can be used (for the joint significance of the country and time dummy variables) to determine whether the OLS model is rejected in favor of the fixed-effects model. A Lagrange multiplier test for the random-effects models devised by Breusch and Pagan (1980), based on OLS residuals, can be used to examine whether the panel GLS model is more appropriate than the strict OLS model. A Hausman test [Hausman (1978)] of the fixed-effects model against the random-effects can be carried out to test the independence assumption of the random-effects model. If this assumption is not valid, then random-effects models produce biased coefficient estimates and the fixed-effects model may be preferable. If these effects are uncorrelated with the regressors, then there is a gain to be made by adopting the random-effects model instead of the fixed-effects model. In addition to using statistical tests to choose between fixed-effects, random-effects and strict OLS, there are important conceptual issues that bear on this choice. The fixed-effects model may be more appropriate when the sample constitutes all or most of the population of interest. Random-effects would be more appropriate if the sample is drawn from a substantially larger population. This factor would seem to favor the fixed-effects model in the case of the OECD countries. Greene (1990, p. 85) states that the fixed-effects model is a reasonable approach when one can be confident that the differences between units (countries) can be viewed as parametric shifts of the regression function. If differences between countries are not due to parametric shifts, but are more related to variation across countries in the regressors, then fixed-effects models are less attractive [for thorough discussions of panel data see Hsiao (1986), Baltagi (1995); see also Pesaran and Smith (1995)].

3.1.2. Gerdtham (1992) – panel data and error-correction models

Gerdtham (1992) used data for 22 OECD countries for the period 1972–1987, exploring different panel data models and issues of lags and dynamic adjustment of *HE* to movements in exogenous variables. A reduced number of explanatory variables were specified: *GDP*, inflation, fraction of public financing, and the fraction of the aged in the population. Both static and restricted error-correction models were specified and tests were carried out using five different panel data models, i.e. two-way country and period fixed and random-effects models, one-way fixed and random country effects models and strict OLS without country and time dummies as well. An important conclusion was that country or time-specific effects, and whether these were treated as fixed or random variables, had important implications for the results. Indeed, permanent non-identified country and time-period effects were found to influence *HE* and had an important impact on the income elasticity of demand. These effects appeared to be fixed, as random-effects models were clearly rejected by the data. Major statistical results regarding the influence of variables were: first, the estimated elasticity of *HE* with respect to *GDP* was

0.74 in static equilibrium models (using both country- and time-period dummies) (Column GTH3) but the remaining variables were insignificant. In dynamic specifications (Column GTH4), the short-run effect of income on *HE* was 0.18 and a unitary long-run income elasticity with respect to *HE* of 1.0 was not rejected. Second, the short-run elasticity for inflation was -0.17, suggesting that when inflation increased, per capita *HE* grew less rapidly. There was a short-run elasticity of -0.21 with respect to the fraction of public financing, but there appeared to be no long-run effect.

3.1.3. Hitiris and Posnett (1992) – replications of Newhouse and Leu

Hitiris and Posnett (1992) re-analyzed the models of Newhouse and Leu using panel data for 20 OECD countries for the period 1960–1987. The specified models assume constant regression coefficients but allow for differing intercepts across groups of countries and for cross-sectionally heteroscedastic and time-wise autoregressive residuals. All models were estimated both in linear and log-linear form. Their results re-confirmed the importance of *GDP* as a major determinant of *HE*, with an elasticity of about one (1.026 with an exchange rate adjustment and 1.16 with a PPP adjustment). The importance of some non-income variables was also confirmed, although the direct effect of such factors appeared to be small.

3.1.4. Viscusi (1994a) – risk–risk analysis

Health promoting policies intended to reduce mortality risks may actually increase mortality risks since they also reduce citizens' disposable income, which in turn increases mortality risks [see Viscusi (1994a, 1994b), Keeney (1997)]. This implies that it is important to estimate the marginal expenditure per statistical life lost, and Viscusi (1994a) proceeds from the following relationship between the expenditure that will generate the loss of a statistical life and the marginal value of life:

$$\text{Marginal expenditure per statistical life lost} = \frac{\text{Marginal value of life}}{\text{Marginal propensity to spend on health}}.$$

If the marginal propensity to spend out of income on mortality-reducing activities such as health care is 1.0 then the marginal expenditure per statistical life lost will equal the marginal value of life. Since not all of individuals' additional income is devoted to such activities, then the marginal expenditure per statistical life will exceed the marginal value of life [Viscusi (1994a)]. One approach to estimating the marginal expenditure per statistical life lost is, first, to estimate the marginal propensity to consume health care out of income using international OECD data, and then to use this figure as a denominator in the equality above in conjunction with a value-of-life range of $3 million to $7 million. In the estimation of the marginal propensity to spend, Viscusi used panel data for 24 OECD countries for the years 1960–1989 and a log-linear weighted least squares model of *HE* (weights were country populations by year) including *GDP* and unemployment rates with and without 29 year dummies and 23 country dummies (two-way fixed-effects models) and also with and without unemployment rates. *HE* and *GDP* were converted by both the current exchange rates and PPPs. In accord with previous studies, the results show that *GDP* alone has a extremely high explanatory power, as is evident from the very high adjusted R^2, and that the unemployment rate was insignificant. The estimated income elasticity in the two-way fixed-effects models is about 1.10 irrespective of whether *HE* and *GDP* are converted by exchange rates or PPPs. The estimated marginal propensity to spend was around 0.1, which implies that the marginal expenditure that will lead to the loss of one statistical life ranges from $30 million to $70 million dollars, with a mid-point of $50 million dollars.

3.1.5. Gerdtham et al. (1998) – effects of institutional variables

Gerdtham et al. (1998) used data for 22 OECD countries for the period 1970–1991 to examine the effects of different sorts of institutional arrangements on *HE*. The *HE* model of the ith OECD country in year t were written in a log-linear form where all continuous variables were defined in natural logarithms:

$$HE_{it} = b_0 + \sum_{j=1}^{14} b_j X_{itj} + \sum_{j=1}^{10} c_j Z_{itj} + \sum_{i}^{22} \mu_{0i} d_i + \sum_{t}^{22} \theta_{0t} d_t + e_{it}, \tag{3.1}$$

where the variables are defined as: $X_1 = GDP$; $X_2 = \%$ of population 75 years and over (*POP*75); $X_3 = \%$ of population 4 years and under (*POP*04); $X_4 =$ Female labor force participation ratio, % of active population (*FPR*); $X_5 =$ unemployment rates, % of labor force (*UNR*); $X_6 =$ alcohol intake, liters per person (*ALCC*); $X_7 =$ tobacco consumption per capita (*TOBCC*); $X_8 = \%$ of beneficiary's health bills normally paid by a public insurer or fund (*COPAY*); $X_9 = \%$ of in-patient care expenditure in total expenditure (*TEXMC*); $X_{10} = \%$ of public in-patient care beds in total in-patient care beds (*PUSH*); $X_{11} = \%$ of the population covered by public insurers (*COVERO*); $X_{12} =$ renal dialysis per million of population (*REND*); $X_{13} =$ number of physicians per 1000 population (*DOCT*); $X_{14} =$ interaction of number of doctors and dummy for fee-for-service system (*DOCT* · *FFSA*); $Z_1 =$ dummy for public reimbursement systems (*PUBREIMB*); $Z_2 =$ dummy for public integrated systems (*PUBINTEGR*);[11] $Z_3 =$ dummy for budget ceilings in the ambulatory sector (*BUDCEILA*); $Z_4 =$ dummy for budget ceilings in the hospital sector (*BUDCEILI*); $Z_5 =$ dummy for gatekeeping

[11] The following classification was used: (public) reimbursement: Australia, Belgium, France, Italy (up to 1978), Japan, Luxembourg, Switzerland and the United States. Public contract: Austria, Canada, Germany, Greece (until 1983), the Netherlands, Portugal (until 1977), Spain (until 1983) and Turkey. Public integrated: Denmark, Finland, Greece (from 1983), Iceland, Ireland, Italy (from 1979), New Zealand, Norway, Portugal (from 1978), Spain (from 1984), Sweden and the United Kingdom.

systems (*GATEKEEP*);[12] Z_6 = dummy for direct payment by patient before reimbursement by insurer (*REIMBMOD*); Z_7 = dummy for capitation (*CAPITA*); z_8 = dummy for wage and salary (*WAG&SALA*);[13] Z_9 = dummy for "overbilling" where there are no "official" or agreed price schedules set (*OVERBILL* = 1); Z_{10} = dummy for fee-for-service or payment by bed days in in-patient care (*FFSI*). The 22 (22) d_i (d_t) are dummy variables with the value one for each of the 22 observations corresponding to country (time) i (t), ordered alphabetically (chronologically), and zero elsewhere (two-way fixed-effects models). b_0 is an overall constant as well as a "country" effect for each country and a "time" effect for each period. The problem of multicollinearity, when the time and country dummy variables both sum to one, is avoided by imposing the restriction: $\sum_i^{22} \mu_0 = \sum_i^{22} \theta_0 = 0$.

Table 4, Column 1, shows the estimated impact of non-institutional (*GDP* and *TOBCC*) and all the "institutional" variables on *HE*; Column 2 shows the results after the elimination of insignificant institutional variables. Columns 3–15 show estimated regressions for various sub-groups of institutional variables added to a number of non-institutional variables: *GDP*, *POP*75, *POP*04, *FPR*, *UNR*, *ALCC* and *TOBCC*.

Amongst the non-institutional variables, only *GDP* and tobacco consumption (*TOBCC*) have generally a significant impact on health expenditure; it appears that the estimated coefficients are rather stable over the variables excluded. The elasticity on tobacco consumption indicates that health expenditure would increase by about 1.3 percent if tobacco consumption increased by 10 percent. Tobacco consumption may, in part, be a proxy for other behaviour that leads to higher health expenditure. Some studies have shown that smoking does not increase health expenditure over the life cycle, for example Leu and Schaub (1983). The income elasticity is lower than unity (0.74), which indicates that health expenditure may be better characterized as a "necessity" rather than as a "luxury" good, and corresponds with results of micro-level studies.

Considering next the institutional variables, unexpected results were found for the dummy variables representing the dominant type of institutional arrangement. In contrast to the evidence of Hurst [OECD (1992)], public reimbursement (*PUBREIMB*) appeared the least expensive, with public integrated arrangements (*PUBINTEG*) about as costly as public contract (benchmark). Countries with budget ceilings on inpatient care (*BUDCEILI*) appeared to have higher total expenditure, while larger numbers of doctors

[12] Countries with primary physicians as gatekeepers: Austria, Canada, Denmark, Germany, Iceland, Ireland, Italy, the Netherlands, New Zealand, Norway, Portugal, Spain and the United Kingdom. Countries where this does not appear to be the case: Australia, Belgium, Finland, France, Greece, Japan, Luxembourg, Sweden, Switzerland, Turkey and the United States.

[13] Countries were classified as follows: fee-for-service: Australia, Austria, Belgium, Canada, France, Germany, Greece, Ireland (up to March 1989), Italy (up to 1977), Japan, Luxembourg, New Zealand, Norway, Switzerland and the United States. Capitation: Denmark, Iceland, Ireland (from March 1989 on for the publicly financed system), Italy (1978 on), the Netherlands, Spain (up to 1983 and then falls gradually) and the United Kingdom. Wage and salary: Finland, Portugal, Spain (gradually increasing after 1984), Sweden and Turkey.

Table 4
Estimated coefficients using the two-way fixed effects model for health expenditure. Coefficients (country and time-effects) are not presented

Variables/Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
GDP	0.76^{a}	0.74^{a}	0.67^{a}	0.74^{a}	0.66^{a}	0.75^{a}	0.78^{a}	0.68^{a}	0.80^{a}	0.71^{a}	0.79^{a}	0.71^{a}	0.79^{a}	0.82^{a}	0.73^{a}
POP75			−0.04^{c}	−0.02	−0.04	−0.03	−0.03	−0.04^{c}	−0.01	−0.02	−0.02	−0.02	−0.03	−0.03	−0.04^{c}
POP04			0.03	0.10^{b}	0.10^{c}	0.10^{b}	0.06	0.07	0.02	0.02	0.04	0.04	0.06	0.04	0.03
FPR			0.04	−0.05	0.00	−0.06	−0.07	−0.01	−0.04	0.03	−0.05	0.02	−0.02	−0.03	0.03
UNR			0.00^{c}	0.00^{b}	0.00^{b}	0.00^{c}	0.00	0.00^{c}	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALCC			−0.01	−0.03	−0.02	−0.02	0.01	0.01	0.01	0.00	0.02	0.02	−0.04	−0.04	−0.04
TOBC	0.12^{a}	0.13^{a}	0.11^{a}	0.09^{a}	0.10^{a}	0.09^{a}	0.11^{a}	0.12^{a}	0.12^{a}	0.13^{a}	0.11^{a}	0.13^{a}	0.11^{a}	0.11^{a}	0.11^{a}
COPAY	−0.08		−0.07		−0.07			0.01		−0.08		−0.06			−0.09
TEXMC	0.05^{c}	0.06^{b}	0.05		0.05			0.04		0.07^{c}		0.07^{c}			0.04
PUSH	−0.34^{a}	−0.32^{a}	−0.16^{b}		−0.21^{a}			−0.28^{a}		−0.22^{a}		−0.22^{a}			−0.14^{a}
COVERO	0.05		0.13^{c}		0.06			0.16^{b}		0.07		0.06			0.12^{c}
REND	0.01		0.03^{b}		0.02			0.02		0.01		0.01			0.03^{b}
PUBREIMB	−0.11^{a}	−0.07^{b}		0.07^{b}	0.01										
PUBINTEG	−0.03			0.11^{a}	0.08^{a}	0.07^{a}									
BUDCEILA	−0.01						0.00	0.00							
BUDCEILI	0.03	0.04^{a}					0.05^{b}	0.08^{a}							
GATEKEEP	−0.19^{a}	−0.18^{a}							−0.20^{a}	−0.19^{a}	−0.20^{a}	−0.19^{a}			
REIMBMOD	−0.10^{c}	−0.08^{c}											−0.16^{a}	−0.08	−0.10
CAPITA	−0.21^{a}	−0.17^{a}											−0.22^{a}		
WAG&SALA	−0.10												−0.10		
CAPITA+WAG&SALA														−0.23^{a}	−0.22^{a}

continued on next page

Table 4, *continued*

Variables/Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
OVERBILL	0.03										−0.05[c]	−0.05		−0.09[c]	−0.06
FFSI	−0.02														
DOCTCA	−0.10[c]	−0.14[a]							−0.08	−0.09[c]	−0.07	−0.09[c]			
DOCTCA*FFSA	0.18[a]	0.20[a]							0.18[a]	0.19[a]	0.20[a]	0.20[a]			
Constant	0.07	0.02	−0.51	−0.97	0.17	−0.95	−1.28[c]	−0.49	−1.59[b]	−0.47	−1.50[b]	−0.46	−1.36[c]	−1.57[b]	−0.78
R^2	0.985	0.984	0.981	0.980	0.981	0.980	0.980	0.982	0.982	0.983	0.982	0.983	0.980	0.981	0.982
Hausman $\chi 2(k-1)$	29.54[c]	167.07[a]	27.87[a]	203.51[a]	29.94[a]	194.96[a]	180.79[a]	28.99[a]	176.64[a]	30.01[a]	121.43[a]	29.52[a]	29.76[a]	157.10[a]	29.14[a]
F-test 2–15 against 1	–	0.49	15.28[a]	11.53[a]	18.64[a]	11.02[a]	11.88[a]	14.91[a]	9.83[a]	13.45[a]	7.84[a]	11.90[a]	12.31[a]	–	–
F-test against 1-FEM,C	6.15[a]	11.43[a]	5.89[a]	12.59[a]	6.52[a]	12.29[a]	12.26[a]	6.73[a]	12.62[a]	6.29[a]	8.10[a]	4.82[a]	11.36[a]	11.82[a]	5.99[a]
F-test against 1-FEM,P	47.58[a]	70.78[a]	62.10[a]	68.30[a]	63.98[a]	71.21[a]	69.96[a]	67.11[a]	73.51[a]	65.81[a]	78.17[a]	71.17[a]	56.39[a]	62.78[a]	53.65[a]
F-test against 0-FEM	27.03[a]	47.52[a]	33.68[a]	40.38[a]	30.38[a]	41.84[a]	41.18[a]	35.05[a]	43.16[a]	35.64[a]	45.60[a]	38.58[a]	33.54[a]	39.03[a]	29.29[a]

[a,b,c] Represent 1%, 5% and 10% levels of significance.

Abbreviations: Hausman $\chi 2(k-1)$ = test of the 2-way random effects model against the 2-way fixed effects model. The test is asymptotically distributed as a chi-squared variable with $k-1$ degrees of freedom. F-test 2–15 against 1 = F-test of model 2–15 against model 1; F-test against 1-FEM, C = F-test of the 1-way fixed country effects model (not presented) against the 2-way fixed effects model. F-test against 1-FEM, P = F-test of the 1-way fixed period effects model (not presented) against the 2-way fixed effects model. F-test against 0-FEM = F-test of the 0-way fixed effects model without country and period specific effects (not presented) against the 2-way fixed effects model.

(*DOCTCA*) appeared to be related to lower expenditure. A number of model specifications were tested in order to evaluate whether these results reflected inter-relationships between various institutional variables and to assess the robustness of the results more generally. Some additional information about the robustness of the coefficients for public integrated (*PUBINTEG*) and public reimbursement (*PUBREIMB*) systems was obtained by examining the interaction of these variables with the in-patient proportion variable (*TEXMC*) and the dummy for gatekeeping (*GATEKEEP*). In general, in-patient expenditure was more costly than ambulatory care and pharmaceutical expenditure, and countries with integrated systems appeared to have higher fractions of inpatient care. Furthermore, additional tests indicated that fewer countries with integrated systems have gatekeeper arrangements than in public reimbursement and public contract setups. On both these grounds, one might expect that public integrated systems could show up as being more expensive, on balance, than public contract systems. In equations with the non-institutional variables and these two system dummies alone (*PUBINTEG* and *PUBREIMB*), this appears to be the case. However, once the in-patient care variable (*TEXMC*) or the gatekeeper dummy (*GATEKEEP*) were introduced, the positive coefficient of the integrated model disappears; including these factors therefore appears to control for influences which tend to make integrated systems more expensive.

The results for public reimbursement systems are more difficult to explain. In the equation reported above (including the non-institutional variables and only the two dummies) public reimbursement systems were more expensive than public contract systems. However, once the additional variables, including the in-patient proportion variable and the gatekeeper dummy, were introduced such systems appeared to be the least expensive of the three, even though one would expect an even larger positive coefficient for *PUBREIMB* once the influence of their lower proportion of in-patient care and the (possible) tendency to have more gatekeeping is controlled for, i.e. once the influences tending to hold down health expenditure in reimbursement systems are controlled for. In this context, the approximate nature of the dummies for the three systems needs to be stressed. It remains difficult to judge to what degree country dummies, for example for the United States, may be picking up part of the variance which should be attributed to institutional differences. Further investigations suggested that the variables representing coverage of the population (*COVERO*) do not add anything to the explanatory power of the model if the reimbursement and integrated system variables (*PUBREIMB* and *PUBINTEG*) are included in the regression. The variable for budget ceilings on in-patient care (*BUDCEILI*) captured the possible impact of budget ceilings in ambulatory care (*BUDCEILA*) on *HE*. Finally, they found that the negative impact of the number of doctors per capita (*DOCTCA*) on *HE* appeared to result from the interaction between several variables, including the non-institutional variables, the gatekeeping dummy, and the interactive variable for doctor numbers and fee-for service arrangements ($DOCTCA \cdot FFSA$).

Bringing the results together, it seems to be the case that:

- The proportion of in-patient care expenditure (*TEXMC*) tends to be positively related to health expenditure.

- A higher proportion of public coverage of medical care billing (*COPAY*), and of public beds to total beds (*PUSH*), tends to generate lower health expenditure, contrary to hypotheses in previous work, e.g., Leu (1986). However, this impact was not robust to the sample.
- Public reimbursement systems (*PUBREIMB*) tend to be less expensive than public contract systems, although the significance was not robust over different country groups and time periods. There is no evidence which supports the hypothesis that public integrated systems (*PUBINTEG*) are less expensive than public contract systems, once the effects of the proportion of in-patient care and gatekeeping arrangements are allowed for.
- Systems with budget ceilings in ambulatory care (*BUDCEILA*) do not appear to be less expensive than systems without budget ceilings and, similarly, budget ceilings in in-patient care (*BUDCEILI*) do not seem to lower health expenditure.
- Countries with primary physicians as gatekeepers for in-patient care (*GATEKEEP*) have consistently lower health expenditure, a finding that was robust in different samples.
- Countries with more doctors have lower health expenditure (*DOCTCA*). This result was sensitive to the variables included in the equation. Although unexpected, this result is consistent with earlier studies [see, for example, Gerdtham et al. (1992a, 1992b)]. One possible explanation may be that an increase in doctor numbers generally drives down income levels (as appears to be the case, for example, in Belgium). However, it seems that the number of doctors increases health expenditure in systems which reimburse their physicians by fee-for-service ($DOCTCA \cdot FFSA$). This latter finding was robust to the sample.
- Countries which reimburse their physicians by capitation (*CAPITA*) appear to have lower health expenditure, but there was no evidence in favor of the hypothesis that countries reimbursing physicians by means of salaries (*WAG&SALA*) have lower health expenditure than those using a fee-for-service approach.
- Health expenditure does not appear to be higher in countries with payments by bed-day or fee-for-service in in-patient care (*FFSI*).
- Countries where the patient pays the provider and then seeks reimbursement (*REIMBMOD*) tend to have lower health expenditure.

One particular problem in using the fixed-effects model is that "too much" of the cross-section variation may be attributed to the dummy variables representing specific countries and/or time-periods, rather than to the regressors which attempt to capture the influences of economic and institutional factors. For example, the relatively high health care expenditure in the USA, or the relatively low health care expenditure in Japan (even after controlling for *GDP*), may appear to be "explained" by the dummy variable for these two countries rather than by their particular mix of institutional arrangements. Such findings may to some extent be valid: influences which are unique to particular countries (such as social and cultural factors) or particular time-periods (e.g., cyclical downturns) may well account for some of the variations observed in health care expenditure. Nevertheless, the concern is that this estimation method may further weaken the scope for finding significant differences in health expenditure as a result of institutional factors. This risk of the fixed-effects models needs however to be balanced against that arising from specification errors if relevant dummy variables are not included.

Another important problem concerns how the individual effects should be interpreted; for example, budget ceilings may be endogenous to health expenditure because policy makers in different countries may respond to higher expenditure by implementing spending caps. If this is the case, then it is possible that the direction of causality between health expenditure and budget ceilings runs two ways, and estimated effects of the dummy variable for budget ceilings on expenditure will be biased by assuming one-way causality from budget ceilings to expenditure.

A third problem is that several variables appear to be closely related, such as the two budget ceiling variables, and also the dummies for gatekeeping, the fraction of in-patient care, and public integrated systems. It is possible under these circumstances that one variable will turn out to be insignificant, even if it has contributed to a significant effect found for the related variable. Multicollinearity may be severe, tending to confound the measurement of separate effects of individual regressor variables on expenditure and making it harder to obtain significant results, and rendering the coefficient estimates highly sensitive to the addition or deletion of other regressor variables.

3.1.6. Barros (1998) – levels and growth rates

Barros (1998) dealt with the same issue as in the previous studies but in a different way. He looked at differences in growth rates (averaged across decades) rather than in levels of health expenditure. He used data for 24 OECD countries from the period 1960–1990. The explanatory variables were as follows:

- Initial health expenditure, i.e. health expenditure per capita (in levels) in the first year of the period;
- Square of initial health expenditure;
- Gatekeeper dummy [as in Gerdtham et al. (1998)];
- Public reimbursement and integrated system dummies [as in Gerdtham et al. (1998)];
- GDP growth rate;
- % of population aged over 65 years;
- Two time decade dummies (1970–1980 and 1980–1990).

Barro's results indicated that the health system dummies (gatekeeper, public reimbursement and integrated systems) were clearly insignificant and the only significant variables were initial health expenditure, the square of initial health expenditure and growth of GDP. Not even the decade dummies were significant. The effect of initial health expenditure was negative, implying that a higher initial health expenditure would lead to lower growth rate in the next decade, which indicates convergence among countries. The effect of the square of initial health expenditure was positive, indicating that the absolute effect is stronger for heavier spenders but at a decreasing rate. In contrast to Gerdtham et al. (1998), Barros concluded that the existence of gatekeepers and the type of health system (public reimbursement and integrated systems) have played no significant role in containing health expenditure growth. He also concluded that aging and the relative size of public financing have not contributed to the growth of health expenditure. In accordance with Gerdtham et al. (1998), Barros concluded that the income elasticity was lower than but close to one. This result is also consistent with Gerdtham (1992), Gerdtham et al. (1998) and Blomqvist and Carter (1997). Barros also re-estimated his results with five-year average growth rates and year-to-year average growth rates and found similar results. Barros further re-estimated his models in *level* form and found about the same qualitative results as Gerdtham et al. (1998), at least with regard to the negative effect of public reimbursement systems on health expenditure.

3.1.7. Roberts (1998a) – heterogeneous models and average effects

Roberts (1998a) considered three issues in *HE* modeling: heterogeneous relationships across countries; dynamic relationships, and relationships including non-stationary variables (this issue will be discussed separately in the next subsection). In the paper, she used new methods to estimate the *average* effects of explanatory variables in heterogeneous dynamic and non-stationary relationships. Four alternative estimation methods were considered [see Pesaran and Smith (1995), Pesaran et al. (1995)]. One of these was the homogeneous fixed-effects estimator with common slopes which has been used by, for example, Gerdtham (1992). The remaining three estimation methods were: (1) the mean group (country) estimator, in which separate time series regressions are estimated for each group and the coefficients are averaged across groups; (2) time series regressions with the data averaged across countries; and (3) cross-section regressions with the data averaged over time. In the static case, if the coefficients differ randomly, these methods provide consistent (and unbiased) estimates of coefficients means but this is not true in the dynamic case, i.e. the fixed-effects estimator tends to underestimate short-run effects and overestimate long-run effects (p. 10). Roberts therefore suggested that fixed-effects error-correction results should be treated with caution.

Roberts used data from 20 OECD countries over the period 1960–1993 and she estimated static and dynamic models by use of the above mentioned estimation methods (except for the time series regression). The exogenous variables included in the *HE* models were: *GDP*, % of public finance of health expenditure, % of population above 65 years and the relative price of health care. The results indicated that the estimated mean group dynamic long-run income elasticity was significantly higher than unity and higher compared with Gerdtham's fixed-effects estimates. This appears to run against theory which predicts that the fixed-effects estimator should overestimate the long-run effects (see above). The estimated mean group short-run effect of income was -0.221 and similar to Gerdtham. The estimated mean group static long-run income elasticity was lower compared to the dynamic estimate but was still significantly higher than unity. Roberts further obtained a positive and significant long-run elasticity of public financing which suggested that a 10 percent increase in the fraction of public financing increases health expenditure by 7 percent. This effect was similar to Leu (1986) but differed from Gerdtham (1992) and Barros (1998). In accordance with Gerdtham (1992) and Barros (1998), the effects of aging of population was not significant. The relative price of health care was also not significant. Roberts concluded overall that the estimated average effects based on the dynamic and static fixed-effects estimator were similar to those derived from the mean group estimator. However the long-run mean group elasticities were extremely sensitive to exclusion of a time trend in the estimated relationship, i.e. the long-run income elasticity exceeded one if the time trend was included but was approximately one if the time trend was excluded. Roberts tested further several restrictions on the dynamic *HE* model and found, in accordance with Gerdtham, that the static model could be rejected (for all countries). She also found, like Gerdtham, that the long-run unitary income elasticity restriction could not be rejected (for any country). In contrast to Gerdtham, she rejected the joint hypothesis of a unitary long-run income elasticity and zero restrictions on the long-run elasticity of other explanatory variables (for nearly all countries).

3.2. Unit root and cointegration analyses

3.2.1. Methods

A further concern arises from the presence of several non-stationary variables in panel data regressions, e.g., *HE* and *GDP*. A major finding in econometrics during the past decade is that regressions involving non-stationary variables may lead to spurious results showing apparently significant relationships even if the variables are generated independently [Phillips (1986), Engle and Granger (1987)]. Non-stationarity in the data can arise from deterministic trends or stochastic trends in the data. One important difference between these kinds of non-stationarity is that variables with deterministic trends would be stationary after detrending (computing residuals from a regression on time) while variables with stochastic trends should be differenced to achieve stationarity; variables which are stationary after they have been differenced are said to be integrated of degree one $I(1)$, i.e. they contain one unit root. This implies that it may be important to discriminate between deterministic and stochastic trends before proceeding with estimation to avoid misleading inferences, and a number of alternative tests are available for testing [see Perman (1991)].

The standard test for non-stationarity of an observed time-series $\{y_t\}$ observed over T time periods is to estimate an augmented Dickey–Fuller regression (here including a time trend):

$$\Delta y_t = \alpha + \delta t + \beta y_{t-1} + \sum_{j=1}^{p} \rho_j \Delta y_{t-j} + \varepsilon_t, \quad t = 1, \ldots, T, \tag{3.2}$$

where $\Delta y_t = y_t - y_{t-1}$. The number of included lags p should be large enough to make the residuals serially uncorrelated. The unit root null hypothesis $H_0 : \beta = 0$, that the data generating process (DGP) for the series can be characterized as a non-stationary $I(1)$ process, is tested against the stationary alternative $H_1 : \beta < 0$ based on the t-statistic of the β estimate (see, e.g., Hamilton (1994) and Campbell and Perron (1991) for thorough treatments of the univariate unit root tests). An alternative test approach is suggested by Phillips (1987), Perron (1988) and Phillips and Perron (1989). This approach tests for unit root based on a non-parametric correction of the ordinary statistics obtained from a simple Dickey–Fuller regression without added lags of differenced variables as in the ADF regression. Asymptotically, the tests have the same limiting distribution. One argument for the use of the ADF tests is that several Monte Carlo simulation studies have found that the Phillips and Perron tests do not always have the correct size, even in fairly large samples, whereas the ADF tests in general are more robust. See, Banerjee et al. (1993) for discussion and references on this issue.

Im et al. (1997) proposed an approach to performing non-stationarity unit root tests for panel data of a sample of N cross-sectional units (industries, regions or countries) observed over T time periods $\{y_{it},\ i = 1, \dots, N,\ t = 1, \dots, T, \}$. These authors proposed that a panel unit root test can be based on the average of the N individual ADF t-statistics as:

$$\bar{t}_{NT} = \frac{1}{N} \sum_{i=1}^{N} t_{iT}(p_i), \tag{3.3}$$

where $t_{iT}(p_i)$ is the individual ADF t-statistic unit root test based on the inclusion of p_i lags in the individual ADF regression:

$$\Delta y_{it} = \alpha_i + \delta_i t + \beta_i y_{i,t-1} + \sum_{j=1}^{p_i} \rho_{ij} \Delta y_{i,t-j} + \varepsilon_{it}, \quad i = 1, \dots, N; t = 1, \dots, T. \tag{3.4}$$

The null hypothesis of unit roots for the panel unit root test is given by:

$$H_0 : \beta_i = 0 \quad \text{for all } i,$$

against the stationary alternative:

$$H_1 : \beta_i < 0, \quad i = 1, 2, \dots, N_1, \quad \beta_i = 0, \quad i = N_1 + 1, N_1 + 2, \dots, N.$$

This alternative allows β_i to differ between groups and only a fraction N_1/N of the individual series to be stationary. If the null hypothesis cannot be rejected it is concluded that the panel data series are $I(1)$, or difference stationary, around a linear trend. Im et al. showed that the proposed panel test (specified below) is consistent under the alternative hypothesis that the stationary fraction of the individual processes is non-zero. That is, as T and $N \to \infty$, the test is consistent as long as $\lim_{N\to\infty} N_1/N = \delta$, with $0 < \delta \leqslant 1$.

Assuming that the cross-sections are independent, Im et al. proposed using the following standardized t-bar statistic:

$$\Psi_{\bar{t}} = \frac{\sqrt{N}(\bar{t}_{NT} - E(\bar{t}_{NT}))}{\sqrt{\mathrm{Var}(\bar{t}_{NT})}}, \tag{3.5}$$

where

$$E(\bar{t}_{NT}) = \frac{1}{N}\sum_{i=1}^{N} E(t_{iT}(p_i)|\beta_i = 0), \quad \text{and}$$

$$\mathrm{Var}(\bar{t}_{NT}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}(t_{iT}(p_i)|\beta_i = 0),$$

assuming the individual ADF tests $t_{iT}(p_i)$ (estimated with p_i lagged differences) are independent. The means $E(t_{iT}(p_i)|\beta_i = 0)$ and variances $\mathrm{Var}(t_{iT}(p_i)|\beta_i = 0)$ of the individual ADF t-statistics can be obtained from Monte Carlo simulations. Im et al. provided the relevant mean and variance for a selection of sample size and individual lag-structures, for models including an intercept and an intercept and time trend, respectively. The authors conjecture that the standardized $\Psi_{\bar{t}}$ statistic converges weakly to a standard normal distribution. Hence the panel data unit root inference can be conducted by comparing the obtained $\Psi_{\bar{t}}$ statistic with critical values from an $N(0, 1)$ distribution.[14]

It might further be the case that the failure to reject the unit root null hypotheses for the variables is due to the fact that the variables can be characterized by a higher order of non-stationarity, i.e. the series might need to be differenced more than once to attain stationarity. This means that one should also apply unit root tests on the differenced variables, i.e. $I(2)$ hypothesis tests. If this hypothesis can be rejected, then one may conclude that the variables are $I(1)$.

If the non-stationarity tests fail to be rejected, then one can difference the variables to achieve stationarity as in the Box–Jenkins methodology and estimate the relevant coefficients using only differenced variables. While this is acceptable, differencing may result in a loss of information concerning long-run relationships between variables. One way out of this is that non-stationary variables (which are integrated of the same order) may be cointegrated, i.e. that a linear combination of non-stationary variables is itself stationary. Granger (1981) has shown that if this set of variables are cointegrated the OLS

[14] Im et al. note that the test procedure is not longer applicable if the disturbances are correlated across groups (countries). To allow for the possibility of correlated errors in the case where the correlation arise as a result of a time-specific effect θ_t common to all countries, Im et al. propose to use a de-meaning procedure where the time-effect θ_t is removed by subtracting cross-section means from both sides of the ADF regression.

estimator will still produce consistent estimates and the OLS residuals of the possible cointegrating regression can be used to test for cointegration.

The most straightforward approach to test for cointegration is the two-step approach suggested by Engle and Granger (1987). Based on the static cointegration regression model with an intercept and a time trend:

$$y_t = \alpha + \delta t + x_t' \beta + \varepsilon_t. \tag{3.6}$$

The null hypothesis of no-cointegration is performed based on the ADF residual ADF regression:

$$\Delta \widehat{\varepsilon}_t = \rho \widehat{\varepsilon}_{t-1} + \sum_{j=1}^{p_i} \varphi_j \Delta \widehat{\varepsilon}_{t-j} + v_t, \tag{3.7}$$

where $\widehat{\varepsilon}_t$ are least squares residuals from (3.6). The null hypothesis $H_0 : \rho = 0$, that the data generating process for the residuals can be characterized as a non-stationary $I(1)$ process (and hence that the series y and x are not cointegrated), are tested against the stationary alternative $H_1 : \rho < 0$ based on the t-statistic of the ρ estimate. There are many alternative test approaches which can also be extended to panel data [see Johansen (1991), Levin and Lin (1993), McCoskey and Kao (1997)].

3.2.2. Empirical results

Four recent articles have focused on non-stationarity and cointegration of *HE* and *GDP* and reached partly different conclusions:

(1) Hansen and King (1996) used data for 20 OECD countries for the period 1960–1987 and presented individual country-by-country Augmented Dickey–Fuller (ADF) unit root and Engle–Granger cointegration tests. The unit root hypothesis could (generally) not be rejected for either *HE* or *GDP*. Nor could the hypothesis of no-cointegrating relationship among *HE*, *GDP* and a range of other variables (generally) be rejected. Hansen and King consequently suggest that panel data estimations of the *GDP*/*HE* relationship may be spurious;

(2) Blomqvist and Carter (1997) used data for 18 OECD countries for the period 1960–1991 and reported various unit root and cointegration test results of an *HE* model including *GDP*, an intercept and a time trend. In accordance with Hansen and King, the country-by-country results of Blomqvist and Carter, based on the Phillips and Perron (1989) test, reject the unit root hypothesis in only one case (Finland) for *GDP* and in no case for *HE*. Consequently Blomqvist and Carter proceeded by country residual-based cointegration tests based on static as well as dynamic models. The results differ from Hansen and King in that the null of no-cointegration is rejected for all countries by the Phillips and Perron test and the null of cointegration by the Shin (1994) test cannot be rejected for any country;

(3) Roberts (1998b) used data for 10 EC countries for the time period 1960–1993 and reported individual country-by-country and panel data unit root tests and individual country-by-country cointegration tests. In the ADF regression, she included an intercept and time trend and found that the null hypothesis of a unit root cannot be rejected for *HE* and *GDP* among a number of other variables. Roberts reported contradictory results based on various country-by-country cointegration tests: Engle–Granger and Pesaran et al. (1996) tests failed to reject the null of no-cointegration but the Johansen (1991) test provided evidence for the existence of at least one cointegrating vector for most of the countries studied;

(4) McCoskey and Selden (1998) used the same data as Hansen and King and applied the recently developed heterogeneous panel unit root test by Im et al. (1997). In contrast to both Hansen and King, Blomqvist and Carter, and Roberts, McCoskey and Selden (1998) rejected the null hypothesis of unit roots for both *HE* and *GDP* and suggested "that researchers studying national health care expenditure need not be as concerned as previously thought about the presence of unit roots in the data" (McCoskey and Selden, p. 8).

McCoskey and Selden declared that it is not surprising that Hansen and King could not reject the unit root hypothesis, since they relied on low-powered country-by-country unit root tests (augmented Dickey–Fuller tests; ADF). To reduce this problem, McCoskey and Selden employed the recently developed panel unit root test by Im et al. (1997). However, Blomqvist and Carter and Roberts also used panel unit root tests by Levin and Lin (1993) and the same panel test as McCoskey and Selden, respectively, and none of them could reject the unit root hypothesis. So it is not likely that it is the low statistical power of the country-by-country tests that principally explain the conflicting results between Hansen and King and McCoskey and Selden. Apart from the new panel test, McCoskey and Selden argued that the ADF regression should not include time trends, despite that both *HE* and *GDP* are trended and despite the fact that McCoskey and Selden (footnote 9) were conscious of the fact that their results and conclusions are conditioned upon whether time trends are included or not. McCoskey and Selden (p. 6) motivated the exclusion of the time trend by the argument that it is not needed since the intercept term by itself allows the series to drift over time. This indicate that the reason for the differing results on non-stationarity is the omission of the time trend in McCoskey and Selden. However, the cointegration results also differ between Hansen and King, Blomqvist and Carter and Roberts, i.e. Hansen and King says "no-cointegration", Blomqvist and Carter says "cointegration" and Roberts says "it depends…". One possible, and the most likely, explanation for the differing results is the fact that different methods for estimation of cointegration relationships are used; for example, Hansen and King and Roberts tests the null hypothesis of "no-cointegration" on static models while Blomqvist and Carter tests the same hypothesis on dynamic models, although it is still open to considerable question which test is the most reliable.

4. Summary and concluding remarks

This chapter has reviewed the literature on international comparisons of health expenditure which has attracted considerable interest inside and outside health economics during the last three decades. One reason for this interest is the large cross-country differences in health expenditure and the opportunity for analyzing institutional arrangements influencing the demand, funding and delivery of health services in different countries. It is intriguing to ask: Does the organization of the health care system have any impact on health expenditure, and in that case, how large is this impact? Over the years, international comparisons of health expenditure have also become an active research area with a growing range of participants. But, in spite of this intellectual activity, research is still in its infancy and has raised more questions than it has answered. However, this should not come as a surprise in view of the lack of theoretical guidance and the numerous data and measurement problems. Before we discuss some issues for future research we briefly summarize the empirical results, which are frequently contradictory and inconclusive.

4.1. Summary of empirical results

4.1.1. Estimated effects of explanatory variables

The more significant results on the estimated effects of various explanatory variables are summarized below. We begin with the effects of non-institutional variables, and then consider effects of institutional variables. With regard to the institutional variables, we focus mainly on the results of Gerdtham et al. (1998); they tested an extensive range of explanatory variables on health expenditure which relate to different health system characteristics. One should bear in mind, however, that many of the variables tested in that paper have only been tested once and in one data set, which means that it is important to validate these results with updated data and methods. Moreover, many of the variables used in the estimation, such as those representing the public fraction in health care financing, overbilling, and the use of high cost procedures, are at best rough approximations of the underlying influences of interest. The distinctions between institutional arrangements of different countries are not usually as simple and clear-cut as implied by the use of dummy variables. Thus all results at this stage must be treated with considerable caution.

Non-institutional variables

- A common and extremely robust result of international comparisons is that the effect of per capita GDP (income) on expenditure is clearly positive and significant and, further, that the estimated income elasticity is clearly higher than zero and close to unity or even higher than unity. This result appears to be robust to the choice of variables included in the estimated models, data, the choice of conversion factors and methods of estimation. However, it is still unclear if the income elasticity is unity or higher. Most recent studies indicate that the income elasticity is about one, with the exception of Roberts (1998a).

- The effects of population age structure and unemployment rate are usually insignificant. Age of population is included in almost every estimation, while the unemployment rate is included in only a few studies. The same insignificant result usually also holds for female labor force participation, which has been used as a measure of the substitution of informal care [see Gerdtham et al. (1992a, 1992b, 1998)]. Gerdtham et al. (1998) also found that the effect of per capita tobacco consumption on health expenditure was positive and significant.

Institutional variables

Gerdtham et al. (1998) tested a number of variables related to different OECD countries' health systems, and six results appeared to be reasonably strong and in the "expected" direction.

First, the use of primary care "gatekeepers" seems to result in lower health expenditure. The estimated coefficient suggests that countries with gatekeepers have expenditure which is about 18 percent lower than those without gatekeeper. It should be noted, however, that the effect of gatekeeper was not significant on decade health expenditure growth rates as in Barros (1998).

Second, significantly lower levels of health expenditure appear to occur in systems where the patient first pays the provider and then seeks reimbursement, compared to other systems. The expenditure was about 9 percent lower in systems with patient reimbursement.

Third, the method of remunerating physicians in the ambulatory care sector appears to influence health expenditure. Capitation systems tend to lead to lower expenditure on average than fee-for service systems by around 17 to 21 percent.

Fourth, there are indications that in-patient care is more expensive than ambulatory care. The ratio of in-patient expenditure to total health expenditure is positively related to health expenditure.

Fifth, there is some evidence that public sector provision of health services (proxied by the ratio of public beds to total beds) is associated with lower health expenditure. This result was inconsistent with Leu (1986). The validity of this result was tempered by the poor quality of this proxy, given the fact that many "private" beds are in the voluntary sector, are quasi-integrated into the public sector, or face fixed reimbursement rates.

Finally, the total supply of doctors may have a positive effect on health expenditure, and this also appeared to be the case for countries where doctors practice under fee-for-service arrangements. However, this needs to be balanced against the result of a negative effect of the supply of doctors.

Turning to results which differed from expectations, there were indications that budget ceilings on in-patient care are associated with a higher health expenditure. One possible explanation for this result is that it reflects reverse causality: countries with high expenditure may be more inclined to introduce budget ceilings or to introduce them earlier. Of course other explanatory variables may also be endogenous, and in particular other institutional variables. Newhouse (1977) suggested that the organizational form and financing of health care are endogenous and do not exert independent effects on health expenditure. He suggested that centralized control of, or influence on, health budgets is itself a response to low income and a desire to control costs.

Contrary to the evidence given by Hurst [OECD (1992)], health systems characterized by the public reimbursement approach appear to have a lower health expenditure on average than public contract systems, and expenditure in public integrated systems is broadly the same as in public contract arrangements. Further tests suggest that public integrated systems might be even more costly than public contract systems, possibly because countries in this group also tend to have higher fractions of high cost in-patient care and fewer gatekeeping arrangements. It remains difficult to explain the result for public reimbursement systems.

The conclusion drawn from this analysis was that the organization of ambulatory care – the first contact point with the health system for most people – appears to be of particular importance for the containment of health expenditure. This conclusion was suggested by the relatively robust findings regarding the gatekeeper role, capitation-based remuneration systems in ambulatory care, and up-front payments by patients (which may later be reimbursed by insurers). As noted above, the findings above need to be interpreted cautiously and should be validated in future studies. Nevertheless, building in better incentives for doctors and patients at the ambulatory care level, through these and other methods, may help to counter some of the influences which tend to expand the supply and demand for health services – particularly as a result of asymmetric information in the doctor-patient relationship and moral hazard arising from health insurance coverage. Furthermore, the suggested importance of improving incentives at a microeconomic level stands in contrast with the relatively weak, and at times unexpected, results found for measures reflecting broader system characteristics and restraints (such as the use of integrated, contract, or reimbursement approaches, and the effects of overall budget caps).

4.1.2. Unit root and cointegration analysis

The results of unit root and cointegration analysis can be summarized as follows:

(1) Hansen and King (1996) performed individual country-by-country unit root and cointegration tests. Their results did not (generally) reject the unit root hypothesis for *HE* or *GDP*. Nor could they reject the hypothesis that *HE* and *GDP* are not cointegrated, which provides no support for the existence of equilibrium cointegrating relationships between *HE* and *GDP*. Hansen and King consequently suggested that panel data estimations of the GDP/HE relationship may be spurious;

(2) Blomqvist and Carter (1997) performed individual country-by-country and panel data unit root tests and individual country-by-country cointegration tests. They reached the same results as Hansen and King concerning the unit root hypothesis but they rejected the no-cointegration hypothesis;

(3) Roberts (1998b) performed individual country-by-country and panel data unit root tests and individual country-by-country cointegration tests and found clear evidence of non-stationarity (unit roots) in the series. The evidence regarding the existence of equilibrium cointegrating relationships between *HE* and *GDP* is not conclusive.

(4) In opposition to Hansen and King, Blomqvist and Carter and Roberts, McCoskey and Selden (1998) rejected the null hypothesis of unit roots for both *HE* and *GDP* based on individual country-by-country and panel unit root tests.

The conflicting results regarding unit root tests are principally due to the fact that the recent work by McCoskey and Selden omits the time trends in the ADF regressions while Hansen and King, Blomqvist and Carter and Roberts include time trends in the unit root tests. Hansen and King (1998) argued that the omitted time trends raise doubts about the validity of the results by McCoskey and Selden, since both health expenditure and *GDP* are clearly trended. If, and only if, this argument is valid then we can conclude that both health expenditure and *GDP* are non-stationary since all three studies which have included a time trend in the ADF regression could not reject the unit root hypothesis. However, cointegration results in previous studies appear also confused. The most likely explanation for the differing results is difference in methods, and it is an open question which test is most reliable.

4.2. Issues for the future

Some issues for future research are listed below:

- We need more theory of the macroeconomics of health expenditure, at least relative to the macroeconometrics of health expenditure. Some approaches have been suggested previously in the literature. McGuire et al. (1993) and Roberts (1998a) have suggested that future work on aggregate health expenditure can build on Dunne et al. (1984) and Selvanathan and Selvanathan (1993), which analyze health expenditure within a macroeconomic framework of expenditure.
- Empirical studies in recent years have been remarkably unwilling to test “new” variables as regressors in their models. One possible such “new” candidate is government budget deficits, which is likely to be a strong constraint on public health expenditure [Jönsson (1996)]. Another possible candidate is tax subsidy of private health insurance, which may be expected to increase health expenditure since a higher tax subsidy reduces the relative price of health insurance which in turn increases insurance coverage and health expenditure [Pauly (1986)]. It may also be important to replicate Gerdtham et al. (1998) with extended data sets and also with respect to growth rates of health expenditure as in Barros (1998), and new methods as in Roberts (1998a).
- It is important to continue the analysis of the role and effects of relative health care prices on health expenditure and the effects of income on the relative prices since the existing literature on this subject is ambiguous. It may further be interesting to discuss effects of institutional factors on quantity of health care and the relative price of health care, separately, and eventually also investigate possible endogeneity between health care prices and quantity. It may, for example, be the case that an increase in the number of physicians drives down salary levels, which implies that countries with more quantity may have lower relative health care prices.
- One final issue that merits further attention is the inconclusive results in the area of testing for non-stationarity and cointegration of health expenditure relationships.

Why do the results differ between different studies and what conclusions can be drawn? The tests need also to be developed further to account for possible heteroscedasticity across time and dependent t-statistics across countries. Yet another important extension is to consider tests for the validity of the various forms of homogeneous model restrictions commonly imposed in studies of determinants of health expenditures [see, e.g., Gerdtham (1992), Hitiris and Posnett (1992), Viscusi (1994a), Gerdtham et al. (1998)].

References

Abel-Smith, B. (1967), An International Study of Health Expenditure (WHO, Geneva).

Andersen, R., and L. Benham (1970), “Factors affecting the relationship between family income and medical care consumption”, in: H. Klarman, ed., Empirical Studies in Health Economics (John Hopkins University Press, Baltimore and London).

Baltagi, B.H. (1995), Econometric Analysis of Panel Data (Wiley, Chichester).

Banerjee, A., J. Dolado, J.W. Galbraith and D.F. Hendry (1993), Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data (Oxford University Press, New York).

Barr, N. (1992), “Economic theory and the welfare state: a survey and interpretation”, Journal of Economic Literature 2:741–803.

Barros, P.P. (1998), “The black-box of health care expenditure growth determinants”, Health Economics 7:533–544.

Berk, M.L. (1992), “The concentration of health expenditures: an update”, Health Affairs 11:145–149.

Besley, T., and M. Goevia (1994), “Alternative systems of health care”, Economic Policy 19:200–258.

Bird, R.M. (1970), “The growth of government spending in Canada”, Canadian Tax Paper, no. 51 (Canadian Tax Foundation, Toronto).

Blomqvist, Å., and R.A.L. Carter (1997), “Is health care really a luxury?”, Journal of Health Economics 16:207–229.

Box, G.E.P., and D.R. Cox (1964), “An analysis of transformations”, Journal of the Royal Statistical Society 26(ser B):211–243.

Breusch, T.S., and A.R. Pagan (1980), “The Lagrange multiplier test and its application to model specification in econometrics”, Review of Economic Studies 47:239–254.

Buchanan, J.M. (1965), The Inconsistencies of the National Health Service (Institute of Economic Affairs, London).

Campbell, J.Y., and P. Perron (1991), “Pitfalls and opportunities: what macroeconomists should know about unit roots”, National Bureau of Economics Research, Macroeconomics Annual.

Carlsen, F., and J. Grytten (1998), “More physicians: improved availability or induced demand”, Health Economics 7:495–508.

Cromwell, J., and J.B. Mitchell (1986), “Physician-induced demand for surgery”, Journal of Health Economics 5:293–313.

Culyer, A.J. (1988), “Health expenditures in Canada: myth and reality; past and future”, Canadian Tax Paper, no. 82 (Canadian Tax Foundation, Toronto).

Culyer, A.J. (1989), “Cost containment in Europe”, Health Care Financing Review (Annual Supplement) 21–32.

Dunne, J.P., P. Parshardes and R. Smith (1984), “Needs costs and bureaucracy: the allocation of public consumption in the UK”, Economic Journal 94:1–15.

Enggard, K. (1986), “Sundhedssektoren i Danmark er ikke billigere”, Indenrigsministeriet 2 Okt. 1986.

Engle, R., and C. Granger (1987), “Co-integration and error correction: Representation, Estimation and Testing”, Econometrica 35:251–276.

Engle, R.F., and C.W.J. Granger (1991), Long-Run Economic Relationships: Readings in Cointegration (Oxford University Press, New York).

Evans, R.G. (1974), "Supplier-induced demand: some empirical evidence and implications", in: M. Perlman, ed., The Economics of Health and Medical Care (Macmillan, London, John Wiley, New York).

Fuchs, V. (1972), "The basic forces influencing costs of medical care", in: V. Fuchs, ed., Essays in the Economics of Health and Medical Care (Columbia University Press, New York and London).

Gbsemete, K., and U.-G. Gerdtham (1992), "The determinants of health expenditure in Africa: a cross-sectional study", World Development 20:303–308.

Gerdtham, U.-G. (1992), "Pooling international health expenditure data", Health Economics 1:217–231.

Gerdtham, U.-G., and B. Jönsson (1991a), "Health care expenditure in Sweden, an international comparison", Health Policy 19:211–228.

Gerdtham, U.-G., and B. Jönsson (1991b), "Conversion factor instability in international comparisons of health care expenditure", Journal of Health Economics 10:227–234.

Gerdtham, U.-G., and B. Jönsson (1991c), "Price and quantity in international comparisons of health care expenditure", Applied Economics 23:1519–1528.

Gerdtham, U.-G., and B. Jönsson (1992), "International comparisons of health care expenditure: conversion factor instability, heteroscedasticity, outliers and robust estimators", Journal of Health Economics 11:189–197.

Gerdtham, U.-G., and B. Jönsson (1994), "Health care expenditure in the Nordic countries", Health Policy 26:207–220.

Gerdtham, U.-G., J. Sögaard, F. Andersson and B. Jönsson (1988), "Econometric analyses of health care expenditures: a cross-section study of the OECD countries", Center for Medical Technology Assessment (CMT) 1988:9, University of Linköping, Sweden.

Gerdtham, U.-G., J. Sögaard, F. Andersson and B. Jönsson (1992a), "Econometric analysis of health expenditure: a cross-sectional study of the OECD countries", Journal of Health Economics 11:63–84.

Gerdtham, U.-G., J. Sögaard, B. Jönsson and F. Andersson (1992b), "A pooled cross-section analysis of the health expenditure of the OECD countries", in: P. Zweifel and H. Frech, eds., Health Economics Worldwide (Kluwer Academic Publishers, Dordrecht).

Gerdtham, U.-G., B. Jönsson, M. MacFarlan and H. Oxley (1998), "The determinants of health expenditure in the OECD countries", in: P. Zweifel, ed., Health, The Medical Profession, and Regulation (Kluwer Academic Publishers, Dordrecht). See also "Factors affecting health spending: a cross-country econometric analysis", in OECD (1995b) New Directions in Health Care Policies: Improving Cost Control and Effectiveness.

Getzen, T.E., and J.P. Poullier (1992), "International health spending forecasts: concepts and evaluation", Social Science and Medicine 34:1057–1068.

Granger, C.W. (1981), "Some properties of time series data and their use in econometric model specifications", Journal of Econometrics 16:121–130.

Greene, W. (1990), Econometric Analysis (Macmillan Publishing Company, New York).

Greene, W. (1993), Econometric Analysis, 2nd edn. (Macmillan Publishing Company, New York).

Grossman, M. (1972), The Demand for Health: a Theoretical and Empirical Investigation (Columbia University Press, New York).

Ham, C., R. Robinson and M. Benzeval (1990), Health Check: Health Care Reforms in an International Context (The King's Fund Institute, London).

Hamilton, J.D. (1994), Time Series Analysis (Princeton University Press, Princeton, NJ).

Hansen, P., and A. King (1996), "The determinants of health care expenditure: a cointegration approach", Journal of Health Economics 15:127–137.

Hansen, P., and A. King (1998), "Health care expenditure and GDP: panel data unit root test results – comment", Journal of Health Economics 17:377–381.

Hausman, J. (1978), "Specification tests in econometrics", Econometrica 46:1251–1271.

Hitiris, T., and J. Posnett (1992), "The determinants and effects of health expenditure in developed countries", Journal of Health Economics 11:173–181.

Hsiao, C. (1986), Analysis of Panel Data (Cambridge University Press, Cambridge).

Im, K.S., M.H. Pesaran and Y. Shin (1997), "Testing for unit roots in heterogeneous panels", mimeo (Department of Applied Economics, University of Cambridge).

Johansen, S. (1991), "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models", Econometrica 59:1551–1580.

Jönsson, B. (1996), "Comprende la réforme des systémes de santé" (Making sense of health care reform), in: OECD, La Réforme des Systémes de Santé- La Volunté de Changement, Études de Politique de Santé (OECD Health Policy Studies 8, OECD).

Jönsson, B., and P. Musgrove (1997), "Government financing of health care", in: G.J. Schieber, ed., Innovations in Health Care Financing, Proceedings of a World Bank Conference (March 10–11) 41–64.

Keeney, R.L. (1997), "Estimating fatalities induced by the economic costs of regulation", Journal of Risk and Uncertainty 14:5–23.

Kendix, M., and T.E. Getzen (1994), "US health services employment: a time series analysis", Health Economics 3:169–181.

Kleiman, E. (1974), "The determinants of national outlay on health", in: M. Perlman, ed., The Economics of Health and Medical Care (Macmillan, London and Basingstoke).

Leu, R.E. (1986), "The public-private mix and international health care costs", in: A.J. Culyer and B. Jönsson, eds., Public and Private Health Services (Basil Blackwell, Oxford).

Leu, R.E., and T. Schaub (1983), "Does smoking increase medical care expenditures?", Social Science of Medicine 17:1907–1914.

Leviatan, L.I. (1964), "Consumption patterns in Israel, Jerusalem", Falk Project for Economic Research in Israel.

Levin, A., and C.-F. Lin (1993), "Unit root test in panel data: new results", mimeo (University of California, San Diego).

Lewis, M. (1994), The Organization, Delivery and Financing of Health Care in Brazil: Agenda for the 90s (World Bank, Latin America and the Caribbean Department 1, Human Resources Division, Washington, DC).

Manning, W.J., J. Newhouse, N. Duan, E. Keeler, A. Leibowitz and M. Marquis (1987), "Health insurance and the demand for medical care: evidence from a randomised experiment", American Economic Review 773(June).

Maxwell, R.J. (1981), Health and Wealth (Lexington books, Lexington).

McCallum, B.T. (1972), "Relative asymptotic bias from errors of omission and measurement", Econometrica 40:757–758.

McCoskey, S., and C. Kao (1997), "A residual-based test of the null of cointegration in panel data", Econometric Reviews, forthcoming.

McCoskey, S.K., and T.M. Selden (1998), "Health care expenditure and GDP: Panel data unit root test results", Journal of Health Economics 17:369–376.

McGuire, A., J. Henderson and G. Mooney (1988), The Economics of Health Care: an Introductory Text (Routledge and Keegan, London).

McGuire, A., D. Parkin, D. Hughes and K. Gerard (1993), "Econometric analyses of national health expenditures: can positive economics help to answer normative questions?", Health Economics 2:113–126.

Milne, R., and H. Molana (1991), "On the effect of income and relative price on demand for health care: EC evidence", Applied Economics 23:1221–1226.

Murthy, V.N.R. (1992), "Conversion factor instability in international comparisons of health care expenditure: some econometric comments", Journal of Health Economics 11:183–187.

Muurinen, J.M. (1982), "Demand for health: a generalised Grossman model", Journal of Health Economics 1:5–28.

Newhouse, J.P. (1977), "Medical care expenditure: a cross-national survey", Journal of Human Resources 12:115–125.

Newhouse, J.P. (1987), "Cross-national differences in health spending: what do they mean?", Journal of Health Economics 6:159–162.

Newhouse, J.P. (1992), "Medical care costs: how much welfare loss?", Journal of Economic Perspectives 6:3–21.

Newhouse, J.P., and C.E. Phelps (1974), "Price and income elasticities for medical care services", in: M. Perlman, ed., The Economics of Health and Medical Care (Macmillan, London and Basingstoke).

OECD (1987), "Financing and delivering health care: a comparative analysis of OECD countries", OECD Social Policy, No. 4 (OECD, Paris).

OECD (1992), "The reform of health care: a comparative analysis of seven OECD countries", Health Policy Studies No. 2 (OECD, Paris).

OECD (1994), "The reform of health care systems: a review of seventeen OECD countries", Health Policy Studies, No. 5 (OECD, Paris).

OECD (1995a), "Internal markets in the making", Health Policy Studies, No. 6 (OECD, Paris).

OECD (1995b), "New directions in health policy", Health Policy Studies, No. 7 (OECD, Paris).

OECD (1996), "Health care reform: the will to change", Health Policy Studies, No. 8 (OECD, Paris).

OECD (1998), OECD Health Data, A software package for the international comparison of health care system, User's Manual, Version x.06, Paris.

Okunade, A.A. (1985), "Engel curves for developing nations: the case of Africa", Eastern Africa Economic Review 1:13–22.

Parkin, D., A. McGuire and B. Yule (1987), "Aggregate health expenditures and national income: is health care a luxury good?", Journal of Health Economics 6:109–127.

Parkin, D., A. McGuire and B. Yule (1989), "What do international comparisons of health expenditures really show?", Community Medicine 11:116–123.

Pauly, M. (1968), "The economics of moral hazard", American Economic Review 58:531–537.

Pauly, M. (1986), "Taxation, health insurance, and market failure in the medical economy", Journal of Economic Literature (June):629–675.

Perman, R. (1991), "Cointegration: an introduction to the literature", Journal of Economic Studies 18:3–30.

Perron, P. (1988), "Trend and random walks in macroeconomic timeseries: further evidence from a new approach", Journal of Economic Dynamics and Control 12:297–332.

Pesaran, M.H., and R. Smith (1995), "Estimating long-run relationships from dynamic heterogeneous panels", Journal of Econometrics 68:79–113.

Pesaran, M.H., Y. Shin and R. Smith (1996), "Testing for the existence of a long-run relationship", DAE Working Papers 9622 (University of Cambridge).

Pesaran, M.H., R. Smith and K.S. Im (1995), "Dynamic linear models for heterogeneous panels", in: L. Matyas and P. Sevestre, eds., The Econometrics of Panel Data: Handbook of Theory With Applications, 2nd edn. (Kluwer Academic, Amsterdam).

Pfaff, M. (1990), "Differences in health care spending across countries: statistical evidence", Journal of Politics and Law 15:1–67.

Phillips, P.C.B. (1986), "Understanding spurious regressions in econometrics", Journal of Econometrics 33:311–340.

Phillips, P.C.B. (1987), "Time series regression with a unit root", Econometrics 55:277–301.

Phillips, P.C.B., and P. Perron (1989), "Testing for a unit root in time series regression", Biometrika 75:335–346.

Poullier, J.-P. (1989), "Health data file: overview and methodology", Health Care Financing Review (Annual Supplement):111–194.

Rice, T. (1983), "The impact of changing medicare reimbursement rates on physician-induced demand", Medical Care 21:803–815.

Roberts, J. (1998a), "Sensitivity of elasticity estimates for OECD health care spending: analysis of a dynamic heterogeneous data field", Paper prepared for the Seventh European Workshop of Econometrics and Health Economics, STAKES, Helsinki, Finland, 9–12 September 1998.

Roberts, J. (1998b), "Spurious regression problems in the determinants of health care expenditure: a comment on Hitiris", Applied Economics Letters, forthcoming.

Sahn, D.E. (1992), "Public expenditures in sub-saharan Africa during a period of economic reforms", World Development 20:673–693.

Selvanathan, S., and E.A. Selvanathan (1993), "A cross country analysis of consumption patterns", Applied Economics 25:1245–1259.

Shin, Y. (1994), "A residual-based test of the null of cointegration against the alternative of no cointegration", Econometric Theory 10:91–115.

Spitzer, J.J. (1982), "A primer on Box–Cox estimation", Review of Economics and Statistics 64:307–313.

Stoddart, G.L., and R.J. Labelle (1985), Privatisation in the Canadian Health Care System (Ministry of Supply and Services, Ottawa).

Ståhl, I. (1986), Can Health Care Costs be Controlled? (Universitetsförlaget, Lund University, Sweden).

van de Ven, W.P.M.M., F.T. Schut and F. Rutten (1994), "Forming and reforming the market for third party purchasing of health care", Social Science and Medicine 39:1405–1412.

Viscusi, W.K. (1994a), "Risk-risk analysis", Journal of Risk and Uncertainty 8:5–17.

Viscusi, W.K. (1994b), "Mortality effects of regulatory costs and policy evaluation criteria", Rand Journal of Economics 25:94–109.

Wagstaff, A. (1986), "The demand for health: some new empirical evidence", Journal of Health Economics 5:195–233.

Weisbord, B.A. (1991), "The health care quadrilemma: an essay of technological change, insurance, quality of care, and cost-containment", Journal of Economic Literature 29:523–552.

Zarmebka, P. (1974), "Transformation of variables in econometrics", in: P. Zarembka, ed., Frontiers in Econometrics (Academic Press, New York).

Zeckhauser, R. (1970), "Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives", Journal of Economic Theory 2:10–26.

This Page Intentionally Left Blank

Chapter 2

AN OVERVIEW OF THE NORMATIVE ECONOMICS OF THE HEALTH SECTOR*

JEREMIAH HURLEY

McMaster University, Hamilton, ON, Canada

Contents

*Helpful comments were received from R. Cookson, A.J. Culyer, J.P. Newhouse, T. Rice, T. Scott and G. Stoddart. I would like to acknowledge financial support of the Ontario Ministry of Health to the Centre for Health Economics and Policy Analysis.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This chapter provides an overview of normative analysis in the health sector in recent decades. It surveys two distinct, but related, literatures. The first is normative analysis of the operation of health care and health care insurance markets, market failure, and the scope for non-market institutional arrangements to improve the efficiency and equity of the financing, funding, organization and delivery of health care. The second is the debate about the most appropriate normative framework within which to carry out normative analysis in the health sector, focusing on the welfarist and extra-welfarist frameworks. This is a debate about assumptions and methods. Although the rival frameworks share the broad conclusion that market failure pervades the health sector, the diagnoses regarding nature of that failure sometimes differ and, more importantly, the prescriptions to improve efficiency and equity often differ. Because it is not always clear what writers mean by "welfare economics" and "extra-welfarism," I briefly summarize key concepts of efficiency and key assumptions and elements of each framework. The three subsequent sections then analyze the nature of health care as an economic commodity and the implications of these characteristics both for the operation of health care and health care insurance markets and for the methods of normative economic analysis. Section 4 surveys prominent approaches to analyzing equity in health care. Section 5 examines the methods of normative analysis as applied to evaluate individual health care services. Finally, I end with some observations on recent discussions of the role of normative economic analysis in policy making and of health economists as policy advisors.

Keywords

welfare economics, extra-welfarism, health care efficiency, health care equity,
economic evaluation

JEL classification: I10, I11, J18, D6 (D61, D62, D63)

1. Introduction

As the sub-title of a recent book on health economics underscored,[1] the hand (or more correctly, “the hands”) that guides resource allocation in the health sector is neither hard to see nor necessarily as magical as the metaphorical invisible hand of Adam Smith. In virtually every advanced economy the majority of health care expenditures is financed from the public purse either explicitly or through tax expenditures such as the exemption of health care benefits from taxable income [OECD (1998)]. Regulations and other non-market institutions play a major role in guiding behaviour and the attendant resource allocations in the private sector. If the role for the invisible hand working through markets is in practice quite circumscribed in the health sector, how do we know what institutional designs will produce an efficient and equitable allocation of health care resources? It is one thing to demonstrate what does not work; it is quite another to demonstrate what will. Normative economics is precisely about attempting to rank, from better to worse from an economic perspective, resource allocations and the policies that generate them. Ranking a policy requires positive analysis that correctly describes the effect that the policy will have on resource allocation and ethical criteria regarding what constitutes a “better” resource allocation. The market failures that pervade the health care sector create an important role both for non-market institutional arrangements (i.e., the visible hand) and for normative economic analysis to help sort “good” policies from “bad.”

Historically, much of this normative work in the health sector has been done within the neo-classical tradition and specifically within Paretian welfare economics. Health economics, however, is increasingly divided over the most appropriate framework for normative analysis in the health sector. In the introduction to his now classic text on welfare economics de V. Graaff observed that:

> “...whereas the normal way of testing a theory in positive economics is to test its conclusions, the normal way of testing a welfare proposition is to test its assumptions... the interest attaching to a theory of welfare depends almost entirely upon the realism and relevance of its assumptions, factual and ethical, in a particular historical context” [de V. Graaff (1967, p. 3)].

This observation captures well the central features of the recent debate regarding the normative economics of health care, a debate rooted in different views regarding the “relevance and realism” of various factual and ethical assumptions, particularly those associated with neo-classical welfare theory.

The conventional assumptions of welfare economics are challenged in the health sector for many reasons. Many health economists argue that some of the characteristics of health care as a commodity that cause markets to fail also cause aspects of conventional neo-classical economic methods (both positive and normative) to “fail” [Evans (1984),

[1] Donaldson, C., and K. Gerard (1993), The Economics of Health Care Financing: The Visible Hand (Macmillian, London).

Culyer (1989), Rice (1992, 1998)]. Informational asymmetries in the market for health care, it is argued, jeopardize the status of the standard neo-classical demand curve as a normative construct. Further, unless the potential for supply-side demand inducement is incorporated into models, even positive predictions based on demand analysis can be inaccurate. The applied, interdisciplinary nature of much health economic analysis exposes health economists to non-economist researchers who challenge the realism and relevance of the assumptions and methods of mainstream economics. It is interesting in this respect that a survey of health economists found that those based in economics departments were more likely to believe in the applicability of standard neo-classical models for health sector than those based on other settings [Feldman and Morrisey (1990)].[2] Self-selection may also be at work. Until recently health economics has been a relatively small sub-discipline that has perhaps attracted those intrigued by an area that poses so many challenges to traditional analytic economic approaches.[3] Regardless of the specific reasons, while Paretian welfare economics has proven invaluable in understanding the nature of market failure in the health care sector, it has proven a less durable basis for normative economic analysis of alternative policy responses to these failures. The health sector has proven fertile ground for extra-welfarist approaches that emphasize non-utility, and in particular health information, in evaluating resource allocations. Extra-welfarist methods for many years have largely supplanted welfare-economic approaches to the economic evaluation of individual health care services and procedures.

This chapter provides an overview of normative analysis in the health sector in recent decades, dating roughly since Arrow's seminal article on the economics of medical care [Arrow (1963)]. The chapter surveys major developments in normative analysis during this period, identifying key issues and the relationship among rival approaches, emphasizing both points of shared understanding and points of debate. Health economics has developed remarkably as a specialty area of economics during this period (witness this *Handbook*) and normative analysis has been a particularly vigorous area of inquiry in recent years. Even with a generous space allocation, it has been necessary to treat only selected material and to treat it often at a fairly general level.

The overview attempts to survey two distinct, but related, literatures. The first is normative analysis of the operation of health care and health care insurance markets, market failure, and the scope for non-market institutional arrangements to improve the efficiency and equity of the financing, funding, organization and delivery of health care. As noted, historically much of this analysis has been conducted within the Paretian welfare economics [Culyer (1989)]. The conclusion that the characteristics of health care as a commodity lead to pervasive market failure, however, is shared by adherents to both Paretian welfare economics and alternative frameworks. Regardless of whether it

[2] This observation is based on the responses to questions pertaining to demand inducement and the applicability of models of perfect competition.

[3] Self-selection may also explain why economists who subscribe more strongly to the applicability of the standard neo-classical model tend to be based in economics departments.

is utility or health in the objective function, the basic conclusion that a system of private markets leads to an inefficient allocation of resources remains intact.

The second literature surveyed is the debate about the most appropriate normative framework within which to carry out normative analysis in the health sector. This is a debate about assumptions and methods. Although the rival frameworks share the broad conclusion that market failure pervades the health sector, the diagnoses regarding nature of that failure sometimes differ and, more importantly, the prescriptions to improve efficiency and equity often differ. The literatures on the nature of market failure in the health sector and on the methods of normative analysis obviously intertwine, though for expositional clarity I try to keep them separate. My hope is that the overview will set a broad context for normative analysis within which to place later chapters that delve more deeply into specific issues.

The plan for the overview is as follows: because it is not always clear what writers mean by "welfare economics" and "extra-welfarism" and because much of the discussion makes little sense without an understanding of the central differences among them, in the next section I briefly summarize key concepts of efficiency and key assumptions and elements of each framework. The three subsequent sections then analyze the nature of health care as an economic commodity and the implications of these characteristics both for the operation of health care and health care insurance markets and for the methods of normative economic analysis. Section 4 surveys prominent approaches to analyzing equity in health care. Section 5 then shifts gears to examine the methods of normative analysis as applied to evaluate individual health care services, an area commonly referred to as the "methods of economic evaluation." Finally, I end with some observations on recent discussions of the role of normative economic analysis in policy making and of health economists as policy advisors.

A few caveats on language are required before I begin. I use the term "market failure" to refer in general to a situation in which a freely operating market results in an inefficient allocation of resources, where that inefficiency could be defined from the perspective of either welfare economic theory or extra-welfarism. "Welfare" refers to well-being as assessed specifically in utility terms; in contrast "well-being" is used more generally and can be assessed in terms other than utility. Where important, I have also tried to be specific about the relevant efficiency concepts. For each of these terms I have tried to be consistent, though I have undoubtedly slipped in places. I apologize in advance for any resulting confusion.

2. Efficiency and normative frameworks

Efficiency is a purely instrumental concept. It is meaningful to discuss the efficiency of a service, good or activity only if an explicit objective has been articulated against which efficiency can be assessed. Economists generally distinguish three concepts of efficiency. The first two concern supply-side efficiency. *Technical efficiency* is achieved when production is organized to minimize the inputs required to produce a given output.

This is a purely physical, engineering-based notion of efficiency that depends solely on the physical production function. Technical efficiency coincides with being on an isoquant (and so there are many technically efficient input combinations for a given production function). *Cost-effectiveness efficiency* is achieved when production is organized to minimize the cost of producing a given output.[4] It is determined by both the production function and prevailing input prices. It coincides with the tangency of the isoquant and the isocost line (and so, under standard convexity assumptions, in a given setting there is normally only one cost-effective input mix). The third efficiency concept, *allocative efficiency*, incorporates demand-side, or consumption factors: allocative efficiency is achieved when resources are produced and allocated so as to produce the "optimal" level of each output and to distribute the outputs in line with the value consumers place on them. Within allocative efficiency there exist alternative ways to define "optimal" and to assess "value." Within welfarist approaches value is assessed using utility; within extra-welfarism value is assessed using subjective health measures. And within either, optimality can be assessed using the Pareto criterion (i.e., an allocation of resources is allocatively efficient only if it is not possible to increase one person's utility (health) without decreasing another person's), or a maximization criterion (i.e., an allocation of resources is allocatively efficient if it maximizes the sum of utility (health)). There is a hierarchical relationship among the concepts – technical efficiency is a necessary condition for cost-effectiveness, and both technical efficiency and cost-effectiveness are necessary conditions for allocative efficiency.

2.1. Neo-classical welfare economic framework

Four tenets of neo-classical welfare economics are of particular importance for understanding the development of normative analysis in the health sector: utility maximization, individual sovereignty, consequentialism and welfarism. The first of these, utility maximization, is essentially a behavioral assumption; the latter three of these are normative assumptions regarding who is in the best position to judge welfare and the types of information relevant to judging the goodness of a resource allocation. Utility maximization holds that individuals choose rationally – that is, given a set of options, an individual can rank the options and choose the most preferred among them according to defined notions of consistency. Without consistency, one could infer little from observed behaviour. Individual sovereignty asserts that individuals are the best judges of their own welfare; that any assessment of individual welfare should be based on a person's own judgement. It rejects paternalism, the notion that a third party may know better than the individuals themselves what is best for them. Consequentialism holds that any action, choice or policy must be judged exclusively in terms of the resulting, or consequent, effects. Outcome, not process, matters. Welfarism is the proposition that the "goodness" of any situation (e.g., resource allocation) be judged solely on the basis

[4] This is sometimes referred to as "production efficiency."

of the utility levels attained by individuals in that situation. It excludes all non-utility aspects of the situation.

Group welfare is defined in terms of an individualistic social welfare function: overall welfare is a function only of the levels of welfare (utility) attained by members of the group. The social welfare function and the associated ranking criterion in classical [e.g., Mill (1994 [1848])] and early neo-classical [e.g., Marshall (1961)] welfare economics were utilitarian: utility was assumed to be cardinally measurable and interpersonally comparable, so the optimal policy was that which maximized the sum of utilities in the group. With the development of ordinal utility theory, which dropped the assumptions that utility was cardinally measurable and interpersonally comparable, so the criterion of maximizing the sum of utilities was replaced by the criterion of Pareto Optimality. A resource allocation is Pareto Optimal (i.e., allocatively efficient) if and only if it is impossible to increase one person's utility without simultaneously decreasing another's. Hence, although the basic nature of the social welfare function remained the same (individualistic, utility-based), the understanding of utility and the decision criterion for identifying the optimal allocation changed.

For applied welfare economics, this change came with a heavy price. For a given set of resources, each of many possible allocations of those resources can be Pareto Optimal; the Pareto criterion does not lead to a single, best allocation. The assumption that utility is not interpersonally comparable severely limits (and indeed effectively precludes) the analysis of distributional issues. Because nearly all policy changes make someone worse off, strict application of the Pareto criterion leads to policy paralysis.

In an effort to overcome this latter limitation, attention shifted to the criterion of a Potential Pareto Improvement.[5] A policy is said to produce a Potential Pareto Improvement if benefits that accrue to the gainers are sufficiently large to enable them (hypothetically) to compensate the losers, making the losers no worse off than they were before the policy, while still retaining some net benefit for gainers. Most applied, empirical welfare analysis is based on this criterion which harkens back to the utilitarian roots of welfare economics in which the goal is to maximize utility. In applied welfare economics, utility (benefit) is normally measured in a money metric. The measure of benefit is the area under the demand curve.[6] For valuation of non-marketed goods, which are common in the health sector, willingness-to-pay is often assessed using contingent valuation methods.

Two "fundamental" theorems of welfare economics have been influential in setting market allocation as the reference standard in normative economic analysis and in justifying a near exclusive focus on efficiency concerns over distributional equity. The first theorem states that the allocation of resources generated by a perfectly competitive market process is Pareto optimal (i.e., achieves all three levels of efficiency). The

[5] Also called the compensation test or the Kaldor–Hicks criterion in honor of the two individuals who first proposed variants of it [Hicks (1939, 1941), Kaldor (1939)].

[6] For Marshallian demand curves, the consumer and producer's surplus; for compensated demand curves, the compensating or equivalent variation.

second theorem states that any Pareto optimal allocation can be achieved through a perfectly competitive economy.[7] The theorems provide the rationale within welfare economics for taking a market allocation as the reference standard. The only efficiency rationale for non-market arrangements is market failure caused by the violation of one or more of the model's assumptions. The burden of proof falls on advocates for non-market institutional arrangements to show why a market will not produce an efficient allocation of resources.[8] Second, given the above noted problem of analyzing distributional issues under the assumption that utility is ordinally measurable and interpersonally non-comparable, the second theorem provides economists with a rationale for separating efficiency and distributional concerns (or, some would argue, for ignoring distributional concerns [Reinhardt (1998)]). Because any Pareto optimal allocation can be reached through a competitive market process given the right initial distribution of resources (income), economists have felt free to analyze only questions of efficiency, leaving questions of the right distribution of resources to the political process. In the absence of costless, lump-sum transfers (i.e., in the real world), however, efficiency and distributional concerns obviously cannot be separated [Reinhardt (1992)].

In summary, key elements of the welfare-economic framework include individual sovereignty, welfarism, willingness-to-pay as a monetary metric for utility in applied analysis, market allocation as a reference standard, and a separation of efficiency and equity with an almost exclusive focus on efficiency. This brief summary admittedly does not do full justice to the field of welfare economics[9] but rather it has highlighted some of the controversial aspects whose relevance and realism within the health sector have been questioned.

2.2. Critiques of welfare economics within the health sector and extra-welfarism

There is a wide variety of critiques of neo-classical welfare-economics, many of which have a long history in economics. This section cannot comprehensively discuss them, or even all those specific to the health sector. Rather, it tries to identify key concerns and alternative approaches prominent in the health sector.

Some health economists do not necessarily reject the philosophical bases of welfare-economic theory but believe that important assumptions of the model do not hold [Rice (1998)]. It is commonly held, for example, that the assumption of individual

[7] See standard texts on welfare economics for a full discussion of the theorems and their associated assumptions [Koopmans (1957), Bator (1957), Ng (1979), or Boadway and Bruce (1984)].

[8] Demonstrating that a market fails is not sufficient justification for government intervention. Government can also fail. In a second-best world, a non-optimal market allocation may be preferred to the best possible allocation under government intervention. The vast majority of welfare economics, however, has focused on problems of market failure. See Chalkley and Malcomson (2000) for a more extended discussion of government failure.

[9] For in-depth treatments see, for example, Bator (1957), de V. Graaff (1967), Ng (1979), or Boadway and Bruce (1984).

sovereignty is violated in the health sector. Consequently, measurement techniques such as willingness-to-pay, as represented by the area under a demand curve, lose their normative relevance [Evans (1984), Rice (1992, 1998)].

Others reject at a philosophical level important assumptions of the welfare-economic framework, such as the ethical proposition that the value of a health care service to an individual is accurately represented by the person's willingness to pay for the service [Williams (1981)]. Because health care is necessary at times for a person's very existence, its value, or benefit, should not be linked to the economic resources of an individual. This also negates the normative interpretation of the demand curve; indeed, it invalidates virtually any willingness-to-pay metric. Consequently, there has been a major effort in health economics to develop benefit measures that do not depend directly on a person's income and wealth.

More fundamentally, extra-welfarists argue that utility is not the only relevant argument, or indeed even the most important argument, in the social welfare function. They argue that health, not utility, is the most relevant outcome for conducting normative analysis in the health sector [Culyer (1989, 1990), Culyer and Evans (1996)].[10] Debate concerning the nature of the relevant objective function for normative economic analysis in the health sector, and in particular the place of health in that function, has erupted as a major point of controversy within the field in recent years [Labelle et al. (1994a, 1994b), Pauly (1994a, 1994b, 1996), Culyer and Evans (1996)], though its roots extend far back. As early as 1963, for instance, Feldstein asked "... should not health care be allocated to maximize the level of health of the nation instead of the satisfaction which consumers derive as they use health services?" [Feldstein (1963, pp. 22–23), quoted in Culyer (1971, p. 190)].

Culyer has attempted to develop an alternative, extra-welfarist framework that embodies the centrality of health as the outcome of concern [Culyer (1989, 1990)]. Building on Sen's notion of extra-welfarism, Culyer argues that normative evaluation should focus on the characteristics of people, including non-utility characteristics: "if the characteristics of people are a way of describing deprivation, desired states, or significant changes in people's characteristics, then commodities and their characteristics are what is often *needed* (emphasis in the original) to remove their deprivation..." [Culyer (1990, p. 12)]. The most relevant characteristic in evaluating alternative policies in the health sector is health. Ill-health creates a need for health care, which restores a person's health (or forestalls a worsening of health).[11] By this reasoning, extra-welfarism integrates two

[10] In this respect, it is part of a broader intellectual tradition that questions the place of preferences in social evaluation [Sen (1985), Sagoff (1994), Scanlon (1975)].

[11] With its emphasis on need, extra-welfarism has affinity with an earlier tradition of the materialist school, which gave central place in normative analysis to meeting basic human material needs, and which was displaced by the "new" welfare economics [Cooter and Rappoport (1984)].

key concepts that do not fit easily in a welfarist framework: the concept of need (as opposed to demand) and health (as opposed to utility) as a final outcome of concern.[12]

More generally, rejection of the welfarist individualistic social welfare function has led to the development of a "decision-maker" approach to cost-benefit analysis [Sugden and Williams (1978), Williams (1993)] and a call for a more communitarian approach to evaluation [Mooney (1998)]. Under the decision-maker approach, the relevant arguments in the objective function are defined by the decision-maker commissioning the analysis. Hence, the role of the analyst is limited to identifying the most efficient way to achieve the decision-maker's objectives. The decision-maker approach does not preclude either a welfarist or an extra-welfarist objective function. The relevance of one or the other depends on what the decision maker specifies as the objective. Extra-welfarists have argued that, in fact, decision makers have declared that producing health is the primary objective of the health care system [Culyer et al. (1991)].

Mooney's recent communitarian critique of the individualism that underlies normative health economics (both welfarist and extra-welfarist) explored the role of individual preferences, the relationship between individual preferences and the preferences of a community as a community rather than as simply the aggregation of individual preferences,[13] and the question of the value of a community as a community.[14] The work is exploratory, probing the nature of individual preferences that should be included, the nature of what is evaluated (the specific services in health care system or the system *per se*), and the ways in which preferences are aggregated.

Table 1 summarizes the relationships among the alternative normative approaches discussed. The rows of the table represent different conceptions of the outcome of interest (welfarist and extra-welfarist). The columns represent different conceptions of the form of the social welfare function. In the welfarist row, the first three cells represent perspectives found in welfare economics, all of which are based on individualistic social welfare functions. The first two cells, sum-maximizing social welfare functions (cell 1) and those based strictly on Pareto optimality (cell 2) are insensitive to distributional issues. The sum-maximizing version of welfarism corresponds to traditional utilitarianism and to applied welfare economics (cost-benefit analysis) based on the Kaldor–Hicks potential-Pareto-improvement criterion, which seeks to identify programs that generate positive net benefit. The second cell corresponds to strict Pareto optimality concepts for ranking allocations. Cell 3 includes Bergson–Samuelson-type social welfare functions

[12] Extra-welfarism, in principle, does not supplant utility information with non-utility information. Rather, it extends the relevant information set by including both utility and non-utility information. In practice in the health sector, however, it has often placed a near exclusive focus on health. See Culyer (1989, 1990) and Hurley (1998) for further discussion of the issue.

[13] Shiell and Hawe (1996) explore this issue in the context of the economic evaluation of community-level health promotion programs.

[14] Mooney is welfarist, rejecting the extra-welfarist emphasis on health outcomes [Mooney (1994)], especially when combined with its consequentialism that disallows consideration of such factors as process utility associated with receiving health care [McGuire et al. (1988)] or the value of access to health care *per se*, independent of whether it is consumed [Mooney et al. (1991)].

Table 1
Alternative normative approaches: specifying the objective function

	Social welfare functions				
	Individualistic			Non-individualistic	
	Sum-maximizing	Paretian	Bergson–Samuelson type	Communitarian	Decision-maker
Welfarist (utility)	(1) • Classical utilitarianism • Kaldor–Hicks criterion	(2) • Pareto optimality	(3) • Functional form that reflects preferences over the both the level and distribution of utility among members of society	(4) • Societal welfare is more than the sum of utility across individual programs and individuals	(5) • Decision makers specify the objective in terms of utility • The functional form of objective function could be as in (1)–(3) or something else
	• See standard welfare economics texts [e.g., Boadway and Bruce (1984)]	• See standard welfare economics texts [e.g., Boadway and Bruce (1984)]	• See standard welfare economics texts [e.g., Boadway and Bruce (1984)]	• Mooney (1998)	• Not aware of empirical examples
Extra-welfarist (health)	(6) • Health maximizing extra-welfarism	(7) • Pareto optimality defined with respect to health outcomes	(8) • Functional form that reflects preferences over the both the level and distribution of health among members of society	(9) • N/A	(10) • Decision makers specify the objective in terms of health • The functional form of objective function could be as in (1)–(3) or something else
	• Culyer (1989, 1990)	• Possibility discussed in Culyer (1995b)	• Wagstaff (1991)		• Culyer et al. (1991)

that have utility arguments and that incorporate distributional concerns through their functional specification. The last two cells in this row represent approaches that reject individualistic social welfare function but which see a central place for utility information in the social welfare function. In the case of the communitarian perspective, individualism is rejected because it is argued that social welfare is more than the aggregation (no matter what method is used) of individual preferences over individual programs. In the case of the decision-maker approach, the objective function is defined by the decision maker. Although in principle it could be welfarist (with any functional form), I am unaware of examples based on a decision-maker approach focused on utility outcomes. The second row corresponds to extra-welfarist approaches. Cell 6 represents the most dominant form of extra-welfarism in applied work to date, where the objective is specified to be maximizing the health of the population. Though he does not necessarily advocate it, Culyer (1995b) discusses the possibility of applying standard Pareto concepts within an extra-welfarist perspective (Cell 7).[15] Cell 8 corresponds to the use of social welfare functions that incorporate both efficiency and distributional concerns but which include only health outcomes as arguments. Wagstaff (1991) first explored such an approach. I am unaware of any extra-welfarist communitarian approaches (Cell 9). Cell 10 corresponds to the frequent justification for extra-welfarist approaches based on the stated objectives of decision-makers.

This chapter concentrates on the welfarist and the extra-welfarist distinction, with reference to specific variants of each as is appropriate to the context, with particular emphasis on the differing abilities of the variants to accommodate equity concerns. Welfarism and extra-welfarism represent the two most prominent approaches to normative economic analysis in the health sector and have been the focus of the most intellectual groundwork and debate. They also represent, in important respects, non-reconcilable frameworks for normatively assessing health policies and their attendant resource allocations. They derive from distinct conceptual foundations: welfare economics is utility-based and gives primacy to satisfying preferences; extra-welfarism is health-based. But it is also the case that, structurally, the two approaches share key elements, including strong consequentialist reasoning, a (near) exclusive focus on a single outcome, and an ability to accommodate only a limited range of equity concerns. Hence, both are subject to some of the same recent criticisms of normative economic analysis in general [Hausman and McPherson (1993), Sen (1979, 1987), Hurley (1998)].

With this background, then, let us turn to what the frameworks have to say about implications of the nature of health care as an economic commodity for efficient and equitable resource allocation in the health sector.

[15] Allocation A is preferred to allocation B only if the health of at least one person is greater and no one's health is worse.

3. Health care as an economic commodity

The question, "Is health care different?" has been a refrain since economists first focused on the health care sector in the 1950s [Mushkin (1958), Arrow (1963), Klarman (1963), Culyer (1971), Pauly (1978), Pauly (1988), Folland et al. (1996)]. The consensus is that yes, health care is different in ways that generate market failure and which are therefore important for formulating public policy in the health sector. Its distinctiveness is rooted in four characteristics of health care: (1) demand for health care is a derived demand (for health); (2) externalities; (3) informational asymmetries between providers and patients; and (4) uncertainty with respect to both the need for and the effectiveness of health care. Individually, each of these features can be found in other commodities, but no other commodity shares all of these features to the extent found in health care. It is the combination of these features that poses such a challenge for sound economic analysis and sound health policy.

Health care is generally defined to encompass those goods and services whose primary purpose is to improve, or prevent deterioration in, health.[16] It includes a heterogeneous set of goods and services that vary in the extent to which they share these distinctive features. The informational asymmetries faced by a consumer in deciding whether to take an aspirin normally pale in comparison to those faced in deciding whether to undergo neurosurgery; the uncertainty in a given year regarding the need for drugs to treat chronic arthritis is considerably less than that regarding the need for repairing a broken bone; the uncertainty regarding the effectiveness of repairing a broken bone is normally much less than the uncertainty regarding the effectiveness of chemotherapy for cancer; and the externalities generated by ensuring access to cosmetic surgery may be considerably less than those generated by ensuring access to life-saving appendectomies. Amidst this heterogeneity, however, lies the basic truth that, as a class of goods and services whose primary purpose is to improve health (and thereby, well-being), health care shares these features and sound economic analysis of the health care sector must be built on this insight.[17]

Consensus also does not imply unanimity and among health economists there are gradations of belief regarding the extent to which health care differs from standard, textbook commodities, the relative importance of various features of health care, and

[16] Definitions of health range from narrow conceptions based solely on abnormal physiological function to broad definitions such as the World Health Organization's, which defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease and infirmity" [WHO (1947)]. The problem with the latter for an economist is that it conflates health and utility. The working definition of health in this paper emphasizes physical and mental function, encompassing more than a purely physiological definition but falling short of the WHO definition [see Evans and Stoddart (1990) for further discussion of health concepts]. In addition, strictly interpreted, this definition would include safety interventions such as crash barriers whose primary purpose is to reduce injuries. Such interventions, however, are not conventionally considered health care.

[17] In the analysis of a specific service, of course, one or more these characteristics may not have an important bearing.

the implications of these differences for both public policy and methods of economic analysis in the health sector [Pauly (1978, 1988), Feldman and Morrisey (1990)]. Differences in such beliefs have even generated a rough typology of health economists into "broads" – those who emphasize the distinctiveness of health care and believe it has important implication for the operation of health care markets and modes of economic analysis – and "narrows" – those who believe that health care is not so distinctive and that health care markets can be fruitfully analyzed using standard neo-classical economic models [Evans (1976a)]. Perhaps not surprisingly, there is a strong correlation between whether one is a broad or a narrow and the preferred normative framework.

Let us examine each of health care's features in turn.

3.1. Derived demand for health care

Health care is one of many determinants of health and, from an economic perspective, it is simply an input into the production of health.[18] Consequently, unlike most consumer goods, which are consumed for their direct utility generating properties, health care is consumed to produce health, which is the desired good. In fact, health care itself is often a "bad," whose direct effects decrease utility (e.g., it is often painful). Most of us would be happy never to consume health care. But, conditional on being ill, health care becomes a "good" because of its ultimate effect on our health, the benefits of which outweigh health care's short-term direct negative effects. Demand for health care is derived from our demand for health itself [Grossman (1972)].[19]

The implications of this insight for normative analysis can be illustrated within a simple consumer framework. Following Evans (1984), suppose an individual's utility depends on general goods and services, X; health status (HS), which is produced by health care (HC) and other determinants of health (Z); and health care:

$$\mathrm{U} = \mathrm{U}\big(\mathrm{X}, \mathrm{HC}, \mathrm{HS}(\mathrm{HC}, \mathrm{Z})\big). \tag{1}$$

The effect of health care on welfare then depends upon:

$$\partial \mathrm{U}/\partial \mathrm{HC}, \tag{2}$$

the direct effect on welfare of consuming health care; and

$$(\partial \mathrm{U}/\partial \mathrm{HS})(\partial \mathrm{HS}/\partial \mathrm{HC}), \tag{3}$$

[18] The determinants of health include genetics, the social environment, the physical environment, and individual responses to these determinants [Evans and Stoddart (1990)]. Evidence suggests that these non-health care determinants may, on average, be more important than health care in determining the health of populations.

[19] Grossman provided the first formal treatment of this within an intertemporal human capital framework. Subsequent efforts to analyze the demand for health care within such a framework can be found in Muurinen (1982) and Reid (1998). See Grossman (2000) for an overview of this human capital framework. In this model, health care has both consumption and investment properties.

the contribution of health care to health status, combined with the contribution of health status to welfare.

The first term, $\partial U/\partial HC$, is the direct effect of health care on utility just like that found for standard consumer goods. This is often negative and, though important in some contexts to the decision to consume health care (e.g., the strong negative side effects associated with some chemotherapies), it is generally of less analytic importance. Of more general analytic importance is the effect of health care on welfare through its effect on health status. This depends on two factors: (1) the marginal contribution of improvements in health status to utility, $\partial U/\partial HS$, which is subjective and known only by the individual; and (2) the marginal productivity of health care in producing health, $\partial HS/\partial HC$, which is a technical relationship that can, in principle, be established by scientific research and is knowable by a third party. To the extent that a health care service is consumed to improve health, a positive marginal product of health care in producing health is a necessary condition for a service to improve welfare.

This physical production relationship provides a foundation for a class of third-party normative judgements in health care that are not possible for standard consumer goods whose effects on welfare depend solely on subjective assessments known only by the consumer. If a health care service has been demonstrated to be ineffective in improving health, a third-party can often speak with confidence in stating that it is will not improve well-being to consume it.[20] Because the effects on welfare of consuming health care depend in part on a production relationship, health economists can use technical and cost-effective efficiency concepts in making assessments of consumption decisions.

This can lead to considerable confusion (particularly for non-economists) when discussing efficiency in health systems. Efficiency concepts can apply at three levels: (1) efficiency in the production of health care services; (2) efficiency in the use, or consumption, of health services; and (3) efficiency in choosing a level of health. At the first level, producing health care services, only the supply-side notions of technical and cost-effectiveness efficiency are relevant. At the second level (consuming health care) both supply- and demand-side efficiency concepts are relevant. To the extent that health care is consumed to produce health, technical and cost-effectiveness efficiency are relevant in assessing both the mix of health care services consumed and the use of health care versus other inputs to produce health. However, because health care also has direct effects on welfare, demand-side, or allocative efficiency concerns also arise. A consumer may trade-off efficiency in the production of health for direct utility effects by choosing a less effective treatment that also has fewer negative side-effects. Finally, at the third level, allocative efficiency is relevant in choosing the optimal level of health for the consumer, where health is traded off against other goods and services to maximize welfare.

[20] While health care is sometimes consumed for reasons other than improving health, such as a diagnostic test that provides information valued by the patient even if it will not alter treatment decisions or health [Mooney and Lange (1993)], the above observation holds true for a large portion of health care consumption.

Derived demand and the associated relevance of supply-side efficiency concepts in assessing patients' consumption of health care services also serve as a basis for a useful concept of need in the health sector [Boulding (1966), Culyer (1995a), Williams (1978)]. There is general agreement that a necessary (though perhaps not sufficient) condition for a need for a good or service to exist is that the good or service be effective in attaining a desired objective. Hence, the technical effectiveness relationship can serve to operationalize the concept of need in health care: a need can exist only where there is an effective service to improve health [Williams (1978)].[21] Again, because this relationship can be established by clinical research (at least at a population level) and is knowable by a third party, need can serve as a basis for normative analysis in the health sector.

Health economists therefore distinguish between need and demand. Need depends on the ability to benefit from health care; demand depends on preferences backed by ability to pay. Normative approaches rooted in neo-classical welfare economics emphasize demand, with its foundation in allocation according to preferences supported by ability and willingness-to-pay. Extra-welfarist approaches posit a central role for need as a normative standard in assessing the efficiency and equity of alternative systems of finance and delivery.

The derived nature of the demand for health care, with the associated relevance of both supply-side and demand-side concepts of efficiency in the analysis of health care consumption, therefore has profound implications for normative analysis in the health sector. It underlies both the usefulness of the concept of need as a basis for normative analysis of health care utilization,[22] which is unavailable to economists in other sectors and which economists more generally reject, and the disputes over the nature of the appropriate objective function to guide normative analysis in the health sector.

3.2. Externalities

Externalities remain one of the most discussed, if least empirically studied, aspects of health economics. Except for physical health externalities, most arguments regarding the presence and nature of externalities for health care services are based on introspection and the broad public support in most countries for subsidies to increase citizens' access to health care. The first attention by economists to externalities associated with health care services arose in the early and middle 1960s in the context of a debate regarding the potential efficiency of heavy public involvement in health care finance and delivery, particularly as represented by the British National Health Services [see, e.g.,

[21] A fuller discussion of the concept of need and alternative definitions of need is provided in Section 4.

[22] The large literature on small-area variations, for example, makes little sense in the absence of the technical concept of need [see, e.g., Anderson and Mooney (1990), Paul-Shaheen et al. (1987), Folland and Stano (1990), Phelps and Parente (1990)].

Lees (1960, 1962, 1967), Buchanan (1965), Klarman (1965a)].[23] The work, however, was intended as much to be an effort in positive economics to provide models of why such public programs might be efficient as it was to be a normative assessment of such arrangements.

Culyer and Simpson (1980) document three phases in the evolution of economic analyses of externalities in the health sector. In the first phase, economists argued that external effects were small or non-existent for general health care services such as physician and hospital care. Policy-relevant externalities were limited to physical health effects associated with interventions targeted at communicable diseases, passed either directly among humans (e.g., small pox, syphilis) or indirectly through the physical environment (e.g., tuberculosis, polio) [e.g., Weisbrod (1961), Lees (1960, 1962, 1967)]. An action taken by one person (e.g., ensuring clean, safe water; immunizing oneself against, or seeking treatment for, a communicable disease) generates direct health benefits for other individuals (i.e., reduced rates of disease). Market exchange, which ignores such positive external effects, yields less than socially optimal levels of such activities. Because such interventions fall so clearly within the classic concept of public goods and externalities, some of the earliest health economics work focused on analyzing the economic efficiency of such efforts [e.g., Klarman (1965b)].

Clean water and air, as well as aspects of sanitation, are pure public goods. Others cannot be excluded from enjoying the benefits of providing them. The potential for free-riding, which can threaten their provision at any level without collective action, is most pronounced for pure public goods. Immunization against communicable diseases confers appreciable private benefit, especially at low levels of population coverage where the disease is still prevalent. But, as coverage rises, herd immunity causes immunization to approach a pure public good: if virtually everyone has been immunized, an unimmunized person is effectively as well protected as if immunized though he has not incurred the cost of the immunization [Musgrove (1996), Phillipson (2000)].

Goods with external effects therefore call for collective action to ensure their provision at efficient levels. Pure public goods often require both collective financing and public provision. For goods that produce positive externalities for which exclusion is possible, the standard corrective policy is price subsidy. In some cases in the health sector, however, additional action may be justified. For asymptomatic communicable diseases requiring treatment, such as some sexually transmitted infections, people may not realize that they are infected and demand would be too low even if the care were free. Hence, in addition to providing a price subsidy, ensuring optimal level may call for education and/or mandatory programs. Within welfarist approaches mandatory programs can be Pareto optimal under certain conditions [Brito et al. (1991)]. Alternatively, within an extra-welfarist approach, mandatory programs have been supported by merit-good arguments, which are more paternalistic.

[23] At this time for example, universal Medicare was introduced in Canada, Medicare and Medicaid were introduced in the United States, and many European countries, which previously had social insurance systems, expanded coverage.

In the second phase emphasis shifted to general health care services as a source of policy-relevant external effects. Externalities were modeled as being generated by good-specific utility interdependencies in which others' consumption of health care services enters a person's utility function. The interdependency was modeled in a variety of ways, but generally the interdependency related to either the absolute level of health care consumption by others [e.g., Pauly (1970)] or the relative levels of consumption, with a particular concern for the extent of inequality in health care consumption [e.g., Lindsay (1969)]. Health care was generally treated as any standard commodity in the utility function except for the argument pertaining to others' consumption that generated an external effect. In particular, consumers of health care were assumed to be able to judge the value of health care and health care entered the utility function only as a direct argument (rather than via a demand for health). Consequently, the welfare-maximizing policies derived from such frameworks were price subsidies to encourage the consumption of health care services, such as is derived from standard analyses of policies to correct for consumption externalities.[24]

Weisbrod (1964) argued for an additional type of externality associated with some health-care services, an externality he labeled "option value." Current period market demand for goods and services that are purchased infrequently (including some health care services), for which there is uncertainty in demand, and for which there are high costs to resume production if it is curtailed, will reflect only the value accruing to the users. It ignores the "option value" that accrues to non-users for whom such a service will be available should it be needed in the future (e.g., emergency services, hospital services). The optimal level of production may therefore call for public subsidy.

The third phase marks a period in which externalities in the health sector were predominately seen to derive from concern over others' health status. Others' consumption of health care *per se* is not the object of concern; rather, it is their health status. And because health care is one determinant of health, often the most important determinant when ill, ensuring access to health care services is one policy response to the externality [Culyer and Simpson (1980), Evans and Wolfson (1980)].

This formulation of the external effects generates policy implications that do not necessarily follow from the previous formulations. External effects associated with others' health status may call for policy interventions outside the health care sector that have important health effects but which do not generate physical health externalities (e.g., occupational safety). Because health care services generate externalities to the extent that they affect health status, people derive benefit from knowing that others receive not just any health care services but *needed* health care services. Interdependencies associated with utilization of needed health care may justify a larger public role in financing, organization and delivery. Public financing and price subsidies (for either health care insurance or health care itself) may be necessary for ensuring widespread access

[24] Models that focused on equality of consumption potentially also had implications for reducing the consumption of high-users (as well as increasing it for low-users). See [Lindsay (1969)].

to needed health care, but they are seldom sufficient. Where need is a pivotal concern, purely demand-side, price-based policies are inadequate. Ensuring widespread access to and utilization of needed health care may call for regulation of the supply and distribution of providers, initiatives to ensure the appropriate delivery of services and, at times, public delivery itself where private interest is not sufficient to bring forth appropriate supply of services [Culyer and Simpson (1980)].

3.3. Informational asymmetry

Informational asymmetry occurs when one party to a transaction has more information pertinent to the transaction than does the other party, which may allow the better-informed party to exploit the less-informed party. Informational asymmetries pervade the health sector and cause market failure in both health care and health care insurance markets.[25] In this section we focus on health care; health care insurance is discussed in the next section.

The principal informational asymmetry in the health care sector is that between a provider and a patient. This informational asymmetry is pivotal because most health care resources are allocated through decisions made in the provider-patient encounter. It is also one of the most inescapable asymmetries: information is often the good patients most seek from providers when they perceive themselves to be ill. Patients normally seek two types of information. The first is diagnostic information – what is wrong with me? Is the pain in my chest indigestion that will pass or is it angina that presages a heart attack? Because patients are unable to self-diagnose many types of illness or injury, they rely on a health care professional for such information. The second type of information patients seek from providers is treatment information – given the diagnosis, what should I do to restore my health? If the diagnosis is angina, what will be the most effective treatment – drugs, coronary-artery bypass surgery, angioplasty, etc.? The provider has the technical information regarding the treatment options that can form a basis for a decision as to what health care to consume.

Optimal health care consumption depends on utilizing effective health care to improve health to the extent that health is valued by an individual (relative to other activities to improve well-being). Within the simple framework set out earlier, patients know best how improvements in health affect their well-being ($\partial U/\partial HS$), while providers have better information regarding both the causes of ill-health and the effectiveness of alternative health care services in restoring health or preventing the further deterioration of health ($\partial HS/\partial HC$).

[25] Though it has received the most attention, informational asymmetry is only one type of informational problem found in the health sector. Problems of symmetric but incomplete information, for instance, have been analyzed as a source of variations in the level of health care utilization that can not be explained by either needs or preferences for care in the population. Health economists are drawing more heavily on developments in the economics of information to understand both the behavior of actors in the health sector and the welfare implications of information problems [e.g., Harris (1977), Dranove and White (1987), Phelps and Parente (1990), Phelps (1992), Gaynor and Polachek (1994)].

The asymmetry between the patient and the provider regarding both the nature of the illness and the effectiveness of alternative treatments causes market failure. Because of informational asymmetries, patients may fail to purchase care they would if well-informed, they may purchase care they would not have purchased if well-informed, they may purchase care of a differing quality, etc. "Well-informed" does not mean "perfectly informed," but simply as informed as a knowledgeable provider. More generally, the informational asymmetry confers an advantage on the provider that can be exploited for the provider's gain by manipulating one or more of the quantity, quality and price of health care services in a way not easily detected by consumers.

The problems caused by informational asymmetry are exacerbated by the context in which many health care consumption decisions are made. Patients often have very limited time to shop around or seek out information[26] and health care is often consumed at times of extreme vulnerability and sometimes cognitive impairment for individuals, compromising the ability of individuals to process information. Finally, even where patients do not face such problems, the opportunities for learning from experience in health care may be limited. Learning from experience is most feasible for repeat purchases. Although much primary care may reasonably fall into such a category, much secondary and tertiary care does not (one only has one's gall bladder removed once!). This is where the majority of health care resources are consumed. It is estimated, for instance, that among beneficiaries in the US Medicare program 28% of health care expenditures are for those in their last year of life [Newhouse (1992) and references therein]. Weisbrod (1978) emphasized that learning is inhibited by a further problem: it is often impossible for the patient to know the counterfactual – what would have happened in the absence of treatment. Many ailments are self-limiting and would resolve themselves in the absence of intervention. The inherent uncertainty associated with medical treatment at the individual level makes it difficult for consumers to judge quality. Even a well-provided, appropriate treatment may still fail.[27]

There is considerable heterogeneity in the extent of informational asymmetry in health care services, and corresponding variability in the scope for using market forces. It is generally acknowledged, for instance, that for optometry services advertising and associated competition can have beneficial effects [Benham (1972), Kwoka (1984)]. Similarly, while informational problems require that patients have access to prescription drugs only through a licensed health care provider, there may be considerable scope for competitive processes in the market for dispensing prescriptions [Evans and Williamson (1978)]. Although important, such examples make up a small portion of all health care consumed.

[26] Although many individuals face similar informational asymmetries with respect to auto repair, as Musgrove (1996) points out, one can rent a car while shopping around for the automobile repair shop that provides the best value for money; one can not rent another body to shop around for alternative providers.

[27] This contrasts with automobile repair, a sector with analogous informational problems, where there is far less uncertainty regarding the effectiveness of a repair procedure, and hence, greater scope for learning about the quality of a particular supplier over time.

Demand-side policies attempt to correct the market failure by providing consumers with relevant information. Such policies are generally advocated by economists with a stronger allegiance to neo-classical methods, who do not perceive asymmetries to be severe, who judge health care to be "not that much different" than "standard" commodities and who generally favor market-oriented approaches to resource allocation [e.g., Pauly (1978, 1988), Feldman and Sloan (1988)]. Initiatives to provide consumers with information upon which to make choices in market contexts are most developed in the United States, which, among OECD countries, relies most heavily on private markets in health care. Most of these efforts, however, are targeted not at patient-provider choices regarding treatment. Rather, they are targeted at the choice of health care plan or provider organization through which to receive care. The efforts have been spearheaded less by typical consumers than by large purchasers such as employers (who provide health insurance to employees as an employment benefit). The objective is not, therefore, to address the informational gap in a given provider/patient clinical encounter, but to inform choices in the context of a competitive market in health care insurance among multiple health care plans.

Even for the circumscribed context of choosing a provider/insurance organization, efforts thus far have had mixed results [Schneider and Epstein (1998), Hibbard and Jewett (1996, 1997), Hibbard et al. (1996)]. Considerable progress has been made in collecting more standardized data upon which to base measures of quality. Its use, however, has been limited to large purchasers. Individual consumers often find it difficult to understand the meaning of quality and performance measures presented, they have difficulty relating the measures to aspects of care of value to them, they find the volume of data on multiple dimensions of performance difficult to process and there are important questions as to what consumers find most important regarding quality [Hibbard and Jewett (1996, 1997)].

Independent of these efforts to provide purchasers/consumers with information to foster conditions conducive for markets are initiatives in what may be termed the "shared decision-making movement." The objective of these efforts is to provide information to patients relevant to their specific clinical situations so that they can participate more fully in their treatment choices [Charles and Demiao (1993), Levine et al. (1992)]. Creating more informed consumers working with health professionals in making treatment decisions increases the likelihood of generating efficient allocations that reflect both technical information on effectiveness and the preferences of the patients.

A complementary approach to informational asymmetry is agency. As agent for their patients, a provider is expected to act in the interest of patients, not self-interest [Evans (1984), Mooney and Ryan (1993)]. This notion of agency is distinct from that encountered in the literature on principal-agent problems. The principal-agent literature focuses on contexts in which both the principal and the agent are assumed to be wholly self-interested, there are informational problems that preclude perfect monitoring of the agent by the principal, and the problem is to design efficient, incentive-compatible arrangements between the principal and the agent [for a survey of this literature, see Sappington (1991)]. Although the provider-patient context shares some of these elements, agency in the provider-patient relationship is crucially different in that the provider is not expected to act self-interestedly. The agency relationship has been incorporated into economic models of physician behaviour in a variety of ways, either in the objective function of the physician or through the constraint [Evans (1976b), Woodward and Warren-Boulton (1984), Dionne and Contandriopoulus (1985)].

Agency is fostered through two strategies. One is to create a professional culture that emphasizes agency, that socializes health care providers (or, in economic jargon, that modifies provider preferences) to act differently than a prototypical supplier of a good who is assumed to pursue profit in a wholly self-interested fashion. As the informed agent for the poorly informed patient, a provider is to act in the interest of the patient, providing those services which the patient would demand if she had the same information as the provider. But if providers operated in a competitive environment, survival would demand that they act self-interestedly. So the second strategy is to reduce competitive market pressures that might induce providers to give primacy to self-interest rather than patient interest. Providers, and physicians in particular, have been protected from competitive pressures through supply-side regulations such as licensure (which restricts entry), limitations on advertising, and other professional norms that reduce competition among providers.

These supply-side policies also have another rationale – to prohibit low levels of quality. Under certain conditions, it is welfare-improving to prohibit low qualities that few well-informed consumers would voluntarily choose rather than providing information to consumers and allowing them to choose their own quality levels [Pauly (1988)]. Licensure of health care providers, accreditation of health care facilities, drug approval regulations, and so on attempt to prohibit quality of care below a specified level.[28] Chalkley and Malcomson (2000) provides a more extended treatment of these issues.

Supply-side approaches to correcting market failure induced by informational asymmetry attempt to balance their counteracting effects. On the one hand, it can be welfare-enhancing to specify minimum quality levels and to foster agency relationships. On the other hand, it can be welfare-reducing to grant monopoly powers to providers and to create work contexts that allow providers discretionary scope to pursue professional and other objectives not consistent with patient needs, preferences or an efficient use of resources. Within traditional public-utility approaches to regulation, the balance is to be maintained by the regulator who monitors the behaviour of the regulated. Here again, however, we bump up against an informational problem: in general, non-professionals do not have the requisite expertise to judge when health care providers are acting in

[28] An alternative explanation for licensure, especially of physicians, is self-interested advocacy to garner higher incomes by limiting entry [Kessel (1958)]. Physicians (and others) have undoubtedly used licensure and regulation to advance their own interests, and such "capture" presents real challenges for advancing the public interest through such policies. These issues do not disallow that licensure represents an attempt by society to ensure that all providers are above a minimum level of quality. Licensure has also been criticized as ineffective because it has not required continuous re-certification throughout a physician's career. Again, while this may be true and more effective programs may be required, the basic point still holds.

ways consistent with the public interest, particularly with respect to the content of clinical care.[29] Consequently, in most settings governments have granted health care providers, and physicians in particular, broad powers of self-regulation [Tuohy and Wolfson (1978)].[30]

Providers have not, nor probably could they, act as perfect agents, either in individual encounters or as self-regulators. It is likely impossible both to ascertain patient preferences perfectly and to repeatedly ignore self-interest. Perfect agency is therefore the ideal. Mounting evidence of the extent of the apparent deviation from perfect agency has spawned a host of regulatory initiatives that challenge provider autonomy, including managed-care approaches to the organization and delivery of health care – utilization review, pre-certification programs, practice guidelines, and other initiatives to reduce variation in practice across providers that cannot be linked to differing needs for care.

Asymmetries of information underlie the design of recent reforms to exploit competitive forces through the creation of "internal markets" with purchasers and providers [e.g., Shackley and Healy (1993)]. Equity concerns associated with health care (discussed in Section 4 below) demand that competition be introduced in a way that continues to provide universal access to needed health care (hence, an "internal" market, within a public system) and asymmetry of information precludes extensive demand-side competition at the level of individual patients. Instead, competition is restricted to competing providers (hospital trusts, laboratories, long-term care facilities, launderers, and so on) who must sell their services to "purchasers," either regional health authorities who procure services on behalf of their residents or providers such as GP fundholders in the UK, who purchase services on behalf of their rostered patients. Providers retain the agency relationship with patients and suppliers, dealing with knowledgeable purchasers, face competitive pressures. Because there are often only one or two potential suppliers of many services in a region and because many of the services purchased (e.g., surgical operations) are complex in nature, must be custom-designed to the specific individual, and have high asset-specificity for the producer, the market is very different from typical textbook competitive markets. Rather, the internal markets rely heavily on contestability (the threat of entry by another) among suppliers and long-term relational contracting between purchasers and providers.[31] Successful contracting arrangements depend on effective agency relations and overcoming informational problems associated with health care.

[29] This asymmetry is directly analogous to that found at the patient level.

[30] In a related but distinct analysis, recent work modeling physician behaviour has also tried to model the dual agency role of physicians – agent for their patients and agents for the funder, especially within publicly financed systems [Bloomqvist (1991)].

[31] Williamson (1986) provides a general transaction-cost analysis of the economic issues involved in such contracting arrangements; Shackley and Healy (1993) provide an economic analysis of the NHS internal market.

3.3.1. Supplier-induced demand

Lastly, informational asymmetries are at the root of what has been one of the most prominent debates within health economics: supplier-induced demand. Providers, acting as agents for their patients, have a major influence in "demanding" the services they will supply to the patient. This influence violates the assumption of neo-classical economic theory (both positive and normative) that the demand and supply sides of a market are independent. For a fuller discussion of supplier-induced demand, particularly the myriad statistical and econometric challenges faced in empirically testing for inducement see Feldman and Sloan (1988, 1989), Rice and Labelle (1989), Labelle, Stoddart and Rice (1994a, 1994b). However, a few observations about supplier inducement that bear directly on normative analysis are in order.

At its most general level, supplier inducement refers to a phenomenon whereby a provider shifts the demand curve for health care by patients. More commonly in the literature, inducement refers to a situation in which a provider violates the agency relationship out of financial self-interest by recommending services of questionable benefit to a patient. In this definition, motive is an operative element which, in a strict sense, makes it virtually untestable. Although controversy continues as to whether providers have such power [e.g., Pauly (1994a)], there is a broad consensus among health economists that they do.[32] Much less clear, however, is the extent to which, and the contexts in which, they are most likely to induce demand. Feldman and Sloan have argued that the key policy issue associated with supplier-induced demand is "not whether physicians have ever induced consumers to purchase a service that they would not have purchased had they been fully informed; rather, it is whether such care is quantitatively important and whether the amount of induced demand varies systematically with variables such as physician supply" [Feldman and Sloan (1988, p. 240), see also Folland et al. (1996)].

Labelle et al. (1994a) have argued that for policy purposes, the concept of supplier-inducement should drop the question of motive and be broadened to include consideration of both the integrity of the agency relationship (in recommending a service, is the provider reflecting patient preferences?) and the effectiveness of the induced service (does the service improve health status?). Because both providers and patients operate in a world of imperfect information, even some services provided under perfect agency may create policy concerns (e.g., if they are not effective or if they are not cost-effective). Within this broader framework, allocative efficiency implies that inducement may be a concern even if it is not correlated with changes in other system variables. A policy concern arises whenever a provider recommends a service that violates the agency relationship or whenever an ineffective service is provided.

[32] Specifically, in a survey of health economists, 82.6 percent agreed with the statement that, "Within broad limits, physicians generate demand for their services in response to economic incentives" [Feldman and Morrisey (1990, pp. 640–641)].

The consequences of supplier-induced demand for welfare analysis in the health sector are two-fold. Normative analyses are valid only if the positive analyses that underlie them are valid. To the extent that positive economic models assume that supplier-induced demand will not occur when in fact it does, the predictions of the underlying model will be false, as will any normative analyses that flow from it. This concern, for example, leads many to question the estimates of the aggregate welfare loss associated with "excessive" health insurance coverage (and, by implication, the welfare gains to increasing patient cost-sharing) based on demand elasticities from the Rand health insurance experiment (e.g., Manning et al. (1987), Feldman and Dowd (1991)]. In the experiment, because only a small proportion of any provider's patients faced increased cost-sharing, the effects of their reduced utilization on the provider's income was minimal. If substantial cost-sharing were to be implemented system-wide, however, a large reduction in utilization by patients would have large income effects for providers, and potentially generate demand inducement by providers.[33] Hence, aggregate analysis based on elasticities from the experiment would be misleading. Valid normative analysis of the effects of broadly-based cost-sharing must incorporate expected behavioral responses of providers. Inducement therefore undermines the positive economic analysis upon which the welfare analysis is built.

More importantly, asymmetry of information, agency and supplier inducement vitiate the assumption of individual sovereignty, which transforms a demand curve from a purely positive construct (the relationship between the quantity of a service demanded, prices, income, etc.) into a normative construct that can be used to measure consumer welfare. If consumer ignorance and provider influence pervade the markets for health care services, then the area under the demand curve for health care will not represent a valid measure of consumer welfare. Neither demand based on poorly informed consumer judgements nor demand based on provider influence in the presence of imperfect agency accurately represent the welfare associated with the service. In either case, the traditional assumption is violated and the normative significance of the demand curve is undermined, which in turn undermines traditional welfare analysis [Evans (1983, 1984), Rice (1992)].

3.4. Uncertainty

Arrow (1963) identified two important types of uncertainty associated with health care: uncertainty in the demand for health care and uncertainly regarding the effectiveness of treatment. Because illness and injury at the individual level are to a great extent random events, individual demand for health care and the associated expenditure have a large random component. Although clinical research can demonstrate whether, on average, a treatment is effective for a given condition (i.e., whether there is a scientific basis for

[33] McGuire and Pauly (1991) analyze the crucial role of income efforts in predicting physicians responses to payments policies.

offering a treatment in a particular context), in the end, it can not demonstrate whether a treatment will be effective for a particular individual with a particular condition: prior to treatment there is uncertainty as to a service's effectiveness. Both of these types of uncertainty at the individual level – uncertainty in demand and uncertainty in the effectiveness of treatment – are inherent.

The economic efficiency of market arrangements therefore depends on the ability of a competitive system to create a full set of risk-bearing (i.e., insurance) markets. If a full set of markets is not created, market failure results and non-market arrangements may improve economic efficiency. Missing markets in risk-bearing may explain a number of the non-market institutional forms observed in the health sector [Arrow (1963)]. Although health care insurance exists for many types of health care expenditures, insurance markets themselves suffer from market failure.

3.4.1. Welfare improving effects of insurance

In the presence of uncertainty risk-averse individuals, each of whom is at risk for a negative event such as illness, can often make themselves better off by risk-pooling. Risk-pooling reduces risk because, although an event is unpredictable for any single individual, the number of such events that will occur in a large group of individuals can be predicted.[34] Risks can be pooled only for events (phenomena) that can be traded. It is not possible to pool the risks of illness *per se*, as one can not trade health (i.e., give up a little bit of health at the start of the year for a guarantee that no severe illness will occur). But the financial risks associated with illness can be pooled.

The welfare gain of insurance for a risk-averse individual seeking to maximize expected utility can be illustrated as follows. Let W_0 be the individual's level of wealth if healthy. The individual will become ill with probability p, and experience a financial loss of L. The individual's expected wealth is: $p(W_0 - L) + (1 - p)(W_0)$, which equals $W_0 - pL$, where pL is the expected loss. Figure 1 depicts their level of utility under the different health states. The vertical axis represents total utility, the horizontal axis represents wealth, and the total utility curve is concave to reflect risk aversion (or equivalently, diminishing marginal utility for wealth). If the individual remains healthy, they attain $U(W_0)$ [point C]; if they become ill, they attain $U(W_0 - L)$ [point A]. Their expected utility is therefore $pU(W_0 - L) + (1 - p)U(W_0)$, which can be found on the chord connecting points A and C, in this case, at point B. Suppose, however, that the individual could buy an insurance contract with an actuarially fair premium (i.e., equal to the expected loss (pL)) that provided coverage for the financial loss in the event of illness. The individual would face a certain wealth level of $(W_0 - pL)$, and they would achieve utility of $U(W_0 - pL)$ [point D]. Purchasing such an insurance contract increases the individual's expected utility by the amount BD.

[34] Phelps (1992) derives the relationship between the average risk faced by individuals and both the number of individuals pooling their risks and the independence of the risks across individuals.

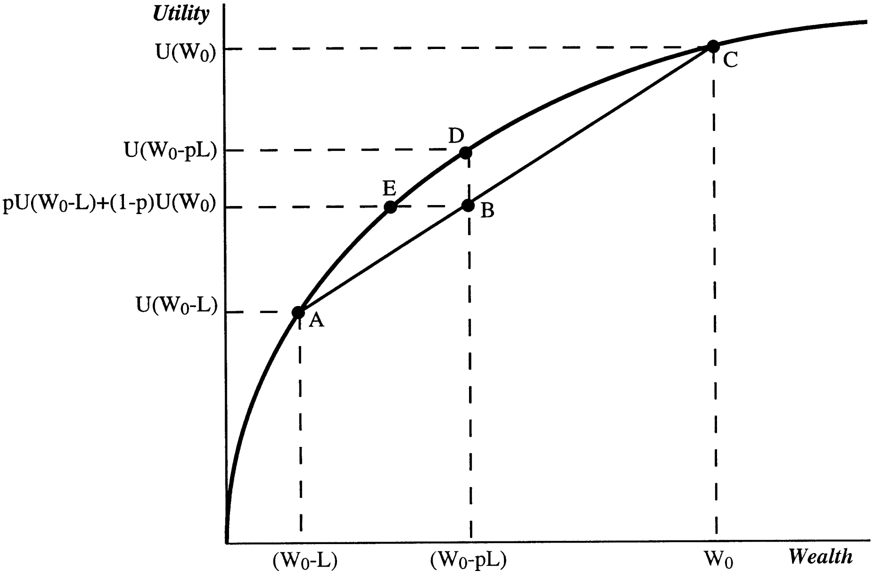


Figure 1. Welfare effects of insurance under risk aversion.

Because actuarially fair premiums provide an insurance organization no revenue to cover administrative costs, premiums must include loading charges to cover such costs. As long as the loading charges do not exceed EB in Figure 1, the individual is still better off purchasing insurance. Under certain conditions, loading charges cause consumers to prefer policies with a deductible, particularly if the loading factor is larger for small claims than for large, as would be expected [Arrow (1963)]. An individual is better off self-insuring for small losses through a deductible in a policy that provides full coverage for large losses. All of the above analysis assumes that the utility of wealth is independent of health status (healthy or ill). If the utility of wealth is health-state-dependent, then the optimal level of insurance is unknown [Shavell (1978)].

Given that insurance is welfare-improving for individuals, the critical issue from a policy perspective is how best to organize insurance markets to provide such insurance. This is particularly challenging because insurance markets are subject to a number of types of market failure, the most prominent of which arise from economies of scale, adverse selection, and moral hazard.

3.4.2. Economies of scale

The fixed costs associated with establishing the insurance pool and with calculating a full set of risk-adjusted premiums generate economies of scale. Depending on the overall size of the markets, attempting to sustain competitive markets with numerous small firms creates technical inefficiencies as each firm operates at output levels below minimum average cost. On the other hand, if only one or a small number of firms operate, one risks inefficiency associated with monopolistic practices. In both cases, some individuals willing to purchase insurance at an actuarially fair premium that reflects their risk status plus a load factor associated with lowest costs production will not be able to purchase insurance, reducing allocative efficiency. Hence, economies of scale can create market failure from both technical and allocative inefficiencies.

Single-payer, tax-financed public systems of insurance, one possible response to economies of scale, can generate technical efficiencies. Evidence suggests that such systems can reduce administrative costs by avoiding a separate infrastructure to collect revenue (it is integrated into the existing tax system), eliminating the need to set premiums altogether and for advertising among competing firms, as well as reducing the resources required for providers to collect reimbursement. Woolhandler and Himmelstein (1991), for example, estimated that administrative costs accounted for 19–24% of health care spending for the multi-payer US health care system but only 8–11% for the single-payer Canadian system.[35]

3.4.3. Risk selection

Risk selection arises from informational asymmetries between the insured and insurers. Adverse selection, a process whereby low-risk individuals drop out of the insurance pool leaving only high-risk individuals, arises when the individuals purchasing insurance have better information regarding their risk status than does the insurer. An insurer that can not distinguish low- and high-risk individuals must base the premium on a risk-pool that includes both high- and low-risk individuals. Low-risk individuals (who know they are low-risk) will not purchase insurance because the premium does not reflect their risk status. This leaves only high-risk individuals in the pool and the premium revenue of the insurer is insufficient to cover expected losses. If the insurer raises premiums to reflect the increased risks remaining in the pool, another segment of lower-risks will exit, again leading to losses. In the limit, adverse selection can make insurance markets unsustainable. Even short of the market disappearing altogether, individuals who would be willing to purchase insurance contracts that reflect their risk status are not able to because the insurer does not have the information required to offer them such a policy. The market cannot offer a full set of insurance contracts, reducing allocative efficiency. The most prominent strategy to combat adverse selection is to define risk pools in ways that retain individuals from all risk levels, such as through compulsory public insurance

[35] The optimal level of administrative costs is unknown. While the US system clearly has administrative waste, it can reasonably be argued that the Canadian system has historically underspent on administrative and management functions. And an overall evaluation of alternative systems of finance would have to assess allocative as well as technical efficiency.

or by basing risk-pool membership on a group, such as employee-sponsored plans, that requires all members to participate.[36]

A second risk selection problem is cream-skimming, which occurs when insurers have better information on an individual's risk status than does the individual. Under cream-skimming, an insurer generates higher profits by purposefully selecting low-risk individuals for coverage whose expected losses are below the premium charged.[37] Insurers can cream-skim in a number of ways including designing policies with deductibles and co-insurance provisions that prompt individuals to self-select into risk categories, selling insurance in settings where low-risks predominate, and other creative strategies [Giacomini et al. (1995), Neuman et al. (1998), Newhouse (1996, 1998)]. Cream-skimming is normally combated through either regulatory approaches to control risk selection behaviours or through the development of risk-adjusted premiums, which reduce the incentive to risk-select by better matching the premium to an individual's risk-status. Risk-adjustment however, remains a rather crude science at this point [Giacomini et al. (1995), Newhouse (1996, 1998), van den Ven and Ellis (2000)].[38]

3.4.4. Moral hazard

Moral hazard refers to the tendency for insurance coverage to induce behavioral responses that raise the expected losses that are insured, because it increases either the likelihood of a loss or the size of a loss. Those with health insurance coverage may take less care to avoid illness or injury knowing that they will not have to bear the associated financial consequences. In general, this is probably not a large source of moral hazard in the health sector, as the financial consequences are only a portion of the total "costs" associated with illness or injury, which often include pain and suffering.

Of more importance in the health sector is moral hazard associated with the fact that once an insurable event occurs, because an insured individual does not have to pay for the full cost of treatment, the individual may incur higher total costs than in the

[36] On average, workers are healthier than the general population, but such plans include all risk levels among workers.

[37] In the context of integrated systems with capitation payment, it arises when the provider/insurance organization enrolls low-risk individuals whose costs are below the capitation payment received.

[38] It is important to distinguish two streams in the risk-adjustment capitation payment literature. One stream focuses on risk-adjustment in the context of competitive health insurance markets. Here the goal is to ensure that capitation payments accurately represent the expected utilization of an enrollee, with risk adjustment often based in part on past utilization. The criterion for successful risk-adjustment is financial. In the second stream, developed predominantly in publicly funded systems (e.g., UK, Canada), the goal is to adjust for relative need for care from a population perspective. Hence, whether it is intended to fund a geographically defined population or an enrolled population, the criterion by which to judge a capitation formula is the extent to which it captures variation in relative need, assessed independent of previous patterns of utilization (that often reflect many factors other than need). See, e.g., Hutchison et al. (1999), Birch et al. (1993), Mays and Bevan (1987).

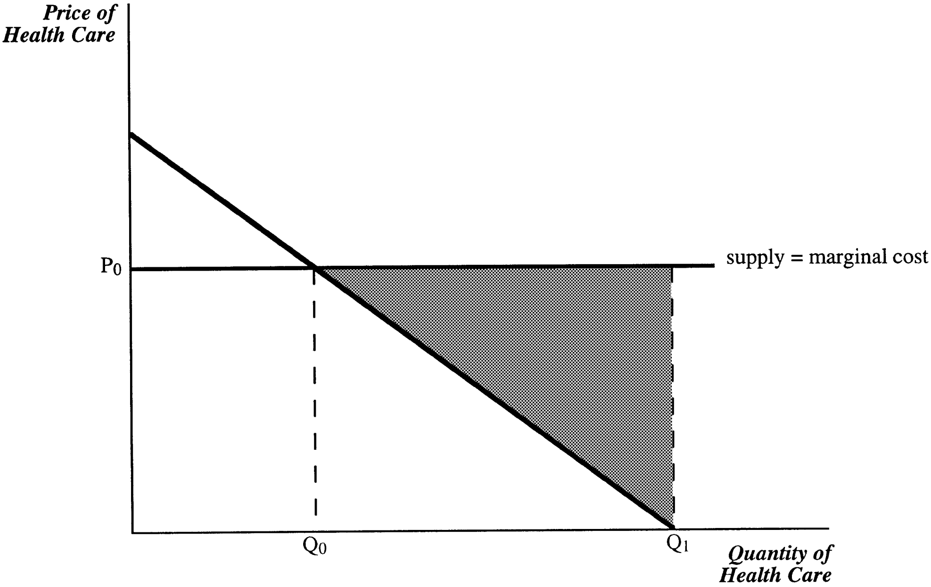


Figure 2. Neo-classical analysis of moral hazard.

absence of insurance. The increased expenditures associated with such moral hazard result from the behavioral responses of either patients or providers: patients, whose care is now subsidized may (and would be expected to) demand a greater quantity of services; providers, knowing that patients do not bear the full cost of services, may increase the quantity of treatments recommended and/or the prices of those services.

Moral hazard has the potential to limit the range of insurance contracts that can be offered, decreasing allocative efficiency. To remain in business, an insurance organization has to set a premium based on *ex post* losses in the presence of insurance, but individuals may make their consumption decisions on the basis of *ex ante* expected losses. Individuals willing to purchase an insurance contract based on *ex ante* losses, find such contracts unavailable. Hence, moral hazard can lead to missing, or at least incomplete markets, for risk-bearing [Evans (1984)].

A second type of allocative efficiency loss arises from the "excess" utilization generated by insurance, which creates an excess burden [Pauly (1968)]. The argument is as follows. Assume the health care market is as depicted in Figure 2. Price P_0, equals the long run (constant) marginal cost of care.[39] In the absence of insurance, Q_0 care will be consumed; under full insurance that provides first-dollar coverage, Q_1 care will be consumed. For each unit of increased consumption under insurance $(Q_1 - Q_0)$, the

[39] The argument does not depend on constant marginal cost, but the elasticity of supply does affect the size of the excess burden associated with a given increase in utilization, *ceteris paribus*.

marginal cost exceeds the marginal benefit, generating an excess burden for the economy. Moral hazard can be eliminated by increasing the price that the consumer faces to P_0, but, of course, this completely eliminates insurance coverage.

This analysis provides the foundation for the argument that optimal insurance coverage must balance the competing welfare consequences of insurance. On the one hand, insurance increases welfare by reducing risk for individuals, with (subject to some caveats) the welfare gain directly related to the extent of coverage. On the other hand, insurance creates a welfare burden through moral hazard. Hence, optimal insurance must balance these competing welfare effects by including patient cost-sharing provisions [Zeckhauser (1970)].

The positive and normative basis of the analysis, however, remains controversial. From a positive perspective, the analysis assumes that health care is produced in a perfectly competitive market by profit-maximizing firms supplying care at a price equal to its long-run marginal cost. Health care, however, is dominated by highly regulated nonprofit and not-only-for-profit providers, so it is not clear that the supply curve represents the true opportunity cost of the resources used to produce the care provided.

Normatively, the analysis rests on an standard welfare interpretation of the demand curve. "Excess" or "inefficient" is defined solely with reference to the market demand derived from preferences backed by willingness-to-pay. Even if one accepts welfarism, as we saw above, informational problems may invalidate the assumption of consumer sovereignty, which in turn invalidates the normative interpretation of the demand curve. Because cost-sharing selectively reduces utilization on the basis of ability and willingness-to-pay rather than on the basis of need for health care, cost-sharing may reduce care that is effective and needed. Hence, from an extra-welfarist perspective, care between Q_0 and Q_1 is not necessarily wasteful or inefficient when viewed against the standards of need and health improvement (indeed, by these standards care to the left of Q_0 may well be inefficient). In fact, studies of cost-sharing demonstrate that it reduces both necessary and unnecessary care [Lohr et al. (1986), Rice (1992), Stoddart et al. (1994)].

Rather than demand-side cost-sharing policies to address moral hazard, an alternative is to intervene on the supply-side to reduce selectively ineffective or inappropriate utilization. Because of their informational advantage, providers are in the best position to judge what utilization cannot be expected to improve health. Such efforts vary from instilling a culture of evidence-based practice, regulatory initiatives associated with managed care (utilization review, pre-authorization programs, and practice guidelines), and designing funding models that attempt to align the incentives of providers with issues of efficiency. In fact, this is much of what health reform in the 1990s has been about.

This "standard" model of insurance may be seriously incomplete as a basis for policy prescriptions in the context of health care markets. Nothing about the model is specific to health care – simply by re-labeling the axes it would be just as suitable for analyzing the welfare improving effects of house insurance, automobile insurance, flight insurance, and so on. The sole effect of insurance in the model is to lower the price of a good that enters individual utility functions directly and that is produced and exchanged in competitive markets through arms-length relationships among well-informed buyers and sellers, all of which we know to be uncharacteristic of most of health care.[40]

Active, interventionist insurers have a much larger impact in the health care market than do insurers in other markets such as housing or automobiles. Under universal house insurance, the proportion of all housing transactions (purchases or renovations/repairs) covered by the house insurance contract is small; the same is true for automobile repair (though a higher proportion of the automobile repairs is probably covered in some way by insurance). For both of these insured goods (and many others), a large proportion of purchases of the insured good happen in the absence of an insured loss, outside any insurance contract. In contrast, the vast majority of health care purchases occur only in the presence of an insurable loss (i.e., ill-health). Hence, insurers and insurance play a much more dominant role in the dynamics of health care markets.

Weisbrod (1991), for example, argues that the static analysis of the welfare loss associated with excess utilization induced by insurance is incomplete. He posits that, in a dynamic analysis, the level and extent of health care insurance and the development of health care technologies are endogenous. Because insurance coverage affects the expected returns to R&D investments in health technology, the spread of insurance is an important factor in explaining the post-war growth of technology. The development of new technologies, however, also affects the demand for health care insurance. Extensive insurance coverage combined with retrospective, cost-based reimbursement encouraged the development of costly technologies that offered minimal increases in quality. The combination may even encourage the development of technologies we would not *collectively* be willing to pay for, inducing potentially negative welfare effects. In contrast, prospective reimbursement, which dominates today in many countries, encourages the development of cost-reducing technologies that have minimal negative effects on aspects of quality that can easily be monitored by patients, but that may have negative effects (especially combined with behavioral incentives facing providers under prospective reimbursement) on aspects of quality not easily monitored by patients. Weisbrod's analysis is, by his own admission, more speculative than definitive, but a key message is simply that at present we do not have well-developed models with which to explore the behavioral and normative aspects of the dynamics between insurance, health care and technological development, though these issues are of crucial importance for the design of health care systems.

Evans (1983) argues that one cannot fruitfully understand the rationale for, or the welfare effect of, universal, first-dollar public insurance using the standard insurance model. The potential welfare implications of universal, first-dollar public insurance (with capital financed separately), such as that found in Canada, can be understood only by simultaneously considering asymmetry of information, the attendant agency relationship and potential for the supply-side to influence resource allocation; externalities; the dynamics between insurance, providers and technological development and diffusion; and

[40] Pauly (2000) analyzes some additional differences between health care markets and other insurance markets.

broader social goals concerning income redistribution (generally favoring redistribution from the healthy wealthy to the sick poor). All of these potential effects of single-payer public insurance fall outside the standard insurance model, and can be understood only in light of the full nature of health care and health care markets.[41]

More generally these analyses highlight why analyzing each feature of health care in isolation provides only limited guidance to policy. Health care is a classic second best world in which one cannot be sure that prescriptions to fix one source of inefficiency, based on models that do not reflect the other distinctive features of health care, will in fact improve resource allocation. Jointly analyzing the features of health care, and the markets for health care and health care insurance in particular, can lead to policy prescriptions quite different than may be derived considering each in isolation.

4. Equity in the health sector

Equity concerns fairness and justice, the idea of balancing legitimate, competing claims of individuals in society in a way that is seen as impartial or disinterested.[42] Distributional equity, which concerns the fair distribution of some good or service of interest, has been the dominant equity concern both of normative economic analysis and of health policy makers. Most of the analytic arguments justifying a focus on distributional equity in the health sector draw on one or more of the following lines of reasoning. Health is a critical component of well-being, a basis for a person's ability to function and, where a person's life is at stake, the ability to achieve anything at all. Ill-health and the need for health care have large random components; people suffer the misfortune of ill-health for reasons beyond their control and should not have to suffer excessively because of fate. Justice therefore dictates that those in ill-health should receive treatment on the basis of their need for care, not on the basis of non-health-related attributes (such as ability-to-pay, as is the case for most commodities).

[41] This is not to say that Evans has provided an unassailable argument supporting such a system of public insurance, or that there are no concepts from the standard insurance framework that are useful in a welfare analysis of such a system; it is simply to say that any welfare analysis that relies primarily on the standard insurance framework will be incomplete.

[42] Equity arguments, which often serve as a basis for redistribution of resources within society, can be distinguished from arguments for redistribution based on caring externalities or even compassion. Each may serve as a legitimate basis for distributional concerns, but in the case of caring externalities, the argument rests on efficiency concerns and the nature of the utility functions. If utility functions are interdependent then efficiency dictates that these interdependencies be taken into account in assessing the optimal allocation of resources. It ultimately rests on the standard economic arguments of respecting preferences and allocating resources efficiently. In the absence of such interdependencies, there is no case for distributional concerns. In contrast, because it is grounded in notions of fairness and justice, equity appeals more explicitly to reasoned arguments about what is right and just, and therefore what ought to be done as a matter of principle. Equity concerns may underlie utility interdependencies, but they need not. See Culyer (1989), van Doorslaer et al. (1993), Dolan (2000), and Williams and Cookson (2000) for a discussion of these distinctions.

Empirically, there is strong support for equity in health care. Van Doorslaer et al. (1993), for instance, found that among the 10 OECD countries included in their analysis, official policy statements place great emphasis on equity both in financing health care and in its use. These statements are backed by extensive efforts by governments to achieve these objectives through public systems of finance, funding and delivery. These policies enjoy extensive support among the public; so much support that governments act contrary to them at their own risk. A growing experimental and survey literature documents the extent to which individuals care about distributional equity in the health sector. When given the task of allocating resources with health-improving effects among individuals who can benefit, rather than allocate resources to maximize the total health benefit generated, respondents consistently opt for allocations that provide for a more equal distribution of the benefits [Yaari and Bar-Hillel (1984), Kahneman and Varey (1991), Nord et al. (1995), Ubel et al. (1996), Ubel and Loewenstein (1996)]. Individuals display a willingness to sacrifice total benefit for a more equitable distribution in the face of trade-offs between total health and its distribution, even when they are in the group who may be "hurt" by the more equal distribution. In a typical result, for example, Kahneman and Varey (1991) found that more than three-quarters of respondents allocated a fixed supply of pain relief medication between two individuals, identical in all respects (including level of pain) except their ability to metabolize the pain medication, so as to equalize the pain experienced by each rather than to maximize the amount of pain relieved. The decisions do not appear to result from a misunderstanding of the effects of alternative allocations – the researchers went to great lengths to ensure that the participants understood the consequences of their decisions. The equity concerns appear to be specific to health-related effects. Yaari and Bar-Hillel (1984) found that participants choose very different allocations for goods perceived to have important health consequences compared to non-health-related goods. Participants chose different allocations for the *same* good depending on whether the good is described as generating important health effects (which creates notions of need) or as simply desired as a consumer good (which is based simply on tastes/preferences).

Agreement on the importance of equity concerns, however, does not translate into agreement on the relevant notion of equity. Sen (1992) has argued that virtually every theory of justice that has withstood reasoned argument has had at its core, the proposition that justice demands equality in the distribution of something (which Sen calls the "focal variable"). Different theories differ on what the focal variable is. The choice of focal variable is critical because, given the diversity of human beings, achieving equality with respect to the focal variable implies inequalities in other dimensions (and in particular, in other, competing focal variables). In the pill example cited above, achieving an equal distribution of pain relief required an unequal distribution of the pills. Horizontal equity calls for equal treatment of equals – those who are similarly situated with respect to the focal variable. Hence, horizontal equity in financing health care may call for those with the same income to pay the same amount; horizontal equity in the allocation of health care resources may call for equal treatment for equal need. Vertical equity calls for unequal treatment of unequals – those who are differentially situated with respect to the focal variable. Specifically, it calls for unequal treatment in accord with the extent to which they are unequal. Vertical equity in finance may call for a person with a higher income to pay greater tax than a person with lower income, just as it might call for greater resources for those with greater needs. Many different focal variables have been proposed for the health sector – access, utilization or expenditure, resources, met need, health, etc. Those that have received the most sustained attention are variants that fall within three broad distributional equity principles: (1) allocation according to need; (2) allocation to ensure equality of access; and (3) allocation to equalize the distribution of health [see Wagstaff and van Doorslaer (2000)].[43]

4.1. Equality of access

Equality of access is an often heard standard for health care [Olsen and Rogers (1991)]. Access has been defined as "freedom or ability to obtain or make use of" [Merriam-Webster (1986)]. Equal access, then, implies that everyone in society is equally able to obtain or make use of health care. It pertains to the ability or capacity to do something, and not to whether it is actually done; it is independent of demand or utilization. Hence, as Olsen and Rogers (1991) and Mooney et al. (1991) emphasize, it can not be assessed by examining consumption patterns: equality of access does not imply equality of consumption. The ethical basis for equality of access does not derive from any necessary relation with its ultimate effects on the distribution of health care or health. It is intimately linked to the notion of equal opportunity or a fair chance. Nord et al. (1995), for example, found in their survey that whatever level of resources are available in a system, people want to know that the system will provide them the same opportunity as all others for treatment.

Equality of access has perhaps more affinity to process notions of equity than to strictly consequentialist notions. The principle has at times been coupled with the notion of need, so that it has been stated as equal access for equal need. Here the principles of horizontal and vertical equity become important, as this does not imply strictly equal access to all health care. Rather, those with equal needs should have equal access to services, while those with unequal needs should have differential access. To cite a common example, access to an emergency room physician is not based purely on a first-come, first-serve basis. Although everyone has access to emergency care, priority is given to those most in need, so that, *ceteris paribus*, the person in cardiac arrest does not have to wait as long as a person with a sprained ankle. The principle may also justify equal access to primary care but unequal access (through the filter of a primary care provider on the basis of need) to higher levels of care.

[43] Space does not allow discussion of broad theories of justice that underlie specific conceptions of equity [e.g., Rawls (1971), Dworkin (1981), Nozick (1974, 1989)]. Nor do I treat some approaches to equity that have received considerable attention in economics (e.g., envy-free allocations [Varian (1974, 1975), Baumol (1986)], rank reversal measures of horizontal equity [Plotnick (1981, 1982)]. See Pereira (1993) for a broader survey of equity principles and theories of justice in health; see also Williams and Cookson (2000).

LeGrand (1982) defined equal access to a good as a situation in which individuals face the same price (both monetary and non-monetary) for the good. This definition has been criticized, however, as falling short of equality of access because two individuals with different income or wealth would have different abilities to pay for a good even if they faced identical (positive) prices. Equal prices are therefore not sufficient to ensure equal "ability to make use of."

LeGrand (1987, 1991) proposed an alternative definition based on the notion that individuals have equal access only if they face the same feasible choice set. This requires that they have the same budget space, where again this is interpreted to include monetary and non-monetary factors. Individuals would have the same opportunities to trade-off different goods and services at the same rate so differences in consumption would reflect nothing but differences in preferences.[44]

Olsen and Rogers (1991) argue that this definition is too broad, that it does not correspond to a common concern for equal access to a particular good whose distribution is of special interest, such as health care. They define equal access to a good as a situation in which everyone is able to consume the same quantity of the good (i.e., the budget constraint with respect to the good in question is identical for everyone). This definition, they argue, is consistent with the literal meaning of access and is not obviously inconsistent with a concern for equal access to a particular good. It allows individuals differential ability to purchase other goods and services and implies that if two people with different incomes choose to consume the same quantity of the good in question, they will not be able to consume the same quantity of other goods.

They develop the welfare implications of this definition using a two-person, two-good model that assumes a linear production possibilities frontier, and that limits government intervention to lump-sum taxes and grants and to per-unit taxes and subsidies on the good for which there is a concern about access. Among other things, they demonstrate that if both individuals care about equal access to one of the goods and the situation in the absence of government action involves unequal access, then: (i) this situation in inefficient; (ii) all states preferred by both individuals to the initial situation involve greater equality of access; (iii) all *efficient* states preferred by everyone to the initial situation involve greater equality of access to the good; (iv) it is not necessarily the case that all efficient states involve greater equality of access to the good; and (v) it is not necessarily the case that all efficient states preferred by everyone to the initial situation involve greater equality in *consumption* of the good.

4.2. Allocation according to need

The idea that health care resources should be allocated in line with health care needs has a strong intuitive appeal. If those most in need are also those who can most benefit from

[44] Note that it does not imply equal rates of trade-offs or sacrifices among goods in terms of utility, which of course depends on preferences. But there is no *a priori* reason to believe that these will be systematically correlated with socio-economic characteristics.

health care, then under the efficiency objective of maximizing health gain, equity and efficiency are not in conflict: the same allocation of resources advances both efficiency and equity [Culyer (1989, 1990)]. Hence, the principle of allocation according to needs had received considerable attention within health economics, particularly among extra-welfarists.

Need is often not explicitly defined, though doing so is obviously essential to judge whether allocations of resources are consistent with the principle. Differing definitions of need can have important effects on what would be judged to be an equitable allocation [Culyer and Wagstaff (1993a)]. Three definitions have received the most attention. The first equates need for health care with ill-health and the degree of need with the severity of illness – those most severely ill have the greatest need. This definition, however, ignores the fact that there may be no effective treatments for some types of ill-health. No matter how ill a person is, if there is no effective treatment there is no need for health care (though there may be a need for other types of care or services).

The second, and perhaps most prevalent, definition of need is strongly consequentialist and centers on effectiveness. It argues first that a need can be defined only with respect to a specific objective: "Y is needed to achieve X." A need exists only when Y has been demonstrated to be effective in achieving X. Finally, within public systems of funding, normally not just any X will do; X must be an objective that the broader community endorses as being meritorious or worthwhile, such that "needs" can be distinguished from mere "wants" [Williams (1978), Culyer and Wagstaff (1993a)]. Some would add the further proviso that Y must not only be effective, but it must also be the cost-effective way to achieve X.

Although this definition establishes when a need exists, it does not establish how much health care is needed. Culyer and Wagstaff (1993a) proposed an alternative definition for need: the expenditure required to effect the maximum possible health improvement, or equivalently, the expenditure required to reduce the individual's capacity to benefit to zero. This definition is consequentialist, because it links need to the outcome (health), and it is quantifiable in a metric that forms a direct link to resource allocation (expenditure). A potential disadvantage is that it conflates two concepts, the extent of need and the amount of resources required to meet that need. By this definition a person suffering from a severe allergic reaction to a bee sting, who requires a simple and inexpensive anti-toxin to prevent sure death, would have less need than a person with a moderate cataract who requires eye surgery to exhaust benefit. Although the latter person needs more health care (as measured by expenditures), would we say that they have a greater need for health care? This distinction becomes relevant when priorities must be set regarding the use of health care resources either at the individual or population level. At the individual level, the principle of triage dictates that the person with the allergic reaction receive priority. An analogous principle holds at the population level when allocating resources among regions if the system is constrained so that each region must leave some needs unmet. One region's needs may be more urgent (or serious) in some sense than another's, even if it requires fewer resources to meet them.

Priority would be given to meeting all of its needs before funding lesser needs in the second region.

Under the principle of allocating resources according to need, horizontal and vertical equity call for equal treatment for equal need, and unequal treatment in proportion to unequal need.[45] Some have objected to this formulation of the principle at the individual level as being too coercive: strictly interpreted it implies that a person should receive health care even if they do not want it [Mooney (1986), Mooney et al. (1991)]. It rides roughshod over personal autonomy, over heterogeneity of preferences for health care and for health improvements.

This principle is widely used, however, at the population level to allocate resources among defined populations on the basis of relative need for care. Within regional systems of governance, for example, the share of the budget allocated by the central authority to each region is commonly based on each region's relative need [Birch et al. (1993), Mays and Bevan (1987)]. The average expenditure on residents of the region therefore corresponds to the need for care in the region (compared to other regions) but no individual is forced to consume care. Heterogeneity may still therefore exist in the extent to which those with the same needs within a region utilize services.[46]

4.3. Equality of health

Culyer and Wagstaff (1993a) have argued that the relevant equity principle is equality of health. Their argument is as follows. Health care is consumed to produce health; that is, for purely instrumental reasons. Hence, the equitable distribution of health care can be judged only in relation to the ultimate good toward which health care is consumed: health. Because good health is necessary for individuals to "flourish," and any position other than one in which everyone has the same opportunity to flourish is hard to defend, a just distribution of health is an equal one. Given that health care is consumed to produce health, it follows that an equitable allocation of health care is that which gives rise to an equal distribution of health. They make two qualifications, however: (1) health care is not the only determinant of health, so it is not expected that health care alone can lead to an equal distribution of health; and (2) equalizing the distribution of health is not to be achieved by intentionally, as an act of policy, reducing the health of some members of society.

[45] A variant of this is the principle of equalization of marginal met need [Steele (1981), Mooney (1986), Culyer (1995b)]. It has been criticized as really being an efficiency criterion, as equalization of marginal met need is a necessary condition for maximizing health in a population.

[46] Whether this is a problem depends on the source of the variation. If the source is on the demand-side, it may not be perceived as a problem (especially if those with equal needs had the opportunity to consume the same services); if it is on the supply-side because of poor system design or performance, it may represent a problem. This highlights the close relation between this principle at the population level and the principle of equality of access – allocation by need at the population level may be an important ingredient in creating equal access.

One way to capture a concern for the distribution of health is through an extra-welfarist social welfare function (SWF) in which the health of the members of society are the arguments [Wagstaff (1991)]. Ideally, the social welfare function should be flexible enough to reflect both the strength of aversion to inequality and allow for different weights to be attached to the health of different members of society. Wagstaff explores the properties of the following SWF:

$$W = (\tau - 1)^{-1}\big[(\alpha h_a)^{1-\tau} + (\beta h_b)^{1-\tau}\big], \quad \tau \neq 1, \tag{4}$$

where W indicates the level of social welfare, h_a and h_b are the levels of health for two individuals A and B, τ indicates the degree of aversion to inequality in the distribution of health between A and B ($\tau > 0$ indicates some aversion to inequality), α indicates the weight attached to A's health and β indicates the weight attached to B's health. This SWF is increasing in the level of health attained by A and B, it accommodates a range of concern for inequality as τ varies ($\tau = 0$ implies lack of concern for inequality; as $\tau \to \infty$ it approaches a Rawlsian SWF in which overall welfare depends only on the health of least healthy individual) and the parameters α and β allow for differential concern for the health of A and B.

The question of differential aggregation weights that reflect differential levels of concern among the population has received considerable conceptual and empirical attention. The question has arisen most forcefully in the context of the economic evaluation of health care interventions and programs, where the effects of the intervention must be aggregated across affected individuals. The standard methods for doing so (see Section 5 below) call for equal weights for each individual (total benefits is an unweighted sum), which ignores any distributional issues. Society, however, may care about who is affected by a program. To the extent that these distributional concerns are linked to observable characteristics of individuals, a system of differential aggregation weights may be able to reflect theses concerns, an idea that can be traced back at least to Weisbrod (1968). Harberger (1971) argued vigorously against such weights in the context of cost-benefit analysis, arguing that any set of weights would be arbitrary. (His own proposal for unitary weights is as arbitrary as unequal weights.) Strictly speaking, such weights also do not fit easily into the welfare economic framework, which calls solely for utility information when ranking resource allocations. In contrast, the notion of weights linked to the characteristics of an individual (current health status, age, income, etc.) fits easily into the extra-welfarist approach. Hence, they have received considerable attention among extra-welfarists. Culyer (1989) argued that through such weights, one could reconcile efficiency and distributional equity concerns by allocating resources so as to maximize the weighted sum of health in society.[47]

Harberger's basic question still stands, of course: On what basis can such weights be justified and how can they be estimated? A number of approaches have been suggested.

[47] He has since modified his views, emphasizing the importance of equality of the final distribution of health as discussed above.

One is to base the weights on the preferences or values of members of society. That is, elicit the α's and β's from the general public. This is consistent with the economic tradition of respecting individual preferences. Research in this vein has focused on eliciting the value individuals place on producing health among individuals in different age and occupational categories. Findings consistently give priority to individuals responsible for children and the young [Williams (1988), Charney (1989), Nord et al. (1995)]. Such efforts are exploratory at this stage, but this approach runs up against the well-established problem of building a social welfare function from individual preferences in the face of heterogeneity of preferences and preferences that might be judged to be morally repugnant.[48]

Murray and Lopez (1996) took a different tack in developing disability-adjusted life-years for use in estimating the level of health in a country. They based their aggregation weights on the expected productivity of a member of society. Hence, working-age individuals received the highest weight and the elderly and children receive lower weights. Many object to such weights because they link the value to society exclusively to a person's economic productivity and appear motivated more by efficiency concerns than equity concerns.

Finally, Williams has recently proposed basing the weights on an ethical principle he terms the "fair-innings" approach [Williams (1997)]. The fair-innings approach is based on the premise that everyone in society is entitled to some "normal" span of health. Those who fall short of this have been "cheated" in some sense, and those who exceed it are living on borrowed time. Hence, the "normal" span might be taken as quality-adjusted life-expectancy. This principle can be used to derive weights to be attached to generating health (quality-adjusted life-years) for individuals at different stages of their life [see Williams (1997) for an attempt to do this].

These represent only three justifications for a system of weights (social preferences, economic productivity, and derivation from an ethical principle). Others are obviously possible. The critical point is the imperative to assess the source and rationale of any weights that are used. Unequal weights are often motivated by a concern for equality in a particular dimension. In the fair-innings approach, the unequal weights are motivated by a desire that individuals be given a chance for an equal amount of health over the lifetime (which is one possible interpretation of Culyer and Wagstaff's call for an equal distribution of health). In the previously discussed pill experiment, the unequal distribution of pills was motivated by a desire for an equal distribution of pain relief across individuals. Hence, unequal weights can be motivated by egalitarianism in a different domain and any justification for the weights must be made on the basis of egalitarianism in this other domain.

[48] In empirical work the mean or some other measure of central tendency would likely be used in a social welfare function, which would ignore heterogeneity. Still, the mean is a summary measure based on the distribution of individual preferences.

4.4. Rival notions of equity

Given that striving for equality in one focal variable necessarily means tolerating inequality in rival variables, it is perhaps not surprising that the above equity principles are mutually incompatible: each of them would lead to a different distribution of health care resources [Culyer and Wagstaff (1993a), Culyer (1995b)]. Policy making therefore requires that we choose among them. Unfortunately, there is no scientific basis for choosing among them – they are, by definition, normative principles.

Each of the equity principles articulated strives in a sense to be a general, universalistic principle to guide resource allocation throughout a health care system. This raises a more fundamental issue: must we choose a single, over-arching equity principle, and if we do, what does it mean to do so? Resources are allocated through myriad decisions in a multiplicity of contexts throughout the health care system, ranging from cabinet decisions at the national level, through the deliberations of regional and institutional boards, all the way down to each individual clinical encounter. If coherence demands that at the population level the system have a single over-arching equity goal, what does it demand at these other levels, where a host of contingent factors bear directly on the (real and perceived) claims of specific individuals or groups that impinge on allocation decisions? One of the strongest and most consistent messages from the empirical research on moral and ethical reasoning of people is the context-specific nature of such judgements [Walzer (1982), Yaari and Bar-Hillel (1984), Elster (1992), Miller (1992), Mannix et al. (1995)].

As one changes decision contexts, factors beyond distribution emerge such as notions of procedural fairness, duty, obligation, due process, informed consent, non-coercion, or rule of rescue. An equitable or just allocation is one that conforms to the relevant principle. Regardless of the health impact, for example, a person can not be coerced into receiving medical treatment, except in the most extreme cases involving serious public risks. Public hospitals have an obligation (subject to capacity constraints) to treat all those in need who walk in their doors. Because these principles are often posed as ethical imperatives regarding the behaviour of individuals or organizations within the health care system, to the extent that they enter economic analysis, they often enter as constraints in the choice problem. They can often, therefore, be important for understanding behaviour.[49] They have received less attention in normative economic analyses because they tend to apply at lower levels of decision making rather than with respect to system-level issues and because they do not conform well to consequentialist reasoning.

Although the work of economists analyzing equity at a conceptual level cannot, by definition, provide guidance as to what equity principle(s) should guide decisions in a given context, health economists have a vital role to play in explicating the differences among the principles, identifying the implications of alternative principles, and demonstrating the relationship among them. Although today there is no greater consensus than

[49] Recall the extensive effort noted above to develop models of physician behaviour that reflect the ethical dimensions of a physician's professionalization.

before on the appropriate equity principle to guide allocation in the health sector, the work of health economists has advanced the discussion considerably by carefully analyzing the rival conceptions of equity.

5. Evaluation of programs and interventions

This section shifts focus from normative analysis of system-level issues (e.g., such as the operation of health care and health care insurance markets) to the normative economic analysis of individual services, interventions and programs.[50] The development and application of such "methods of economic evaluation," as they are commonly called, comprise a large part of health economics because reliance on non-market allocation mechanisms generates extensive need for the explicit evaluation of the efficiency and equity effects of policies, programs and services. The role of evidence produced by such evaluations, for example, has figured prominently in the methods proposed for setting priorities for resource allocation within health systems [e.g., Oregon Health Services Commission (1991), Fox and Lichter (1993), Coast et al. (1966), Maynard and Bloor (1998)].

The different methods derive from the two normative frameworks emphasized thus far – neo-classical welfare theory and extra-welfarism – which posit different ways to measure, value and aggregate the costs and consequences. Detailed discussion of the methods of economic evaluation are contained in a number of chapters in this Handbook, including Dolan (2000) and Garber (2000). This section outlines some of the important questions/controversies in the development of such methods in recent decades and highlights how these developments relate to broader developments in normative analysis in the health sector.

Paretian welfare economic theory provides the conceptual foundation for the methods of cost-benefit analysis, although it must be recognized that there are more gaps than is commonly appreciated between the foundation of formal Paretian welfare economics and the edifice of empirical cost-benefit analysis.[51] Critical elements of cost-benefit analysis drawn from welfare economic theory include the centrality of individual utility (preferences) in valuing resource allocations and the proposition that under certain assumptions regarding the nature of individual utility functions for members of society: (1) utility can be measured in a money metric (compensating variation, equivalent variation, consumer's surplus); (2) such monetary measures can be summed across individuals to obtain an aggregate benefit measure; (3) the sign of this aggregate benefit measure can indicate whether the hypothetical compensation test is passed (indicating a potential Pareto improvement); and (4) that a potential Pareto improvement represents an increase in welfare. The set of assumptions required for this chain to hold together is

[50] The focus is on the evaluation of "small" programs and interventions, where small means that they are not expected to generate changes in price and the analysis does not therefore require general equilibrium approaches.

[51] See, for example, Blackorby and Donaldson (1990).

large and includes assumptions unlikely to be met in the real world, especially for the types of interventions being evaluated in the health sector, which generate a wide range of effects beyond price and income effects.[52] Nonetheless, welfare economic theory provides the intellectual pretext for the practice of assessing programs and services by measuring the costs and benefits in monetary units, calculating the net benefit (benefits-costs), and ranking the allocative efficiency of those programs and services on the basis of net benefit.

A second tradition, emanating from the decision sciences and systems analysis, emphasizes assessments of technical and cost-effectiveness efficiency, and is exemplified by cost-effectiveness analysis and cost-utility analysis. In cost-effectiveness analysis, costs are measured in monetary units but the benefits are measured in natural units of outcome for the programs being evaluated. In the health sector, these may be life-years gained, cases prevented, cases detected, etc., depending on the nature of the intervention. The result is summarized in a cost-effectiveness ratio, which represents additional cost per additional unit of outcome achieved. In cost-utility analysis, costs are again measured in monetary units, but the outcome is measured in terms of quality-adjusted life-years (QALYs) that reflect both the quantity and quality of life years gained as a result of the intervention. As in cost-effectiveness analysis, the results are summarized by a ratio that indicates the cost per QALY achieved.[53]

Although not initially developed in reaction against cost-benefit analysis, cost-effectiveness and cost-utility analysis were embraced by economists within the health sector because of the difficulties (conceptual, ethical, practical) in monetarily valuing life-years gained, as well as by extra-welfarists who emphasize health as the primary outcome for normative analysis in the health sector. Their ancestry within decision science, and its emphasis on seeking the best way to achieve an objective defined by those who commission the analysis, also spurred the development of the decision-maker approach to economic evaluation (see Section 2 above).

Adherents of both welfare economic and extra-welfarist approaches agree that opportunity cost, not accounting or financial cost, is the relevant cost concept for economic evaluations. Hence, at a conceptual level, the cost side has not been a source of controversy. There are, of course, a number of problems associated with empirically estimating opportunity costs. Because health care markets are generally heavily regulated and decidedly non-competitive, prices cannot be assumed to represent opportunity cost. Shadow pricing is therefore often necessary though seldom straightforward.[54]

[52] The assumptions can be found in Boadway and Bruce (1984) or similar welfare economic texts.

[53] Some consider cost-utility analysis as a special case of cost-effectiveness analysis and therefore refer to both as cost-effectiveness analysis. Because constructing a QALY as part of a cost-utility analysis involves valuation of a health outcome (unlike CEA – more on this below), and because recent work has attempted to provide a welfare economic foundation to cost-utility analysis (but not CEA – again, more on this below), I retain the distinctive labels.

[54] There is one context in which welfarists and extra-welfarists may differ in the approach to opportunity cost. Extra-welfarists accept price as measuring the opportunity costs of a resource because it represents

There has been long-standing controversy regarding the inclusion of certain costs. Debate has recently erupted, for example, as to whether treatment costs incurred during the additional life years attributed to an intervention should be included as a cost of the intervention (Klarman (1982) discusses early work on this question; for a more recent discussion, see Johannesson and Meltzer (1998)), or whether productivity gains should be included as a negative cost in cost-effectiveness analyses [Williams (1981), Johannesson and Meltzer (1998)]. Including such productivity effects would reflect a person's earnings. This raises some difficult issues for extra-welfarists who advocate for the use of cost-effectiveness and cost-utility analysis in an attempt to avoid linking the evaluation of health interventions to individuals' economic resources. With the exception of some issues like these, however, costing has not been the focus of the most intensive methodological developments and of sustained controversies in the economic evaluation of health care programs and interventions.

Far more contentious, and the locus of much more methodological development, has been the outcome, or benefit side of the equation, where vigorous debate continues regarding the outcomes to be included, how they are to be measured, valued and aggregated, and the nature of the social welfare function. A defining element in the historical development of outcome measures is the fact that the primary "output" of many health care interventions is life-years, and in particular, life years of varying quality. How should one value the extension of a person's existence, without which nothing else is possible and which cannot be traded (intra-personally over time or interpersonally among individuals)? Health economists have led in the development of methods for valuing life-years gained. Within the welfarist tradition methods have been developed to value life in terms of monetary units and through Paretian, non-monetary outcome measures that reflect individual preferences over both the quantity and quality of life gained. Within the extra-welfarist approach methods have been developed to value life using non-Paretian subjective health measures that reflect the quantity and quality of life-years gained. Advance, particularly in the non-monetary measures, have arisen from extensive collaborations between health economists and researchers from other disciplines (e.g., psychology, decision science, medical science, statistics, epidemiology) and have been influential in a broad range of evaluative health research (e.g., clinical trials).

The work of health economist pioneers such Weisbrod (1961) and Fein (1958) was solidly within the cost-benefit approach then being developed within welfare economics. The value of life years gained was assessed using the human capital approach. The economic value of additional life years was the value of the economic production associated with those years – the expected increments in earnings of those whose lives had been extended. The human capital approach is biased towards those who work in

the resource's value in its next best use in the economy overall (i.e., outside the health sector), for which willingness-to-pay is an appropriate measure. In contexts, however, where the opportunity cost is borne fully within the health sector, consistency within extra-welfarism demands that opportunity cost be assessed in terms of health itself rather than monetary terms.

market settings and who earn high wages (in western societies, generally white, middle-aged males); it discriminates against those not in the workforce or those who receive lower wages (the elderly, women, children). Modifications such as imputing the value of housework partially address these concerns, but run up against the more fundamental objection (often from outside economics, but also from within) that is not appropriate to link the value of additional life years exclusively to economic production (whether market or non-market).[55]

The fatal blow to the human capital approach, however, came from within welfare economic theory. Schelling (1968) distinguished the value of a livelihood, which the human capital approach measured, from the value of a statistical life, which can be measured by the amount an individual is willing to pay to achieve a specified reduction in the probability of death. The value of a statistical life, he argued, is the relevant measure of economic benefit for programs that "saved lives." His insights represented two important advances. The first is that the outcome measure should reflect the probabilistic nature of the outcomes. Health interventions generally affect the probability that an individual will die during a given period, and therefore their benefits should indicate the value an individual places on this reduced probability of death. The second is aligning the valuation of life-years gained more clearly with welfare economic theory, for which the relevant measure of benefit is a person's willingness-to-pay. These insights were formalized by Mishan (1971), who demonstrated that this is the only measure "of life and limb" consistent with Paretian welfare theory.

These advances were, ironically, a mixed blessing for applied cost-benefit analysis in the health sector. On the one hand, they clarified and firmly established the theoretically "correct" approach to assessing the value of life for cost-benefit studies within the Paretian welfare framework. On the other hand, they presented an obstacle for applied cost-benefit analysis because it was not clear how one could measure willingness-to-pay for reductions in the probability of death. Cost-benefit analysis generally estimates willingness to pay by the area under a demand curve. But there are no markets in which individuals trade chances of death and hence no relevant demand curves. Economists attempted to ascertain such values indirectly from contexts such as the labor market, in which individuals may voluntarily accept jobs with greater risk of death in return for a wage premium [Viscusi (1992, 1993) provide recent reviews]. Such estimates, however, suffer from being based on several strong and, to some implausible, assumptions regarding the competitiveness of labor markets, the information workers have regarding job-related risks, and the extent to which other (often unobservable) job characteristics influence wage levels. The relevance of such estimates for the health sector is also limited by the fact that they reflect

[55] Many non-economists, particularly those in the health professions, argued that it was impossible to put a dollar value on a life, that a life was "priceless." Regardless of whether such individuals truly understood what economists were striving to capture (many did not), because the users of the results of cost-benefit studies were usually not economists (particularly in the early days), such arguments have had considerable influence in how health economists have approached the task of valuing life-years gained.

only the types of mortality and morbidity associated with work, which often do not correspond directly to those associated with health interventions.

More recently monetary valuation of health outcomes has shifted toward contingent valuation methods, which employ hypothetical scenarios to elicit stated preferences (rather than preferences revealed through actual choices). Contingent valuation requires that the health effects associated with a health care intervention be described to individuals and that they imagine there is a market for these effects. It then elicits how much the individual would be willing to pay to obtain them. Operationalizing this requires a host of assumptions and decisions regarding the outcome being valued (e.g., health only, health and non-health benefits, benefits measured under certainty or uncertainty, etc.), and the specific methods employed to elicit willingness-to-pay. The exact design chosen can have important influences on the values obtained, and much of the current work on contingent valuation is to understand better the effects of alternative designs on the values elicited. Recent discussion of many of these issues can be found in Jones-Lee (1989), Gafni (1991), Kahneman and Knetsch (1992), Johannesson (1996), O'Brien and Gafni (1996), Johansson (1996), Drummond et al. (1997).

Qualms regarding monetary measures in the health sector led many health economists away from cost-benefit analysis to cost-effectiveness analysis [Klarman (1982)]. From an ethical point of view, CEA provided for systematic analysis and planning while obviating the need to assess benefits in monetary terms. This was seen as particularly consistent with the objectives of many public health care programs, whose primary purpose was to ensure access to needed services for all. At a pragmatic level, it avoided the thorny problem of trying to measure the economic value using willingness-to-pay. This made cost-effectiveness studies easier and less costly to carry out than cost-benefit studies.[56] By measuring the effects in natural units, the results could be more intuitively understood by non-economists, particularly individuals with medical backgrounds, who were often responsible for using the results in making decisions. Finally, cost-effectiveness analysis also accorded more closely with the common understanding of efficiency, which is to get the most out of the resources used. Although it was recognized early that CEA could not address questions of allocative efficiency, most of those making spending decisions were not wont to question whether improved health was a worthy objective.[57]

CEA also carried some important disadvantages. Because different programs can generate different health effects, measuring consequences in natural units limits cross-program comparisons. Only programs that generate identical outcomes can be compared. It is not possible, for example, to compare the efficiency of allocating resources

[56] Klarman (1982) notes that one of the ironic effects of Acton's early attempt to apply the advances of Schelling and Mishan [Acton (1973)] was to demonstrate that one could travel much of the desired distance with CEA while avoiding much of the most difficult terrain required to do a full cost-benefit study, convincing many (even those who did not particularly reject the welfare roots of CBA) that CEA was the way to go.

[57] In this respect, the results of CEA do not carry the normative implications often attributed to them. CEA can never address the question of whether an objective is worth seeking; it can only identify the lowest-cost way to attain an objective. A low cost-effectiveness ratio does not imply that something is worth doing.

to a program to treat ulcers, for which this outcome is "ulcers healed" with a program to provide coronary or lung bypass surgery, for which the outcome is life-years gained. In addition, many programs generate multiple effects, intended and unintended, positive and negative. Because each effect is measured in (different) natural units, one must select a primary outcome for the analysis, limiting the ability to consider all the effects simultaneously.

The Quality-Adjusted Life-year (QALY) was intended to overcome some of these deficiencies by capturing a health care intervention's effects on both the quantity and quality of life. The concept of a QALY appears to have been first raised by Klarman et al. (1968) in his economic evaluation of renal dialysis but formal work on developing the QALY as an outcome measure occurred independently in the 1970s in the US [Fanshel and Bush (1970), Weinstein and Stason (1977)], Canada [Torrance et al. (1972)], and the UK [Rosser and Kind (1978)]. See Dolan (2000) for a fuller discussion of QALYs. The QALYs associated with a health profile that consists of a series of health states between now and death, can be written:

$$\text{QALYs} = \sum w_h * t_h, \tag{5}$$

where h indexes different health states, w_h is a quality weight associated with each health state (normally scaled so that death equals 0.0 and perfect health equals 1.0), and t_h is the length of time spent in each health state. Hence, the QALY represents the number of years in full health that is equivalent to an actual health profile that includes periods of less than full health. Suppose a 50 year-old individual faced the following health profile following an intervention: total life expectancy of 20 years where the first 8 years are characterized by an ability to function normally but with chronic and persistent pain, the next 9 years in a wheel chair in chronic and persistent pain, and the final 3 years in pain restricted to bed in an institution. Assuming the quality weights associated with each health state are 0.8, 0.6 and 0.3 respectively, then the QALYs associated with this 20-year health profile are $0.8 * 8 + 0.6 * 9 + 0.3 * 3 = 12.7$.

Because a QALY is a general health measure that captures changes in both the quality of life (morbidity), as well as quantity of life (mortality), it can serve as the outcome measure for a wide range of health interventions (any health intervention that can be linked to a final health outcome). This allows direct comparison across a variety of health programs and interventions. Provided that they are obtained from individuals, the weights incorporate the value individuals as a group place on different health states. The value is not directly linked to a person's economic resources.[58] Hence, when used in cost-utility analysis, QALY-based outcome measures are able to capture a wide array

[58] The values may be indirectly linked if the level of a person's economic resources influences the value they place on health. For a further discussion of this point, and other ways that the level of a person's economic resources may influence economic evaluations that use non-monetary outcome measures see Donaldson, Birch and Gafni (1998).

of health effects, are broadly comparable across a wide array of health programs, explicitly value the health outcome and have no direct dependence on a person's economic resources.

Within the genus of QALYs there are many species, each distinguished by the methods used to estimate the weights. Unfortunately, although some go by different labels (e.g., Disability-adjusted Life-years (DALYs) [see Murray and Lopez (1996)]; Euro-Qual [The EuroQol Group (1990)], in many cases the differences can be ascertained only by investigating the methods used to estimate the weights. Two species that are important to distinguish are those whose weights are estimated using psychometric principles and those whose weights are estimated using utility theory. Psychometrically estimated weights are often derived in the context of certainty, in which an individual is asked to rate how they would value being in a particular health state on a scale with designated anchors for death and perfect health [e.g., Rosser and Kind (1978)]. Utility-based approaches use choice-based exercises in the context of uncertainty to elicit the von Neumann–Morgenstern (vNM) utilities associated with each health state.[59] Because psychometrically based weights are not linked in any way to utility, they are by definition extra-welfarist. QALYs constructed using utility weights, however, have been variously interpreted as preference-based measures of subjective health (or health-related quality of life) or as utilities themselves, depending on the assumptions one makes regarding the nature of individuals' utility functions. That is, although the weights are preference-based utilities, the QALY itself is a utility score that would accurately represent preferences over health states only under quite restrictive assumptions regarding the utility function,[60] assumptions that are known to be commonly violated in the real world. Hence, utility-based QALYs are used as both extra-welfarist measures of health and as utility measures within the welfarists tradition, a point we will return to below. Because they are preference-based, derived under uncertainty from tradeoffs (even if hypothetical), and can be derived rigorously from axioms of rational behaviour, economists have tended to favor QALYs constructed using utility weights.

The potential for the QALY measure to represent individual preferences over health states inaccurately has led to the development, within the general framework of quality-adjusted life years, of non-monetary, Paretian outcome measures that are intended to represent patient preferences over health states accurately under less restrictive assumptions regarding the utility function. The most prominent example is the healthy-year equivalent [Mehrez and Gafni (1989)]. Once again assume an individual faces a particular lifetime health profile, Q_T. Let $U(Q_T)$ represent the vNM utility function over the lifetime health profile for T years. The healthy years equivalent (H) of the lifetime

[59] See Shoemaker (1982) and Machina (1987) for the axioms that underlie vNM – Neuman–Morgenstern utility functions as well as a discussion of the empirical evidence regarding the extent to which individuals commonly violate both axioms. Feeny and Torrance (1989a) provide a discussion of the vNM utilities in the context of QALYs.

[60] Constant proportional trade-off and risk neutrality [see Pliskin, Shepard and Weinstein (1980), Loomes and McKenzie (1989), Feeny and Torrance (1989a), Johannesson et al. (1993), Drummond et al. (1997)].

health profile Q_T is the number of years such that the utility of living in full health (Q_H) for that number of years just equals the utility associated with the lifetime profile Q_T. That is, the H such that $U(Q_H, H) = U(Q_T, T)$. Because the measure stems directly from the utility function and conceptually makes no particular assumptions regarding the nature of the utility function, it is argued that the health years equivalent (HYE) accurately represents individual preferences over health states while retaining an intuitively meaningful outcome measure akin to the quality-adjusted life-year.

Since the HYE was first proposed as an alternative Paretian outcome measure to the QALY, a large literature has grown assessing the properties of the HYE, particularly under the measurement method proposed by Mehrez and Gafni (1989), and the relationship between a utility-based QALY and the HYE [see, e.g., Johannesson, Pliskin and Weinstein (1993), Mehrez and Gafni (1993), Bleichrodt (1995), Johannesson (1995), Culyer and Wagstaff (1993b), Buckingham (1993), Gafni and Birch (1993)]. For details of that debate, consult the original literature and summaries provided in recent overviews such as Drummond et al. (1997).

At times the debate proceeds as if the two measures were intended to measure the same construct and, in particular, that they are both intended to represent preferences, over health states. But they are not necessarily so intended: the QALY is intended by many as a subjective measure of health and the rationale for its use in normative economic analysis emanates directly from an extra-welfarist perspective. As Culyer has argued, although it uses utility theory to inform its construction (which he sees as a strength), for extra-welfarists a QALY it is not meant to be a utility score itself [Culyer (1989)]. Hence, the fact that it does not map perfectly with preferences is not necessarily a flaw. In contrast, the HYE emanates directly from a Paretian framework, attempting to retain the centrality of preferences within a non-monetary outcome measure that can be used in cost-utility analysis.

There has been considerable recent effort to discover whether cost-utility analysis can be given a Paretian welfare economic foundation. Phelps and Mushlin (1991) argued for the near equivalence of CBA and CEA on the basis that each requires one to place a value on life-years gained. The former does it as part of the analysis (using either human capital, risk studies, or contingent valuation) while the latter does it at the end of the analysis when it must be decided if a particular cost-per-life-year-gained is acceptable. But this superficial similarity masks deeply different philosophical bases. The individualistic foundation of CBA calls for eliciting the amount *each individual* is willing-to-pay for a health gain. Two programs that produced the same health gains in two populations that differ only with respect to their wealth and income could be judged efficient [(net-benefit) $>$ 0] in the wealthier group (who would have greater ability and willingness-to-pay) and not efficient in the poorer group. In contrast, to rank programs CEA relies on a *social* judgement as to the willingness-to-pay for a given health outcome. The two programs that serve the two distinct populations in the above scenario would be judged identically. Extra-welfarists do not deny that society must make tradeoffs that place a value on heath gains; they argue that such tradeoffs and judgements should be done at the societal level rather than the individual level.

Garber and colleagues have explored the possibility of building a welfare-economic foundation for cost-utility analysis (CUA) [Garber and Phelps (1997), Garber et al. (1996)].[61] They demonstrate that if one restricts attention to variants of CUA that use the QALY as the measure of outcome, assumes individual utility functions of the type such that a QALY is also a utility score, and assumes that individual utility in a given period depends only on utility derived from health-related quality of life and utility derived from material consumption, then it is possible to build a welfare-economic foundation for CUA such that basing decisions on individual-level CU ratios is equivalent to applying the Kaldor–Hicks criteria for potential Pareto improvements. That is, CBA and CUA are equivalent. A crucial element in this finding is that the "CUA threshold" (i.e., the dollar value of cost-per-QALY that determines whether an intervention is acceptable) must be allowed to vary among individuals. In particular, the threshold must reflect the fact the wealthy would in general have a higher threshold than the poor because the wealthy are willing to sacrifice a greater absolute level of material wealth (though not necessarily utility) for a given health improvement. Hence, we are back to individual willingness-to-pay which, of course, is the basis for CBA. (See Garber (2000) for a fuller explication of these ideas.)

These developments highlight the current state of ambiguity regarding the normative framework informing the evaluation methods commonly applied in the health sector. It is ironic that the effort to clarify the relationship among the various outcome measures has, in many respects, created greater confusion for the average user and interpreter of study results. It turns out that a QALY is *not* just a QALY. Sometimes it is a utility score and sometimes it is a measure of health (health-related quality-of-life) and the interpretation is in the eyes of the beholder, depending on what assumption one is willing to make. Although some methods are unambiguously derived from welfarist or extra-welfarist approaches, others are not: CBA is clearly derived from welfare theoretic foundations; CEA that does not use QALY- or HYE-based measures of effectiveness or which uses QALYs incorporating non-utility weights are clearly extra-welfarist. For a CUA that uses utility-based QALYs as the measure of outcome, which constitutes a large and growing share of economic evaluations, there is nothing observable about the methods themselves to indicate whether the analysis is intended to be welfarist or extra-welfarist – it all depends on the assumptions one is willing to make about the nature of the utility function. Nothing observable distinguishes which approach is being invoked.

5.1. Equity and the methods of economic evaluation

Although the methods of economic evaluation have historically been intended primarily to assess the efficiency of alternative health care interventions, they embody a number of assumptions and procedures that have equity implications. So questions have arisen

[61] The titles of their work refers to CEA, but in fact they are concerned only with CUA that uses utility-based QALYs.

regarding the correspondence between the equity principles articulated for the health care system (a number of which were discussed in Section 6) and the equity principles embodied in the methods of economic evaluation. Critical attention has focused most closely on the methods for measuring and valuing outcomes, the methods of aggregation employed (both over time and over people) and the associated maximizing decision criterion.

The most obvious example of an equity principle embodied within a measurement technique is willingness-to-pay, which links the value of a health effect to a person's economic resources. This point has been emphasized throughout this chapter and so there is no need to elaborate on it here. But non-monetary measures, such as the QALY, which were developed in part to avoid monetary valuation, incorporate their own equity assumptions. The techniques used to measure the utility weights for QALYs, for example, are intended to reflect egalitarianism in the health domain in the sense that, "... the difference in utility between being dead and being healthy is set equal across people... that is, each person's health is counted equally" [Torrance (1986, p. 17)] or, in a reformulation of this, "... the utility of a full healthy life from birth is set equal for each individual" [Feeny and Torrance (1989b, p. 193)]. In actual practice, the way in which analysts empirically estimate utilities often violates this assumption, so that utilities are not consistently scaled, either across individuals within a study or across studies [Gafni and Birch (1991)]. Bleichrodt (1997) has recently shown that consistently scaled utilities are a necessary condition for incorporating distributional equity principles into aggregation procedures.

Aggregation methods inherently contain distributional equity principles. Because health care programs often generate effects into the future, the methods of aggregating effects that occur at different times embody intergenerational equity principles. Standard methods of economic evaluation [see, e.g., Drummond et al. (1997)] call for discounting the costs and consequences that occur in the future. The rate of discount chosen implies a value to be placed on costs and benefits that accrue to future generations compared to those presently living and hence carries with it implications for intergenerational equity. Two philosophical bases for selecting the discount rate are notable: the argument that the discount rate should equal the market rate of interest, as this represents the opportunity cost of capital diverted from private investment to public investment; and the argument, based on individual sovereignty, that the discount rate should equal the social rate of time preference of members of society at the time the public investment is undertaken. Contemporary economists tend toward the latter argument, though the question continues to be much disputed. Robinson (1990) provides a succinct account of the arguments regarding the discount rate.

The simple unweighted aggregation of QALYs (or monetary units) which underlies the maximization criterion of CUA and CBA has strong equity consequences. On the one hand, unweighted aggregation is argued to be egalitarian because each person's valuation has equal weight [Williams (1985)]. Although there is a sense in which everyone is treated equally, it is more true to say that each QALY (or unit of money) gained is treated equally. On the other hand, the (unweighted) maximizing criterion focuses only on the total amount of the outcome and not on its distribution. An intervention that generates 1000 QALYs is judged the same whether it does this by generating 20 QALYs for 5 people, 5 QALYs for 20 people, 0.5 QALYs for 2000 people, or 0.05 QALYs for 20,000 people. Yet, intuition suggests that we would judge quite differently an intervention that had a large impact on the length and quality of a small number of individuals' lives and an intervention that had a negligible impact on the lives of many.

Closely allied with unweighted aggregation is the principle of anonymity: that it does not matter who gains or loses – that a life year-gained, for example, is the same no matter to whom it accrues. Anonymity can be argued to be fair because it is impartial. To the extent, however, that distributional equity demands recognition of the differing moral claims of individuals to health care, which may be linked to their characteristics, anonymity impedes distributional equity. As we saw in Section 4, distributional equity may call for differential weights attached to health benefits that accrue to individuals based on identifiable characteristics (age, family status, etc.).

Bleichrodt (1997) provides perhaps the most rigorous treatment of the scope for formally incorporating equity considerations into cost-utility studies. For the standard unweighted QALY maximization procedure to have meaning, utility must be cardinally fully measurable (i.e., ratio scale) with fully consistent scaling across individuals to allow for interpersonal comparability (e.g., utility of full health life is the same for everyone). In addition, four conditions must hold: (1) individual preferences satisfy vNM axioms; (2) social preferences satisfy vNM axioms; (3) anonymity; and (4) a condition that, if from the standpoint of every individual, two alternative QALY allocations are judged to be indifferent, then they are also indifferent from a social point of view.

He then demonstrates that to incorporate either *ex ante* equity (which concerns the fairness of the process of resource allocation) or *ex post* equity (which concerns the final distribution of the outcome of interest), condition (4) must be relaxed. Consider the following simple example from Bleichrodt, in which there are two individuals, three interventions, and two possible outcomes (X and Y) associated with each intervention, each with probability of 0.5 of occurring (Table 2). Under simple utility maximization, all three programs would be judged to be equal; there would be social indifference among them. If we are concerned with fair process we might prefer either program 2 or 3 over program 1, as each of them at least gives person 2 a chance at benefit while program 1 does not even provide such a chance. This *ex ante* equity consideration can be incorporated by replacing condition 4 with a condition that states that when two individuals taken together have the same probability of receiving a particular outcome in each of two programs (e.g., programs 1 and 2), a preference is given to the program in which the probabilities are more equally distributed between the two individuals.

On the other hand, program 3 may be preferred to program 2 because the programs have the same expected outcome but program 3 guarantees an equal distribution of the heath outcome. Again, this can be incorporated by replacing condition 4 with one whereby a preference is given to programs in which the outcomes are more equal between affected individuals. More generally, to incorporate *ex post* equity considerations, Bleichrodt shows that social choice must depend on more than just individual prefer-

Table 2

Programs	State X	State Y	Expected utility
1	(1,0)	(1,0)	1
2	(1,0)	(0,1)	1
3	(1,1)	(0,0)	1

ences; it must allow for complementarity between individual outcomes. Multi-attribute utility approaches, with choice functions with attributes reflecting efficiency concerns (the total amount of health produced) and equity concerns (*ex ante* or *ex post*) are one potential way forward. This endeavor starts to move beyond examining the equity assumptions embodied in the methods of economic evaluation to specifying the social welfare function that is to guide choice.

Although the results of economic evaluations can support allocation decisions based on both efficiency and equity criteria, in most cases the results of economic evaluations will have no direct implication for whether a program or service conforms to particular equity principles such as those discussed in Section 4. The results of economic evaluations, for example, say nothing about how to achieve equality of access, but only about the interventions to which there ought to be equal access. Although such evaluations provide necessary information for making allocations to achieve a more equal distribution of health, achieving equal distribution through health care depends not only on the effectiveness and efficiency of alternative interventions but also the initial levels of health of the recipients in the absence of the program. The equity principle focuses on health levels; the evaluations focus on health gains. A less effective or efficient program that improves the health of those in relatively poor health may be preferred to one that improves the health of those already relatively healthy.

6. Concluding observations: health economists as policy advisors

Division over the most appropriate normative frameworks for the health sector, the recommendations that flow from alternative approaches, and the role of health economists as policy advisors in shaping health care reform have prompted critical self-reflection on how economists practice normative economic analysis in the health sector and how they explain findings to the public and policy makers [Reinhardt (1989, 1992, 1998), Fuchs (1996), Evans (1998), Hurley (1998), Mooney (1998), Rice (1998)]. Although economists often conceive of normative economic analysis as "objective science," it is inescapably a form of social ethics and ought to be treated as such. Hume observed over 200 years ago that one cannot derive "ought" from "is," and debates about the "new," Paretian welfare economics in the 1930s, 1940s and 1950s firmly established that there is no such thing as a "value-free" welfare economics [Robbins (1935), Myint (1948), Little (1957), Baumol (1965)].

This creates considerable tension for practitioners of normative economics. The fact that Paretian welfare economics and its associated ethical assumptions commands the assent of a large portion of professional economists does not, *ipso facto*, give it privileged ethical status in public policy; nor does an extra-welfarist near-exclusive focus on health in the social welfare function. We are back to de V. Graaff's opening observation – the relevance of assumptions for public policy depends critically on their realism and acceptability to broader society. The case has to be made. The Pareto criterion may appear weak and innocuous to many economists (who could argue that a situation in which at least one person is better off and no one worse off is to be preferred to the status quo?), but it is a restrictive assumption that appears to be contradicted by a wide variety of observed behaviours [Frank (1985), Rice (1998), Evans (1998)].

A clear social objective for health policy is to improve health. Health care has been singled-out as a policy concern because its primary objective is to produce health. Even if health is a primary concern, however, the public and policy makers clearly care about more than health [Mooney (1994), Hurley (1998)]. We care neither exclusively about utility nor exclusively about health as an outcome. In some situations, non-health related utility effects appear important (e.g., benefits of information); in others, we readily discount utility-effects (e.g., extremely risk averse individuals demanding high cost tests for rare diseases). Extra-welfarism, in principle, does not reject utility (or other non-health measures of well-being); compared to welfarism it adds health. As developed thus far, however, extra-welfarism provides no guidance for when health concerns should predominate and when utility concerns should predominate.

The question then is whether we can identify, on the basis of both analytic reasoning and empirical analysis, principles that can identify contexts in which each should be given prominence. The principles employed might reflect, for example, alternative levels of analysis (e.g., system-wide issues, programmatic issues, interventions), the nature of the alternatives under examination (e.g., clinical intervention; non-medical organizational issues); the nature of the groups affected by the policies under consideration, or the ultimate source of the benefit (e.g., what underlies a utility benefit, satisfaction of some basic need or merely the satisfaction of some preference). The debate has thus far tended to polarize. Such principles may offer a middle ground for sound reasoning that reflects the real world of social values, for the flexibility to respond to the particularities of different decision contexts; and for enough methodological rigor for results to be meaningful.

The inherent limitations of any normative analysis within a pluralistic society leads some to argue that economists should treat their work, even their ostensibly normative analyses, as positive economics and forgo the normative aspiration for their work. Mishan, for example, a leading figure in both theoretical and applied welfare economics, drew the following conclusion in the most recent edition of his text on cost-benefit analysis:

> "... I [now] virtually forswear earlier endeavors to base the Pareto criterion of economic efficiency on a consensus... the growing fragmentation of such a consensus – arising chiefly though by no means solely, from frequent rejection of the economist's basic maxim – [meant that] there was nothing for

> it but a retreat from the ambitious forward position I once occupied to far less impressive and, to that extent, far more defensible terrain. The $\sum$CV figure for any project is now more modestly to be regarded as the result of an exercise in positive economics, one having no normative overtones. And the economist's findings may or may not be received by the public or government as a contribution toward the decision-making process... I do not see how it can legitimately aspire to assume any normative status. The economist's expertise may be able to produce a correct figure for the $\sum$CV, but he can claim nothing more for it than that it is the $\sum$CV" [Mishan (1988, p. xiv)].

Although analytically a logical position to hold in the face of pluralism, it is likely untenable in the real world of policy making, where one cannot divorce language, even that which is intended to be purely descriptive, from its normative overtones.

This is part of a broader concern emerging over the use of language by economists in their role as policy advisors [Reinhardt (1989, 1992, 1998), Evans (1998), Williams (1993)]. The terms "efficiency," "optimal," "welfare," "net social gain," etc. have specific technical meanings within economics that do not correspond to general usage. Policy makers and the general public are likely to think that "optimal" means "best" in some overall sense. So when Sloan and Feldman (1988, p. 258) write that "Price controls definitely cannot lead to socially optimal levels of both quality and quantity; only competition can," non-economist readers may mistakenly be led to the conclusion that price competition in the physician sector will definitely make society best off. In fact, all that Sloan and Feldman meant was that, given the assumptions of their model, competition would lead to an efficient allocation of resources based on the Pareto criterion. Of course, many Pareto optimal allocations can be judged to be socially inferior to non-Pareto optimal allocations. Within economics, optimal means Pareto efficient, not best. Similarly, the concept of efficiency in general usage refers to "not wasting resources," in which case it is hard to be against efficiency. Even when economists use such jargon in a purely descriptive way in conversations with policy makers and the general public ("on the basis of our analysis, policy X results in a more efficient allocation of resources than policy Y") the professional jargon constitutes persuasive, emotive language in general usage that has clear, unavoidable and often misunderstood normative overtones.

To some, these difficulties are a counsel of despair, for they seem to rob our sophisticated, normative methods (and economists) of much of their punch, to generate a certain nihilism about what our analysis can say. Alternatively, they can be seen as a healthy development that should bring with it a greater self-consciousness by economists of the methods and language we employ and perhaps a shift of reference for policy-oriented economists from an internal professional focus on our own highly refined models and frameworks, to an external focus on the values and perspectives found in the society. As long as the intended audience for normative work is other than our professional colleagues, the interest in the work will depend on the relevance and reasonableness of its assumptions, particularly its ethical assumptions. Recognizing that normative economics is social ethics may foster greater interdisciplinary collaboration with ethicists, philosophers, social psychologists and others who have contributed to our analytic and empirical understanding of social values and social ethics.

The last forty years of normative analysis have generated a wealth of insights regarding the nature of heath care as a good, the normative implications of its production and consumption, the operation of health care and health care insurance markets, the merits and demerits of alternative approaches to financing, funding, organization and delivery of health care. If much contested ground remains, we have a much better picture of that ground – what is being contested, why it is being contested and what the terms of the debate are.

References

Acton, J.P. (1973), Evaluating Public Programs to Save Lives: The Case of Heart Attacks (Rand Corporation, Santa Monica).

Anderson, T.F., and G. Mooney, eds. (1990), The Challenges of Medical Practice Variations (Macmillan, London).

Arrow, K. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53(5):940–73.

Bator, F.M. (1957), "The simple analytics of welfare maximization", American Economic Review 47(1):22–59.

Baumol, W. (1965), Welfare Economics and the Theory of the State (Harvard University Press, Cambridge, MA).

Baumol, W.J. (1986), Superfairness: Applications and Theory (MIT Press, Cambridge, MA).

Benham, L. (1972), "The effects of advertising on the price of eyeglasses", Journal of Law and Economics 15:421–77.

Birch, S., J. Eyles, J. Hurley, B. Hutchison and S. Chambers (1993), "A needs-based approach to resource allocation in health care", Canadian Public Policy 19(1):68–85.

Blackorby, C., and D. Donaldson (1990), "The case against the use of the sum of compensating variations in cost-benefit analysis", Canadian Journal of Economics 23(3):471–94.

Bleichrodt, H. (1995), "QALYs and HYEs: under what conditions are they equivalent?", Journal of Health Economics 14:17–37.

Bleichrodt, H. (1997), "Health utility indices and equity considerations", Journal of Health Economics 16:65–91.

Bloomqvist, A. (1991), "The physician as a double-agent", Journal of Health Economics 10(4):411–22.

Boadway, R., and N. Bruce (1984), Welfare Economics (Basil Blackwell, Oxford).

Boulding, K. (1966), "The concept of need for health services", Milbank Memorial Fund Quarterly 44 (October):202–23.

Brito, D., E. Sheshinski and M. Intriligator (1991), "Externalities and compulsory vaccinations", Journal of Public Economics 45(1):69–90.

Buchanan, J. (1965), The Inconsistencies of the National Health Service (Institute of Economic Affairs, London).

Buckingham, K. (1993), "A note on HYE (Healthy Years Equivalent)", Journal of Health Economics 11:301–9.

Chalkley, M., and J.M. Malcomson (2000) "Government purchasing of health services", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 15.

Charles, C., and S. DeMaio (1993), "Lay participation in health care decision making: a conceptual framework", Journal of Health Politics, Policy and Law 18(4):881–904.

Charney, M.C. (1989), "Choosing who shall not be treated in the NHS", Social Science and Medicine 28:1331–38.

Coast, J., J. Donovan and S. Frankel (1996), Priority-Setting: The Health Care Debate (John Wiley and Sons, Chichester).

Cooter, R., and P. Rappoport (1984), "Were the ordinalists wrong about welfare economics?", Journal of Economic Literature 22(2):507–30.

Culyer, A.J. (1971), "The nature of the commodity 'health care' and its efficient allocation", Oxford Economic Papers 23:189–211.

Culyer, A.J. (1989), "The normative economics of health care finance and provision", Oxford Review of Economic Policy 5(1):34–58.

Culyer, A.J. (1990), "Commodities, characteristics of commodities, characteristics of people, utilities, and quality of life", in: S. Baldwin, C. Godfrey and C. Propper, eds., Quality of Life: Perspective and Policies (Routledge, London) 9–27.

Culyer, A.J. (1995a), "Need: the idea won't do – but we still need it", Social Science and Medicine 40 (6):727–730.

Culyer, A.J. (1995b), "Equality of what in health policy? Conflicts between the contenders", Discussion Paper 142 (University of York, Centre for Health Economics).

Culyer, A.J., and R.G. Evans (1996), "Mark Pauly on welfare economics: normative rabbits from positive hats", Journal of Health Economics 15(2):243–251.

Culyer, A., and H. Simpson (1980), "Externality models and health: a Rückblick over the last twenty years", Economic Record 56:222–230.

Culyer, A.J., and A. Wagstaff (1993a), "Equity and equality in health and health care", Journal of Health Economics 12(4):431–57.

Culyer, A.J., and A. Wagstaff (1993b), "QALYs and HYEs", Journal of Health Economics 11:311–23.

Culyer, A.J., E. van Doorslaer and A. Wagstaff (1991), "Comment: utilization as a measure of equity by Mooney, Donaldson and Gerard", Journal of Health Economics 11(1):93–98.

de Graaff, J.V. (1967), Theoretical Welfare Economics, 2nd edn. (Cambridge University Press, Cambridge).

Dionne, G., and A.P. Contandriopoulos (1985), "Doctors and their workshops: a review article", Journal of Health Economics 4:21–23.

Dolan, P. (2000), "The measurement of health-related quality of life for use in resource allocation decisions in health care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 32.

Donaldson, C., and K. Gerard (1993), Economics of Health Care Financing: The Visible Hand (Macmillan Press, London).

Donaldson, C., S. Birch and A. Gafni (1998), "The 'distribution' problem in economic evaluation: income and the valuation of costs and consequences of health care programs", Analysis Working Paper 98-10 (McMaster University, Centre for Health Economics and Policy).

Dranove, D., and W.D. White (1987), "Agency and the organization of health care delivery", Inquiry 24(4):405–415.

Drummond, M., B. O'Brien, G. Stoddart and G.W. Torrance (1997), Methods for the Economic Evaluation of Health Care Programs, 2nd edn. (Oxford University Press, Oxford).

Dworkin, R. (1981), "What is equality?", Philosophy and Public Affairs 10:185–246, 283–345.

Elster, J. (1992), Local Justice (Russell Sage Foundation, New York).

EuroQol Group (1990), "EuroQol – A new facility for the measurement of health-related quality of life", Health Policy 16(3):195–208.

Evans, R.G. (1976a), "Book review of the economics of health and medical care, M. Perelman (ed.)", Canadian Journal of Economics 9(3):532–37.

Evans, R.G. (1976b), "Modelling the economic objectives of the physician", in: R.D. Fraser, ed., Health Economics Symposium: Proceedings of the First Canadian Conference (Queen's University Industrial Relations Centre, Kingston) 33–46.

Evans, R.G. (1983), "The welfare economics of public health insurance: theory and Canadian practice", in: L. Soderstrom, ed., Social Insurance (North Holland, Amsterdam) 71–103.

Evans, R.G. (1984), Strained Mercy: the Economics of Canadian Health Care (Buttersworth, Toronto).

Evans, R.G. (1998), "Toward a healthier economics: reflections on Ken Bassett's problem", in: M. Barer, T. Getzen and G. Stoddart, eds., Health, Health Care and Health Economics: Perspectives on Distribution (John Wiley and Sons, Toronto) 465–500.

Evans, R.G., and G.L. Stoddart (1990), "Producing health, consuming health care", Social Science and Medicine 31(12):1347–63.

Evans, R.G., and M.F. Williamson (1978), Extending Canadian National Health Insurance: Options for Pharmacare and Denticare (University of Toronto Press, Toronto).

Evans, R.G., and A.D. Wolfson (1980), "Faith, hope and charity: health care in the utility function", Discussion Paper 20-46 (University of British Columbia, Vancouver).

Fanshel, S., and J. Bush (1970), "A health status index and its application to health services outcomes", Operations Research 18(6):1021–66.

Feeny, D., and G. Torrance (1989a), "Utilities and quality adjusted life-years", International Journal of Technology Assessment in Health Care 5(4):559–75.

Feeny, D., and G. Torrance (1989b), "Incorporating utility-based quality of life assessment in clinical trials", Medical Care 27:S190-S204.

Fein, R. (1958), Economics of Mental Illness (Basic Books, New York).

Feldman, R., and B. Dowd (1991), "A new estimate of the welfare loss of excess health insurance", American Economic Review 81:297–301.

Feldman, R., and M. Morrisey (1990), "Health economics: a report from the field", Journal of Health Politics, Policy and Law 15(3):627–46.

Feldman, R., and F. Sloan (1988), "Competition among physicians, revisited", Journal of Health Politics, Policy and Law 13(2):239–261.

Feldman, R., and F. Sloan (1989), "Reply from Feldman and Sloan", Journal of Health Politics, Policy and Law 14(3):621–625.

Feldstein, M. (1963), "Economic analysis, operational research, and the national heath service", Oxford Economic Papers 15(March):19–31.

Folland, S., and M. Stano (1990), "Small area variations: a critical review of propositions, methods and evidence", Medical Care Review 47(4):419–65.

Folland, S., A. Goodman and M. Stano (1996), The Economics of Health and Health Care (Prentice Hall, Upper Saddle River, NJ).

Fox, D., and H.M. Lichter (1993), "State model: Oregon", Health Affairs 12(2):66–70.

Frank, R.H. (1985), Choosing the Right Pond: Human Behaviour and the Quest for Status (Oxford University Press, New York).

Fuchs, V. (1996), "Economics, values and health care reform", American Economic Review 86(1):1–24.

Gafni, A. (1991), "Using willingness to pay as a measure of benefits: what is the relevant question to ask in the context of public decision making about public health care programs", Medical Care 29:1246–52.

Gafni, A., and S. Birch (1991), "Equity considerations in utility-based measures of health outcomes in economic appraisals: an adjustment algorithm", Journal of Health Economics 10(3):329–42.

Gafni, A., and S. Birch (1993), "Economics, health and health economics: HYEs vs QALYs", Journal of Health Economics 12(2):325–39.

Garber, A.M. (2000), "Advances in cost-effectiveness analysis of health interventions", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 4.

Garber, A.M., and C.E. Phelps (1997), "Economic foundations of cost-effectiveness analysis", Journal of Health Economics 16:1–31.

Garber, A.M., M.R. Weinstein, G.W. Torrance and M.S. Kamlet (1996), "Theoretical foundations of cost-effectiveness analysis", in: M.R. Gold, J.E. Siegel, L.B. Russell and M.C. Weinstein, eds., Cost-Effectiveness in Health and Medicine (Oxford University Press, New York) 25–53.

Gaynor, M., and S. Polochek (1994), "Measuring information in the market: an application to physician services", Bell Journal of Economics 60(4):815–831.

Giacomini, M., H. Luft and J.C. Robinson (1995), "Risk-adjusting community rated health plan premiums: a survey of risk assessment literature and policy applications", Annual Review of Public Health 16:401–30.

Grossman, M. (1972), "On the concept of health capital and the demand for health", Journal of Political Economy 80:223–55.

Grossman, M. (2000), "The human capital model", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 7.

Harberger, A.C. (1971), "Three basic postulates for applied welfare economics: an interpretive essay", Journal of Economic Literature 9:785–97.

Harris, J.E. (1977), "The internal organization of hospitals: some economic implications", Bell Journal of Economics 8(Autumn):467–482.

Hausman, D., and J. McPherson (1993), "Taking ethics seriously: economics and contemporary moral philosophy", Journal of Economic Literature 31(2):671–73l.

Hibbard, J.H., and J.J. Jewett (1996), "What type of quality information do consumers want in a health care report card?", Medical Care Research and Review 53(1):28–47.

Hibbard, J.H., and J.J. Jewett (1997), "Will quality report cards help consumers?", Health Affairs 16(1):218–228.

Hibbard, J.H., S. Sofaer and J.J. Jewett (1996), "Condition-specific performance information: assessing salience, comprehension and approaches for communicating quality", Health Care Financing Review 18(1):95–109.

Hicks, J. (1939), "The foundation of welfare economics", Economic Journal 49:696–712.

Hicks, J. (1941), "The four consumer surpluses", The Review of Economic Studies 11:31–41.

Hurley, J. (1998), "Welfarism, extra-welfarism, and evaluative economic analysis in the health sector", in: M. Barer, T. Getzen and G. Stoddart, eds., Health, Health Care and Health Economics: Perspectives on Distribution (John Wiley and Sons, Toronto) 373–96.

Hutchison, B., J. Hurley, S. Birch, J. Lomas, S.D. Walter, J. Eyles and F. Stratford-Devai (1999), "Needs-based primary medical care capitation: development and evaluation of alternative approaches", Health Care Management Science, forthcoming.

Johannesson, M. (1996), Theory and Methods of Economic Evaluation of Health Care (Kluwer, Dordrecht).

Johannesson, M. (1995), "The ranking properties of healthy-years equivalent and quality-adjusted life-years under certainty and uncertainty", International Journal of Technology Assessment 11(1):40–48.

Johannesson, M., and D. Meltzer (1998), "Some reflections on cost-effectiveness analyses", Health Economics 7:1–8.

Johannesson, M., J. Pliskin and M. Weinstein (1993), "Are healthy-years equivalent an improvement of quality-adjusted life-years?", Medical Decision Making 13(4):281–86.

Johansson, P.O. (1996), Evaluating Health Risks (Cambridge University Press, Cambridge, MA).

Jones-Lee, M. (1989), The Economics of Safety and Physical Risk (Basil Blackwell, Oxford).

Kahneman, D., and C. Varey (1991), "Notes on the psychology of utility", in: J. Elster and J. Roemer, eds., Interpersonal Comparisons of Well-being (Cambridge University Press, Cambridge) 127–63.

Kahneman, D., and J. Knetsch (1992), "Valuing public goods: the purchase of moral satisfaction", Journal of Environmental Economics and Management 22:57–70.

Kaldor, N. (1939), "Welfare propositions and interpersonal comparisons of utility", Economic Journal 49:549–52.

Kessel, R.A. (1958), "Price discrimination in medicine", Journal of Law and Economics 1:20–53.

Klarman, H.E. (1963), "The distinctive economic characteristics of health services", Journal of Health and Human Behaviour 44:44–9.

Klarman, H.E. (1965a), "The case for public intervention in financial health and medical services", Medical Care 3:59–62.

Klarman, H. (1965b), "Syphilis control programs", in: R. Dorfman, ed., Measuring the Benefits of Government Investments (The Brookings Institute, Washington, DC) 367–410.

Klarman, H. (1982), "The road to cost effectiveness analysis", Milbank Memorial Fund Quarterly 60(4):585–603.

Klarman, H., J. Francis and G. Rosenthal (1968), "Cost-effective analysis applied to the treatment of chronic renal disease", Medical Care 6:48–54.

Koopmans, T. (1957), Three Essays on the State of Economic Science (McGraw Hill, New York).

Kwoka, J.E. (1984), "Advertising the price and quality of optometric services", American Economic Review 74:211–16.

Labelle, R., G. Stoddart and T. Rice (1994a), "A re-examination of the meaning and importance of supplier-induced demand", Journal of Health Economics 13(3):347–368.

Labelle, R., G. Stoddart and T. Rice (1994b), "Editorial: response to Pauly on a re-examination of the meaning and importance of supplier-induced demand", Journal of Health Economics 13(4):491–494.

Lees, D.S. (1960), "The economics of health services", Lloyds Bank Review 56:26–41.

Lees, D.S. (1962), "The logic of the British National Health Service", Journal of Law and Economics 5:111–18.

Lees, D.S. (1967), "Efficiency in government spending: social services", Public Finance 22:176–89.

LeGrand, J. (1982), The Strategy of Equality (Allen and Unwin, London).

LeGrand, J. (1987), "Equity, health and healthcare", Social Justice Research 1(3):257–74.

LeGrand, J. (1991), Equity and Choice (Harper Collins Academic, London).

Levine, M., A. Gafni, B. Markham and D. MacFarlane (1992), "A bedside instrument to elicit a patient's preferences concerning adjuvant chemotherapy for breast cancer", Annals of Internal Medicine 117:53–58.

Lindsay, C.M. (1969), "Medical care and the economics of sharing", Economica 36(144):531–7.

Little, I.M.D. (1957), A Critique of Welfare Economics, 2nd edn. (Oxford University, Oxford Press).

Lohr, K.W., et al. (1986), "Effect of cost-sharing on use of medically effective and less-effective care", Medical Care 24:S31–S38.

Loomes, G., and L. McKenzie (1989), "The use of QALYs in health care decision making", Social Science and Medicine 28:299–308.

Machina, M. (1987), "Choice under uncertainty: problems solve and unsolved", Journal of Economic Perspectives 1(1):124–54.

Manning, W., et al. (1987), "Health insurance and the demand for medical care: evidence from a randomized experiment", American Economic Review 88(3):251–77.

Mannix, E., M. Neale and G. Northcraft (1995), "Equity, equality or need? The effects of organizational culture on the allocation of benefits and burdens", Organizational Behaviour and Human Decision Processes 63(3):276–86.

Marshall, A. (1961), Principles of Economics, 9th edn. [1890] (Macmillan, London).

Maynard, A., and K. Bloor (1998), Our Certain Fate: Rationing in Health Care (Office of Health Economics, London).

Mays, N., and G. Bevan (1987), Resource Allocation in the Health Services: A Review of the Methods of the Resource Allocation Working Party (Bedford Square Press, London).

McGuire, T.G., and M.V. Pauly (1991), "Physician response to fee changes to multiple payers", Journal of Health Economics 10(4):385–420.

McGuire, A., T. Henderson and G. Mooney (1988), The Economics of Health Care: An Introductory Text (Routledge and Kegan Paul, London).

Mehrez, A., and A. Gafni (1989), "Quality adjusted life-years, utility theory and healthy years equivalent", Medical Decision Making 9(2):142–49.

Mehrez, A., and A. Gafni (1993), "HYEs vs QALY: in pursuit of progress", Medical Decision Making 13(1):142–149.

Merriam-Webster (1986), Webster's Third New International Dictionary (Merriam-Webster, Inc., Springfield, MA).

Mill, J.S. (1994 [1848]), Principles of Political Economy (Oxford University Press, Oxford).

Miller, P. (1992), "Distribution justice: what people think", Ethics 102(3):555–93.

Mishan, E. (1971), "Evaluation of life and limb: a theoretical approach", Journal of Political Economy 79:687–706.

Mishan, E. (1988), Cost-Benefit Analysis, 4th edn. (Unwin Hyman, London).

Mooney, G. (1986), Economics, Medicine and Health Care (Wheatsheaf Books, Brighton).

Mooney, G. (1994), "What else do we want from our health services?", Social Science and Medicine 39:151–54.

Mooney, G. (1998), “Economics, communitarianism, and health care”, in: M. Barer, T. Getzen and G. Stoddart, eds., Health, Health Care and Health Economics: Perspectives on Distribution (John Wiley and Sons, Toronto) 397–415.

Mooney, G., J. Hall, C. Donaldson and K. Gerard (1991), “Utilization as a measure of equity: weighing heat”, Journal of Health Economics 10(4):475–80.

Mooney, G., and M. Lange (1993), “Ante-natal screening: what constitutes a benefit”, Social Science and Medicine 37(7):873–78.

Mooney, G., and M. Ryan (1993), “Agency in health care: getting beyond first principles”, Journal of Health Economics 12(2):125–135.

Murray, C., and A. Lopez (1996), The Global Burden of Disease (World Health Organization).

Musgrove, P. (1996), “Public and private roles in health care”, World Bank Discussion Paper No. 339 (World Bank, Washington).

Mushkin, S. (1958), “Toward a definition of health economics”, Public Health Reports 73(9):785–93.

Muurinen, J.M. (1982), “Demand for health: a generalized Grossman model”, Journal of Health Economics 1:5–28.

Myint, H. (1948), Theories of Welfare Economics (Harvard University Press, Cambridge, MA).

Neuman, P., M.E. Marback, K. Dusenbury, M. Kitchman and P. Zupp (1998), “Marketing HMOs to medicare beneficiaries”, Health Affairs 17(4):132–39.

Newhouse, J.P. (1992), “Medical care costs: how much welfare loss”, Journal of Economic Perspectives 6(3):3–21.

Newhouse, J.P. (1996), “Reimbursing health plans and health providers: efficiency in production versus selection”, Journal of Economic Literature 34:1236–63.

Newhouse, J.P. (1998), “Risk adjustment: where are we now?”, Inquiry 35:122–31.

Ng, Y.K. (1979), Welfare Economics: Introduction and Development of Basic Concepts (MacMillan, London).

Nord, E., J. Richardson, A. Street, H. Kuhse and P. Singer (1995), “Maximizing health benefits vs egalitarism: an Australian survey of health issues”, Social Science and Medicine 41:1429–37.

Nozick, R. (1974), Anarchy, State and Utopia (Basic Books, New York).

Nozick, R. (1989), The Examined Life: Philosophical Meditations (Simon and Schuster, New York).

O’Brien, B., and A. Gafni (1996), “When do the ‘dollars’ make sense? Toward a conceptual framework for contingent valuation studies in health care”, Medical Decision Making 16:288–99.

OECD (1998), OECD Health Data 98: Comparative Analysis of 28 Countries (OECD, Paris).

Olsen, E.O., and D.L. Rogers (1991), “The welfare economics of equal access”, Journal of Public Economics 45(1):91–105.

Oregon Health Services Commission (1991), Prioritization of Health Services: A Report to the Governor and Legislature (Oregon Health Services Commission, Salem).

Paul-Shaheen, P., J.D. Clark and D. Williams (1987), “Small area analysis: a review and analysis of the North American literature”, Journal of Health Politics, Policy and Law 12(4):741–809.

Pauly, M.V. (1968), “The economics of moral hazard”, American Economic Review 58(3):231–37.

Pauly, M.V. (1970), “The efficiency in the provision of consumption subsidies”, Kyklos 23:33–57.

Pauly, M. (1978), “Is medical care different?”, in: W. Greenberg, ed., Competition in the Health Care Sector (Aspen Systems, Germantown, MD) 11–35.

Pauly, M. (1988), “Is medical care different? Old question, new answers”, Journal of Health Politics, Policy and Law 13:227–37.

Pauly, M.V. (1994a), “Editorial: a re-examination of the meaning and importance of supplier-induced demand”, Journal of Health Economics 13(3):369–372.

Pauly, M.V. (1994b), “Reply to Roberta Labelle, Greg Stoddart and Thomas Rice”, Journal of Health Economics 13(4):495–496.

Pauly, M.V. (1996), “Reply to Anthony J. Culyer and Robert G. Evans”, Journal of Health Economics 15(2)253–254.

Pauly, M.V. (2000), "Insurance reimbursement", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 10.

Pereira, J. (1993), "What does equity in health mean?", Journal of Social Policy 22(1):19–48.

Phelps, C.E. (1992), Health Economics (Harper Collins, New York).

Phelps, C.E., and S.T. Parente (1990), "Priority setting in medical technology and medical practice assessment", Medical Care 28(8):703–723.

Phelps, C.E., and A.J. Mushlin (1991), "One the (near) equivalence of cost-effectiveness and cost-benefit analysis", International Journal of Technology Assessment in Health Care 7(1):12–21.

Phillipson, T. (2000), "Economic epidemiology and infectious diseases", in: A.J. Culyer and J. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 33.

Pliskin, J.S., D. Shepard and M.C. Weinstein (1980), "Utility functions for life years and health status", Operations Research 28:206–24.

Plotnick, R. (1981), "A measure of horizontal equity", Review for Economics and Statistics 63(2):283–8.

Plotnick, R. (1982), "The concept and measurement of horizontal equity", Journal of Public Economics 17(3):373–91.

Rawls, J. (1971), A Theory of Justice (Harvard University Press, Cambridge).

Reid, W. (1998), "Comparative dynamic analysis of the full Grossman model", Journal of Health Economics 17(4):383–425.

Reinhardt, U. (1989), "Economists in health care: saviors or elephants in a porcelain shop?", American Economic Review 79(2):337–42.

Reinhardt, U. (1992), "Reflections on the meaning of efficiency: can efficiency be separated from equity?", Yale Law and Policy Review 10(2):302–15.

Reinhardt, U. (1998), "Abstracting from distributional effects, this policy is efficient", in: M. Barer, T. Getzen and G. Stoddart, eds., Health, Health Care and Health Economics: Perspectives on Distribution (John Wiley and Sons, Toronto) 1–52.

Rice, T. (1992), "An alternative framework for evaluating welfare losses in the health care market", Journal of Health Economics 11(1):85–92.

Rice, T. (1998), The Economics of Health Reconsidered (Health Administration Press, Chicago).

Rice, T.H., and R.J. Labelle (1989), "Do physicians induce demand for medical services?", Journal of Health Politics, Policy and Law 14(3):587–600.

Robbins, L. (1935), An Essay of the Nature and Significance of Economic Science, 2nd edn. (Macmillian, London).

Robinson, J.C. (1990), "Philosophical origins of the social rate of discount in cost-benefit analysis", Milbank Quarterly 68(2):245–65.

Rosser, R., and P. Kind (1978), "A scale of valuations of states of illness: is there a social consensus?", International Journal of Epidemiology 7(4):347–58.

Sagoff, M. (1994), "Should preferences count?", Land Economics 70(2):127–44.

Sappington, D.E. (1991), "Incentives in principal-agent relationships", Journal of Economic Perspectives 5(2):45–66.

Scanlon, T. (1975), "Preference and urgency", Journal of Philosophy 72(19):655–69.

Schelling, T. (1968), "The life you save may be your own", in: S.B. Chase, ed., Problems in Public Expenditure Analysis (The Brooking‘s Institute, Washington) 127–62.

Schneider, E.C., and A.M. Epstein (1998), "Use of public performance reports: a survey of patients undergoing cardiac surgery", Journal of the American Medical Association 279(20):1638–1642.

Sen, A. (1979), "Personal utilities and public judgements: or what's wrong with welfare economics", Economic Journal 89(September):537–58.

Sen, A. (1985), Commodities and Capabilities (North Holland, Amsterdam).

Sen, A. (1987), On Ethics and Economics (Blackwell, Cambridge).

Sen, A. (1992), Inequality Re-examined (Harvard University Press, Cambridge).

Shackley, P., and A. Healey (1993), "Creating a market: an economic analysis of the purchaser-provider model", Health Policy 25:153–168.

Shavell, S. (1978), “Theoretical issues in medical malpractice”, in: S. Rottenberg, ed., The Economics of Medical Malpractice (American Enterprise Institute, Washington, DC).

Shiell, A., and P. Hawe (1996), “Health promotion community development and the tyranny of individualism”, Health Economics 5:241–47.

Shoemaker, P. (1982), “The expected utility model: its variants, purposes, evidence and limitations”, Journal of Economic Literature 20:529–63.

Steele (1981), “Marginal met need and geographical equity in health care”, Scottish Journal of Political Economy 28(2):186–95.

Stoddart, G.L., M. Barer and R.G. Evans (1994), User Charges, Snares and Delusions: Another Look at the Literature (Premiers Council on Health, Well-being and Social Justice, Toronto).

Sugden, R., and A. Williams (1978), The Principles of Practical Cost-Benefit Analysis (Oxford University Press, Oxford).

Torrance, G. (1986), “Measurement of health state utilities for economic appraisal”, Journal of Health Economics 5:1–30.

Torrance, G., W. Thomas and D. Sackett (1972), “A utility maximization model for evaluation of health care programs”, Health Services Research 7(2):118–33.

Tuohy, C., and A. Wolfson (1978), “Self -regulation: who qualifies?”, in: P. Slayton and M.J. Trebilcock, eds., The Professions and Public Policy (University of Toronto Press, Toronto) 111–22.

Ubel, P., and G. Loewenstein (1996), “Distributing scarce livers: the moral reasoning of the general public”, Social Science and Medicine 42(7):1049–55.

Ubel, P., M. Dekay, J. Baron and D. Asch (1996), “Cost-effectiveness analyses in a setting of budget constraints: is it equitable”, New England Journal of Medicine 334(18):1174–77.

van de Ven, W.P.M.M., and R.P. Ellis (2000), “Risk adjustment in competitive health plan markets”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 14.

van Doorslaer, E., A. Wagstaff and F. Rutten (1993), Equity in the Finance and Delivery of Health Care: An International Perspective (Oxford University Press, Oxford).

Varian, H. (1974), “Equity, envy, and efficiency”, Journal of Economic Theory 9(1):63–91.

Varian, H. (1975), “Distributive justice, welfare economics, and the theory of fairness”, Philosophy and Public Affairs 4(3):223–47.

Viscusi, K. (1992), Fatal Tradeoffs (Oxford University Press, Oxford).

Viscusi, W.K. (1993), “The value of risks to life and health”, Journal of Economic Literature 31:1912–46.

Wagstaff, A. (1991), “QALYs and the equity-efficiency trade-off”, Journal of Health Economics 10(1):21–42.

Wagstaff, A., and E. van Doorslaer (2000), “Equity in health care finance and delivery”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 34.

Walzer, M. (1982), Spheres of Justice (Basic Books, New York).

Weinstein, M.C., and W.B. Stason (1977), “Foundations of cost-effectiveness analysis for health and medical practices”, New England Journal of Medicine 296:716–21.

Weisbrod, B. (1961), The Economics of Public Health (University of Pennsylvania Press, Philadelphia).

Weisbrod, B.A. (1964), “Collective-consumption services of individual-consumption goods”, Quarterly Journal of Economics 78:471–74.

Weisbrod, B. (1968), “Income redistribution effects and benefit-cost analyses”, in: S.B. Chase, ed., Problems vs. Public Expenditure Analyses (Brookings Institute, Washington) 395–428.

Weisbrod, B.A. (1978), “Comment on paper by Mark Pauly”, in: W. Greenberg, ed., Competition in the Health Sector: Past, Present and Future (Bureau of Economics, Federal Trade Commission, Washington) 49–56.

Weisbrod, B.A. (1991), “The health care quadrilemma: an essay on technological change, insurance, quality of care, and cost containment”, Journal of Economic Literature 29:523–552.

Williams, A. (1978), “Need: an economic exegesis”, in: A.J. Culyer and K.G. Wright, eds., Economic Aspects of Health Services (Martin Robertson, London).

Williams, A. (1981), “Welfare economics and health status measurement”, in: J. van der Gaag and M. Perlman, eds., Health, Economics and Health Economics (North Holland, Amsterdam) 271–81.

Williams, A. (1985), “The value of QALYs”, Health and Social Service Journal.

Williams, A. (1988), “Ethics and efficiency in the provision of health care”, in: J.M. Bell and S. Mendus, eds., Philosophy and Medical Welfare (Cambridge University Press, Cambridge) 111–126.

Williams, A. (1993), “Cost-benefit analysis: applied welfare economics or general decision aid?”, in: A. Williams and E. Giardina, eds. (Edward Elgar Publishing Co., Brookfield) 65–82.

Williams, A. (1997), “Intergenerational equity: an exploration of the ‘fair innings’ argument”, Health Economics 6:117–32.

Williams, A., and R. Cookson (2000), “Equity in health”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 35.

Williamson, O.E. (1986), Economic Organization: Firms, Markets and Policy Control (New York University Press, New York).

Woodward, R.S., and F. Warren-Boulton (1984), “Considering the effects of financial incentives and professional ethics on ‘appropriate’ medical care”, Journal of Health Economics 3(3):223–37.

Woolhandler, S., and D.U. Himmelstein (1991), “The deteriorating efficiency of the U.S. health care system”, New England Journal of Medicine 324:1253–58.

World Health Organization (1947), “The Constitution of the World Health Organization”, WHO Chronicles 1:29.

Yaari, M., and M. Bar-Hillel (1984), “On dividing justly”, Social Choice and Welfare 1(1):1–24.

Zeckhauser, R. (1970), “Medical insurance: a case study of the tradeoff between risk spending and appropriate incentives”, Journal of Economic Theory 2(1):10–26.

Chapter 3

MEDICAL CARE PRICES AND OUTPUT*

ERNST R. BERNDT

MIT Sloan School of Management

DAVID M. CUTLER

Harvard University, Department of Economics

RICHARD G. FRANK

Harvard Medical School

ZVI GRILICHES[†]

Harvard University, Department of Economics

JOSEPH P. NEWHOUSE

Harvard Medical School, Harvard School of Public Health, and John F. Kennedy School of Government

JACK E. TRIPLETT

The Brookings Institution

Contents

*Research support to the authors is gratefully acknowledged from the US Bureau of Economic Analysis (Berndt, Cutler, Griliches), US Bureau of Labor Statistics (Berndt, Cutler, Griliches), Eli Lilly and Co. (Berndt, Frank, Griliches, Triplett), National Institute of Mental Health (Frank), National Institute on Aging (Cutler), National Science Foundation (Berndt, Griliches), Hans Sigrist Stiftung (Newhouse) and the Alfred P. Sloan Foundation (Cutler, Frank, Newhouse). We have benefited from conversations with numerous colleagues and friends, and in particular thank Dennis Fixler of the Bureau of Labor Statistics, Brent Moulton of the Bureau of Economic Analysis, and Richard Zeckhauser, Harvard Kennedy School of Government for comments on earlier drafts. A substantial portion of this paper also appears in Berndt, Cutler, Frank et al. (1998).

[†] Deceased, November 4, 1999.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

We review in considerable detail the conceptual and measurement issues that underlie construction of medical care price indexes in the US, focusing in particular on the medical care consumer price indexes (MCPIs) and medical-related producer price indexes (MPPIs). We outline salient features of the medical care marketplace, including the impacts of insurance, moral hazard, principal-agent relationships, technological progress and organizational changes. Since observed data are unlikely to correspond with efficient outcomes, we discuss implications of the failure of transactions data in this market to reveal reliable marginal valuations, and the consequent need to augment traditional transactions data with information based on cost-effectiveness and outcomes studies.

We describe procedures currently used by the US Bureau of Labor Statistics in constructing MCPIs and MPPIs, including recent revisions, and then consider alternative notions of medical care output pricing that involve the price or cost of an episode of treatment, rather than prices of fixed bundles of inputs. We outline features of a proposed new experimental price index – a medical care expenditure price index – that is more suitable for evaluation and analyses of medical care cost changes, than are the current MCPIs and MPPIs. We discuss the ways in which medical care transactions enter national economic accounts, including inter-industry flows and national health accounts, as well as aggregate economy implications of possible mismeasurement of prices in the medical sector. We conclude by suggesting future research and measurement issues that are most likely to be fruitful.

Keywords

price indexes, MCPIs, MPPIs, outcomes, episode

JEL classification: C43, I11

> "Statistics on medical prices should be improved; indexes of medical productivity should be developed; and the search for an understanding of the determinants of medical price and cost behavior should be developed"

Report to the President on Medical Care Prices, US Department of Health, Education and Welfare (1967, p. 11)

1. Introduction

The measurement of the output of the medical care system is necessary to assess the productivity levels and growth of a country's economy and of course its medical care system. This is true in countries with universal health care coverage or incomplete coverage, and regardless of the mix of public and private provision of medical care.

For most industries in most countries, real output measurement is accomplished by dividing data on revenues or sales by a price index to obtain a measure of real output. Reliable output measurement for an industry therefore requires correspondingly reliable revenue data and a price index. A number of conceptual difficulties and institutional characteristics of medical care markets, however, make reliable price measurement of medical goods and services particularly difficult and challenging.

For countries where medical care goods and services are provided by the government without direct charge, or with only nominal direct charges, data on revenue or receipts for medical care may not be available or may not be relevant. For these countries, the problem of measuring the output of medical care goes well beyond the inherent difficulty of measuring medical care prices. In such cases the difficult problem of measuring prices and output of medical care is combined with the equally formidable problem of measuring the output of the government sector.[1]

Medical price indexes have uses other than those involving output and productivity measurement. In the US, both within the health sector and more generally, contracts occasionally contain provisions that depend on growth of the medical Consumer Price Index (CPI).[2] Medical CPIs and medical Producer Price indexes (PPIs) are also employed in updating of fee schedules for certain administered pricing schemes and payments to some health plans. Medical CPIs and PPIs are also employed by public policy analysts in projecting the impacts of changes in public policy.

Although medical CPIs and PPIs play prominent roles in private and public sector transactions and analyses, both the US Bureau of Labor Statistics (BLS) and its critics have acknowledged that current BLS practices for tracking price changes in the medical care industries, industries characterized by dynamic technological and organizational

[1] For a discussion of measurement issues in public sector output, see Kendrick (1991) and Griliches (1992, pp. 18–19). Murray (1992) and the Swedish Ministry of Finance (1997) contain empirical analyses of publicly provided health sector output and productivity growth.

[2] For general discussion of CPI use in escalation clauses, see Triplett (1983).

changes, are likely to be inaccurate and in need of substantial improvement and overhaul.[3]

Several aspects of the medical care industry make the BLS' task of constructing accurate and readily interpretable medical CPIs and PPIs particularly difficult. Output measurement of the health care system is inherently difficult when mortality is but one possible outcome from treatment. Mortality is particularly inappropriate as an output measure for treatments of a variety of acute conditions that are not life-threatening, and for many increasingly prevalent chronic illnesses. Additional attributes, such as morbidity, pain and suffering, functional and emotional impairment, and quality of life are each highly valued aspects of treatment response.

Another output measurement challenge that is rather unique to the medical care sector arises from the moral hazard caused by health insurance which causes marginal private and social costs to diverge. As emphasized by, among others, Newhouse et al. (1993), the existence of demand side moral hazard or administratively set prices make it inappropriate to attribute the usual normative properties to medical CPIs that are commonly associated with other such price indexes. The provision of medical care services also involves a principal-agent relationship: in choosing treatment, patients typically rely considerably on the advice and counsel of their physician, whose incentives and financial interests may or may not align well with those of the patient. Any misalignment of interests may result in inefficient outcomes.

A third dimension of medical care that poses significant price measurement challenges relates to technological progress. While not unique to medical care, technological progress is nevertheless of great significance in this sector of the economy. New treatment technologies are continuously emerging and being introduced into common clinical practice. This creates many of the problems of new goods that economists interested in index number and productivity measurement have struggled with for many years.[4]

Finally, organizational changes have been dramatic in the medical care sector. The manner in which medical technologies are rationed, delivered and even priced has evolved rapidly during the last decade. Managed care arrangements have resulted in changes in the locus of care, the organization of medical practice, contractual relations between buyers and sellers, and the manner in which inputs are combined to create treatment [see Glied (2000)]. Thus the way in which typical treatment for an illness such as depression is organized and provided has been remarkably altered in just a few years. Even given a known set of treatment technologies, important qualitative differences have emerged in the supply of treatment and in the way care is experienced by patients.

In this chapter we review in considerable detail the measurement issues that underlie construction of medical care price indexes, we describe procedures employed by

[3] See, for example, US Senate Finance Committee (1996), US Department of Labor, Bureau of Labor Statistics (1997b), and Abraham, Greenlees and Moulton (1998).

[4] See, for example, the chapters and references in Bresnahan and Gordon (1997).

the BLS in the construction of its medical CPIs and PPIs (including recent revisions and changes), we discuss alternative notions of medical care output that involve the price of an episode of treatment rather than the prices of fixed bundles of inputs, we outline salient features of a new medical care expenditure price index, we consider interactions between national economic accounts and national health accounts, and we suggest future research initiatives that are likely to be most fruitful. We begin with a description of the market environment underlying medical care CPIs and PPIs in the US.

2. The market environment underlying medical care CPIs and PPIs

Viewed by an economic statistician, the medical care sector is large and intimidating. As with an elephant, one can employ several approaches in cautiously observing, walking around and measuring it. We begin by describing the principal actors, characteristics and incentive structures that must be taken into account in providing a foundation for the measurement of medical care prices.

2.1. Distinguishing features of the US medical care marketplace

Economists generally presume some form of consumer optimization and efficiency in the purchase of goods and services. As in other markets, consumers of medical services are envisaged as maximizing some notion of utility, buying goods and services that generate direct utility, and using some of these goods and services as intermediate goods to produce utility. In the medical care marketplace, however, this optimization and efficiency is exceedingly complex; it involves behavior based on the use of asymmetric information and personnel who act as imperfect agents for consumers, under rationing constraints that are not nearly as pervasive as in other consumer markets.

The medical care industry provides goods and services in a number of specific subsectors: hospitals (including hotel and cafeteria services), physician practices, laboratories, pharmaceuticals, clinics, medical devices, nursing homes, home health agencies, and so on. These services are provided to consumers, but consumers typically do not value these services *per se*. Rather, they value the health outcomes resulting from medical interventions provided by the medical care industry.[5] These impacts on health are conceptually the composite good that we want to price. But the nature of transactions in this industry is exceedingly complex.

As in any industry, market structure affects the industry's price level, and perhaps the rate of price growth, particularly if production effiency is affected over time. Licensing, reputation, the regulatory environment and intellectual property rights provide suppliers of medical services with varying amounts of market power, particularly since some medical service suppliers such as hospitals and physicians face limited competition outside rather narrow geographical market boundaries. In many cases fixed costs are high,

[5] For further discussion, see Triplett (1998a, 1998b).

to a great extent consumers arrive at random times, and price is greater than short-run marginal cost.

Buyers also have market power. Although the federal government has long been a major purchaser of medical services (providing funds for about 39% of personal health care expenditures in the US in 1996),[6] within the last decade there has been much consolidation of buying power among health maintenance and managed care organizations. Thus on both the supply and demand sides of the medical care marketplace, market power is present. Moreover, since most medical care services are not resellable, price dispersion is not easily eliminated by arbitrage, price discrimination is prevalent, and thus the "law of one price" typically does not hold.

There are several other features of the market structure of the medical sector that, while present to some extent in other sectors, are particularly pervasive in medical goods and services. First, the vast majority of medical care payments are not made directly by consumers. Indeed, in the US in 1996, out-of-pocket payments by consumers accounted for only about 19% of total personal health care service expenditures.[7] The remainder of medical care is largely paid for by insurers.[8] Insurance programs may be run publicly, as with Medicare, Medicaid, and other federal state and local funds, who together accounted for 53% of personal health care service expenditures in 1996; or the sources of funds may be private, which in 1996 made up 37% of personal health care service expenditures, primarily for the non-elderly. Ultimately, the insurance payments not paid directly by individuals are passed back to individuals, in the form of higher taxes or reduced other government spending when the insurance payments are by the public sector, or in the form of an adjusted employee compensation package when insurance is provided by employers.[9]

The predominance of the indirect nature of payments creates several difficulties for constructing and interpreting consumer price indexes. The most significant of these is moral hazard. If consumers pay for only, say, 20% of medical care at the margin, they will seek to consume medical care until its marginal value is only about twenty cents per dollar of spending. This is true even though people *on average* must pay for the full dollar of medical care. Individuals will therefore tend to overconsume medical resources – resources will be consumed that cost society $1 (less if there are rents) but are worth less than that at the margin.

The second important feature of the medical care market is that consumers do not always know what services they want. Patients tend to rely on physicians both to provide them services and to recommend the services they need. As a result, there is a principal-agent problem: patients would like physicians to act in the patients' best interests, but physicians might not always have an interest in doing so.

[6] Levit, Lazenby, Braden et al. (1998, Exhibit 3, p. 39 and Exhibit 4, p. 43).

[7] *Ibid.*

[8] In the US, however, about 4% derives from other philanthropic sources.

[9] For a discussion of the incidence of employer-provided health insurance, see Gruber (1994, 1997) and Pauly (1997).

In traditional US health insurance arrangements, physicians and patients both had incentives for excessive medical care. Patients were well-insured at the margin and physicians were paid on a fee-for-service basis – earning more when they did more since fees were generally above marginal cost. The result was an incentive structure on both the demand and supply side that induced excessive care. Today's environment in the US is much changed, and increasingly involves more complicated rationing. Health plans now often operate under fixed budgets, whereas before they typically passed costs through to the employer or government. Thus they have begun to employ administrative mechanisms and financial incentives to control health care spending. The result is that, with increasing frequency, patient demand incentives are at odds with those of their health plans or physicians.

The implication of both of these pervasive features of the medical care industry is that revealed consumer purchases are *not* a reliable guide to the marginal *value* of medical care. This is in contrast to other markets, such as that for, say, compact disks, where consumers' marginal valuations are likely to be well-reflected in prices and expenditures. Consumers may receive too much medical care, as they likely did under traditional insurance arrangements, or too little care, as some allege they do under managed care or capitated insurance (one price per patient per year, independent of the amount of services the patient actually receives). In many markets, it is eminently reasonable to relate relative prices to marginal rates of substitution in consumption, but in medical care this assumption is simply not tenable. As a practical matter, this inability to employ the assumptions underlying traditional revealed preference theory severely hampers the ability of economic statisticians to construct accurate and readily interpretable price indexes for medical care.

The extent to which medical care services differ from other services can be illustrated by considering a hypothetical transaction in a restaurant. Suppose an individual places an order for a particular set of items on the menu, and then leaves. Another person enters the restaurant, sits down at a table, eats the meal that was ordered, and then leaves. Finally, a third person comes in and pays for the meal. In medical care, these three persons are the physician, the patient and the insurer. Whose valuation shall one measure?

As with many other services such as ATM banking services, the production function for medical care involves interdependent efforts of suppliers and consumers. This interdependent aspect of medical care production makes it more difficult to distinguish between producer and consumer price indexes. Moreover, for consumers, medical care, health and utility are quite different. This occurs in part because the production function for health has a number of arguments, other than medical care. One formulation of the production function is as follows:

$$\text{Health} = H(\text{medical care, knowledge, time, lifestyle, environment, etc.}).$$

It is useful to consider these other inputs into the production of health.[10] Many of the

[10] For a more complete discussion, see Grossman (1972a, 1972b).

inputs have been shown to contribute more to health than medical care.[11] Knowledge, for example, mediates between medical care and health. Medical treatments must not only be produced, but they must be used as well, and knowledge about how to use them changes over time. To the extent knowledge is non-rivalrous in nature and has public good properties, its existence together with the interdependent nature of production makes it difficult to assess uniquely the impacts of changes in knowledge on prices for suppliers vs. that facing consumers.

As an example of the importance of medical knowledge, suppose that medical research discovers that a particular pharmaceutical agent is just as effective when taken in half a dosage strength as when taken at full strength; something like this occurred for contraceptives several decades ago. Has the price of medical care changed? From the consumers' perspective, the answer is likely yes, for the cost of achieving a particular health state has fallen.[12] From the pharmaceutical manufacturers' perspective, however, the answer may be no, for the marginal cost of producing a milligram of the medication may not have changed. From the vantage of the family practitioner physician, whether the price has fallen depends in part on how one views physician services. The advice provided by the physician to the patient may still take the same amount of billable time, but if the physician is part of a staff model health maintenance organization with total pharmaceutical coverage, the price of providing family planning services may have fallen.

Knowledge, of course, is just one form of technological change. The reason we distinguish knowledge from other forms of technological progress is that knowledge is often envisaged as *disembodied* technological progress, while most new technologies are *embodied* in a particular service or product. Although the absorption of knowledge is not without cost, to some extent knowledge has public good properties and is non-rivalrous, quite unlike say, a piece of medical diagnostic equipment.

These two types of technological change overlap, yet it is useful to distinguish between them. Significant quality changes are often embodied in new medical care-related goods and services, but use of the new good may require additional knowledge. For example, a new non-invasive operation is typically performed by physicians using novel inputs (endoscopic instruments), and knowledge about such new treatments can be usefully employed by patients and clinicians alike. In any case, it is appropriate to envisage knowledge as an input into the production of health, distinct from medical treatments.[13]

Another input into the production of health is time. Producing better health requires time inputs from households as well as providers: time is spent in seeking and receiving

[11] See Fuchs (1974, 1983).

[12] Even here, matters are complex. For many brand name pharmaceuticals in the US (but not in Europe), the price per tablet is the same regardless of strength. Moreover, in the example here it is implicitly assumed that the consumers' cost for a contraceptive medication is not fully covered by insurance. Until recently in the US, unlike the case for most medications for which the patient makes a copayment, for contraceptives consumers have generally borne the entire direct cost of the prescription.

[13] For an exchange of views on this, see Gilbert (1961, 1962) and Griliches (1962).

treatment, in recovery, and in assisting others. Some medical innovations, for example, new anaesthetics, have shortened the recovery time for patients.[14] This too will reduce the consumers' cost of better health, but may well leave the producer's costs unchanged (say, if the new anaesthetic cost the hospital as much as the old).

An individual's lifestyle is another input into the production of health. Eating habits, drinking patterns, exercise regimens, and the pursuit of risky behavior all affect an individual's health. In some cases greater use of medical care and unhealthy behavior occur simultaneously to maintain a given health state. For example, with the introduction of over-the-counter H_2-antagonists such as Pepcid AC, individuals can preemptively take a heartburn prevention medication and then eat a high calorie, highly spiced meal.

Yet one more input affecting health is the environment. Environmental changes may improve or retard health. For example, new diseases such as AIDS may develop and be discovered, while other diseases such as smallpox may be eradicated. Changes in air and water pollution, in climate and weather, as well as in rates of criminal activity, will also have important impacts on health. It is important that such environmental changes be envisaged as primarily affecting the quantity or quality of medical services provided, not their price.

The age distribution of the population might also be envisaged as an environmental input affecting health care spending of populations. As people age, typically more inputs of medical care are required to maintain health, or to mitigate deterioration. Increased medical care expenditures, or increased health insurance premiums that reflect impacts of an aging population are appropriately viewed as quantity rather than price increases; in such cases medical expenditures rise due to increased quantity consumption, not price changes.

Finally, it is important to emphasize that while the marginal utility of health is positive, health is not the only argument in an individual's utility function. Thus a utility function might be envisaged as follows:

$$\text{Utility} = U(\text{health, lifestyle, leisure time, other consumption goods and services,}\\ \text{environment, etc.}).$$

Note that some factors such as lifestyle and environment not only have a direct impact on utility, but also have an indirect impact via health. One important implication of this, as has been emphasized elsewhere by Triplett (1998a, 1998b), is that the output of the medical care industry is not something like the *average* health of the population, but rather is best viewed in *marginal* terms as the health implication of a medical intervention, conditional on lifestyle, environment and other inputs affecting health.

The production function for health, as well as the utility function, has intertemporal aspects. Some current consumption goods affect future health states as well as current

[14] Since the largest cost to receiving medical care is often the patient's time cost, there are substantial incentives to develop innovations that conserve on time, particularly when the cost of these innovations is covered by insurance.

utility, while some medical interventions impact future consumption possibilities and patterns. Here we put these complications aside, but see Grossman (1972a, 1972b) and Meltzer (1997) for further discussion.

2.2. Pricing medical care services

With this discussion of salient characteristics of the medical care marketplace as background, we now consider approaches to price measurement. A representative consumer can be envisaged as an individual who is making decisions before knowing what diseases he or she might eventually experience. Extensions to heterogeneous consumers complicate matters, but for our purposes it is sufficient initially to work here with the simpler representative consumer framework.[15] Let this representative consumer have a utility function that depends on consumption of goods and services (other than medical care) and health. For concreteness, assume there is only one disease, which everyone contracts; extending the analysis to multiple diseases and probabilities of having each disease is straightforward. Denote Y as exogenous income, H as the health state, M as the quantity of medical care and p_M its (normalized) price, I and p_I as the quantity and price (premium) of a constant-quality insurance policy, K as medical care knowledge, E as the state of the environment, L as leisure time, and T_M as time allocated to receiving medical treatments. For simplicity, M and P_M include both consumers' direct health expenditures and indirect medical services obtained through health insurance in a competitive, actuarially fair insurance market where changes in costs of medical services to insurers are passed on to consumers via insurance premium changes. In this context, I and P_I are associated only with pure insurance services. The utility function is then written as:

$$U = U\big[Y - P_M M - P_I I,\ H(M, K, E),\ L - T_M\big]. \tag{1}$$

The first term is non-medical care consumption (non-medical expenditures divided by numeraire price), the second is health, and the third is non-medical care time.[16] Although Equation (1) embeds a multi-year framework, for simplicity we assume but one time period. Notice also that Equation (1) makes no assumption about how medical treatment decisions are made, or how medical prices are set.[17]

Over time, medical care and its price may change, or there may be changes in knowledge, the environment, and time devoted to medical care. For concreteness, consider

[15] For an extension to heterogeneous consumers, see Pollak (1980, 1998) and Fisher and Griliches (1995).

[16] For simplicity, time at work is omitted.

[17] The relationship between utility maximization and index numbers relies critically on a number of assumptions. In the present context, such assumptions might well be that the consumer chooses M, I, K and T_M, given Y, P_M, P_I, P_K (which could be zero at the margin if knowledge is non-rivalrous), E and L, so as to maximize U in each time period. As we point out at various times in this paper, these assumptions are likely to be particularly untenable in the medical care marketplace.

changes between periods 0 and 1. The question posed is: What is the correct price index for changes between periods 0 and 1, assuming that consumers optimize in each time period? We can define the *cost of living* in one of several natural ways – the change in the cost of living between periods 0 and 1 is the additional funds the individual needs in period 1 to be just as well off as he or she was in period 0. This amount may be positive, in which case the cost of living has increased, or it may be negative, in which case the cost of living has fallen.[18] This hypothetical is associated with the Laspeyres, base period utility notion of cost of living. An alternative, associated with the Paasche notion, uses the current period utility as the point of reference, and asks: What is the change in funds the individual needs in period 0 to be just as well off as he or she is in period 1? We consider this distinction in further detail below. A third index, the Fisher Ideal, is the geometric mean of the Laspeyres and the Paasche.

Consider the amount C of additional money the consumer requires in period 1 to make him or her indifferent between living in periods 0 and 1 (the Laspeyres notion):

$$\begin{aligned} &U\bigl[Y - P_{M1}M_1 - P_{I1}I_1 + C,\ H(M_1, K_1, E_1),\ L - T_{M1}\bigr] \\ &\quad = U\bigl[Y - P_{M0}M_0 - P_{I0}I_0,\ H(M_0, K_0, E_0),\ L - T_{M0}\bigr]. \end{aligned} \tag{2}$$

C is the change in the cost of living – a positive C implies an increase in the cost of living, and a negative C a decrease.

To form a price index, one can scale C by the income required to produce utility in period 0, i.e., Y. The cost of living index could therefore be:

$$\text{Cost of Living Index} \equiv 1 + C/Y. \tag{3}$$

Using a first order difference approximation, we differentiate and rearrange Equation (2), yielding

$$\begin{aligned} C \cong\ &[\mathrm{d}(P_M M + P_I I)/\mathrm{d}t - (U_H/U_X)\bigl\{H_M(\mathrm{d}M/\mathrm{d}t) + H_K(\mathrm{d}K/\mathrm{d}t) \\ &\quad + H_E(\mathrm{d}E/\mathrm{d}t)\bigr\} + (U_L/U_X)(\mathrm{d}T_M/\mathrm{d}t), \end{aligned} \tag{4}$$

where U_H is the marginal utility of health, U_X is the marginal utility of non-medical consumption ($X \equiv Y - P_M M - P_I I$), U_L is the marginal utility of leisure, and H_M, H_K and H_E are partial derivatives of H with respect to M, K and E.[19] Several comments are worth noting.

[18] The issues under discussion here involving measurement of the cost of living are very different from those raised by the Boskin Commission, who recommended that the BLS move from a Laspeyres price index formula to a superlative index such as the trailing Tornquist, the latter more closely approximating a much more narrow notion of a cost of living index. See US Senate Finance Committee (1996), Boskin et al. (1998), Abraham, Greenlees and Moulton (1998) and Persky (1998).

[19] We assume here that $\mathrm{d}C/\mathrm{d}t = U_X$, and that the marginal price of non-rivalrous additional knowledge is zero.

The first term on the right hand side of Equation (4), $\mathrm{d}(P_M M + P_I I)/\mathrm{d}t$, is additional spending on medical care and insurance services over time. A spending increase may be due to increased quantities of medical services provided (direct or via health insurance), increases in the prices paid for those medical services by consumers/insurers, increases in the carrying cost of insurance or in the quantity of pure insurance services provided. Thus it is clear that an increase in the cost of medical services, *ceteris paribus* (in particular, health outcomes and environment assumed constant), increases the cost of living index. Notice that if the medical environment changes, e.g., a new disease such as AIDS appears, medical expenditures will likely increase, but this is not properly viewed as a change in the cost of living, for the latter assumes an unchanged environment. As Griliches (1997) has pointed out, price index computations assume an average, unaging, unchanging individual living in a world in which nothing changes except prices. When a country's population becomes more aged, medical expenditure and the quantity of medical resources consumed increases, but as stated earlier, this is properly viewed as an expenditure and quantity increase, not a price increase. Similarly, since outcomes are being held fixed in the Laspeyres type hypothetical, if bacteria develop drug resistance and low priced antibiotics are replaced by more expensive drugs, the price index should increase, reflecting the reduced efficacy (quality deterioration) of the older antibiotic.

The second set of terms in the first line of Equation (4), $-(U_H/U_X)\{H_M(\mathrm{d}M/\mathrm{d}t) + H_K(\mathrm{d}K/\mathrm{d}t) + H_E(\mathrm{d}E/\mathrm{d}t)\}$, is the dollar value of change in health over time. Health may change because the quantities of M, K and/or E change, or because of, say, changed efficacy of a given medical treatment (change in H_M). The $-U_H/U_X$ term multiplying $\{\cdot\}$ is the marginal rate of substitution between health and all other goods. Multiplying the health change by this amount expresses health in dollars. Note that an improvement in health through any of these three channels, *ceteris paribus*, reduces the cost of living index, i.e., $C < 0$.

The final term in Equation (4), $(U_L/U_X)(\mathrm{d}T_M/\mathrm{d}t)$, is the change in the time cost of receiving medical care. Hours are converted into dollars by multiplying hours by the marginal rate of substitution between leisure and goods (in most cases, equal to the after-tax wage). If more efficient delivery of medical care reduces patient travel and waiting time for medical care, or if recovery time from orthopedic surgery is reduced due to increased use of arthroscopic surgery, *ceteris paribus*, the cost of living falls.

Our discussion to this point on cost of living is that for a representative consumer. There are various ways in which group or aggregate cost of living measures and price indexes can be constructed, even when consumers' preferences are diverse and income (or total expenditure) has an unequal distribution. As discussed by, among others, Pollak (1980, 1998) and Fisher and Griliches (1995), a common aggregation procedure is to weight each person's utility in each of the two time periods by his/her dollar share in total expenditures; the share weights can be base period, current period, or some average of the two (the Tornquist index). The aggregate cost of living indexes, analogous to Equation (4), then include terms that represent share-weighted averages of various expenditures. Notice also that such aggregate cost of living indexes are conditional on the distribution of income and demographic composition of the population.[20]

Several other issues merit attention. First, it is useful to consider the $P_M M + P_I I$ term in Equation (4) further. One possible price index to compute would involve asking, what price would the consumer pay in two adjacent time periods for an actuarially fair medical care insurance policy to keep on the same expected level of utility, *ceteris paribus*? Note that this is not the same as a disability insurance policy, for with that the beneficiary only recovers lost income and medical care costs, and is not compensated, for example, for lost utility due to loss of vision. The consumer may have an expected life pattern in mind, with age-related probabilities of experiencing certain diseases. Thus the price index would be based on the price of contracting for a year of medical costs, given expected disease susceptibility, technology, efficacy, environmental factors and so forth.

Realizations over the ensuing time period could well change the market price of such an insurance contract, for a variety of reasons, with differing implications for price and quantity. If the consumption of more medical related goods is induced by the expansion of technological opportunities (new artificial hips) or changes in the environment (increased sensitivity to allergens), the change is appropriately viewed as one of quantity or quality, not price. The premium paid on repriced insurance policies might increase as a result, but that is because of the changed technology or environment, not because of a price change. Note also that because of moral hazard, when improvements in technology occur, they may be difficult to value properly within the medical marketplace.

Earlier we noted that a Laspeyer's type of cost-of-living index uses the base period utility as reference, whereas the Paasche employs the current period utility as the reference point. Suppose that in the time interval between the base and current period, the individual experiences a deterioration in health state so severe that it no longer is biologically feasible for the individual in period 1 to maintain the period 0 level of utility, e.g., the individual loses eyesight or develops an illness such as AIDS. In such a case, there may not be any feasible answer to the Laspeyres question, but one might still be able to answer the Paasche question, i.e. what would the necessary change in funds be in period 0 to make the individual as well off as in period 1?

A still deeper problem occurs when unexpected changes take place, such as those that result in the unanticipated lengthening of life expectancy. An individual might want to alter considerably his/her lifetime optimization plan given a change in information, yet he/she may lack the resources to modify consumption to a new path that has now become optimal. Hence it is possible for the cost of living *per year* to decrease, for the cost of living *a lifetime* to increase, all as a result of this unanticipated *benefit* of increased life expectancy. This raises difficult issues, and mixes up changes in cost of living indexes with technological progress. Cost of living indexes typically refer to the

[20] Pollak (1980, 1998) therefore calls these aggregate price indexes "plutocratic", and contrasts them with ones he names "democratic".

cost of a flow of services over a relatively short time period. Converting from, say, a lifetime stock to an annual flow may be reasonable if the population is assumed to be ageless or has a fixed age composition, if there are no unexpected changes, and if *ex ante* decisions are still correct *ex post*. If these conditions are not met, paradoxes may well emerge.

2.3. Forming a price index

A fundamental issue is how one estimates the values of the variables in cost of living index equations such as those in Equations (3) and (4). Suppose we focus attention just on how changes in the medical sector affect the cost of living index. Although current procedures used by the BLS in its medical care related CPIs and PPIs are discussed in detail below, here we briefly consider several alternative procedures.

One approach used in other settings is hedonic price analysis. If one estimates a regression model where the price of a medical service is the dependent variable, and where attributes of the medical procedure, the patient and the provider are explanatory variables, then one can decompose price changes over time into changes in the value of services to patients and pure changes in price. An example of this type of hedonic price measurement, for the treatment of acute phase depression, is in Berndt, Busch and Frank (1998).[21]

There are a number of problems with using hedonic analysis in this market. At the level of individual diseases, hedonic prices are not necessarily equal to consumers' marginal valuations. Since consumers are insured, the price they pay for medical care at the margin is different from the cost of medical care to society. Further, providers have their own incentives in recommending treatment decisions, which may reinforce or contradict consumer preferences. Thus, both because of moral hazard and principal-agent issues, we would not necessarily expect treatment decisions to be made optimally. Estimated parameters in hedonic price equations could therefore be based on data points reflecting socially (and privately) inefficient actions by consumers/physicians/insurers. This raises difficulties in placing any social welfare interpretations on movements in hedonic price indexes over time.

Alternatively, one might perform hedonic analysis at the level of the insurance plan, as was recommended by Reder (1969) and has been implemented by Jensen and Morrisey (1990). One could estimate an hedonic model for the price of insurance, using the attributes of the insurance policy as regressors, and thus infer the residual price increase. The difficulties here are both theoretical and practical.

At the theoretical level, a theory about how consumers choose health insurance plans is required that incorporates consumers' self selection, moral hazard (augmented by

[21] For an introductory discussion to the hedonic method, see Griliches (1988, chapters 7 and 8), and Berndt (1991, chapter 4). Other applications in the medical context include Trajtenberg (1990), Berndt, Cockburn and Griliches (1996), and Cockburn and Anis (1998).

tax subsidies), and preferences for compensation. Since most private health insurance is provided through employment, this involves a link between workers and employers, and between different workers within a firm. Our knowledge about how insurance decisions are made in firms is very limited.[22] Hedonic analysis also presumes that consumers are fully aware of the attributes of the good they are buying. But with health insurance, there are often fundamental parts of the insurance contract that consumers do not know – indeed, cannot know – in advance.[23]

At the practical level, we probably are unable to control for many of the other factors that influence plan costs. For example, plan premiums will depend on the health status of people who are enrolled in the plan as well as the benefits offered. But plan enrollment reflects adverse selection. When premiums change, we need to be able to decompose them into changes in the cost of a given set of benefits, and changes in the sickness of the people enrolled in the plan. Without knowing in detail who is enrolled in each plan and what their expected medical spending would be, we cannot adequately control for the many factors involved in premium variation. Moreover, data to control for these factors are typically unavailable.[24] As with observations on disease-specific treatment costs, the analysis of cross-sectional and/or time series insurance policy data might well be comparing various inefficient equilibria.[25] This is particularly likely with noncontractible aspects of health plan rationing under managed care.

Pauly (1998) has recently revived Reder's proposal to use medical insurance prices as the basis of a price index for medical care. Pauly contends that the development of willingness to pay techniques in economics has become sufficiently advanced that one could now ask respondents to put evaluations on an insurance policy that covered some new medical technique, or a bundle of new medical techniques, compared with an insurance policy that did not cover those techniques. One advantage of this form of pricing insurance policies would be that it would in theory capture behavior towards risk in a way that is typically neglected in studies that address only the *ex post* cost of treating an illness/condition. If one has a disease, the cost of treating the disease matters. If one does not have the disease, then insuring against the risk of a costly medical bill, if the disease is contracted, is important. Though this alternative approach may have advantages over attempting to construct price indexes for the treatment of specific diseases, pricing insurance policies also has significant disadvantages, as discussed above and by Feldstein (1969) many years ago, and willingness to pay techniques remain subject to framing, reference point, and other issues. Moreover, empirical work to implement Pauly's suggestion is not yet available.

[22] See Gruber (1997), Pauly (1997) and Summers (1989).

[23] For example, most consumers do not know the details of who they are allowed to see for cancer care in advance of being diagnosed with cancer. Indeed, the specific benefits may depend on the severity of the cancer of the person and may change with new knowledge about cancer treatment.

[24] This data situation is gradually improving. See Cohen et al. (1996) for a discussion of the Medical Expenditure Panel Survey.

[25] For further discussion, see Feldstein (1969).

Yet another alternative to the hedonic and insurance policy approaches is to make specific assumptions about the way that medical treatment decisions are made. For example, one can assume that consumers have a specified known distribution of preferences for one prescription drug over another, as in Fisher–Griliches (1995) and Griliches–Cockburn (1994), or that consumers are making purchase decisions for goods with a high out-of-pocket share, such as for prescription drugs in Cockburn–Anis (1998). One can then combine this model with observed data on treatment and prices to form a component of the cost of living index. While this approach is reasonable in some applications, it does not work well in markets where consumer information is poor and the share of out-of-pocket costs is low, as occurs in most medical care markets.

A third option involves more direct measurement. Suppose one focuses on a particular disease or condition and estimates empirically the changes in treatment costs and medical outcomes for that disease. If in addition one makes an assumption concerning the dollar worth of health improvements, one can calculate the various individual factors in a cost of living index. This approach has recently been implemented by Cutler, McClellan, Newhouse and Remler (1998a, 1998b). If such an approach were to be followed more generally, it would of course be necessary to undertake such analyses for a representative mix of illnesses where outcomes could be reliably measured. We return to a discussion of this approach later in this chapter.

With this discussion on the difficulties of conceptualizing and implementing price measurement of medical care services as background, we now turn to a review of price measurement procedures currently employed by the BLS in its medical care related CPIs and PPIs. As we shall see, while changes have recently been implemented at the BLS in its medical care CPI and PPI programs, for the most part the BLS still treats medical care in the same way it treats other industry and consumer prices. The combination of inherent difficulties in measuring service industry prices, distinctive features of the medical care industry, and use of traditional index number procedures for measuring prices makes clear interpretation of the BLS' current medical care CPIs and PPIs very difficult.

3. Construction of medical care CPIs and PPIs at the BLS

The US Bureau of Labor Statistics (BLS) constructs and publishes CPIs and PPIs for various components and aggregations of medical care goods and services. Hereafter we designate these medical care CPIs and PPIs with the acronyms of MCPIs and MPPIs, respectively. Although the essential structure and conceptual foundations of these price index measurement efforts have been in place for some time, the BLS has recently announced and undertaken a considerable number of changes in its MCPI and MPPI programs. Here we summarize both continuing and recently changing procedures. We begin with a more general overview of the CPI and the PPI, and then we consider issues particularly important to measuring prices and quantities of medical care goods and services in the MPPI and MCPI programs.

3.1. A brief summary of the CPI

According to the BLS, the CPI is "... a measure of the average change in the prices paid by urban consumers for a fixed market basket of goods and services."[26] It is calculated monthly, and is published about two weeks after the end of the month to which it refers.[27]

From its first regular publication in 1921 until the end of World War II, the CPI was called a "Cost of Living" index. In March 1944 the Chairman of the President's Committee on the Cost of Living appointed a group of technical experts (Wesley Mitchell, Simon Kuznets and Margaret Reid) to examine whether the BLS' cost of living index was properly accounting for war-related quality deteriorations in goods and services, as well as the effects of rationing and shortages. Controversy had emerged in part because in 1942 the "Little Steel Formula" had been adopted which linked permissible wartime wage increases to the index, but representatives from organized labor argued that the cost of living index understated true price inflation.[28]

Along with a Special Committee of the American Statistical Association appointed in 1943, the technical experts concluded that "... the index understated the wartime price rise to some extent because of a number of factors, of which incomplete account of quality deterioration was only one".[29] To avoid confusion with popular notions of cost of living, the President's Committee, as well as the union critique of the index, also recommended that the name be changed to "Consumers' Price Index", a change which was adopted in August 1945.[30]

Since 1945 many changes have occurred involving the BLS' construction of the CPI, but its underlying hierarchical structure has been relatively stable. We now summarize this hierarchical structure.

The identity and number of items sampled, and the weights used in aggregating sampled items into increasingly comprehensive sub-indexes, constitute a hierarchical structure of market baskets that the BLS changes infrequently. Based on data from its Consumer Expenditure Surveys (CEX), the BLS identifies and defines a fixed 'market

[26] US Department of Labor, Bureau of Labor Statistics (1992), *Handbook of Methods*, Bulletin 2414, p. 176.

[27] Here we focus primarily on the CPI for All Urban Consumers (CPI-U), introduced in 1978 and representative of the buying habits of about 80% of the US non-institutional population. An alternative index, CPI-W (wage earners and clerical workers only), was introduced much earlier for use in wage negotiations, and represents but 32% of the US population. The methodology for producing CPI-U is the same as that for CPI-W.

[28] Ethel D. Hoover (1961, p. 1175). Union criticism of the index was written up in a "Meany Report". Also see Persky (1998).

[29] Hoover (1961, p. 1175). Hoover reports that from January 1941 to September 1945, the estimated downward bias was 5 percentage points. Also see Samuel Weiss (1955).

[30] Hoover (1961, fn. 2, p. 1175); also see Weiss (1955, p. 23). Incidentally, the Meany report argued that "To most people, 'cost of living' means the amount of money a family spends. If it buys more food and finer clothes, or moves to a roomier home, its cost of living goes up. That interpretation is so widespread that we think the Bureau's index is misnamed" (Meany Report, p. 18).

basket' of goods, employing a classification system known as the item structure. The item structure has been updated approximately every ten years, the most recent being in January 1998.

For example, based on data from the 1993–95 CEX, the BLS identified eight major product groups of items for representation in the CPI beginning in January 1998: food and beverages, housing, apparel and upkeep, education and communication, transportation, medical care, entertainment, and other goods and services. In turn, these major groups are divided into 70 expenditures classes, which are disaggregated further into 211 item strata. Weights for the 211 item strata are fixed in between major revisions, as are those for the higher level of aggregations of strata into expenditure classes, intermediate aggregates, major groups and all items indexes. The CPI calculations are done separately for 38 geographic areas.[31]

CPI calculations are undertaken based on a modified Laspeyres price index. The Laspeyres price index is a weighted sum of price relatives, where the weights are revenue shares of each of the N item strata in the market basket. For month t, the Laspeyres price index is:

$$L_t \equiv \sum_{i=1}^{N} w_{ib}[p_{it}/p_{ib}], \qquad w_{ib} \equiv p_{ib}q_{ib}/\sum_{i=1}^{N} p_{ib}q_{ib}, \tag{5}$$

where p_{it} is the price of the ith item in time period t, $i = 1, \ldots, N$, p_{ib} are base period prices, q_{ib} are fixed base period quantities, and w_{ib} is the fixed base period expenditure weight. The term p_{it}/p_{ib} is often called the "price relative" of good i. An attractive feature of the Laspeyres index is that it is consistent in aggregation, i.e., one obtains the same composite Laspeyres index by aggregating over all items simultaneously, or first aggregating items into a set of sub-indexes, and then constructing a master aggregate from the weighted sub-indexes.

Because the CPI has 211 item strata, the terms p_{it}/p_{ib} in Equation (5) are in fact price indexes, often called "basic components" or "elementary aggregates". There are thousands, perhaps millions, of "items" in a modern economy. Within each of the 211 item strata, BLS takes a probability sample of the detailed items that are grouped together into each of the item strata. For example, in the medical care component, there are 13 item strata. Based on a nonlinear programming optimization algorithm, the BLS determines the optimal number of price quotes at the expenditure class level.[32] For some of the item strata (e.g., babysitting, car pooling), it is very difficult to obtain sample price data; thus for 27 of the 211 item strata, the BLS does not sample prices, but instead imputes prices from other goods and services in the same expenditure class.

[31] Lane (1996, p. 22). Also see Ford and Ginsburg (1997, 1998). Several of these numbers have been revised since publication of these articles. We thank Dennis Fixler for providing final updates.

[32] This optimization problem and its implementation are discussed in Leaver et al. (1997).

When a detailed item is selected for pricing in the CPI, a price for the exact same item is collected at regular intervals, usually monthly or bi-monthly. These detailed prices are formed into the basic component price indexes, which are the lowest level for which price index information is published in the CPI.

To accommodate practical issues involving the fact that some products are discontinued and cannot be repriced, that consumers' point of purchasing items changes, and that the CEX provide data on expenditures rather than prices, current BLS practice incorporates a number of modifications. The related issues are discussed in further detail in US Department of Labor, Bureau of Labor Statistics (1997b, 1998) and in Moulton–Stewart (1997).[33] Here it is worth emphasizing that while weights may change for elementary items within item strata (due in part to sample rotation), at the item strata level and above the weights are fixed over time between major revisions, and thus for aggregate price indexes at the level of item strata and higher, use of Equation (5) with its fixed weights is essentially what is done by the BLS.

In 1997 the BLS began issuing a monthly experimental measure constructed with use of the geometric mean formula for all index components at levels of aggregation underneath the item strata. The geometric mean index permits limited substitutability among products within the item strata. Provided that commodity substitution is the primary economic behavior that affects these lower level indexes, the difference between geometric and arithmetic mean item strata indexes can be interpreted as a measure of "lower level" substitution bias.[34] This experimental index using geometric means appears to lower the growth of the all-items CPI by approximately one-quarter of one percent per year.[35]

Recently the BLS announced that beginning in January 1999, the aggregating formula for constructing most of the elementary aggregates (comprising approximately 61% of total consumer spending) will be moved over to a geometric mean. Medical care CPI components are largely exceptions, however; all but prescription drugs, and non-prescription drugs and medical supplies will continue to be constructed by the traditional arithmetic mean calculation.[36] Note that for each of the more highly aggregated 211 item strata, fixed quantity weights will still be employed, reflecting the continuing assumption of zero substitutability *between* these strata.

With this overview of the CPI hierarchical structure, weights, and aggregation formulae as background, we now move on to a brief summary of the PPI.

3.2. A brief summary of the PPI

The Producer Price Index (PPI) "measures average changes in selling prices received by domestic producers for their output".[37] Before 1978 the BLS named this price series

[33] Also see Moulton (1996) and Moulton–Moses (1997).

[34] See Pollak (1998) for further discussion.

[35] US Department of Labor, Bureau of Labor Statistics (1997a).

[36] US Department of Labor, Bureau of Labor Statistics (1998), updated in Eldridge (1998).

[37] US Department of Labor, Bureau of Labor Statistics (1992, Bulletin 2414, p. 140).

its Wholesale Price Index (WPI). The change in name to Producer Price Index emphasized that its conceptual foundations were based on prices received by producers from whomever makes the first purchase, rather than on prices paid to wholesalers by retailers or others further down in the distribution chain.[38] At the same time, the structure of the index was changed substantially. The old WPI corresponded, roughly, to the P in the well known quantity theory of value expression, $MV \equiv PT$. In this view of an inflation index, all transactions mattered, so the WPI combined into one index the prices of, e.g., iron ore, steel and the automobile in which the steel was an input.[39] The resulting substantial double counting in the WPI was regarded as a serious problem.[40] In response, the BLS converted the old price index to the concept of an industry output price index.[41] As a result, the basic measurement unit for the PPI has become an industry – in the case of medical care, hospitals, physicians' offices and clinics, and nursing homes are each separate industries. The PPI publishes separate price indexes for the outputs of each of these industries.

The PPI is calculated monthly, and is usually published in the second or third week following the reference month. The PPI involves pricing the output of domestic producers, while the BLS' International Price Program publishes price indexes for both imports and exports.

The PPI program at the BLS takes as its definition of an industry that based on the Standard Industrial Classification (SIC) code.[42] Since its inception in 1902, the PPI has focused heavily on the goods-producing sectors of the US economy, but ever since 1986, in recognition of the growing importance of services in the US economy, the BLS has gradually begun to broaden the PPI's scope of coverage into the service sectors.

Currently the BLS does not calculate and publish an economy-wide aggregate goods and services PPI, although it plans to do so beginning January 2002. Rather, PPIs are published by industry (based on the SIC 4-digit industry code and higher levels of aggregation), by commodity classification (by similarity of end use or material composition, regardless of whether these products are classified as primary or secondary in their industry of origin, for fifteen major commodity groupings), and by stage of processing (according to the class of buyer and the amount of physical processing or assembling the products have undergone), separately for finished goods, intermediate materials, and crude materials, both by commodity and industry classifications. Hereafter we focus primarily on PPIs by industry.

Within each industry, the BLS calculates aggregate PPIs using the Laspeyres price index formulae (see Equation (5) above). At the most disaggregated level of PPI price

[38] *Ibid*, p. 141.

[39] The WPI did not implement this theory completely, however, for it omitted nearly all service prices and also transactions in financial and second-hand assets.

[40] See Council on Wage and Price Stabillity (the "Ruggles Report") (1977).

[41] The implementation of an industry output price index was based on the theoretical model developed by Fisher and Shell (1972), and amplified by Archibald (1977) and Diewert (1983).

[42] Issues concerning how industries are defined and aggregated, as well as economic issues underlying the SIC code system, are discussed in Triplett (1990).

measurement (called the "cell index"), the BLS defines a price as "... the net revenue accruing to a specified producing establishment from a specified kind of buyer for a specified product shipped under specified transactions terms on a specified day of the month".[43] Prices are for output currently being provided or shipped, and not for order or futures prices.[44] Although in general the BLS seeks transactions rather than list prices for its price quotes, responses by firms are less cumbersome when list rather than transactions prices are reported.[45] Participation in the PPI by firms is on a voluntary basis. As of December 1992, the overall PPI "productive" response rate was 63%.[46]

The PPI is also based on a hierarchical system, though as noted above, unlike the case of the CPI, currently there is no economy-wide measure of the PPI. The BLS constructs and publishes aggregate PPIs for the total mining and total manufacturing industries, but apparently because of a lack of sufficient coverage, the BLS does not currently publish an aggregate PPI for total services; the BLS hopes to publish such an aggregate services industry PPI by January 2002.[47]

Price quotes from the most disaggregated cell indexes are aggregated via a Laspeyres weighting scheme, where fixed weights are based on value of shipments data collected primarily by the Bureau of the Census; industry net output weights are employed to take account of intraindustry sales. The net output weights therefore vary with the level of industry aggregation (e.g., four-digit to two-digit); the detailed industry flow data required to distinguish net from gross output are derived for the most part from use of input–output tables compiled by the Bureau of Economic Analysis. Beginning in January 1996, industry price indexes have been calculated primarily with net output weights based on 1987 input-output relationships. The 1992 input-output tables have just recently been released, and the BLS envisages using them by the end of 1998 or early 1999.[48]

With respect to the specific establishments and items sampled by the BLS in its PPI program, the BLS currently draws a sample of items for each industry on average every seven years or so, and then reprices this fixed set of items monthly until an entirely new sample is drawn. Since 1978, the BLS has attempted to employ a sampling procedure that makes the probability of selection be proportional to a product's value of shipments. Because it recognized that in some technologically dynamic industries a seven year time lag between samples could result in a sample of products and services much older and quite unrepresentative of market transactions, in 1996 the BLS announced

[43] US Department of Labor, Bureau of Labor Statistics (1992, ch. 16, p. 141). Net revenue is net of any discounts as opposed to net of production costs.

[44] Problems can emerge for industries in which a great proportion of currently shipped output is covered by long-term price contracts, but for which "spot" prices differ from contracted prices.

[45] US Department of Labor, Bureau of Labor Statistics (1992, ch. 16, pp. 141–142).

[46] Catron and Murphy (1996, Table A-2, p. 31).

[47] US Department of Labor, Bureau of Labor Statistics, "Producer Price Index Coverage Expansion Plan", December 1996, p. 1.

[48] Lawson (1997).

that for certain industries, including pharmaceuticals and electronics, samples would be supplemented at one or two-year intervals.[49]

Issues surrounding the reliability and possibility of biases in price index measurement have recently received much less attention for the PPI than for the CPI. Use of the Laspeyres weighting procedure, accounting for unmeasured quality changes, and discontinuation and exit of sampled goods and services, raise issues which in many respects are similar for the CPI and PPI. On the other hand, a number of significant differences exist between the CPI and PPI medical care components.

First, the lowest level of aggregation is defined differently in the two indexes: the PPI is defined on four-digit SIC industries, and below that, the item detail is defined specifically to each of the medical care industries (DRG major groups for the hospital index, medical specialties for the physicians' index, and so forth). The item strata in the CPI are based on groups that, in principle, should correspond to consumer demand categories.[50]

Second, the frequency and nature of major revisions differ. The CPI has been revised every ten to twelve years, when new weights are assigned based on the consumer expenditure survey. The PPI is normally rebased every five years, with weights drawn from the economic censuses.[51]

For the medical price indexes, another major difference exists between CPI and PPI. In the case of the PPI, revenues and output prices collected from the sampled unit refer to revenues from all sources – government, industry and final consumers. For the CPI, only consumers' out-of-pocket costs are included. Government expenditures made on behalf of consumers and financed by taxes, and health expenditures by insurance companies where employers (but not consumers directly) pay the premiums are out of scope for the present definition of the CPI. We return to discuss this difference in scope in Section 3.4 below.

With these overview discussions of the CPI and PPI as general background, we now move to consideration of issues of particular importance to medical care goods and services. We begin with the MPPI.

3.3. PPIs for medical-related goods and services

As noted above, the BLS does not construct and publish a PPI for an aggregate of services. Nor does the BLS publish a PPI for an aggregate consisting of medical-related goods and services. Indeed, it is only within the last decade that the BLS, as part of its increased effort to measure prices in the various service industries, has begun publishing price indexes for hospital and physician services.

[49] See Kanoza (1996).

[50] See Lane (1996).

[51] Another difference involves the length of time between collection of underlying sales revenue census/expenditure survey data and the introduction of new weights into the Laspeyres index. For the PPI, this is about one to two years, but for the CPI it has been about three to four years.

Among the manufacturing industries associated with health care, the BLS has published PPIs for some time for industries such as pharmaceuticals; hospital beds; medical books; surgical, medical and dental instruments and supplies; ophthalmic goods and others. Among the service industries, separate PPIs for numerous health care related industries are a rather recent development. A PPI for health services was introduced by the BLS effective 1994.12, that for offices and clinics of doctors of medicine in 1993.12, for skilled and intermediate care facilities 1994.12, for hospitals in aggregate and by type in 1992.12, and for medical laboratories in 1994.6.

If the BLS is ever to construct an aggregate MPPI, as with other industry aggregates, it will need to distinguish net from gross output by industry, using some form of an input–output matrix to measure inter- and intra-industry flows. Given the major changes in the health care sectors over the last decade, including impacts from the growth of managed care, it will of course be necessary to employ input–output matrices that are based on much more recent data than the 1987 input–output matrix currently employed by the BLS for defining net output in other industries. While the BLS plans to begin using 1992 input–output data beginning in late 1998, these data will already be six years out of date, and much organizational and technological change has occurred in the health care industries since 1992.

3.3.1. Output measurement in the MPPI

A central measurement issue in the construction of MPPIs involves the specification and implementation of a concept of industry output. Although the PPI program utilizes the four-digit SIC classification system to identify and define industries, this SIC structure does not provide information enabling the BLS to define what is the appropriate real output concept in medical care industries, and on how this output quantity and output price can best be measured.[52] As we shall see, important problems also emerge when medical treatments from distinct SIC industries are substituted for each other in treating an illness or condition.

In the US, medical goods and services were traditionally paid for by fee-for-service arrangements. In a fee-for-service context, a reasonable business procedure involves identifying and separately billing for each particular component of medical care from, say, a physician, a hospital and a pharmacy. The fee-for-service was essentially the price for the inputs to medical care.

In 1983 the Health Care Financing Administration (HCFA) introduced major changes in how general acute care hospitals treating Medicare patients were to be reimbursed. Specifically, beginning in 1983 HCFA implemented a prospective payment system for inpatient hospital care, whereby general acute care hospitals received a fixed payment for almost every Medicare patient admission, regardless of the amount or duration of services actually provided the patient. This prospective payment mechanism represented

[52] For a discussion of the economic foundations underlying SIC definition, see Triplett (1990).

a sharp departure from the retrospective cost-based accounting framework used for many years.

Medicare prospective payment schedules are based on estimates of (average accounting) costs for the resources utilized in providing services for a typical patient in a given geographical area being treated for a particular medical case. As of 1995 payments were distinguished for treatments of 24 major diagnostic categories, which are broken down further into 495 medical and surgical groupings, known as diagnostic related groups (DRGs).[53] The DRG prospective payment schedules have been updated regularly by the Congress utilizing recommendations from HCFA and the Prospective Payment Assessment Commission (now the Medicare Payment Advisory Commission); updates include changes in "medical costs" and case-mix indexing to account in part for secular trends in upcoding, also known as "DRG creep."

DRGs provide one possible output concept, and while DRGs in theory are applicable to all populations, Medicare currently employs DRGs only to reimburse hospitals for inpatient hospital care; many outpatient commodities (e.g., home health care) and services for illnesses of the elderly, and particularly of the non-elderly, are not included in the DRG system.[54]

Classification schemes used for other services include version four of Current Procedural Terminology (CPT4) codes, a list containing thousands of procedures for which physicians and hospitals can bill; these CPT4 codes can be envisaged as inputs into the treatment of an illness or condition.[55]

A systematic structure of diagnostic codes for illnesses and conditions is version nine (now version ten) of the International Classification of Diseases (ICD-9).[56] Relationships among ICD-9, CPT4 and DRG codes are multifaceted. A single DRG encompasses treatment of somewhat arbitrary aggregations of distinct ICD-9 diagnoses, alternative combinations of CPT4 codes can be used in the treatment of a particular ICD-9 diagnosis, and a given CPT4 procedure can be used in the treatment of various ICD-9 diagnoses. Other diagnostic-related systems used in setting risk-adjusted capitation rates include the Ambulatory Care Group algorithm[57] and the Hierarchical Coexisting Conditions model.[58]

DRGs and their offspring represent the beginning of a structure which could facilitate defining, measuring and pricing the output of medical care providers. In particular, the

[53] A number of these 495 DRGs are no longer valid. For a recent list, see Prospective Payment Assessment Commission (1995, Appendix E).

[54] DRG weights have been calculated for non-elderly patients for Maryland, New Hampshire and New York, but no DRG non-elderly weights exist based on national data. However, a limited number of private insurers use DRGs for non-elderly beneficiaries, as do several state Medicaid programs. Also see footnote 61 below.

[55] For a discussion of CPT4, see American Medical Association (1990).

[56] ICD-9 codes are discussed and listed in US Department of Health and Human Services (1980). The ICD-9 system with clinical modifications is called ICD-9-CM, and it has recently been updated to version 10.

[57] Weiner et al. (1996).

[58] Ellis et al. (1996).

output of a particular DRG billing involves the treatments for an episode of hospitalization for a particular condition/diagnosis. Instead of pricing each of the components of a hospitalization, with DRGs the composite bundle of hospital services is given a single *ex ante* price.[59]

Along with the development of CPT4 and ICD codes, the notion of an episode of illness or treatment has expanded far beyond the hospitalization realm, suggestive of yet alternative ways of measuring medical care output. Numerous professional medical associations, as well as the Agency for Health Care Policy and Research (AHCPR) – an agency of the Public Health Service in the US Department of Health and Human Services, have developed clinical practice guidelines and treatment protocols for various illnesses and conditions. These treatment guidelines, which change over time, define *ex ante* medically acceptable and often therapeutically similar bundles of treatment involving medical inputs such as laboratory tests, pharmaceuticals, minutes of service from physicians and other medical personnel, and various other inpatient and outpatient procedures. Health insurance plans, hospitals and pharmaceutical companies have developed programs and protocols for the management of certain diseases. These disease management programs implicitly, and sometimes explicitly, suggest outputs of the medical sector that facilitate the pricing of treatment bundles and the accounting framework for assigning payments to providers.[60]

With this as background, we now turn to a discussion of how the BLS' PPI program has implemented medical care sector output price and quantity measurement, and how it has built on the notion of DRG treatment episodes as output measures in the health care sector. As noted earlier, within the last decade the BLS's PPI program has made major changes in, and introduced many new, health-care related PPIs. We begin with a discussion of medical services – physicians and hospitals, and then we discuss selected medical goods, such as pharmaceuticals.

3.3.1.1. Physicians' services in the MPPI. The PPI program has initiated procedures for constructing medical service PPIs at two rather aggregate levels, physician services and hospital services. Each of these two classes of services in turn encompasses a variety of more detailed physician and hospital service industries. In Table 1 we list the entire set of detailed physician, hospital and medical laboratory industries in SIC 80 for which the BLS is currently constructing health services sub-index PPIs.

With respect to offices and clinics of doctors of medicine ("physician services"), the new BLS procedures distinguish Medicare from non-Medicare treatments Within the non-Medicare treatments, multispecialty group practices are treated separately from one and two physician practices and single specialty group practice, with the latter in turn being broken down into nine specialties. For skilled and intermediate care facilities, public payers are distinguished from private.

[59] The Medicare payment scheme reserves 5% of its payments for outlier or exceptionally expensive cases. At the margin, these are reimbursed on a cost basis.

[60] See Triplett (1998b) for further discussion.

Table 1
Sub-indexes of the health services PPI

Industry	SIC code
Health Services	80
Offices and clinics of doctors of medicine	8011
Primary services	8011-P
Medicare treatments	8011-1
Non-Medicare treatments	8011-3
One and two physician practices and single specialty group practices	8011-31
General/family practice	8011-311
Internal medicine	8011-312
General surgery and other surgical specialties	8011-313
Pediatrics	8011-314
Obstetrics/gynecology	8011-315
Radiology	8011-316
Psychiatry	8011-317
Anesthesiology	8011-318
Other Specialty	8011-319
Multispecialty group practices	8011-331
Skilled and intermediate care facilities	8053
Primary services	8053-P
Public payers	8053-101
Private payers	8053-301
Other receipts	8053-SM
Hospitals	806
General medical and surgical hospitals	8062
Primary services	8062-P
Inpatient treatments	8062-1
Medicare patients	8062-131
All medical diagnosis related groups	8062-13101
All surgical diagnosis related groups	8062-13103
Medicaid patients	8062-171
All other patients	8062-171
Diseases and disorders of the nervous system	8062-17101
Diseases and disorders of the eye	8062-17102
Diseases and disorders of the ear, nose, mouth and throat	8062-17103
Diseases and disorders of the respiratory system	8062-17104
Diseases and disorders of the circulatory system	8062-17105
Diseases and disorders of the digestive system	8062-17106
Diseases and disorders of the hepatobiliary system and pancreas	8062-17107
Diseases of the musculoskeletal system and connective tissue	8062-17108
Diseases and disorders of the skin, subcutaneous tissue and breast	8062-17109
Endocrine, nutritional, and metabolic diseases and disorders	8062-17111
Diseases and disorders of the kidney and urinary tract	8062-17112
Diseases and disorders of the male reproductive system	8062-17113
Diseases and disorders of the female reproductive system	8062-17114
Pregnancy, chilbirth and puerperium	8062-17115

continued on next page

Table 1, *continued*

Industry	SIC code
Newborns and other neonates with conditions originating in the perinatal period	8062-17116
Diseases and disorders of the blood and blood forming organs and immunological disorders	8062-17117
Myeloproliferative diseases and disorders, and poorly differentiated neoplasms	8062-17118
Infectious and parasitic diseases (systemic or unspecified sites)	8062-17119
Mental diseases and disorders	8062-17121
Alcohol/drug use and alcohol/drug induced organic mental disorders	8062-17122
Injuries, poisonings and toxic effect of drugs	8062-17123
Burns	8062-17124
Factors influencing health status and other contacts with health services	8062-17125
Outpatient treatments	8062-3
Medicare patients	8062-311
Medicaid patients	8062-331
All other patients	8062-351
Other receipts	8062-SM
Psychiatric hospitals	8063
Primary services	8063-P
Inpatient treatments	8063-1
Medicare patients	8063-101
Non-Medicare patients	8063-103
State and county hospitals	8063-10301
Private hospitals	8063-10303
Outpatient treatments	8063-2
Other receipts	8063-SM
Specialty hospitals, except psychiatric	8069
Primary services	8069-P
Inpatient treatments	8069-1
Rehabilitation hospitals	8069-101
Children's hospitals	8069-104
Alcoholism and other chemical dependency hospitals	8069-107
Other specialty hospitals except psychiatric	8069-108
Outpatient treatments	8069-3
Other receipts	8069-SM
Medical laboratories	8071
Primary services	8071-P
Pathology and laboratory	8071-1
Urinalysis	8071-102
Chemistry, toxicology, and therapeutic drug monitoring	8071-103
Hematology	8071-104
Pathology	8071-107
Profiles and panels	8071-108
Radiological tests	8071-3

The second principal sub-index within health services is hospital services. As is seen in the bottom panel of Table 1, general medical and surgical hospitals are differentiated from psychiatric hospitals, and specialty hospitals except psychiatric. Both inpatient and outpatient treatments are separated into those involving Medicare patients, Medicaid patients and all other patients; for the non-Medicare and non-Medicaid inpatients, hospital treatments are differentiated involving 23 distinct illnesses/diseases/conditions.

Development of the BLS' PPIs for physician services has benefited considerably from the prior implementation and common usage of the DRG, CPT4 and ICD classification systems by insurers, hospitals, physicians, and other providers.[61] Based on a sampling universe including all physician practices in the US, the BLS employs probability sampling stratified by size and specialty. The size of a physician practice is based on the number of physicians in a given practice (not the number of employees, or revenues); the sample is stratified further into nine single specialty categories and one multi-specialty category. Initially in 1993–4 it was expected that the total number of physician practices sampled would be about 400 and the number of quotes obtained would be about 1150,[62] but by mid-1997 only 158 units remained in sample, yielding 845 quotes.[63]

Given the sampling unit, at the price quote initiation point in time, the BLS randomly chooses a bill that measures the net prices paid to a physician's practice for the entire set of services or procedures provided during an office visit, distinguished by type of payer (cash, third party insurance, Medicaid, Medicare, etc.).[64] The physician's output from this visit is represented by the content of the patient's bill, including all the CPT codes associated with that visit. To ensure that the unique combinations of inputs listed on a bill associate with a particular medical condition or surgical procedure, an association which is critical for repricing, the BLS also employs the ICD system, a coding scheme with which physician offices have considerable familiarity.[65]

It is worth noting that the net transactions price by payer type requested by the BLS represents the actual anticipated revenues, including discounts, and not billed charges based on, for example, a "chargemaster".

With this sample bill, the BLS contacts the sampled physician unit each month, and asks it to reprice what the current net transactions prices would be for that particular bundle/payer of services. Thus items on the sample bill remain fixed over time (between major revisions), but item prices could change. Because transaction prices may vary from private payer to private payer, this may present considerable difficulty in practice.

[61] It is interesting to note, however, that in 1996 the percentage of preferred provider organizations reimbursing hospitals by DRG-based methods was only 31.7% (80.2% used per diem methods), and that only 7.7% of hospitals were reimbursed by PPOs using DRG-based methods. See Hoechst (1997, p. 86).

[62] See US Department of Labor, Bureau of Labor Statistics, "A Description of the PPI Physician Services Initiative", not dated, p. 2.

[63] Dennis Fixler and Mitch Ginsburg (1997, 1998).

[64] According to Fixler and Ginsburg (1997, 1998), in 1996 12% of physician revenues came from Medicaid, 43% from private insurance, 18% from consumers' out-of-pocket, and 27% from Medicare.

[65] How the pattern of comorbidities is allocated in such cases is not clear.

Indeed, some payers pay the physician in part or in whole by capitation, thereby making the price for any specified mix of services arbitrary.

PPIs for physicians' services have been published beginning December 1993. Monthly repricing of physicians' bills presents the BLS with numerous practical difficulties. In some cases, bills are purged from the physicians' accounting system, and therefore cannot be repriced; this has occurred for about 35 (4–5% of all) quotes each year. In other cases, the reporter at the sample unit has refused to provide line by line quotes; this has transpired for about 25 (3% of all) quotes each year.[66]

In addition to facing such repricing difficulties at physician practices continuing to cooperate with the PPI, the BLS is operating in an environment in which the organization of physician practices has undergone dramatic changes in the last few years as practices have been consolidated and sold to larger provider groups. Thus it is not surprising that sample attrition for physicians services has been considerable. The impact of this physician practice and bill repricing attrition on the representativeness of the current sample frame is currently unknown.

Finally, in terms of quality change, serious difficulties remain, even with the use of CPT codes. For example, if a new laboratory test becomes available that is more sensitive, reliable and expensive, yet is used for diagnosis of the same condition and has the same CPT code as its predecessor, it will be considered a price change.[67] In such a case, quality improvements will not be incorporated. On the other hand, if the laboratory tests are read and examined by less experienced technicians having larger error rates but price is constant, quality declines would be overlooked. Currently the BLS makes no quality adjustments for the physician or laboratory services component of the MPPI.[68]

3.3.1.2. Hospital services in the MPPI. We now turn to the PPI for hospital services, which the BLS has published since its December 1992 base period. The hospital services PPI measures anticipated net prices paid to hospitals for the entire bundle of services received during a hospital stay, given the type of payer. The hospital's output is represented by the content of a patient's bill, including all room, medical supplies, drugs and ancillary services provided the patient during a single hospital stay; for an outpatient visit, the hospital output is the anticipated net revenues to be received for medical supplies, drugs and ancillary charges accruing from a single hospital visit.

As with the PPI for physicians' services, the hospital service PPI attempts to be based on patients' bills that specify the purpose of the hospitalization, as recorded by ICD codes; such an association is important so that repricing is based on a unique combination of inputs listed on the bill with a particular medical condition or surgical procedure. This focus on hospitalization episode for a particular treatment is preferable to pricing

[66] Fixler and Ginsburg (1997, 1998).

[67] If the CPT code changes, either a new bill will be constructed and repriced, or the new and old laboratory test will be linked in.

[68] For further discussion of quality adjustments, see Moulton and Moses (1997) and Nordhaus (1998).

based on bed-days, drugs, tests, etc., irrespective of the patient's illness. To take into account the possibility that price per bed-day is increasing along with a reduction in average length of stay, when repricing the BLS' PPI program now explicitly asks whether the there has been a change in average length of stay for the hypothetical price quote. If such a change has occurred, it is treated as a quality change, not simply a price change. As of 1998, the change in average length of stay is the only adjustment the hospital PPI makes for quality change.[69]

In principle, net transactions prices incorporate effects of discounts, and therefore are not "list" or "chargemaster" prices. It is not known what proportion of transactions in hospitals *actually* involve only list prices, list prices less certain adjustments, or capitation, and how this has changed since, say, 1992. Although the BLS clearly seeks to obtain price quotes based on net transactions prices, in a recent GAO report involving the MCPI it was noted that only about 15% of the hospital price quotes obtained by the BLS included discounts.[70] In Catron and Murphy (1996), however, it is reported that with the MPPI, 43.4% of the sampled inpatient price quotes and 64.6% of its outpatient price quotes initially collected in 1992 were based on list prices.[71] As with physician services, capitation for hospitals raises further issues, for it calls into question the whole basis of pricing, since it is based on health plan enrollment rather than use of hospital services by any given patient.

The sampling frame for the hospital services PPI is based on a universe compiled by the American Hospital Association, with the probability of a hospital being sampled being proportional to its revenues.[72] The sample is stratified on the basis of size (measured by number of beds), public vs. private ownership, and type of medical specialty. When initially implemented in 1992, given an expected voluntary response rate of 63% (similar to that for other PPI industries), the expected sample size was 558, and the total number of expected monthly price quotes was 2707. By mid-1997, however, the actual sample size was 42% smaller at 322,[73] while the number of quotes was 15% smaller at 2302.[74]

Once a hospital is identified as a sample unit, at the time of sample initiation the BLS chooses a fixed subset of DRGs, and each hospital is then asked on a monthly basis to report on net transactions prices of a single representative patient bill (typically, the last patient bill on file for that DRG) for each of the randomly assigned DRGs.

[69] Correspondence with Dennis Fixler.

[70] United States Government Accounting Office (1996, p. 58).

[71] Catron and Murphy (1996, Figure 1).

[72] As noted by Catron and Murphy (1996, p. 25), Federal hospitals, such as those associated with the military, Veterans Administration and the National Institutes of Health are excluded from both the CPI and PPI hospital universe, because there are no measurable economic transactions between hospital and patient at these Federal hospitals – many services are rendered free to the patient from a budget allocated to a Federal entity.

[73] Fixler and Ginsburg (1997, 1998). The breakdown of actual vs. expected is 211 vs. 358 for general hospitals, 39 of 75 for psychiatric hospitals, and 72 of 125 for specialty hospitals.

[74] *Ibid*. The breakdown on actual vs. expected quotes is 1602 vs. 1889 for general hospitals, 209 vs. 283 for psychiatric hospitals, and 72 vs. 125 for other specialty hospitals.

The DRGs are selected using selection probabilities proportional to expenditures in each DRG based on HCFA and other data from a number of payer sources. Since the identical treatment bundle is not always observed in subsequent months, BLS reporters construct subsequent hypothetical DRG bundle prices by repricing the identical inputs. BLS notes that when a particular hospital does not perform the targeted DRG service, the hospital can instead provide quotes for several alternative DRGs listed by the BLS on the Quote Assignment Sheet.[75] Attrition in the BLS's hospital repricing program is likely to be affected by movement away from DRG billings by hospitals, particularly for non-Medicare patients, and is therefore an important issue worthy of close scrutiny in the very rapidly changing hospital marketplace.

It is also worth noting that in recent years, as hospital length of stay has fallen, the use of post-acute care services such as skilled nursing facilities and rehabilitation units has increased. Often these treatment centers are owned by and even physically located in the hospital. Pricing a hospital stay may present a substantially biased picture of the price of an episode of treatment.

Finally, as noted earlier, the PPI distinguishes as "industries" the "hospital industry" and the "physicians' office industry," largely because the mixes of production processes observed in these two types of establishments are, if not completely disjoint, at least demonstrably not the identical set of production processes. On its own terms, this is clearly reasonable. However, this industry distinction creates a substitution bias with respect to an index for the *purchasers* of health care. Specifically, the problem that arises is that from the purchasers' vantage, the same "product" or service might be "produced" by different industries or by different production processes. For example, with both the physicians' services and hospital services PPI, the nature of the fixed and itemized components for the price quotes requested by the BLS does not permit major input substitution for the treatment of a condition, such as changing the mix of psychotherapy and psychotherapeutic drugs used for the treatment of acute phase depression. When this occurs, even if the industry price indexes are in some sense measured correctly, the PPI measures will miss the purchasers' gain from shifting between different suppliers.

3.3.1.3. Medical products in the MPPI: pharmaceuticals. To this point we have discussed the services component of health care, rather than the goods or commodities components. Although numerous manufacturing products are related to the provision of health care, here we focus on one industry class that has received considerable treatment to date and is perhaps the most significant medical goods industry, namely, prescription pharmaceuticals.[76]

[75] US Department of Labor, Bureau of Labor Statistics, "A Description of the PPI Hospital Services Initiative", not dated.

[76] Since sampling and disaggregation procedures for prescription pharmaceuticals are very similar to that in other PPI industries, we do not discuss construction of the pharmaceutical PPI in detail here. See Berndt, Griliches and Rosett (1993), and the references cited therein, for further discussion.

Prescription pharmaceuticals is a relatively research-intensive industry characterized by a considerable number of new product introductions, and therefore it creates substantial challenges for accurate price measurement. Not surprisingly, the BLS' treatment of prescription pharmaceuticals has long been the subject of controversy. As Dorothy Rice and Loucele A. Horowitz noted thirty years ago, for many years the BLS sample tended to focus excessively on old products: "Until 1960, only three prescribed drugs – penicillin, a narcotic, and a non-narcotic – were included. In that year the list of prescripted drugs was increased to 16 items."[77] Describing the Stigler Commission's Report of 1961, Rice–Horowitz noted that "The Subcommittee urged more prompt introductions of new products – a matter of particular importance in the case of drugs and prescriptions."[78]

More recently, a detailed audit of the BLS' PPI for prescription pharmaceuticals was conducted by Berndt, Griliches and Rosett (1993), which was updated and extended by Berndt–Greenberg (1995). Although these studies examined transactions at a slightly different point in the distribution chain than does the PPI (transactions from wholesalers to retail drug stores, rather than from manufacturers to their initial customer, typically wholesalers), the Berndt et al. studies raised a number of significant issues. In particular, three important findings from these studies were that: (i) the BLS oversampled older goods and undersampled new and middle-aged pharmaceuticals; and (ii) prices of older products increased more rapidly than those of products earlier on in their life cycle.[79] As a result, (iii) the BLS overstated prescription drug price inflation, by perhaps as much as three percentage points a year over the 1986–91 time period. Corroborating evidence has since been reported by others, including the BLS.[80]

Partly in response to this research, the BLS implemented a new sampling method by which newer products are introduced more rapidly. Specifically, to compensate for the age bias in the BLS prescription pharmaceutical sample, in 1995 the BLS linked in a Supplement I sample of about 49 additional drugs newly approved by the FDA since 1992 (the original 1993 sample had 522 products from 92 manufacturers, but attrition to 1995 reduced the 571 to 544), and included these in their sample effective December 1995. As noted by Kelly (1997), the resulting PPI with supplemental sampling rose 2.1% in 1996. Had this supplement not been introduced, the PPI would have risen 2.7% (based on a BLS research index); in three of the 14 months since the introduction of the supplement, price changes in the published index exceeded that of the research index.

[77] Rice–Horowitz (1967, p. 14).

[78] Rice–Horowitz (1967, p. 15). The Stigler Commission report is found in US Congress, Joint Economic Activity (1961). Also see US Department of Health, Education and Welfare (1967, p. 35).

[79] These findings were essentially anticipated almost thirty years earlier by John Gardner, Secretary of Health, Education and Welfare. In his Report to the President, Gardner stated "It is difficult to adjust the drug component of the CPI for the rapid changes in the character of the drugs prescribed. By the time a prescription item is incorporated into the index, its price may have fallen to a lower level than in previous years." US Department of Health, Education and Welfare (1967, p. 35).

[80] Kanoza (1996), Ristow (1996) and Kelly (1997).

One year later the BLS constructed and linked in a Supplement II sample, bringing the total number of observations to 561 (after additional attrition). As is noted by Kelly (1997, p. 17), "In the 14 months since January 1996, the published index has risen 3.3%. Had Supplements I and II not been introduced, the index would have risen 4.1%." Apparently the BLS now plans to add supplements to this industry on an annual basis.

Another area in which the BLS MPPI has recently made substantial changes involves generic drugs. Until several years ago, the BLS procedures for its pharmaceutical PPIs treated generic drugs as entirely unrelated to their patented antecedents. Griliches and Cockburn (1994) noted that generic drugs were a special case of the more general "new goods" problem facing statistical agencies such as the BLS. Since the US Food and Drug Administration certifies generics as being therapeutically equivalent to brand name versions of the same chemical entity, conventional problems encountered when valuing new goods are much simpler with generic drugs. Griliches and Cockburn illustrated the empirical significance of linking generic drugs to their patented antecedents (based on an assumed uniform distribution of tastes between brands and generics), and contrasted their preferred price index construction procedure with that employed by the BLS at that time. Based on data for two antibiotic drugs, Griliches–Cockburn showed that with a Paasche approximation to the "true" index, using reservation prices based on the uniform distribution yielded a price index 25% lower after two years than a Tornqvist index that introduced generics as quickly as was feasible but treated them as new goods, and was 36% lower than an index that mimicked the procedures then employed by the BLS. Several years later, Berndt, Cockburn and Griliches (1996) extended the Griliches–Cockburn research and showed that for the entire class of antidepressant drugs, the BLS' overstatement of price inflation due to the way it handled generic drugs was more than four percentage points per year from 1986 to 1996.[81]

The BLS has announced major changes in how it treats generic drugs in its PPI; these changes are summarized in Kanoza (1996) and Kelly (1997). In particular, effective January 1996, for drugs in the BLS sample losing patent protection and experiencing initial generic competition, the BLS split the fixed weight for that molecule into two parts – 64.2% for the generic, and 35.8% for the brand. Thus the new BLS procedure treated the composite molecule price change as a pure price change. The 64–36 percentages were arrived at as a result of a BLS literature review on typical generic penetrations following the expiration of patent protection. The percentage splits were the same for all molecules, and were fixed over time. Beginning with the Supplement II sample introduced in late 1996, however, the BLS brand-generic split was based on actual brand-generic dollar sales, using data purchased by the BLS from IMS America.[82]

[81] In both the Griliches–Cockburn and Berndt, Cockburn and Griliches studies, transactions were measured at the point of wholesaler to drug store, and not at the initial point in the distribution chain, which is the focal point for the PPI.

[82] The relative growth rates of the published and research PPIs for the pharmaceutical industry, discussed in several earlier paragraphs, reflect the impacts of incorporating both new generics and new branded products into the sample.

There is one other curiosum involving the prescription pharmaceutical PPI. As noted in Berndt, Cockburn and Griliches (1996), for historical reasons involving preferential federal tax treatment, many US pharmaceutical firms currently manufacture drugs in Puerto Rico; the Puerto Rican value of shipments for prescription pharmaceuticals is roughly 20–25% of that on the mainland, and is likely to be higher for newer molecules. For purposes of its PPI calculations, however, the BLS is mandated to treat Puerto Rico as outside the US, and thus the PPI excludes all Puerto Rican production.

It turns out that how one deals with Puerto Rican economic accounts differs across government statistical agencies, and even within the BLS. For example, the Bureau of Economic Analysis' national income and product accounts exclude Puerto Rican production and that of other dependencies, but in the balance of payments accounts, Puerto Rico is treated as domestic. The Census Bureau defines the US as the US customs territory, which consists of the fifty states, DC and Puerto Rico, plus US foreign trade zones and the US Virgin Islands. Within the BLS' International Price Program (IPP), Puerto Rico is considered as part of the US, and thus currently no IPP price quotes are obtained for Puerto Rican pharmaceutical products shipped to the fifty United States.

The issue of how one treats Puerto Rican production is important to the reliability and interpretation of the prescription drug PPI. If Puerto Rico is to be excluded, as is now the case for the PPI, then to the extent public policy analysts and others seek to track the price growth emanating from US producers (many of whom have chosen to produce significantly in Puerto Rico), it will be necessary to collect and publish "import" price series from Puerto Rico, and then to combine those data with the more narrowly defined "domestic" mainland price series.[83] Of total pharmaceutical shipments "imported" into the US from throughout the world, it appears that about 15% emanate from Puerto Rico.[84]

3.4. Medical care products and services in the CPI and MCPI

Medical components of the CPI and PPI programs at the BLS have rather different heritages. It is only within the last decade that the BLS' PPI has extended coverage to a wide variety of service industries, such as medical care. Thus, construction and design of the recently introduced medical care-related PPIs, such as those for physicians' and hospitals' services, have had the opportunity of benefiting from recent thinking and

[83] One incentive for Puerto Rican production has been Section 936 of the Internal Revenue Code, which has provided tax benefits to firms producing in Puerto Rico. It is worth noting that under the omnibus minimum wage bill enacted by the US Congress in 1996, these tax incentives will be phased out over the next decade. Thus it is possible that the empirical significance of this out of scope Puerto Rican production will gradually decline. It is also worth noting that active ingredients of pharmaceuticals could be manufactured in Puerto Rico, shipped to the domestic US, and then be encapsulated with inert materials into tablets and capsules in the US In such a case, the BLS' PPI program would consider it as within the scope of the PPI.

[84] See Table 2 in US Trade with Puerto Rico and US Possessions on the web site http://www.census.gov/prod/3/98pubs/ft895-97.pdf. We thank Dennis Fixler of the BLS for providing information on this matter.

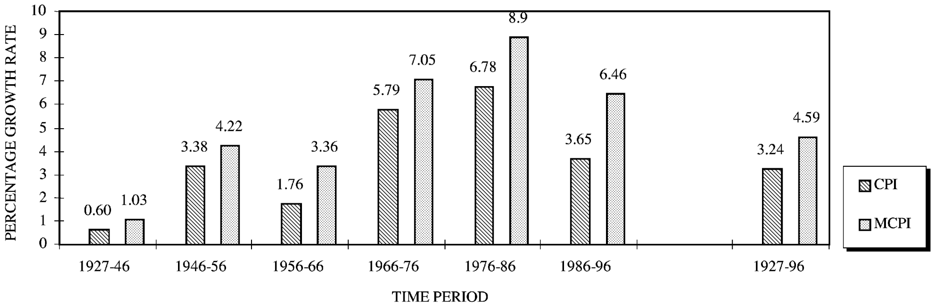

PRICE INFLATION IN THE OVERALL CPI AND IN THE MEDICAL CPI, 1927-96

Figure 1. Source: Getzen (1992) and US Bureau of Labor Statistics.

developments on what in fact are the outputs of the service industries, and how one might measure prices in the context of rapidly changing market structure. By contrast, the medical CPI has been published for a very long period of time, regularly since 1935.[85]

A remarkable fact in the BLS' medical CPI is summarized in Figure 1.[86] Since 1927, the first year for which MCPI data are available, and for each decade since then, measured medical inflation has been greater than that for all goods and services.[87] Over the entire 1927–96 time period, the MCPI has risen at an average annual growth rate (AAGR) of 4.59%, almost half again as large as the 3.24% for the overall CPI.

Beginning with its January 1998 major revisions, the BLS has regularly published an aggregate medical care Consumer Price Index (MCPI), as well as price indexes for nine of the thirteen item strata in the MCPI. Separate MCPIs are also published for two expenditure groups (medical care commodities, and medical care services). The four major sub-indexes of the MCPI, along with their 1993–95 percentage base period weights within the aggregate MCPI, are prescription drugs (15.0%); nonprescription drugs and medical supplies (7.6%); professional medical services (also called physicians' services, although dentists are included, 49.4%); hospital and related services (23.0%); and health insurance (5.0%).[88] Each of these price indexes is based on consumers' out-of-pocket expenditures (OOPs) including employees' contribution to employment-based insurance, and thereby excludes all payments by governments and a portion of that from third party insurers. Any health insurance reimbursements for

[85] For historical discussions, see Langford (1957) and Getzen (1992).

[86] This table is taken from Berndt, Cockburn, Cocks et al. (1998).

[87] However, for several years within the 1927–46 time period, year-to-year changes in the CPI were greater than for the MCPI. See Getzen (1992) for a discussion.

[88] Taken from Ina Kay Ford and Daniel H. Ginsburg (1997), Exhibit 2. By December 1997, these relative importance weights were 14.6%, 7.2%, 50.0%, 23.8% and 4.5%, respectively.

medical services received by a member of the sampled household are netted out to obtain a net out-of-pocket expenditure.[89] Only that portion of third party insurance paid for out-of-pocket by consumers (and excluding employers' contributions to employee health insurance) is included within the scope of the MCPI.[90]

However, in constructing weights for the BLS' MCPI, the OOPs payments for health insurance are in turn distributed into payments by insurers for medical services, medical commodities, and health insurers' retained earnings.[91] Analogous to Equation (4) above, for each MCPI component, OOPs plus the consumer-paid health insurance premium allocation yields a total component weight, which until recently was typically applied to list prices paid by cash-paying customers. Note that over the last decade, actual transactions prices were frequently considerably less than list prices, particularly as discounts to managed care organizations became more common.[92] To the extent this occurred, over that time period it is likely that measured MCPIs overstated true price growth. However, particularly more recently, it is possible that discounts have become smaller and less frequent, in which case use of list prices could understate true price growth.

3.4.1. The item structure of the MCPI

The basic unit of the hierarchical CPI involves definitions of the item strata. Identifying and defining item strata presents considerable difficulties, particularly when markets are undergoing dramatic change during times within the approximately once-each-decade major CPI revisions.

From January 1987 through January 1997, for example, the CPI hierarchical structure distinguished inpatient hospital services as an item stratum separate from outpatient hospital services. Over this same period of time, cost containment efforts by managed care and other health providers resulted in many surgical procedures being transformed from inpatient to outpatient hospitalization. By shifting patients from inpatient to outpatient surgeries, hospitals and insurers were frequently able to cut down on total costs. Moreover, the average length of hospital stays declined over the 1987–97 time period, as skilled nursing facility days and home health visits were substituted for hospital days.

One consequence of this change in place of service was that the case mix severity in both inpatient and outpatient settings increased, resulting in greater costs for the average case in both settings, even as total inpatient plus outpatient costs decreased. The

[89] Cardenas (1996a, p. 36).

[90] We defer additional discussion of OOPs issues to later in this paper.

[91] See Fixler (1996), Daugherty (1964), Ford and Sturm (1988) and Getzen (1992). In Ford (1995), for private insurance the allocation is 39.7% for hospital services, 28.4% for physician services, 5.7% dental services, eyeglasses and eye care services, 0.3%, services by other medical professionals (including home health care) 6.2%, prescription drugs and medical supplies 6.2% and nursing home care 0.6%. For Medicare Part B, there is only a four component breakdown: outpatient hospital services, 27.2%, physicians' services, 56.8%, services by other medical professionals, 9.2%, and supplies and durable medical equipment, 6.8%. The BLS's treatment of pure health insurance has been criticized by the US Senate Finance Committee (1996).

[92] On this see, for example, Dranove, Shanley and White (1991).

mean inpatient severity likely increased, since the less complex and critical surgeries were shifted to the outpatient venue, leaving only the more critical and complex surgeries as inpatient. The mean outpatient severity also likely increased over this time, for outpatient surgeries were now being done on a much larger set of more complex patient cases. Total costs of treatment, taking into account the substitution from inpatient to outpatient, were lowered as a result.

It is illuminating to consider price index measurement implications of this cost-containment approach employed by managed care. Because the BLS treated inpatient and outpatient hospitalizations as distinct item strata, and because the result of the inpatient to outpatient substitution resulted in greater severity/complexity for both inpatient and outpatient services, price indexes for each item strata grew substantially, and given fixed weights for these item strata, the aggregate hospitalization price index also grew, even as total costs were likely to have decreased. Moreover, the CPI, though not the PPI, priced hospital days. There average severity also grew, due to shorter stays. Although empirical evidence is not available, we conjecture that over the January 1987–January 1998 time period, the BLS' measured CPI inflation for hospitalization considerably overstated true hospital inflation, because it failed to account for substitution from inpatient to outpatient, and also failed to account for treatments involving greater severity case mix in each component. This overstatement is consistent with the increased spread of the MCPI over the CPI between 1986 and 1996 (Figure 1) at a time when increased price competition should have decreased the spread.

The BLS has recognized the problem, and in January 1997, one year before its major CPI revisions introduced in January 1998, it began treating the aggregate of hospital inpatient and outpatient services as a single item stratum. It has also shifted to measuring hospital services by the stay rather than by the day, and and it has classified inpatient and outpatient hospital services as substratum indexes, similar to the elementary line items discussed in Section 3.1 above.[93] Information is not available, however, on how linking is implemented when, for example, a shift occurs from inpatient to outpatient surgery. Simple redefinition will not fully address the problem of inpatient-outpatient substitution, unless a satisfactory linking procedure is developed and implemented as well.

A number of other important changes have recently been introduced into the CPI for hospital services, even before the 1997 and 1998 revisions. Until at least 1990, for example, in most cases procedures for the MCPI involved pricing specific input items at list prices, e.g., "chargemaster" fees for X-rays, laboratory tests, and physicians' office visits rather than at the average actual charge for treatment of, say, a child's forearm fracture to a managed care organization obtaining a hospital discount.[94] According to Cardenas (1996b), since 1993, when redrawing outlet and item samples, the BLS has attempted to obtain quotes from hospitals for specific payers, thereby seeking to obtain

[93] See Ford and Ginsburg (1997, 1998).

[94] For further discussion, see Armknecht and Ginsburg (1992), Cardenas (1996b), Daugherty (1964), Ford and Sturm (1988), and Ginsburg (1978).

transactions rather than list prices. Cardenas (1996a, p. 40) reports that "Employing the sample rotation construct as the vehicle for increasing the number of transaction prices in the CPI, however, has yielded slow progress to date." According to the 1996 GAO study cited above, only about 15% of the CPI hospital price quotes obtained by the BLS included discounts.[95]

Obtaining transaction rather than list prices is not an easy task, particularly since with price discrimination and alternative pricing methods currently in the medical marketplace, there are frequently many transaction prices. Consider, for example, hospital services. Some insurers pay for medical care on a *per diem* basis – one price per day to cover all services provided. Other insurers pay on a *DRG* basis – one price per admission, differentiated only by the severity of the admission. And still other insurers pay on a *capitated* basis – one price per patient per year, independent of the amount of services the patient actually receives. Since the market has not settled on one basis of price, appropriate price indexes must be able to handle payments using all of these methods. Obtaining the transaction prices for all three methods will be difficult, particularly since transaction prices are frequently considered highly proprietary and confidential by insurers. Moreover, since health plans have different bargaining power, the same provider may negotiate varying per diem rates with alternative health plans. Note that these problems are not unique to the MCPI, but are relevant for the MPPI as well. Cooperation and joint efforts by the MCPI and MPPI programs in securing price quotes could be very fruitful.

Other recent changes implemented by the MCPI for hospitals involve item descriptions. At one extreme, one can assume zero substitutability among medical care goods and services for treatment of a condition, and simply take quotes of discrete hospital goods and services. An alternative, discussed above, is to employ DRGs. Although the BLS apparently employed non-Medicare DRG prices in three states beginning in 1990, two of those states have since terminated their state-regulated DRG programs. As of September 1992, approximately 6% of the CPI hospital quotes consisted of DRG descriptions.[96] According to Cardenas (1996a, p. 40), use of DRGs is problematic, because "... a DRG treatment path can be wide-ranging, contingent upon the treating physicians' approach", e.g., coefficients of variation range from 0.30 to over 1.5, thereby indicating considerable variation in the treatment strategies used to treat a case as defined by a DRG.[97] Currently the BLS is instead considering use of a "package" treatment, consisting of "highly standardized and tightly defined components and risk factors" for conditions such as appendectomies, tonsillectomies and cataract surgery. Details on how such treatment packages would be defined and how representativeness would be ensured have not been released, nor have any data concerning the composition and nature of hospital quotes being obtained by the BLS MCPI since the major revisions of January 1998.

[95] United States Government Accounting Office (1996, p. 58).

[96] Cardenas (1996b, fn. 16, p. 42).

[97] See Frank and Lave (1985).

Our MCPI discussion to this point has focused on hospital services. We have not seen comparable literature dealing with MCPIs for physician services, although informal conversations with BLS personnel suggest to us that issues of item description, list vs. transactions prices, and lack of quality adjustment are similar for physician and hospital services.

Like the MPPI, the MCPI has a prescription drug component. Issues discussed earlier in the context of the MPPI concerning the linking of prices of newly entering generic drugs, just after branded drugs lose patent protection, to prices of their pioneer antecedents apply here as well.

The MCPI program implemented changes involving generic drugs earlier than the MPPI. Effective January 1995, procedures involving the MCPI prescription drug treatment of generics changed considerably.[98] For branded drugs in the CPI sample losing patent protection, six months after patent expiration the BLS now follows a procedure whereby branded and generic versions of the molecule are randomly selected, where the probability of selection is proportional to the sales of each version of the drug during the sixth month. If a generic substitute is selected, the entire price difference between the original drug and its generic substitute is treated as a price change. Obviously, if the branded version is selected, repricing will continue as before. Drugs entering the CPI sample after their patent has expired would of course not be affected by this new procedure, since during the CPI sample rotation process generic versions would also have had a chance of being selected. Note that use of a six month window is somewhat problematic, for in many cases considerable additional diffusion of generics occurs beyond the six months immediately following patent expiration.[99]

Finally, regarding sample sizes for the MCPI and its components, as of December 1996 the total number of MCPI current price quotes was 7891. This was broken down as follows: prescription drugs, 687; internal and respiratory over the counter drugs, 354; nonprescription medical equipment and supplies, 315; physicians' services, 1304; dental services, 867; eye care, 298; services by other medical professionals, 251; hospital services, 3399; and nursing home services, 416.[100]

3.4.2. Weighting issues in the CPI and MCPI

As noted in Section 3.1 above, from January 1987 until January 1998, the item strata weights employed by the BLS in its CPI program were those based on the 1982–84 CEX; beginning January 1998, the new weights are those based on the 1993–95 CEX. Thus weights used just before the most recent CPI revision were about fifteen years out of date, and the newly introduced "current" weights were already almost four years out of date at the time of unveiling. Up-to-date weights are particularly important in

[98] See Armknecht, Moulton and Stewart (1994), and US Department of Labor, Bureau of Labor Statistics, "Improvements to CPI Procedures: Prescription Drugs", not dated.

[99] See, for example, Berndt, Cockburn and Griliches (1996, Table 2, p. 152).

[100] Ford and Ginsburg (1997, Exhibit 5).

the case of medical care, where technological change may result in substantial shifts across weighting categories. For example, Cutler et al. (1998a, 1998b) compare the old CPI medical methodology (pricing the hospital room rate and other hospital inputs with weights held fixed over a long time interval) with (i) a price index that priced the inputs but reweighted annually, and (ii) a price index that was based on the cost of treating heart attacks. The quantitative impact on the price index from annual reweighting was greater than the impact of moving from pricing medical inputs to pricing the cost of treating heart attacks.

In some goods and services markets characterized by relative tranquillity and stability, it is possible that use of old weights in price index construction would not be problematic. In the health care goods and services markets, however, the last fifteen years – indeed, the entire post World War II era – have been marked by dramatic changes in the number and quality of products offered and consumed, the identity of the payers (cash vs. third party private or government payer), and in how and by whom the services are provided (e.g., from inpatient to outpatient hospitalization, and from fee-for-service to managed care). The pace of both institutional and technological change has been particularly rapid in the health care sector. Moreover, the role of health care expenditures in the overall consumer budget has changed considerably, in part because the BLS' measured MCPI has increased much more rapidly than that for the all-item CPI (6.46% for the MCPI 1986–96, 3.65% for the all-item CPI-U over the same time period).[101] We now examine some of the implications of these changes for CPI and MCPI measurement.

In the CPI hierarchical system used from January 1987 until January 1998, seven major product categories were represented, and in the 1998 revisions an eighth was added. In column (1) of Table 2 we present 1982–84 CEX-based weights for the seven major product categories when they were originally introduced into the 1987 Revision of the Consumer Price Index. As is seen there, when the 1987 basket was introduced, the Medical Care major product category received a weight of 4.80%. Because the BLS' measured price of medical care rose more rapidly than that of the overall CPI, the implicit budget share consistent with fixed 1982–84 base period quantity weights (inflating all base period quantities by CPI measured price changes) increased over time; as is seen in column (2), by December 1995 the implicit relative importance of medical care increased to 7.36%. This raises a number of very important issues.

First, data from other government agencies, such as the Health Care Financing Administration (HCFA), indicate that national health expenditures as a proportion of GDP are much higher than $7+\%$; for example, Levit et al. (1998) report that in 1996, this proportion was 13.6%. Why is the CPI weight for medical care so low?

One important reason for this difference is that the medical care CPI (MCPI) weight reflects only a portion of total medical care outlays; others are discussed in Section 6 below. Specifically, the MCPI weight incorporates only direct out-of-pocket (OOP)

[101] For a discussion of some of these changes, see Berndt, Cockburn, Cocks, Epstein and Griliches (1998).

Table 2
Major product groups of items in the CPI 1982–84 weights, implicit relative importance and 1995 actual budget shares

Product group	(1) 1982–84 Weights in 1987 revision	(2) Implicit relative importance 1995.12	(3) 1995 CEX budget share	(4) Implicit relative importance 1997.12
Food and beverages	17.84%	17.33%	15.57%	16.31%
Housing	42.64%	41.35%	44.37%	39.56%
Apparel and upkeep	6.52%	5.52%	5.57%	4.94%
Transportation	18.70%	16.95%	18.47%	17.58%
Medical care	4.80%	7.36%	5.21%	5.61%
Entertainment	4.38%	4.37%	4.78%	n/a
Recreation	n/a	n/a	n/a	6.14%
Education and communication	n/a	n/a	n/a	5.53%
Other goods and services	5.13%	7.12%	5.74%	4.32%
Total	100%	100%	100%	100%

Sources: (1) US Department of Labor, Bureau of Labor Statistics, *The Consumer Price Index: 1987 Revision*, Report 736, January 1987, Figure 1, All Urban Consumers; (2) US Department of Labor, Relative Importance of Components in the Consumer Price Index 1995, Bulletin 2476, February 1996, All Urban Consumers; (3) United States Department of Labor, Bureau of Labor Statistics, *Consumer Expenditure Survey, 1995*, Table 1300; (4) US Department of Labor, Bureau of Labor Statistics, Relative Importance of Components in the Consumer Price indexes: US city average, December 1997, Table 1 (New Series), CPI-Urban.

cash outlays, plus direct household purchases of health insurance (including Medicare Part B), plus employee contributions to health insurance premiums purchased through work. Significantly, the MCPI excludes employer health insurance premium contributions, treating them as a business expense; MCPI also excludes Medicare Part A, 75% of Medicare Part B (the fraction paid from general government revenues), as well as Medicaid outlays. More generally, the MCPI excludes all government purchased medical services on behalf of its citizens/residents, and weights and prices only those components paid for out-of-pocket by consumers or from payroll deductions borne by employees.[102] Given this conceptual foundation of the MCPI, it is therefore not surprising that the MCPI weight is much smaller than the share of national health expenditures in GDP.[103]

[102] For further discussion, see Armknecht and Ginsburg (1992), particularly pp. 124–142.

[103] The appropriateness of this decomposition into employee out-of-pocket vs. employers' contributions depends in part on the incidence of the income tax, and the extent to which employees are willing to substitute employers' health insurance contributions for other forms of wage and non-wage compensation. While very important, these issues are beyond the scope of this review. For a recent discussion, see Gruber (1997) and Pauly (1997).

Another issue is whether the implicit relative importance of the medical care component in the CPI (column (2) of Table 2) accurately portrayed actual average consumer budget shares in 1995. If the 1982–84 fixed quantity weights provide a poor approximation to actual quantity weights in, say, 1995, then these implicit relative importance percentages could be unreliable and inaccurate as well, thereby compromising the accuracy of the measured CPI and MCPI. Thus it is of interest to compare actual budget shares with implicit relative importance percentages based on fixed weights.

Actual average budget share data based on the 1995 CEX, where budget shares are weighted averages over the various geographical areas comprising the BLS sample, are presented in column (3) of Table 2. As is seen there, the 1995 average budget share for medical care items is 5.21%, which is substantially smaller – 2.15 percentage points, about 29% – than the implicit relative importance of medical care items (7.36%) based on the BLS' fixed 1982–84 quantity weights; alternatively, by 1995 BLS use of the fixed weight index in its CPI resulted in the implicit relative importance of medical care being about 41% larger (7.36 vs. 5.21%) than was warranted.

The implicit relative importance of the eight major CPI components in the recently revised CPI, based on the 1993–95 CEX and updated to December 1997, are given in the final column of Table 2. Interestingly, the new relative importance of medical care is 5.61%. An implication of this is that because updated data from the 1993–95 CEX replaced outdated data from the 1982–84 CEX, with the January 1998 revisions the weight given medical care fell 1.75 percentage points from 7.36% to 5.61%, a relative overstatement of 31%.

This overstatement of the health care relative importance is greater in the 1998 major revision than it was for the major revision eleven years earlier in 1987. Then, as reported by Ford and Sturm (1988), the corresponding overstatement in December 1986 was 5.74% vs. 4.66%,[104] at 23% still substantial but considerably smaller than the 31% overstatement in 1998.

There are at least three reasons why the actual budget shares could diverge so materially from implicit relative importance based on fixed quantity weights. First, the relative quantity weights could have changed over time, reflecting non-zero price substitutability inconsistent with the Laspeyres fixed-weight assumption. For example, it is possible that efforts by managed health care organizations to contain medical expenditures have resulted in physicians and hospitals performing a smaller number of laboratory tests, scheduling fewer specialist physician visits, and shortening lengths of hospital stay. Hence it is possible that as a result of growth in managed care and other cost containment methods, the relative quantities of medical care items for which consumers made out-of-pocket expenditures has fallen since 1982–84.

Second, suppose that demand for health care had a zero price elasticity of demand. In such a case, the divergence would simply reflect overstated medical care price inflation, perhaps from failure to measure transactions prices accurately.

[104] Ford and Sturm (1988, Table 1, p. 19).

Third, if however the demand price elasticity for medical care were greater than unity (say, particularly for those components undergoing dramatic but not fully measured quality change), then the implicit relative importance of these items would be greater than the actual budget share, *ceteris paribus*.

Which of these three reasons, or what weighted combination, contributed to the divergence between the actual 1995 budget shares and implicit relative importance requires additional empirical research. Econometric studies of demand for health care such as those based on the RAND Health Insurance Experiment report modest but price inelastic demand; it is worth noting that the experimental design of that study in effect controlled for quality variations.[105] Additional research that focused on price measures incorporating quality change, and then evaluated the responsiveness of demand to quality changes, would be useful.[106]

These discrepancies between actual budget shares and implicit relative importance values, resulting from the use of outdated CEX surveys, suggest that more frequent weighting could considerably strengthen the reliability of the MCPI. The frequency of such revisions does not necessarily need to be uniform across the entire CPI, but could involve more frequent updatings in some major product groups such as medical care than in others, e.g., housing. For the rapidly changing medical care sector, decennial updates of weights with old weights having been fifteen years out of date before the new revision occurs, results in price indexes whose accuracy and reliability can legitimately be called into question.[107]

4. Related research on medical care price indexes

> "... the average consumer of medical care is not as interested in the price of a visit or a hospital day as he is in the total cost of an episode of illness."
>
> US Department of Health, Education and Welfare (1967, p. 13)

For quite some time now, health economists and government statisticians have made recommendations concerning directions toward which the pricing of medical care services should move, particularly concerning the definition of the item or product that is to be priced. For example, already in 1962 Anne Scitovsky proposed

> "... an index which would show changes, not in the costs of such items of medical care such as drugs, physicians' visits, and hospital rooms, but in the average costs of the complete treatment of individual illnesses such as, for example, pneumonia, appendicitis, or measles."[108]

[105] See Newhouse et al. (1993).

[106] For discussion and references, see Ellison, Cockburn, Griliches and Hausman (1997).

[107] Suggestions for implementing alternative weighting schemes with time – varying weights have been proposed and evaluated by Shapiro and Wilcox (1997).

[108] See Scitovsky (1964), and related discussions in Scitovsky (1967).

In Scitovsky (1967), this approach was implemented on an illustrative basis for five medical conditions. Notably, in the 1950s and 1960s the BLS price indexes appeared to have *understated* medical price inflation, in large part because physicians "customary" pricing in an environment of extensive price discrimination began to change as the proportion of patients covered by insurance increased.[109] Hence, the BLS' alleged upwards bias in measuring medical price inflation has not always been the indictment.

Shortcomings in the BLS' MCPI approach, and preference for the treatment episode-outcomes adjusted approach to price measurement, have appeared steadily since 1967; see, for example, the "Measuring Changes in the Price of Medical Care" chapter in various editions of a well-known health economics textbook by Paul Feldstein (1979, 1983, 1988), as well as the Baxter Foundation Prize Address by Newhouse (1989).

More recently, price indexes for several specific medical treatments, taking outcomes changes into account, have been constructed, thereby demonstrating again the feasibility and importance of the Scitovsky approach. Using one data set of hospital claims from a major teaching hospital and another very large data set consisting of Medicare claims, Cutler, McClellan, Newhouse and Remler (1998a, 1998b), have contrasted input price indexes for the cost of heart attack treatment that rise by 6.7% over 1983–94, with an outcomes adjusted index that takes into account changing treatment regimens and a conservative valuation for the extension of life expectancy attributable to new heart attack treatments; the latter price index increases by only 2.3% per year (in real terms, an annual decrease of 1.1%), implying a net upward bias of 4.4% per year for an MCPI-like index.

Similarly, Shapiro and Wilcox (1996) have constructed a price index for cataract surgery, 1969–93, and find that a CPI-like fixed weight input-based price index increases by a factor of about nine; a preferred alternative price index incorporating realized reduced levels of hospital services (input changes), but ignoring any improvements in the quality of medical outcomes, increases by only a factor of three, implying an annual differential of 4.6%.

A number of other studies, based on retrospective medical claims data, provide additional evidence that implementation of disease or condition-specific measurement procedures that uses treatment episodes of care as a measure of output, is in fact feasible; see, for example Berndt, Busch and Frank (1998) for treatment of depression, Cockburn and Anis (1998) for rheumatoid arthritis, and Shapiro, Shapiro and Wilcox (1998) for cataract surgery.

5. A new medical care expenditure price index based on episode treatment costs

One could envision an ideal medical care price index as providing accurate and reliable measures for use in at least five very important functions: (i) the measurement of quality of life; (ii) the deflation of nominal industry output for the calculation of real output

[109] For further discussion, see Martin Feldstein (1969, 1970).

and productivity growth; (iii) the indexing of health care benefits as a component of employee compensation; (iv) the indexing of payments by health plans to providers of medical care; and (v) the indexing of payments in government transfer programs. Undoubtedly, additional purposes can be envisaged. Unfortunately, these various functions and purposes are very different, and there is no way a single index like the medical CPI (or PPI) can provide an accurate and reliable basis for such diverse needs. The search for a single price index that meets all these purposes is a futile one. But these diverse needs are real and important. We recommend that rather than trying to change dramatically the conceptual foundations and measurement procedures of the MCPI and MPPI in an attempt to accommodate conflicting needs, that government statistical agencies consider constructing and publishing, on an experimental basis, a new price index that we tentatively call a *medical care expenditure price index.*

As we have discussed in considerable detail, the CPI and PPI medical price indexes are very different, they correspond to distinct index number concepts, and thus the appropriate uses to which they are applied must differ as well. The CPI is, in concept, a fixed weight approximation to a cost-of-living (COL) index, where the COL index is defined as follows: What is the minimum change in expenditure necessary to purchase the set of market goods and services yielding the same standard of living as the set of market-purchased goods and services consumed in the base period? The manner in which the BLS has implemented this COL definition in the case of the medical care CPI is to define the scope of the index to apply only to out-of-pocket expenditures. The reasoning is that employer-provided medical insurance is a non-wage part of compensation; BLS does not believe it to be appropriate to add consumption out of non-wage compensation into the consumer expenditures that are defined, implicitly, to be relevant to the wage part of compensation.[110]

Nevertheless, even if the CPI is continued to be defined to include only out-of-pocket expenditures, there are many important purposes for which one needs a price index covering all medical expenditures, no matter who (consumer, employer-provided health insurance, or government) is the nominal payer. This, for example, would be the concept of price change that one would want for most policy analytic purposes, such as containing medical care cost inflation, or examining the impact of new treatment technologies.

The PPI organizes and presents information by medical care *industry*, that is, hospitals, physicians' offices, nursing homes, pharmaceuticals, and so forth. The underlying PPI concept is an industry output price index. This index is useful for a number of purposes, e.g., comparing hospital price movements with the cost of hospital inputs (though one of the great weaknesses of the US statistical system is its inadequacy of information on industry input quantities and input prices). Moreover, the PPI is a price index for *domestic* industries. It provides, for example, information about price movements for domestically produced pharmaceuticals at the manufacturer's level. But the

[110] For discussion of the incidence of these employer-subsidized health benefits, see Pauly (1997) and Gruber (1997).

PPI is not a price index for all pharmaceuticals consumed in the US. It excludes, for example, imported pharmaceuticals and also, because of a definitional oddity in the US national accounts, pharmaceutical production in Puerto Rico. Additionally, the PPI includes pharmaceuticals and medical devices that are produced in the US and sold abroad.

Thus, just as the CPI does not provide a comprehensive price index for health care to US purchasers, neither does the PPI provide this information. Even though the CPI and PPI measures are useful on their own terms (and we are not asserting that these measures are not useful or appropriate ones), there is a great lacuna in medical care price information. The missing part, regrettably, is probably the part that is most vital for medical care policy analysis, namely, the US needs a comprehensive medical care price index for *expenditures* on medical care. Such a *medical care expenditure price index* would apply to all purchases of medical care, and it would take into account, as the present CPI and PPI do not, substitution by buyers or financiers of medical care across providers or industries that produce medical care. In principle separate medical care expenditure price indexes could be constructed for public and private sector expenditures, and for the elderly. The medical care expenditure price index would cover all consumption of medical care goods and services, be the providers/producers domestic or from abroad. And it would, we believe, be profitably structured around determining the costs of treating an episode of a representative set of illnesses or conditions.

As has been emphasized by, among others, Triplett (1998a, 1998b), complementary research efforts on health care outcomes by medical researchers involving cost-effectiveness analyses, as well as the public availability of large retrospective health claims data bases, now allow government statisticians and health economists to build on others' research that defines and identifies episodes of treatment. This research is particularly important were governments to initiate medical care expenditure price index programs. Note that in principle, outcomes research can help somewhat in overcoming the moral hazard problem underlying the failure of revealed preferences as measures of willingness to pay in medical care markets. Together with retrospective claims data, the outcomes studies provide a framework for identifying medical care outputs that incorporate quality change. What Anne Scitovsky proposed in 1962 and illustrated with a small sample of conditions in 1967, and what US Health, Education and Welfare Secretary John Gardner requested more generally in 1967, is clearly possible on a much larger scale today.

Although in market-based economies the usual source of information for output measurement is based on actual market transactions, use of medical outcomes data to define measures of output implies an adjustment in thinking – to look outside of market transactions to consider what medical resources actually do for health.[111] A medical care expenditure price index program should, to as great an extent as is feasible, combine actual transactions data underlying treatment costs of episode of an illness, with outcomes data from cost-effectiveness and related medical studies.

[111] For further discussion, see Triplett (1998a, 1998b).

It is likely that treatment episode price index measurement will need to be done at a very disaggregated level of detail, for a finite number of representative illnesses or conditions. The extent of medical care progress, as well as the underlying increases in medical scientific knowledge, have varied considerably across illnesses and disorders, with spectacular gains in treating conditions such as cataracts, retinal detachment, schizophrenia and cystic fibrosis, but with apparently less progress for other conditions such as rheumatoid arthritis, Alzheimer's and the common cold. While the Hicksian aggregation assumption of common proportional price changes over time across a variety of products may be a useful approximation within a number of other industries, for medical care it is not plausible. As suggested already in 1969 by Martin Feldstein, for government statisticians and health economists to obtain useful measures of medical care output, it would appear to be most useful to obtain a sample of "a representative mix of illnesses".[112] Research that helps identify an appropriate mix of illnesses and their treatments, ones for which outcomes measures and/or published treatment guidelines are available, and ones for which sample sizes in retrospective claims data bases are sufficiently large, would seem to be particularly helpful.

6. Medical care price indexes in the national income and product accounts

Reliable and accurate measurement of medical care price indexes is inherently difficult, as we have seen. We now consider ways in which medical care transactions enter national economic accounts, including inter-industry flows and national health accounts, as well as aggregate economy implications of possible mismeasurement of prices in the medical care sector. We begin with some national accounting definitions and conventions.

6.1. Medical expenditures in national accounts

National income accountants have long defined gross domestic product (GDP) as aggregate final demand. GDP is composed of four components: personal consumption expenditures (PCE), gross private domestic investment, including changes in inventories (GPDI), net exports of goods and services (NEX), and government consumption expenditures and gross investment (G). At the first stage of compilation of national accounts, all of these components are expressed without inflation adjustment, in what is usually termed "nominal" or "current value" GDP.

At the second stage, GDP and its components are adjusted for inflation, using price indexes. After inflation adjustment, the components are referred to as "real", as for example, "real investment". This language is intended to convey the notion that after inflation adjustment, the change in real GDP (or its components) corresponds with a

[112] M. Feldstein (1969, p. 363).

change in quantities of output or of expenditures. To obtain measures of real PCE, for example, national accountants typically deflate detailed components of household expenditures (including medical expenditures) by price indexes, typically drawn from a country's consumer, or retail price indexes. We discuss some examples of this below.

When an economic transaction occurs, there is a buyer and a seller, and expenditures equal receipts. Thus an alternative way of measuring nominal GDP is to focus on the production or selling side of transactions, rather than on the purchasing or expenditure side. National accountants also calculate GDP by aggregating sales by industry (in some countries, including the US, they allocate GDP to industries, but the methodology is similar).

To avoid double-counting, however, care must be taken to exclude from each industry's sales all intermediate purchases. For example, since steel is used in the production of autos, counting the output of the steel mill and of the automobile manufacturing plant would count twice the intermediate input into the automobile industry. The medical care sector is no different from others. It purchases many intermediate inputs (e.g., heat, light, marketing services, diagnostic equipment). For a particular industry, nominal gross product originating by industry (GPO) is calculated as sales (plus net changes in inventories) less purchases of intermediate goods; this GPO calculation is often called value added by industry. When nominal GPO by industry is aggregated across industries (including government), in theory one should obtain a number identical to nominal GDP; in practice, there is typically a modest difference between measured GPO and GDP, and this difference is called "statistical discrepancy".

Matters become more complex once one contemplates conversion of nominal GDP to real, inflation-adjusted GDP, by industry. Here procedures for treating government as a set of industries differ greatly from those used for market-producing industries. For the latter, a procedure called double deflation is commonly applied to the GPO numbers. With double deflation, industry final sales are divided (deflated) by a price index (say, an industry-specific producer price index), and then that industry's intermediate good purchases are also deflated by a price index (say, some other industry producer price indexes). Real GDP by industry is then obtained by deducting deflated intermediate purchases from deflated final sales.[113]

With the double deflation method, creating a real value added measure for the health care sector requires reliable price indexes for health care output, and also reliable price

[113] This step involves some index number complexities. In 1996 the US Bureau of Economic Analysis switched to an aggregation procedure for real GDP known as the Fisher index. Other countries currently retain the Laspeyres index number system for calculating real GDP, in a form similar to the one formerly used in the US. For the Laspeyres system, the language in the text here (which implies addition and subtraction to obtain value added) is descriptive. In the Fisher index number system, aggregations, including value added, cannot be formed by simple additions and subtractions, but must be carried out in more complicated ways. Exploring these index number issues for calculating real GDP takes us too far afield for purposes of this survey. See Yuskavage (1996) for discussion and detailed references.

indexes for health care inputs, such as pharmaceuticals and high-tech medical equipment purchased by the hospital industry. As we have noted above, serious problems surround the construction of price indexes for both medical outputs and medical inputs.

In contrast to private sector expenditures, for government expenditures, including government provided health care, there are normally no sales and prices (when governments do sell items, such as a government parking garage which is paid for from its receipts, it is treated like any other "industry"). In the absence of government sales and prices, national accountants normally value government "output" by government purchases. The implications of this are important for countries in which the health care sector is operated primarily by governments. Before discussing these implications, we digress and comment on the US context.

6.2. The US context

In the US, medical care paid for directly by households (out of pocket expenditures) and care that is paid for by insurance companies from premiums paid by employers, appear in PCE (other health care expenditures are in GPDI and G). PCE accounts for about 65–70% of GDP. To obtain real measures of PCE, the Bureau of Economic Analysis deflates PCE nominal values component by component. Although overall, price indexes for most PCE components are based on the BLS CPI (about 70% of the weights in the PCE employ CPI measures), since 1993 medical care in the PCE has also been deflated by the new PPI medical care price indexes. For earlier years, the CPI medical care price indexes are still used, since other historical price indexes are not available.

In recent years, the BEA's implicit price deflator (IPD, the implicit aggregate deflator obtained by dividing aggregate nominal PCE by aggregate real PCE) for aggregate PCE has grown less rapidly than the BLS' flagship consumer price index, the all-items CPI. As noted by Fixler–Jaditz, for example, from 1992 to 1996 the difference was about 0.35% per year. Medical care accounts for part of this difference.

Fixler and Jaditz (1997) attempt to reconcile these two alternative measures of consumption price growth, and focus in part on the role of medical prices. The PCE, as we noted above, now employs a chained Fisher index procedure, rather than a fixed weight Laspeyres. When the PCE is recomputed as a Laspeyres index with fixed 1992 weights, about one third of the difference between the CPI and the IPD (0.14 of 0.35%) is removed. Thus use of chained rather than fixed weights is empirically significant.

Another source of difference is scope. Recall, for example, that the medical CPI is based on out-of-pocket expenditures, whereas the PCE includes expenditures from third party payers. Thus, physicians account for about 3.9% of total spending in the 1995 PCE, but only 1.9% in the CPI; the sum over all medical items accounted for 6.93% of total 1992 spending in the CPI, but 18.98% in the PCE. Since 1993, the two indexes differ not only in weights assigned to medical care (scope), but also in the way they measure medical care prices. The MCPI appears in the CPI (until early 1997, measured on the old basis), but the new medical care PPIs now enter the PCE.

Fixler–Jaditz find that for medical care items, both price and weight effects contribute substantially to the difference between the IPD and CPI. CPI measures of medical prices have grown more rapidly than the PPI indexes used in the PCE (lowering the IPD relative to the CPI). The weight of medical care is larger in the PCE, influencing the difference in the same direction, since the PCE's larger weight for medical care increases the contribution of the lower PPI medical care price indexes on the IPD.

An alternative way to account for the contribution of medical care to GDP is through GPO. The various detailed industries comprising the health services sector in the US (two-digit industry code 80) have been listed in Table 1. As is noted in Yuskavage (1996, Table 8), for the double deflation of the health services industries, beginning in 1993 the BEA has deflated hospital sales and intermediate purchases by hospital-related PPIs, whereas prior to 1993 they used the MCPI for hospital room, and an index of input prices constructed by HCFA, which in turn were based on reweighted BLS price indexes, as well as other indexes constructed by Data Resources, Inc.[114] For other non-hospital health services, BEA has employed various CPIs and HCFA indexes, although price indexes for nursing homes and certain other health care industries are now available as PPI indexes and have been incorporated into the US national accounts.

It is worth noting, incidentally, that although the importance of the medical care industry to the aggregate US economy is often approximated by analysts who compute medical care expenditures as a percent of GDP, yielding numbers in recent years from 12–14%, such a calculation can be misleading. As we noted above, there are numerous intermediate inputs (heat, light, pharmaceuticals, marketing and accounting services, diagnostic equipment) that are double counted when one merely compares industry sales to aggregate GDP. On a value added basis in current dollars, in 1996 the health services sector was 5.9% of GDP.[115]

Industry accounts can be used to calculate the productivity of the US health care sector. Productivity of any industry is typically calculated as the ratio of the growth of the industry's real output to the growth of its inputs (also deflated to put them into real terms). This is usually called "multifactor productivity" (other productivity concepts also exist, but need not be discussed here).

Because the BEA measure of health services real output is obtained by double-deflation methods relying on medical CPIs and (since 1993) medical PPIs, to the extent that health care price inflation is over (under)-estimated by these price indexes, the real GDP output of the health services sector is under (over)-stated as well. Because real output is the numerator of the productivity ratio, measured productivity growth of the health services sector is affected in the same direction – that is, if true medical care price inflation is lower than measured medical care inflation, then measured medical care productivity is also lower than true medical care productivity.

However, overstatement of input price growth (such as for pharmaceuticals and high-tech medical equipment) operates in the other direction. Inputs are in the denominator of the productivity ratio, so overstatement of their price growth results in understatement

[114] See Health Care Financing Administration (1991).

[115] Lum and Yuskavage (1997, Table 7, p. 28, line 69).

of the growth in the industry's real purchased inputs, and consequently overstatement of industry productivity. Trajtenberg (1990) provides evidence that high-tech medical equipment prices such as those for CT scanners have fallen very rapidly; government price indexes for such equipment do not exist, and accordingly Trajtenberg's research suggests substantial overstatement of at least a portion of the health care sector's capital input prices. In addition, evidence that pharmaceutical price growth was overstated by BLS price indexes was discussed above. Hence it is unclear *a priori* whether multifactor productivity growth for health services is under- or over-stated by possible measurement error in medical-related inputs and outputs.

BLS researchers William Gullickson and Michael Harper (1998) have estimated that in the US health services sector, multifactor productivity growth from 1963–77, and from 1977–93 averaged about -1.25% per year, that is, they estimate that medical care productivity growth has been negative.[116] Economists typically deem negative productivity rates over such an extended period of time as being implausible (although similar negative numbers have been reported by Murray (1992, 1997) for Sweden, using a much different approach, as discussed below). When coupled to the probability that medical care prices are upward biased during much of this period, these numbers might suggest that errors in medical output price measurement might dominate errors in medical input price measurement.

This leads Gullickson–Harper to engage in a hypothetical analysis, using a complicated input-output framework that accounts appropriately for inter-industry flows. Specifically, they ask, suppose that in fact there was zero productivity growth in health services (rather than -1.25% per year) and that price mismeasurement was at fault, what would have been the impact of this mismeasurement on total private business sector multifactor productivity growth? Gullickson–Harper find that if health services had zero rather than -1.25% per year multifactor productivity growth, the corresponding productivity growth of the private business sector in aggregate would have been 0.09% greater per year, 1977–93. Since BLS measured productivity growth of the entire private business sector averaged about 0.25% per year over that time period, zero health services productivity growth would have raised that to about 0.34%.[117]

Gullickson–Harper then repeat the analysis, but instead allow for 1% annual productivity growth in health services; the result is an increase in aggregate private business sector productivity growth of 0.16% per year, from about 0.25% to 0.41%. These impacts of possible mismeasurement in medical-related CPIs and PPIs on economy-wide measures of economic performance are substantial, particularly when cumulated over time.[118]

[116] Gullickson and Harper (1998, Table 4, p. 30).

[117] Private business sector multifactor productivity growth is estimated by Gullickson-Harper as 0.2% per year 1979–1990, and 0.4% between 1990–1994.

[118] Over a fifteen year time span, the cumulative difference is about 67%–6.36% in productivity growth vs. 3.82%.

In summary, measurement errors in medical-related CPIs and PPIs are likely to have had a significant impact on aggregate measures of economic performance in the US, in large part because medical care expenditures are relatively large, although as noted above, on a value added basis, health services is but 5.9% of GDP in 1996.

6.3. National accounts issues outside the US

Outside the US the smaller size of the market health sector diminishes the role of price indexes in economic accounting for health care. Nevertheless, even countries where the predominant form of health care delivery is the public health care system have some form of private health care expenditures. Direct consumer outlays for medical services such as physician visits and non-prescription pharmaceuticals are not trivial, and in some countries with predominantly public health care systems, the private health care portion is growing. For a complete accounting for health care, price indexes for these private purchases must be constructed.

In the UK, for example, the "Chemists' Goods" portion of the Retail Price Index contains non-prescription pain medicines and so forth, and National Health System charges, private health insurance, and certain other health-related items are also included in the index. Moreover, the treatment of UK National Health Service hospitals has recently been changed, so they are now treated in the UK national accounts as government corporations; accordingly, government health expenditures are treated as being purchased from these corporations, which implies that a price index for hospital output is as relevant in the UK as it is in the US. As Berman (1998) has noted, other high per capita income countries, such as Australia, have significant health insurance sectors, and many low and middle-income countries such as those in South America, Southeast Asia and Eastern Europe have emerging private health insurance and private medical care provision. In the future, the need for accurate medical care price indexes for deflation in national accounts is likely to become more urgent, so the research on US medical care price indexes may become more relevant to the needs of other countries' national accounts.

As we have noted, for the US, national accounting for medical care makes extensive use of price indexes because the US medical care system is predominantly one of market provision of health care. Among OECD countries, the US is of course an outlier when one considers the proportionate roles of private and government provision of medical services. In most OECD member countries, health care is provided largely by the government sector. Price indexes for health care have little application for estimating the real value of health care output and expenditures when health care is provided by the government, at no cost or at very low cost. Moreover, use of government budgeted prices and accounts in computing price indexes can introduce serious problems, since transactions are typically not arms length. In the US, for example, this has led to instability in the MPPI estimates, particularly in components where budgeted systems dominate, e.g., public mental hospitals.

In those countries in which governments provide medical coverage, the impact of changes in medical care service production on real GDP depends in large part on the methods employed for deflating government expenditures. In most countries' national accounts, government expenditures are deflated by price indexes for what the government purchases, including wage rates. This carries over to deflation of government health care systems.

If government health care expenditures are deflated by government wage rates and other input prices, this essentially assumes away any productivity growth, because the numerator and the denominator of the productivity ratio are equal. Notice that if true, but unmeasured, government multifactor productivity growth is positive (negative), then real aggregate, economy-wide GDP growth is understated (overstated) when such a deflation procedure is employed. In some countries, an arbitrary allowance for government productivity (1% per year, for example) is inserted into the national accounts, on the grounds that the unknown true government productivity rate, if positive, would lift government output relative to inputs, and so the arbitrary productivity number moves the measure of GDP in a positive direction. Of course, if government productivity growth is in fact negative rather than positive, then a 1% "productivity correction" moves GDP in the wrong direction. There is research that suggests this may be a real possibility. For example, Murray (1992, 1997) reports negative public sector multifactor productivity growth for Sweden, including the provision of medical services, over the 1960–90 time period.

For government provided health care systems, an alternative approach to price index deflation has been employed in several studies. Instead of deflating expenditures by a price index to obtain a quantity (real output) measure, it has long been known that the real output measure could in principle be measured directly by computing a quantity index – weighting up quantities of government "output" activities, with weights derived from the costs of these activities. Applying this alternative approach to medical care requires specification of exactly what are the quantities of medical services, which is symmetric to the problem of specifying what is the price of these services (discussed earlier in this chapter).

One provocative set of studies is that by Murray (1992, 1997) for Sweden. Murray used counts of numbers of patients admitted, inpatient bed days and outpatient visits, and finds negative productivity growth for medical services. He noted that these measures were not totally satisfactory: "Although the measures employed capture some elements of quality like the shortening of hospital stays and the shift of work loads from more costly clinics to less costly, there are shortcomings in the measures of output."[119]

A related study is that by Barer and Evans (1983) for Canada, but unlike that by Murray, it employs price indexes, constructed from historical list fee schedules, actual billing patterns, and other government source data by the authors. Using employment and salary data for hospital personnel, along with data on medical and surgical supplies, drugs, and supplies and other expenses, Barer–Evans compute Paasche price indexes for hospital services. Aggregate expenditure data for hospitals were then divided by this

[119] Murray (1992, p. 534).

hospital price index, to obtain measures of real hospital services. A significant portion of growth in per diems over the 1960–80 time period was driven by increases in real resource inputs per day of care, which to Barer–Evans did not appear to reflect outcomes improvements or changes in case and activity mix, although useful evidence on outcomes and quality was generally unavailable. Barer–Evans conclude on a note similar to that of Murray, stating: "We suggest that only unequivocal evidence of real improvements in patient health outcomes can head off a conclusion of declining productivity in this sector."[120]

The pervasive problem in these various approaches is exactly analogous to the problem surrounding the "old" MCPI, which is that hospital days or visits to a physician's office are taken as the basic measurement unit of medical care quantities, even though what can be done in a hospital day or in a physician's office visit has changed. Advances in medical care that improve patient outcomes, such as shorter recovery times, lower death probabilities, less painful treatments with less severe side effects, are aspects of medical treatment that are not properly captured by such basic quantity measures. Heidenreich and McClellan (1998), for example, show that the average number of days in the hospital following a heart attack has fallen from fifteen to seven over a twenty-year period ending in the mid 1990s in the US. If hospital days (not hospital days per treatment) is the output measure, then the output of hospitals has decreased, when instead one could make a persuasive case that, for this disease, output in a real sense should have increased.

Triplett (1998a, 1998b) has suggested a variant on the direct quantity method for measuring real output of the health care sector. Rather than beginning with expenditures on hospitals and physicians' offices (which is the starting point for present national accounting systems, whether for market or government health care systems), Triplett would begin economic accounting from "Cost of Disease" accounts, which have been constructed for a number of countries [see Hodgson and Cohen (1998)]. Cost of disease accounts assemble the direct costs of treating diseases, and they are organized, not by funder and recipient of funds, but by aggregated categories of the ICD system discussed earlier.

For market health care systems, the cost of treating, say, circulatory disease or heart attacks could be deflated by a price index for heart attacks [such as that constructed by Cutler, McClellan, Newhouse and Remler (1998a, 1998b)], or for circulatory disease (which is now a component of the PPI in the US). For government health care systems, a similar approach could be carried out from the quantity index side. Real output of medical care could be formed from cost of disease accounts by counting quantities of medical procedures (the number of heart bypass operations, say, or of appendectomies, or of influenza shots), and weighting each procedure by its cost. Even if countries do not charge patients directly for health care, national health care systems often do keep track of the numbers of procedures and their costs (though sometimes not in the detail

[120] Barer and Evans (1983, p. 770).

that economists might prefer), and international concern for containment of health care costs is now forcing enhancement of these data accounting systems.

Of course, the method suggested by Triplett does not obviate finding effective measures of medical outcomes. The "quality adjustment" for improvements in medical care enters on the quantity side, rather than (as in the deflation case discussed earlier) on the price index side. Nevertheless, this approach offers advantages over the current alternatives for government provided health care systems, namely the assumption of zero (or +1%) productivity growth, which is an inherent (and generally untenable) assumption when government output is measured by government inputs.

In summary, to measure the output of the medical care sector – be it market or government-based, the challenge is clear – obtain a credible measure of the output of the medical care sector, the health of the population, or at least those who seek care. Unfortunately, as other chapters in this Handbook make clear, this challenge is a difficult one. Health itself is multi-dimensional and changing over time; it is also affected by many factors in addition to medical care. These difficulties ensure that price, output and productivity measurement in medical care will continue to be imperfect. As the share of medical care in GDP continues to grow, however, it will become even more important.

6.4. National health accounts

Before ending this chapter, we comment on the development of national health accounts and their relation to national economic accounts. In addition to national accounts (which measure GDP and its components), a number of countries now produce national health accounts, sometimes referred to as "Satellite Accounts". For example, in the US the Health Care Financing Administration (HCFA) produces the US National Health Accounts (NHA), and in France the Ministry of Health produces the Comptes de la Sante. The World Bank is encouraging developing countries to undertake construction of such accounts.

As part of the National Health Accounts, HCFA has constructed and published its estimates of inflation-adjusted personal health care expenditures, using BLS Laspeyres fixed weight price indexes, and a mix of BLS's medical related CPIs and HCFA constructed input prices for hospitals and nursing homes.[121] Unlike the BLS that weights by consumers' out-of-pocket expenditures, however, HCFA employs as weights the proportion of personal health care expenditures that each component represented in the 1982 base year, where each weight incorporates the sum of direct consumer, private third party payer and government expenditures.

While health accounts resemble national accounts, they are designed for somewhat different purposes. Health accounts provide more detail on health care expenditures than conveniently fits into systems of national accounts, and they are usually more comprehensive in what they count as health expenditures. The US NHA, for example, are the

[121] See, for example, Lazenby et al. (1992), Health Care Financing Administration (1991) and Federal Registers 63FR26290 (May 12, 1998) and 61FR29920 (June 2, 1997).

source for the usual statement that the US spends about 12–14% of GDP on health care, a number which is greater than GPO (value added, about 6%), because it is more comprehensive in what is included, and because the NHA do not deduct, as does GPO, the intermediate purchases of the health care sector from the remainder of the economy.

More importantly, health accounts organize information on health care expenditures around sources of health care financing and recipients of health care expenditures. They are usually designed so that totals from health accounts can be related to totals in national economic accounts, but this principle is also sometimes violated for various reasons.

In the US, a reconciliation project between HCFA and the Bureau of Economic Analysis (BEA) has recently been initiated to explain the differences between the NHA data published by HCFA and health care industry data gathered and published by the BEA. As noted by Sensenig and Wilcox (1998), hospital differences emerge because of varying treatments of government hospitals such as those from the Indian Health Service, possible double counting of nursing homes in the NHA, source data (American Hospital Association annual survey vs. Census of Service Industries for benchmarks), as well as definitions of revenues. The NHAs count revenues of non-health activities to health care organizations (sales of hospital gift shops, for example) as if they were health services, thereby potentially inflating hospital revenue and output. For physician services, most of the difference is attributable to NHA inclusion of osteopaths and medical laboratories that bill independently for their services, which are excluded by the BEA in its PCE computations. Efforts are currently underway to more fully reconcile the NHA and BEA accounts, and to correct inconsistencies.

References

Abraham, K.G., J.S. Greenlees and B.R. Moulton (1998), "Working to improve the consumer price index", Journal of Economic Perspectives 12(1):27–36.

American Medical Association (1990), Current Procedural Terminology (CPT), 4th edn. (American Medical Association, Chicago).

Archibald, R.B. (1977), "On the theory of industrial price measurement: output price indexes", Annals of Economic and Social Measurement 6(1):57–72.

Armknecht, P.A. (1996), "Improving the efficiency of the U.S. CPI in the future", unpublished manuscript (International Monetary Fund, Washington, DC).

Armknecht, P.A., and D.H. Ginsburg (1992), "Improvements in measuring price changes in consumer services: past, present and future", in: Z. Griliches, ed., asist., Output Measurement in the Service Sectors, vol. 56 (University of Chicago Press for the National Bureau of Economic Research, Chicago) 109–156 (see especially 124–132, 139–142).

Armknecht, P.A., B.R. Moulton and K.J. Stewart (1994), "Improvements to the food at home, shelter and prescription drug indexes in the U.S. consumer price index", US Department of Labor, Bureau of Labor Statistics, CPI Announcement-Version I, October 20.

Barer, M.L., and R.G. Evans (1983), "Prices, proxies and productivity: an historical analysis of hospital and medical care in Canada", in: W.E. Diewert and C. Montmarquette, eds., Price Level Measurement: Proceedings from a Conference Sponsored by Statistics Canada (Statistics Canada, Ottawa) 705–777.

Berman, P. (1998), “What can the U.S. learn from national health accounting elsewhere?”, Paper prepared for Health Care Financing Administration Conference on Future Directions of the National Health Accounts, Baltimore, MD, March 12–13.

Berndt, E.R. (1991), The Practice of Econometrics: Classic and Contemporary (Addison Wesley, Reading, MA).

Berndt, E.R., S.M. Busch and R.G. Frank (1998), “Price indexes for acute phase treatment of major depression”, Paper given at the NBER-CRIW Conference on Medical Care Output and Productivity, Bethesda MD, June 12–13.

Berndt, E.R., I. Cockburn and Z. Griliches (1996), “Pharmaceutical innovations and market dynamics: tracking effects on price indexes for antidepressant drugs”, Brookings Papers on Economic Activity: Microeconomics, 133–188.

Berndt, E.R., I.M. Cockburn, D.L. Cocks, A. Epstein and Z. Griliches (1998), “Is price inflation different for the elderly? An empirical analysis of prescription drugs”, in: A. Garber, ed., Frontiers of Health Policy 1:33–75.

Berndt, E.R., D.M. Cutler, R.G. Frank, Z. Griliches, J.P. Newhouse and J.E. Triplett (1998), “Price indexes for medical care goods and services: an overview of measurement issues”, Paper given at the NBER-CRIW Conferences on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Berndt, E.R., Z. Griliches and J.G. Rosett (1993), “Auditing the producer price index: micro evidence from prescription pharmaceutical preparations, Journal of Business and Economic Statistics 11(3):251–264.

Berndt, E.R., and P.E. Greenberg (1995), “An updated and extended study of the price growth of prescription pharmaceutical preparations”, in: R.B. Helms, ed., Competitive Strategies in the Pharmaceutical Industry (American Enterprise Institute, Washington, DC) 35–48.

Boskin, M.J., E.R. Dulberger, R.J. Gordon, Z. Griliches and D.W. Jorgenson (1998), “Consumer prices, the consumer price index, and the cost of living”, Journal of Economic Perspectives 12(1):3–26.

Bresnahan, T.F., and R.J. Gordon (1997), The Economics of New Goods (University of Chicago Press for the National Bureau of Economic Research, Chicago).

Cardenas, E.M. (1996a), “The CPI for hospital services: concepts and procedures”, Monthly Labor Review 119(7):34–42.

Cardenas, E.M. (1996b), “Revision of the CPI hospital services component”, Monthly Labor Review 119(12):4048.

Catron, B., and B. Murphy (1996), “Hospital price inflation: what does the new PPI tell us?”, Monthly Labor Review 119(7):24–31.

Cockburn, I., and A. Anis (1998), “Hedonic analysis of arthritis drugs”, Paper given at the NBER-CRIW Conference on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Cohen, J.W., A.C. Monheit, K.M. Beauregard, S.B. Cohen, D.C. Lefkowitz, D.E.B. Potter, J.P. Sommers, A.K. Taylor and R.H. Arnett III (1996), “The medical expenditure panel survey: a national health information resource”, Inquiry, publication of Blue Cross and Blue Shield Association and Finger Lakes Blue Cross and Blue Shield, vol. 33, Winter 1996/97, 373–389.

Council on Wage and Price Stability (1977), The Wholesale Price Index: Review and Evaluation (Washington, DC).

Cutler, D.M., M.B. McClellan, J.P. Newhouse and D. Remler (1998a), “Are medical prices declining? Evidence from heart attack treatments”, Quarterly Journal of Economics 113(4):991–1024.

Cutler, D.M., M.B. McClellan, J.P. Newhouse and D. Remler (1998b), “Pricing heart attack treatments”, Paper given at the NBER-CRIW Conference on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Daugherty, J.C. (1964), “Health insurance in the revised CPI”, Monthly Labor Review 87(11):1299–1300.

Diewert, W.E. (1983), “The theory of output price index and the measurement of real output change”, in: W.E. Diewert and C. Montmarquette, eds., Price Level Measurement: Proceedings from a Conference Sponsored by Statistics Canada (Statistics Canada, Ottawa) 1049–1113.

Dranove, D., M. Shanley and W.D. White (1991), “Does the Consumer Price Index Overstate Hospital Price Inflation?”, Medical Care 29(August):690–696.

Eldridge, L.P. (1998) , "The role of prices in measuring productivity for the business sector of the US economy", Draft manuscript, July 29 (US Bureau of Labor Statistics, Washington, DC).

Ellis, R.P., G.C. Pope, L.I. Iezzoni, J.Z. Ayanian, D.W. Bates, H. Burstin and A.S. Ash (1996), "Diagnosis-based risk adjustment for Medicare capitation payments", Health Care Financing Review 17(3):101–128.

Ellison, S.F., I. Cockburn, Z. Griliches and J. Hausman (1997), "Price competition among pharmaceutical products: an examination of four cephalosporins", Rand Journal of Economics 28(3):426–446.

Feldstein, M.S. (1969), "Improving medical care price statistics", in: 1969 Proceedings of the Business and Economics Statistics Section (American Statistical Association, Washington, DC) 361–365.

Feldstein, M.S. (1970), "The rising price of physicians' services", Review of Economics and Statistics 52(2):121–133.

Feldstein, P.J. (1979), Health Care Economics, 1st edn. (Wiley, New York).

Feldstein, P.J. (1983), Health Care Economics, 2nd edn. (Wiley, New York).

Feldstein, P.J. (1988), Health Care Economics, 3rd edn. (Wiley, New York).

Fisher, F.M., and Z. Griliches (1995), "Aggregate price indexes, new goods, and generics", Quarterly Journal of Economics 110(1):229–244.

Fisher, F.M., and K. Shell (1972), "The pure theory of the national output deflator", in: F.M. Fisher and K. Shell, eds., The Economic Theory of Price Indexes (Academic Press, New York) 49–113.

Fixler, D. (1996), "The treatment of the price of health insurance in the CPI", unpublished manuscript (US Department of Labor, Bureau of Labor Statistics, Washington, DC).

Fixler, D., and M. Ginsburg (1997), "Health care output and prices in the producer price index", slides from presentation to the NBER Summer Institute, Franco-American Seminar, Cambridge, MA, July 23.

Fixler, D., and M. Ginsburg (1998), "Health care output and prices in the producer price index", Paper given at the NBER/CRIW Conference on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Fixler, D., and T. Jaditz (1997), "An examination of the difference between the CPI and the PCE deflactor", draft manuscript, December (US Bureau of Labor Statistics, Division of Price and Index Number Research, Washington, DC).

Ford, I.K. (1995), "Health insurance allocations", Memorandum to Mary Lynn Schmidt, June 22 (US Department of Labor, Bureau of Labor Statistics, Washington, DC).

Ford, I.K., and D.H. Ginsburg (1997), "Medical care in the CPI", Paper given in presentation to the NBER Summer Institute, Franco-American Seminar, Cambridge, MA, July 23.

Ford, I.K., and D.H. Ginsburg (1998), "Medical care in the CPI", Paper given at the NBER/CRIW Conference on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Ford, I.K., and P. Sturm (1988), "CPI revision provides more accuracy in the medical care services component", Monthly Labor Review 111(4):17–26.

Frank, R.G., and J.R. Lave, "The psychiatric DRGs: are they different?", Medical Care 28(11):1145–1155.

Fuchs, V. (1974), Who Shall Live? Health, Economics and Social Choice (Basic Books, New York).

Fuchs, V. (1983), How We Live (Harvard University Press, Cambridge, MA).

Getzen, T.E. (1992), "Medical care price indexes: theory, construction and empirical analysis of the US Series 1927–1990", in: Advances in Health Economics and Health Services Research (JAI Press) 83–128.

Gilbert, M. (1961), "The problem of quality changes and index numbers", Monthly Labor Review 84(9):992–997.

Gilbert, M. (1962), "Quality change and index numbers: the reply", Monthly Labor Review 85(5):544–545.

Ginsburg, D.H. (1978), "Medical care services in the consumer price index", Monthly Labor Review 101(8):35–39.

Glied, S. (2000), "Managed care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 13.

Griliches, Z. (1962), "Quality change and index numbers: a critique", Monthly Labor Review 85(5):542–544.

Griliches, Z. (1988), Technology, Education, and Productivity (Basil Blackwell, New York).

Griliches, Z. (1992), "Introduction", in: Z. Griliches, ed., Output Measurement in the Service Sectors, Vol. 56 (University of Chicago Press for the National Bureau of Economic Research, Chicago) 1–22.

Griliches, Z. (1997), "The commission report on the consumer price index", Federal Reserve Bank of St. Louis Review 79(3):169–173.

Griliches, Z., and I. Cockburn (1994), "Generics and new goods in pharmaceutical price indexes", American Economic Review 84(5):1213–1232.

Grossman, M. (1972a), The Demand for Health: A Theoretical and Empirical Investigation (Columbia University Press for the National Bureau of Economic Research, New York).

Grossman, M. (1972b), "On the concept of health capital and the demand for health", Journal of Political Economy 80(2):223–255.

Gruber, J. (1994), "The incidence of mandated maternity benefits", American Economic Review 84(3):622–641.

Gruber, J. (1997), "Health insurance and the labor market", unpublished manuscript (Massachusetts Institute of Technology, Department of Economics, Cambridge, MA). Also in: J.P. Newhouse and A.J. Culyer, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 12.

Gullickson, W., and M.J. Harper (1998), "Possible measurement bias in aggregate productivity growth", draft manuscript (US Bureau of Labor Statistics, Office of Productivity and Technology, Washington, DC).

Health Care Financing Administration (1991), "National nursing home input price index", unpublished (Office of the Actuary, Baltimore, MD).

Heidenreich, P., and M. McClellan (1998), "Trends in heart attack treatment and outcomes, 1975–95: literature review and synthesis", unpublished manuscript (Stanford Medical School).

Hoechst, M.R. (1997), Managed Care Digest Series 1997 (Hoechst Marion Roussel, Kansas City, MO).

Hodgson, T.A., and A.J. Cohen (1998), "Medical care expenditures for major diseases", unpublished manuscript (National Center for Health Statistics, Hyattsville, MD).

Hoover, E.D. (1961), "The CPI and problems of quality change", Monthly Labor Review 84(11):1175–1185.

Jensen, G.A., and M.A. Morrisey (1990), "Group health insurance: a hedonic price approach", Review of Economics and Statistics 72(1):38–44.

Kanoza, D. (1996), "Supplemental sampling in the PPI pharmaceuticals index", Producer price indexes detailed price report, January, 8–10.

Kelly, G.G. (1997), "Improving the PPI samples for prescription pharmaceuticals", Monthly Labor Review 120(10):10–17.

Kendrick, J.W. (1991), "Appraising the US output and productivity estimates for government: where do we go from here?", Review of Income and Wealth 37(2):149–58.

Lane, W. (1996), "Changing the item structure of the consumer price index", Monthly Labor Review 119(12):18–25.

Langford, E.A. (1957), "Medical care in the CPI: 1935–1956", Monthly Labor Review 80(9):1053–1058.

Lawson, A.M. (1997), "Benchmark input–output accounts for the US Economy, 1992: make, use, and supplementary tables", Survey of Current Business 78(11):36–82.

Lazenby, H.C., K.R. Levit, D.R. Waldo, G.S. Adler, S.W. Letsch and C.A. Cowan (1992), "National health accounts: lessons from the U.S. experience", Health Care Financing Review 13(4):89–103.

Leaver, S.G., W.H. Johnson, R. Baskin, S. Scarlett and R. Morse (1997), "Commodities and services sample redesign for the 1998 consumer price index revision", unpublished memo (US Bureau of Labor Statistics, Washington, DC).

Levit, K.R., H.C. Lazenby, B.R. Braden and the National Health Accounts Team (1998), "National health spending trends in 1996", Health Affairs 17(1):35–51.

Lum, S.K.S., and R.E. Yuskavage (1997), "Gross product by industry, 1947–96", Survey of Current Business 77(11):220–34.

Meltzer, D. (1997), "Accounting for future costs in medical care cost-effectiveness analysis", Journal of Health Economics 17(4):33–64.

Moulton, B.R. (1996), "Bias in the consumer price index: what is the evidence?", Journal of Economic Perspectives 10(4):159–177.

Moulton, B.R., and K.E. Moses (1997), "Addressing the quality change issue in the consumer price index", in: Brookings Papers on Economic Activity 1997:1 (The Brookings Institution, Washington, DC) 305–349.

Moulton, B.R., and K.J. Stewart (1997), "An overview of experimental U.S. consumer price indexes", unpublished paper (US Department of Labor, Bureau of Labor Statistics, Washington, DC).

Murray, R. (1992), "Measuring public-sector output: the Swedish report", in: Z. Griliches, ed., Output Measurement in the Service Sectors, Vol. 56 (University of Chicago Press for the National Bureau of Economic Research, Chicago) 517–542.

Murray, R. (1997), Public Sector Productivity in Sweden, Vol. 3 (Budget Department, Swedish Ministry of Finance, Stockholm).

Newhouse, J.P., and The Insurance Experiment Group (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge, MA).

Newhouse, J.P. (1989), "Measuring medical prices and understanding their effects", Journal of Health Administration Education 7(1):19–26.

Nordhaus, W.D. (1998), "Quality change in price indexes", Journal of Economic Perspectives 12(1):59–68.

Pauly, M.V. (1997), Health Benefits at Work: An Economic and Political Analysis of Employment-Based Health Insurance (University of Michigan Press, Ann Arbor, MI).

Pauly, M.V. (1998), "Costs, effects, outcomes, and utility: concepts and usefulness of medical care price indexes", forthcoming, in: J.E. Triplett, ed., Measuring the Prices of Medical Treatments (The Brookings Institution, Washington, DC).

Persky, J. (1998), "Retrospectives: price indexes and general exchange values", Journal of Economic Perspectives 12(1):197–206.

Pollak, R.A. (1980), Group cost-of-living indexes", American Economic Review 70(2):273–278.

Pollak, R.A. (1998), "The consumer price index: a research agenda and three proposals", Journal of Economic Perspectives 12(1):69–78.

Prospective Payment Assessment Commission (1995), Report and Recommendations to the Congress (Washington, DC).

Reder, M.W. (1969), "Some problems in the measurement of productivity in the medical care industry", in: V. Fuchs, ed., Production and Productivity in the Service Industries (Columbia University Press, New York).

Rice, D.P., and L.A. Horowitz (1967), "Trends in medical care prices", Social Security Bulletin 30(7):13–28.

Ristow, W. (1996), "IMS presentation to the BLS", IMS America, Plymouth Meeting, PA, July, mimeo.

Scitovsky, A.A. (1964), "An index of the cost of medical care – a proposed new approach", in: The Economics of Health and Medical Care, Proceedings of the Conference on the Economics of Health and Medical Care, May 10–12, 1962 (The University of Michigan, Ann Arbor).

Scitovsky, A.A. (1967), "Changes in the costs of treatment of selected illnesses, 1951–65", American Economic Review 57(5):1182–1195.

Sensenig, A., and E. Wilcox (1998), "National health accounts and national income and product accounts: reconciliation", Paper given at the NBER-CRIW Conference on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Shapiro, M.D., and D.W. Wilcox (1997), "Alternative strategies for aggregating prices in the CPI", Federal Reserve Bank of St. Louis Review 79(3):113–125.

Shapiro, I., M.D. Shapiro and D.W. Wilcox (1998), "A price index for cataract surgery", Paper given at the NBER-CRIW Conference on Medical Care Output and Productivity, Bethesda, MD, June 12–13.

Shapiro, M.D., and D.W. Wilcox (1996), "Mismeasurement in the consumer price index: an evaluation", NBER Macroeconomics Annual 11:93–142.

Summers, L.H. (1989), "Some simple economics of mandated benefits", American Economic Review 79(2):177–183.

Trajtenberg, M. (1990), Economic Analysis of Product Innovation: The Case of CT Scanners (Harvard University Press, Cambridge, MA).

Triplett, J.E. (1983), "Escalation measures: what is the answer? What is the question? in: W.E. Diewert and C.M. Montmarquette, eds., Price Level Measurement: Proceedings from a Conference Sponsored by Statistics Canada (Statistics Canada, Ottawa) 457–487.

Triplett, J.E. (1990), "The theory of industrial and occupational classifications and related phenomena", in: Proceedings of the Bureau of the Census 1990 Annual Research Conference (US Department of Commerce, Washington, DC).

Triplett, J.E. (1998a), "What's different about health? Human repair and car repair in national accounts", unpublished draft manuscript (National Bureau of Economic Research, Cambridge, MA).

Triplett, J.E. (1998b), "Accounting for health care: integrating price index and cost-effectiveness research", forthcoming in: J.E. Triplett, ed., Measuring the Prices of Medical Treatments (The Brookings Institution, Washington, DC).

US Congress, Joint Economic Committee (1961), Government Price Statistics, Hearings before the Subcommittee on Economic Statistics of the Joint Economic Committee, Congress of the United States, Eighty-seventh Congress, First Session, Pursuant to Sec. 5(a) of Public Law 304 (79th Congress), Part 1, January 24 (US Government Printing Office, Washington, DC).

US Department of Health, Education and Welfare (1967), A Report to the President on Medical Care Prices (US Government Printing Office, Washington, DC).

US Department of Health and Human Services (1980), International Classification of Diseases (ICD-9-CM), 2nd edn (US Government Printing Office (PHS)-80-1260, Washington, DC).

US Department of Labor, Bureau of Labor Statistics (1992), Handbook of Methods (US Government Printing Office, Washington, DC), Bulletin 2414.

US Department of Labor, Bureau of Labor Statistics (1997a), "The experimental CPI using geometric means (CPI-U-XG)", Washington, DC, April 10.

US Department of Labor, Bureau of Labor Statistics (1997b), "Measurement issues in the consumer price index", BLS Response to Letter from Jim Saxton, Chairman of the Joint Economic Committee, to Katharine Abraham, Commissioner of the Bureau of Labor Statistics, June.

US Department of Labor, Bureau of Labor Statistics (1998), "Planned changes in the consumer price index formula", Washington, DC, April 16, 14 pp.

US Department of Labor, Bureau of Labor Statistics (not dated), "A description of the PPI physician services initiative", Washington, DC.

US Department of Labor, Bureau of Labor Statistics (not dated), "A description of the PPI hospital services initiative", Washington, DC.

US Department of Labor, Bureau of Labor Statistics (not dated), "Improvements to CPI procedures: prescription drugs", Washington, DC.

US Government Accounting Office (1996), "Consumer price index: cost-of-living concepts and the housing and medical care components", Report to the Ranking Minority Member, Committee on Banking and Financial Services, House of Representatives, GAOO/GGD-96-166, August, Washington, DC.

US Senate Finance Committee (1996), Final Report from the Advisory Commission To Study The Consumer Price Index, updated version, Washington, DC, December 4.

Weiner, J.P., A. Dobson, S.L. Maxwell, K. Coleman, B. Starfield and G. Anderson (1996), "Risk-adjusted medicare capitation rates using ambulatory and inpatient diagnoses", Health Care Financing Review 17(3): 77–100.

Weiss, S. (1955), "The development of index numbers in the BLS", Monthly Labor Review 78(1):20–25.

Yuskavage, R.E. (1996), "Improved estimates of gross product by industry, 1959–94", Survey of Current Business 78(8):133–155.

Chapter 4

ADVANCES IN COST-EFFECTIVENESS ANALYSIS OF HEALTH INTERVENTIONS*

ALAN M. GARBER

Veterans Affairs Palo Alto Health Care System and Stanford University

Contents

*This chapter builds upon earlier joint research with Charles Phelps. The (US) Department of Veterans Affairs, the Agency for Health Care Policy and Research, the National Institute on Aging, and the Henry J. Kaiser Family Foundation provided generous research support for my work in this area. I am grateful to A.J. Culyer, Adalsteinn Brown, Donald Kenkel, David Meltzer, Harri Sintonen, Paul Dolan, and Joseph Newhouse for helpful comments.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

Recent work has clarified the welfare implications of the application of cost-effectiveness analysis to the allocation of health care. Although cost-effectiveness analysis shares many similarities with cost-benefit analysis, it did not develop as an outgrowth of neoclassical welfare economics. Consequently, even though the welfare implications of public decisionmaking based on cost-benefit analysis have long been understood, until recently the conditions under which decisions made on the basis of cost-effectiveness criteria lead to potential Pareto improvement had received little attention.

This chapter describes the welfare economic foundations of cost-effectiveness analysis and how such foundations can be applied to resolve controversies in the application of the technique. It also discusses procedures for applying the technique, the circumstances under which decision rules based on cost-effectiveness analysis have desirable welfare economic properties, the appropriate perspective for the analysis, and issues in measuring outcomes. Even when standard welfare economic assumptions are not fully accurate descriptions of the markets and conditions in which health care is delivered, cost-effectiveness analysis can be a useful guide to allocation decisions.

Keywords

cost-effectiveness, welfare economic, cost-benefit, compensation principle, pharmacoeconomic, quality-adjusted life years, von Neumann–Morgenstern utility, sensitivity analysis

JEL classification: D61, H43, I11, I18

1. Introduction

This chapter discusses the welfare economic foundations of cost-effectiveness (CE) analysis. Although it is not a comprehensive review of the techniques of CE analysis, the chapter addresses application as well as theory because the welfare economic properties of decisions based on CE analysis necessarily depend upon the way that the method is applied. In fact, application has stimulated much of the interest in the theoretical foundations of CE analysis. As government officials, private insurers, health care providers, and others have begun to use CE analysis to inform decisions about the adoption and allocation of specific health interventions, they have revealed the need to improve and standardize its methods.

There is no doubt that CE analysis is potentially useful: by quantifying the tradeoffs between resources consumed and health outcomes achieved with the use of specific interventions, the technique can help physicians, health plans, insurers, government agencies, and individuals to prioritize services and to allocate health care resources. CE analysis aids such decisions by structuring comparisons among alternative interventions. Meaningful comparisons, in turn, require standardization. Without standardization, there can be no assurance that the results of a CE analysis of one set of interventions will be comparable to the results of a study of a different set of interventions. Thus the method must be valid, and it must be applied consistently. Perhaps the most important contribution of an examination of welfare economic foundations is that it can help ensure that any set of standards adopted for CE analysis will be logically consistent, valid, and credible.

Several efforts around the world have sought to move the field of CE analysis forward by strengthening the methodology and promoting standardization. Among these are various governmental guidelines (such as Australian pharmacoeconomic guidelines and those of Ontario), the European Community Concerted Action on the Harmonization of the Methodology for Economic Evaluation of Health Technology (HARMET), and the Panel on Cost-Effectiveness in Health and Medicine. The last group, sponsored by the US Department of Health and Human Services, issued a comprehensive report in 1996 detailing recommendations for the application of CE analysis [Gold et al. (1996)]. The report distinguished between recommendations that had a strong theoretical justification and those that had no firm theoretical grounding, but were made to ensure uniformity, usually based in part upon ease of implementation and other practical considerations.

The advantages of methodological standardization in CE analysis are greatest when the standards are selected with both rigor and transparency. To the extent that standards are chosen arbitrarily, they merely ensure that diverse studies will use consistent – but potentially invalid and misleading – methods. To develop recommendations that could be justified from first principles, the Panel on Cost-Effectiveness in Health and Medicine drew upon recent work on the welfare economic foundations of CE analysis. Since CE analysis evolved largely outside the framework of welfare economics, an exploration of the welfare economic foundations neither recapitulates nor parallels the history of the development of the approach. Yet by relating CE analysis to theoretical foundations it is possible to illuminate the consequences of alternative methodological practices. For example, there has been a longstanding controversy about future costs of health care: Should costs that result solely from living longer, but otherwise are not directly influenced by an intervention, be attributed to that intervention? Some investigators, such as Weinstein and Stason (1977), have recommended always including such "unrelated" future costs of care while others, such as Russell (1986), have urged the opposite. Presumably one of these practices is incorrect, and the persistence of two distinct practices renders the results of different studies non-comparable. Other methodological controversies are no easier to resolve, such as whether to incorporate time costs as dollar costs (hence part of the numerator of the CE ratio), or as a reduction in the health outcome like years of life (in the denominator). In cases such as these, which are discussed below, methodological standardization offers the prospect of replacing a set of inconsistent practices with a single correct method.

An exploration of the welfare theoretic foundations for CE analysis can provide a rationale for selecting specific standards while deepening our understanding of the implications of alternative methodological approaches. However, few attempts to explore the theoretical foundations of CE analysis have been published. Both proponents and critics of CE analysis have been skeptical of the value of some of the traditional standards of welfare economics, at least when applied to health care. To many economists, the forms of market failure common in health care supply much of the rationale for applying a tool like CE analysis or CB analysis. But others are skeptical of the premises and conclusions of welfare economics more generally, and see CE analysis as a method to make policy decisions when market outcomes are unacceptable.

Some proponents of CE analysis have adopted an "extra-welfarist" perspective, arguing that there are fundamental justifications for pursuing CE analysis without reference to welfare economics [see Hurley (2000)]. The assumptions and, some would argue, the values underlying this perspective can be more general than under the typical welfare economic perspective. Proponents of the extra-welfarist perspective claim that improvement of health is a primary goal of social policy, a goal whose value is self-evident and does not depend upon the maximization of individual utility functions. They do not necessarily accept the arguments of social welfare (e.g., the prominence of individual consumption of goods and services) that are typical in formulations proposed by economists, nor do they accept the typical assumptions made. For extra-welfarists, CE analysis offers a way for a social decision maker to learn how to obtain the greatest health effect from a specified expenditure, or to find the lowest-cost approach to achieve a given health effect. It is unnecessary to ask whether an allocation based on CE analysis leads to a potential Pareto improvement or a Pareto-optimal distribution.

Although this perspective makes it possible to analyze the optimal allocation of health resources without accepting the full range of welfare economic assumptions, it has other limitations. By eschewing any claim to justification on the basis of a more fundamental framework, the extra-welfarist perspective requires acceptance of the principle that maximizing quality-adjusted life years or another specific health outcome measure should be the goal of health care provision. Acceptance of a specific measure is much more problematic than accepting the general concept that improvement in health is a social good. Results from a study using QALYs as the health measure may differ from those that measure health in terms of longevity. Usually, the validity of the health outcome measure must be assumed rather than tested. The extra-welfarist approach can determine the best measure of health outcomes by appeal to political processes. But to the extent that it rejects market and personal valuations of health improvements, the extra-welfarist approach cannot appeal to a more fundamental set of principles to resolve whether one measure of health outcomes is more valid than another. Nor is it easy to use this approach to evaluate tradeoffs between health and other social goods, such as education, nutrition, or other aspects of well-being. Finally, it provides no direct mechanism for resolving certain economic issues – such as what constitutes a cost, and how cost should be measured.

In contrast to the extra-welfarist perspective, this chapter uses a welfare economic framework to address questions of standardization. The fundamental question underlying our approach is simple: does decision making based on CE analysis, carried out a specific way, lead to a distribution of resources that has desirable social welfare properties? In other words, does a ranking of alternative uses of health resources based on CE analysis lead to an allocation that improves welfare? The answer depends on the way that CE analysis is performed, the way the results are used, and the definition of *social welfare improvement*.

To economists familiar with cost-benefit (CB) analysis, these questions imply another: Why perform CE analysis, rather than CB analysis, whose economic foundations and social welfare implications are well known? In some circumstances they appear to give nearly equivalent results [Phelps and Mushlin (1991)]. However, in principle, CB analysis is more general than CE analysis [Kenkel (1997)]. Furthermore, CE and CB analysis grew from different historical traditions and have been adopted for different reasons. CB analysis requires placing dollar valuations on the outcomes of any program or intervention. In the context of health and medical care, making that valuation can be equivalent to placing a dollar value on a human life (or, more precisely, on changes in the probability distribution of the length or quality of human life). To many in the worlds of medicine and of public health, any attempt to place a value on a human life – even if it is usually a valuation of a small change in the probability of death or a change in the distribution of expected mortality, rather than an attempt to put a price on an identified individual's life [Schelling (1968)] – is anathema. Thus most "economic" evaluations in health care have applied CE analysis, which limits the analyst's responsibility to providing information about the efficiency with which alternative strategies achieve health effects. The often implicit task of placing monetary valuations on health outcomes falls upon decisionmakers and others who read the analyses.

The fundamental differences between the techniques may also reflect the contexts in which they developed. CB analysis was developed primarily to assist in making decisions about the provision of public goods. Although CE analysis has also been used to evaluate public health measures that are public goods or create externalities (e.g., vaccination programs), it is more often used for the evaluation of private goods and services. The reason to apply formal analysis in this context is that information in health care is imperfect and often asymmetric. Asymmetry is common because the producers of health care, consumers, and payers possess different amounts of information about the benefits, risks, costs, and other characteristics of health services. Although limited and asymmetric information is an issue in some contexts in which CB analysis has been applied, nonexcludability and nonrivalry in consumption are the forms of market failure chiefly responsible for the popularity of CB analysis. CE analysis, in contrast, assists patients and their agents in making decisions about health care, which is generally a private good (with some notable exceptions, such as infectious disease control). Both physicians and insurers can act as agents for patients; although the primary function of insurance is risk-spreading, health insurers reimburse for services used rather than making lump sum payments. Consequently, a health insurer should also assure that optimality is achieved in health care consumption by designing coverage and reimbursement so that the marginal utilities of health care dollars are equated across patients and interventions. CE analysis is a technique for doing so.

Information provided by CE analysis is important in two ways: First, health care is valued insofar as it improves health and well-being, not for intrinsic characteristics of the health services. The relationship between the use of a medical intervention and improved health outcomes may not be known to the individual patient or physician. CE analysis can reveal how much value the patient will obtain for a given expenditure on a health intervention. Second, as Pauly (1968) has noted, nearly all forms of health insurance are subject to moral hazard. Once an enrolled individual has a disease or other health condition, he or she would prefer to consume it to the point at which the marginal benefit equals the marginal cost to his or her patient. Because insurance lowers the patient's share to a small fraction of the full marginal cost (the fraction usually determined by a fixed usage fee, percentage copayment, or deductible), insurance ordinarily results in overconsumption. *Ex ante*, an individual would prefer actuarially fair insurance which guaranteed that care would be provided to the point at which marginal cost (insurance payment and copayment combined) equaled marginal benefit over insurance that was subject to moral hazard. Use of CE analysis to allocate care (usually based on coverage decisions) might help limit moral hazard by overcoming informational limitations.

In theory, the use of CE analysis to address moral hazard is straightforward. Consider a world of (near) perfect information. That is, effectiveness and costs of treatment are known, but information is not sufficiently inexpensive to enable insurers to monitor and overcome moral hazard. What would the ideal health insurance plan attempt to do? Risk-averse individuals desire insurance for the usual reasons. They might also want the insurer to act as their agent in deciding how much and what kinds of health care each should receive (or equivalently, the enrollees would commit to accept levels and types of care that met a net benefit criterion as long as the premiums were actuarially fair). Assume further that every potential subscriber to the insurance plan has the same *ex ante* probability of experiencing each possible stream of health outcomes, so that the prospects of each are equal, as behind the Rawlsian veil of ignorance [Rawls (1971)].

Under these circumstances, if the insurer could act as a perfect agent for the consumer, it would attempt to set the marginal benefits equal to the marginal costs of each intervention, but the marginal cost would be at the point of purchase of the intervention. That is, unless the insurer were a monopsonist, the cost would be the price paid (which in turn would be the sum of the insurer's payment and the copayment). This perspective adds insurer costs to the *patient perspective* that only includes out-of-pocket costs.

The same logic applies to a provider that acts as an insurer, such as a health maintenance organization. However, for services that the provider produces itself, the relevant price is the marginal cost defined over the suitable time horizon. A government program that intended to maximize the welfare of the citizens it serves would use a CE criterion on similar grounds. In each case, it would be optimal to equate the CE ratios of interventions used at the margin, using marginal costs that the program bears – that is, the prices that it actually pays.

To the extent that consensus about specific social welfare criteria is lacking, not everyone will be persuaded by an appeal to welfare economic foundations. Some writers have criticized the utilitarian viewpoint that they believe to be embedded in this approach. The justification for CE analysis on this basis is indeed rooted in the compensation principle (or Kaldor–Hicks criterion) of CB analysis [Hicks (1939), Kaldor (1939)]. This principle states that we should undertake a project if and only if its net benefits are positive, since then those who gain from such a project gain by enough to compensate those who lose. If the losers are compensated, nobody is made worse off by the project, and someone is made better off. Thus the term *potential Pareto improvement* – the project *could* result in an actual Pareto improvement if the winners compensated the losers. Since a precisely compensating reallocation is unlikely to occur, this criterion is less compelling than Pareto improvement, since a project that produces positive net benefit would make people who shared the costs but not the benefits worse off.

The chapter is organized as follows. The first section briefly describes the basics of CE analysis and how it can be applied to aid decisions about the allocation of health resources. The chapter then turns to the potential welfare economic foundations of CE analysis, drawing heavily on my work with Charles Phelps. The chapter then addresses specific issues in carrying out CE analysis, such as which costs to include, whose perspective matters in the analysis, and how health outcomes are measured. It demonstrates how a welfare economic foundation can help resolve ambiguities and uncertainties about the application of CE analysis. The chapter also discusses the limitations of such an approach, which indeed reflect limitations of CE analysis as an analytic framework. Finally, it addresses unresolved issues such as the difficulties in using the results of CE analysis to make health policy at the societal or group level.

2. Cost-effectiveness analysis for decision making

How useful and valid are the results of CE analysis if its purpose is to improve the well-being of a population by guiding the allocation of health care resources? Making this judgment requires choosing a benchmark for well-being and an explicit statement about how CE analysis can be used to achieve the welfare objectives. Major published recommendations for the use of CE analysis in guiding decisions state that it must be weighed with a variety of political, distributional and practical considerations. The information that CE analysis contributes is summarized by the *CE ratio.* The CE ratio is a cost per unit health effect achieved by using a particular health intervention. The CE ratio demonstrates which uses of health resources will provide health most efficiently; by first using interventions that have the lowest CE ratio, i.e., that produce the greatest effect from a specific expenditure, it is possible to obtain the greatest overall health effect from a limited budget for health care. Recent work on welfare foundations of CE analysis has used standard neoclassical welfare economic formulations to examine whether implementation of CE analysis in this way (i.e., using different interventions to the point that their incremental CE ratios are equal at the margin) leads to the same allocations as the ones that result from individual utility maximization subject to income constraints.

To explore these issues further requires knowing precisely what the CE ratio represents and how it is calculated. As one might expect, the closer the connection between the health outcome and individual welfare, the more plausible the claim that allocations based on CE criteria maximize welfare.

Several authoritative textbooks and reviews have described the general approach for performing a CE analysis; see, for example, Drummond et al. (1997), Gold et al. (1996), Weinstein and Stason (1977). I briefly summarize the approach here.

First, the intervention to be studied, along with alternative interventions to which it is being compared, must be defined. One of the alternatives might be "doing nothing," or applying no specific intervention. This has been the principal alternative considered in many CE analyses. Yet a CE analysis based on a comparison with this alternative is not always informative, since the comparison should be between relevant choices, such as two treatments or diagnostic approaches that clinicians or policymakers would consider to be the most promising. Little can be learned from a CE analysis that compares an intervention with placebo when placebo is not considered a reasonable option. The CE ratio for a comparison with placebo can be favorable even when the intervention in question is in every respect inferior to one or more commonly used alternatives. Several medications, for example, are both effective and cost-effective when used to treat adults with moderately elevated blood pressure. The relevant question for a new blood pressure medication is how it compares to another promising medication, or to others that are well-established, rather than how it compares to the abandoned approach of forgoing treatment.

After we choose the intervention and alternative to be studied, we must assemble several elements of the CE analysis to calculate the *incremental* (or marginal) CE ratio. Throughout this chapter, the term CE ratio refers to the incremental CE ratio, unless otherwise specified. The term *incremental* is used rather than *marginal* to avoid confusion with the term *marginal cost*, which is usually the preferred measure of opportunity cost in CE analysis. *Incremental* refers to differences between two interventions; since the comparison does not always involve an infinitesimal change in costs and effectiveness, the term "marginal" can be misleading.

Let the subscripts 1 and 0 denote the intervention under study and the alternative to which it is compared, respectively. If C_1 and C_0 are the net present values of costs that result when the intervention and alternatives are used, and E_1 and E_0 their respective health outcomes, the incremental CE ratio is simply

$$\text{CE ratio} = \frac{C_1 - C_0}{E_1 - E_0}. \tag{1}$$

This ratio, which is a cost per unit incremental health effect, is often used as a measure of value. The CE ratio of the intervention under study is compared to the CE ratios of other commonly used forms of medical care; if it is relatively low, the intervention under study is considered to be a good value. Note that the intervention and alternative can be two different intensities of the same treatment (e.g., dosage of a drug), and that the CE ratio can be defined as an infinitesimal charge. The continuously valued approach to the CE ratio underlies the analysis of Section 3.

The elements of the numerator of the CE ratio, or the incremental cost of the intervention, are discussed below. There is consensus that C_1 and C_0 should represent net present values, but the specific content of these numbers is controversial. Much of the literature has used formulations similar to that of Weinstein and Stason, who stated that net health care costs consist of "all direct medical and health care costs [including] costs of hospitalization, physician time, medications, laboratory services, counseling, and other ancillary services." In addition, the costs include those "associated with the adverse side effects of treatment," the (negative) costs from "savings in health care, rehabilitation and custodial costs due to the prevention or alleviation of disease," and "the costs of treating diseases that would not have occurred if the patient had not lived longer as a result of the original treatment" [Weinstein and Stason (1977, p. 718)]. Many studies have attempted to measure costs by including these categories. Some experts exclude those that arise solely from living longer, as previously noted. Others have included additional costs, such as "indirect" or "productivity" costs (i.e., time costs of treatment and/or disease, lost wages, and so on) and consumption expenditures. The Panel on Cost-Effectiveness in Health and Medicine recommended against including as costs the monetary value imputed for lost life years (i.e., lost earnings; see the chapter on estimating costs by Luce et al. (1996)) and withheld endorsement of including future consumption expenditures, yet many CE studies have incorporated the imputed value of lost years of life in the cost measures.

The denominator of the CE ratio is calculated in an analogous manner; it represents the incremental health effects of using the intervention. Typical measures of health outcomes are either *years of life saved* or *quality-adjusted life years* (QALYs) saved. QALYs were introduced into the literature in the mid-1970s as a way to incorporate the benefits of treatment more fully than could be accommodated with earlier outcome measures. They are intended to serve as a comprehensive measure of health, or health-related well-being. In many respects QALYs are analogous to life expectancy, but give credit to interventions that improve quality of life even when they do not affect survival.

Each year that an individual lives longer contributes an additional year to the life expectancy calculation. The amount that each additional year of life adds to QALYs, in contrast, is a preference weight or utility that takes a value between 0 and 1, varying with health status during the incremental year. Life years marred by functional limitations, pain, and other burdens associated with illness receive less weight than years in good health. Years when health is so bad that it is considered no better than death receive a preference weight of 0; in the usual formulation, death is considered the worst possible health state. A preference weight of 1 corresponds to best health imaginable. Interventions can raise QALYs by lengthening life or improving its quality as reflected in the preference weight. Similarly, an intervention that lengthens life produces more QALYs if it maintains or improves quality of life than if it adds years of life that are impaired by significant morbidity. Both life expectancy and QALYs can be discounted; that is, less weight is given to years of life added in the more distant future.

QALY measurement is most easily understood by extending the measurement of life expectancy. Life expectancy is the sum of the probabilities that an individual will be alive at each age (denoted by i) in the future, up to the maximal life span, or

$$\text{life expectancy} = \sum_{i=\text{current age}}^{\text{maximum age}} F_i, \tag{2}$$

where F_i is the probability that the person who is now at the “current age” will still be alive at age i; this discrete representation is most convenient for working with data such as life table figures, but continuous time representations of life expectancy are also used.

Calculation of QALYs requires the information used to calculate life expectancy and the preference weights. Denote the preference weight for the health characterizing age i by q_i. Each such term is the expected value of quality adjustments for all possible states of health at age i. To illustrate the calculation, imagine that individuals alive at age 60 could be in one of only two possible states of health: perfect health, ($q_h = 1$), occurring with probability 0.5, or suffering from heart disease ($q_d = 0.8$), also occurring with probability 0.5. Then q_{60}, the expected value of the preference weight corresponding to being alive at age 60, is $(0.5 \times 1) + (0.5 \times 0.8) = 0.9$. After estimating the value of q_i for each age i, it is possible to calculate the expected number of QALYs, in the form of present value, according to the formula

$$\text{QALY} = \sum_{i=\text{current age}}^{\text{maximum age}} F_i \delta^i q_i, \tag{3}$$

where δ is a time discount factor whose value is between 0 and 1. As in the formula for life expectancy, F_i is the probability that the person is still alive at age i. If $\delta = 1$, two years of life in which $q_i = 0.33$ contribute the same number of QALYs as one year in which $q_i = 0.66$. If there is no time discounting ($\delta = 1$) and if each year of life has perfect health, or quality adjustment is ignored ($q_i = 1$ for every value of i), then this formula simplifies to the formula for life expectancy.

The mechanical aspects of calculating QALYs are not difficult, but the measurement of the preference weights and the probabilities of alternative states of health is anything but straightforward. The specifics of QALY calculation necessarily account for much of the effort of CE analysis, since the outcome measure is critical to the interpretation of the results. As Section 3 discusses, the outcome measure determines whether the application of CE analysis has desirable welfare-theoretic properties.

2.1. Time horizon

An intervention can alter both costs and health effects long after it is administered. For example, a mammogram uses resources at the time the test is conducted. But if it reveals an abnormality that leads to breast biopsy, mastectomy, and the prevention of morbidity and mortality from breast cancer, it alters the length of life, future morbidity, and future costs of health care. These long-term repercussions are relevant to any evaluation of screening with mammography, so the standard recommendation is that all future costs and health effects should be calculated or estimated in a CE analysis. Measuring these costs and health effects directly – without use of a model that extrapolates these numbers – would require observing until death a large number of women who underwent mammography, along with a number of women who did not have the test. For many treatments and diagnostic or screening strategies, such an approach would require decades of study, yet few randomized clinical trials last for more than five years. Strong beliefs in the credibility of direct clinical trial data, and skepticism about model-based extrapolations beyond the period of the trial, have led some investigators to calculate costs and outcomes for the period of the trial only. Thus, rather than estimate life expectancy or quality-adjusted life years, they calculate survival within the five years of a trial. Similarly, rather than estimate net present value of lifetime health care costs, they measure discounted costs during the period of the trial. Usually, when researchers adopt this approach, they do so in the belief that they have avoided making dubious assumptions needed to extrapolate events and costs that occur beyond the period for which they have valid and reliable data.

This practice is not endorsed by experts on CE analysis. There is no natural interpretation for life-years gained during a finite period of time, and the CE ratios that result from using different time horizons, such as one year and five years, cannot be compared in any meaningful way. In fact, the resulting CE ratios can be understood best by interpreting them as special cases of standard CE ratios. In calculating a standard CE ratio, the time horizon is at least equal to the full span of life. The 5-year CE ratio is the same as a standard CE ratio calculated with an assumption that all individuals die at the end of five years. Thus, in the attempt to avoid the assumptions required for modeling long time horizons, researchers who truncate their analyses have made, perhaps unwittingly, the implausible alternative assumption that study subjects experience neither the costs nor the benefits of living beyond the period of study.

Although it seems intuitive that calculating the CE ratio based on a truncated time period should result in bias, it may not be possible to determine the sign of the bias *a priori*. The bias can only be calculated by making specific assumptions about the costs and health effects that occur after the period of observation. For example, suppose that the intervention in question lowers mortality rates during five years of observation. For individuals surviving the five years, subsequent survival experience and costs are the same for those treated with placebo as for those who received the intervention. Under these assumptions, both the gain in life-years and the increase in costs are greater for the intervention group than would be estimated on the basis of the truncated period of observation. The overall bias in the CE ratio depends upon the relative magnitudes of these omitted costs and health effects.

2.2. *Average CE ratio*

Some CE analyses report an *average CE ratio*, which is simply the ratio of C_1 to E_1. For comparisons among multiple alternatives, a similar practice is common: each intervention is compared to a single alternative. Both approaches are convenient because either they do not require a comparison treatment, or all treatments are compared to a single alternative, rather than to multiple alternatives. Both approaches, however, are misleading. The average CE ratio is equivalent to a standard (incremental) CE ratio in which the alternative is costless and results in immediate death. If such an alternative exists, it is rare for any but the most rapidly and uniformly fatal health conditions. The average CE ratio can deviate greatly from the incremental CE ratio when the intervention under study is a preventive service, which typically would be administered to a relatively healthy population. The members of such a population would be expected to have many years of good health and to generate substantial costs over their remaining lifetimes.

The average CE ratio will not, in general, lead to appropriate rankings of alternative health expenditures [see, for example, Karlsson and Johannesson (1996)], although occasionally it is possible to draw limited inferences about the value of the incremental CE ratio from the average CE ratio. The average CE ratio does not reliably indicate the way to achieve the greatest health benefit from a given expenditure. For example, an intervention that produces more favorable outcomes than one that has a lower average CE ratio could have an acceptable incremental CE ratio but might not be selected on the basis of the average CE ratio; alternatively, the average CE ratio might be considered "acceptable" when the incremental CE ratio was very high.

Comparison of multiple interventions to a single alternative is misleading for nearly the same reason, except that the "baseline" costs and outcomes are not zero, but instead are the costs and outcomes corresponding to the single comparator. It is easiest to understand why this is misleading by comparing it to the incremental approach.

2.3. Incremental CE ratio for multiple alternatives

It is possible to calculate a separate incremental CE ratio for every pair of alternative interventions. When many interventions are considered, the number of such pairs becomes large. However, because most of the incremental CE ratios are irrelevant, the analyst need not calculate all of them. Instead, to determine the incremental CE of a series of different combinations of technologies, the analyst should first rank each alternative by the health effect achieved – e.g., the number of QALYs (or life-years) it produces. Then the analyst should determine whether any interventions are *strictly dominated* (more expensive and less effective than at least one alternative intervention); if any are, they should be eliminated from further consideration. After eliminating all such alternatives, one should calculate the incremental CE ratios between each intervention and the next most expensive alternative. Subsequently, interventions that display *extended dominance* should also be eliminated, and the incremental CE ratios of all remaining alternatives calculated. Extended dominance is defined below.

Figure 1, from Garber and Solomon (1999), illustrates how incremental CE analysis can be applied when multiple alternatives are considered. It shows the costs and health effects of adopting each of several strategies for diagnosing coronary artery disease in 55 year-old women. The first five strategies are exercise treadmill testing (ETT); stress echocardiography (ECHO); planar thallium radionuclide imaging (Thallium); single photon emission computed tomography (SPECT); and positron emission tomography

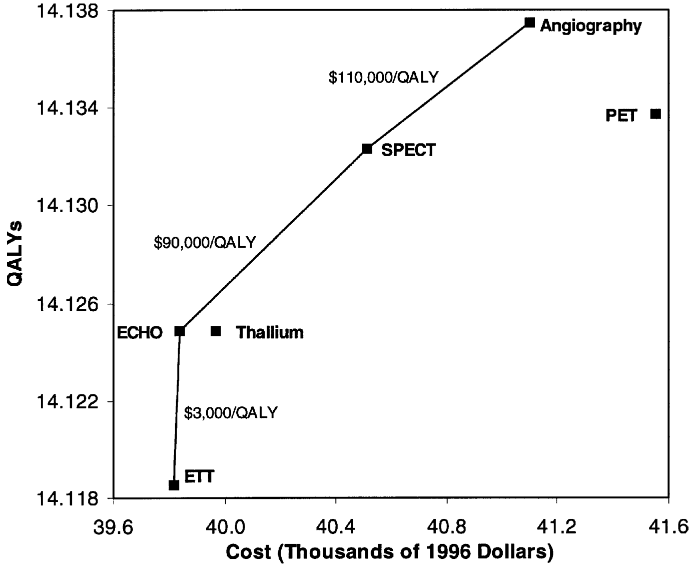


Figure 1. Costs and QALYs with alternative test strategies for coronary artery disease in women, 55 years of age. Reproduced with permission from Garber and Solomon (1999).

(PET). Each of these strategies starts with a noninvasive test for coronary disease. The "gold standard" test for coronary artery disease is cardiac catheterization with coronary angiography; the screening strategies that start with a noninvasive test proceed to catheterization if the test is abnormal. The final strategy shown in the figure (angiography) consists of initial testing with the gold standard test, so that the first test is considered definitive but riskier and more expensive than the other tests.

The costs and outcomes of each of the diagnostic strategies are calculated by modeling the consequences of alternative medical interventions that are pursued on the basis of the test results. For example, if a diagnostic test is positive and leads to the discovery of a severe form of coronary artery disease, it leads to surgical treatment, which in turn may prolong life substantially. A false positive test result has minimal adverse health effects, but leads to substantial expenditures for further testing that is, in retrospect, unnecessary. Figure 1 is a compact representation of results from extensive modeling of alternative strategies that have large but often indirect and complex effects on both costs and health outcomes.

Because each point on the figure represents the overall costs and outcomes in QALYs that result from the use of each test, the incremental CE ratio between any pair of tests is the inverse of the slope of the line drawn between their corresponding points. A point that is above and to the left of another strictly dominates the alternative, i.e., has better outcomes and lower costs. In Figure 1, angiography eliminates PET scanning by strict dominance. Thallium is also eliminated by strict dominance because it produces slightly fewer QALYs than ECHO at greater cost. The incremental CE ratios are calculated for the remaining alternatives.

Figure 2 (also from Garber and Solomon), which shows similar results for 45 year-old men, illustrates extended dominance. For these subjects, unlike 55 year-old women, thallium is not eliminated by strict dominance, since no alternative intervention is both less expensive and more effective in these men. Extended dominance is a somewhat more subtle concept than strict dominance; it occurs whenever a linear combination of two alternatives strictly dominates a third [Keeney and Raiffa (1993), Johannesson and Weinstein (1993), Karlsson and Johannesson (1996)]. Equivalently, the phenomenon occurs when any interventions have "higher incremental C/E ratios than a more effective option" [Siegel et al. (1996)]. Although no alternative is both less expensive and more effective than thallium, it is strictly dominated by at least one point on a line drawn between ECHO and SPECT, so it is eliminated by extended dominance.

Strict dominance and extended dominance are particularly important phenomena because they can identify interventions that should be eliminated from consideration, without making any judgment about what a unit health effect is worth. Strict dominance cannot always be detected without formal analysis, and extended dominance is even harder to discover, unless the analysis includes a systematic approach to incremental CE ratios.

A rational decision maker will never choose an option that can be eliminated under extended dominance, because a more expensive alternative would result in a lower or equivalent CE ratio. Suppose that there are three alternatives under consideration: A, B, and C. Both the costs and the outcomes associated with intervention C are greater than

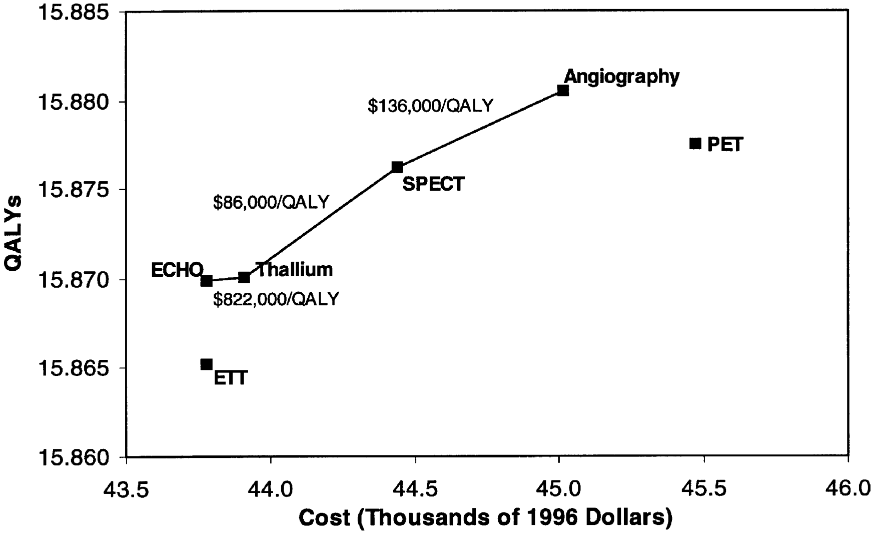


Figure 2. Costs and QALYs with alternative test strategies for coronary artery disease in men, 45 years of age. Reproduced with permission from Garber and Solomon (1999).

those of intervention B, which in turn are greater than those of intervention A. Thus none of the interventions strictly dominates any other. The (incremental) CE ratio of intervention B compared to A is $70,000/QALY, and the CE ratio of C compared to B is $10,000/QALY. If a decision maker would choose B over A, it implies that a gain of a QALY is worth at least $70,000 to him or her. If that is the case, then it must be true that it is worth an additional $10,000 to gain another QALY, so that C would be chosen over B. Thus alternative B is eliminated from consideration by extended dominance.

The CE ratios that result from comparing several interventions to a single alternative, rather than proceeding in this stepwise fashion, can be very different. Usually it is impossible to detect the presence of either strict or extended dominance from such an approach. In fact, the CE ratio produced this way may appear to be "reasonable" even though the intervention under consideration is strictly dominated by another! Suppose that there is an intervention A that generates lower costs than interventions B and C, as in Figure 3. We are interested in choosing among the three. If we calculate cost-effectiveness ratios of B compared to A and C compared to A, it is difficult to determine whether we should choose C over B. If the CE ratio of C compared to A is lower than the ratio of B compared to A, C could eliminate B by extended or strict dominance (points B^1 and B^2 in Figure 3, respectively) or, alternatively, B could have an "acceptable" CE ratio compared to B (point B^3). The only firm conclusion that can be drawn, without further information, is that B does not eliminate C by strict dominance.

Calculation of the incremental CE ratio, then, consists of estimating the QALYs and the present value of costs under the intervention and under its alternatives. The use of

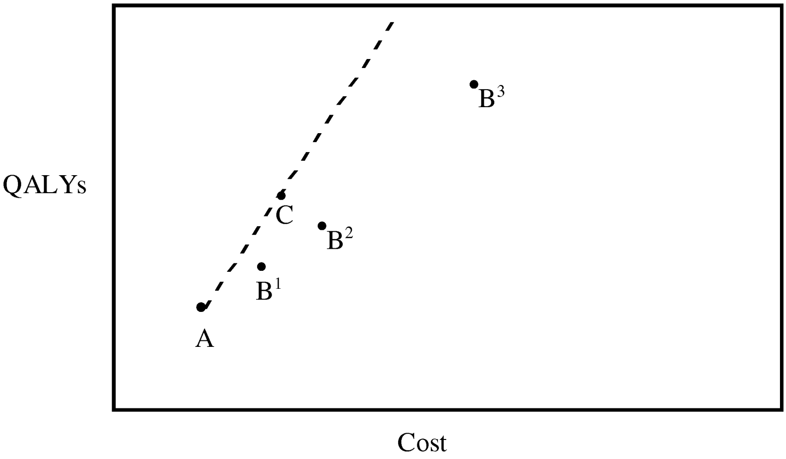


Figure 3. The consequences of comparing two interventions to a third. Intervention A is the lowest cost alternative; the incremental CE ratio of C compared to A is lower than the incremental CE ratio of B compared to A. Interventions B^1, B^2, and B^3 all have the same CE ratio compared to A. C eliminates B^1 by extended dominance and B^2 by strict dominance, while the CE ratio of B^3 compared to C could be "acceptable" (i.e., lower than a CE cutoff). Without further information, it is not possible to determine from the CE ratios of C compared to A and B compared to A which of these three conditions applies.

the average CE ratio or comparison of several interventions with a single alternative is misleading.

2.4. *Sensitivity analysis*

Uncertainty characterizes several components of nearly every CE analysis. Estimates of health effects, whether measured in terms of life-years or quality-adjusted life years, often build upon models that incorporate data from multiple sources. Even if the data are derived primarily from a randomized clinical trial, extrapolations beyond the period of the trial require assumptions about disease course beyond the period of observation. And even if a trial is the sole source of all data used in a CE analysis, sampling variability makes estimates of effect sizes and costs uncertain.

Not all sources of variability are purely random. For example, the costs of an intervention – or of treatments for conditions it prevents – may vary from one setting to another. Thus, for reasons ranging from the usual stochastic nature of experimental information to (possibly non-random) variation in costs and health effects to uncertainty in model structure and specifications, point estimates of CE ratios should ordinarily be considered just that. The variation in possible values around those point estimates may be large.

For this reason, CE analyses are considered incomplete if they do not include some form of sensitivity analysis. *Sensitivity analysis* is an exercise that shows the effects of variation in uncertain parameters on the final results of the analysis (i.e., the CE ratio). Textbooks on CE analysis and decision analysis discuss methods of sensitivity analysis, and most commercial software for CE and decision analysis implements one- or two-way sensitivity analysis. In *one-way* sensitivity analysis, one uncertain parameter is varied at a time, with the values of all other parameters held constant. In *two-way* sensitivity analysis, two parameters are varied simultaneously. When more than two parameters are varied, the presentation of results of multi-way sensitivity analysis can be quite challenging, and creative approaches to graphical presentation are necessary (two-way sensitivity analysis requires three-dimensional plotting, with axes for each of the two parameters being varied and for the CE ratio).

The limitations of traditional sensitivity analysis are most apparent when it is important to display the effects of uncertainty in multiple parameters simultaneously. More powerful alternative approaches, although they are still under development, have been gaining in popularity in part because they are more suitable for complex models with multiple sources of uncertainty. Most are statistical approaches that involve calculating confidence regions around CE ratios and other outcome variables. Briggs and Sculpher's 1995 survey of sensitivity analysis in economic evaluation noted that only one of the 121 CE analyses they reviewed had adopted a "probabilistic sensitivity analysis" approach, whereas 42 used "one way simple sensitivity analysis" and 15 used "multi way simple sensitivity analysis" [Briggs and Sculpher (1995)]. Methods for calculating the range of uncertainty using a probabilistic approach range from the traditional delta method to newer simulation and resampling techniques, such as the bootstrap, which makes it possible to limit parametric assumptions [Mullahy and Manning (1994), O'Brien et al. (1994), Briggs et al. (1994), Wakker and Klaassen (1995)]. But the computational burdens of such approaches remain formidable, and in many cases the statistical theory is not well developed or, like the delta method, require strict distributional assumptions. Furthermore, the patchwork of data used to develop many CE models limits the range of approaches that can be used to gauge the effects of uncertainty.

The welfare theoretical implications of uncertainty in the analysis are important, even if they are indirect. It is not unusual for the range of uncertainty to be great enough to be consistent with different orderings of effectiveness (and costs) of the interventions under consideration. Occasionally differences in costs among alternative interventions are known with a high degree of certainty, but ranges of estimated effectiveness overlap substantially. A common response to this situation is to assume that the effectiveness of each intervention is roughly equal, and to choose the lowest-cost alternative. However, the apparent equivalence of effectiveness may be a consequence either of similar true effectiveness, or of large but highly uncertain differences in effectiveness. In the latter case, further information might alter the ranking of alternatives.

2.5. Interpretation for medical decision making and health policy

After the CE ratios of non-dominated alternatives are calculated, there remains the task of choosing among them. If an intervention improves health at a cost of $80,000/QALY, should it be adopted? Cost-benefit analysis leads to specific recommendations because it places a monetary value on the benefits: any intervention that produces a net benefit generates a potential Pareto improvement. But CE analysis is often preferred precisely because it avoids monetary valuation of health benefits. The next section describes how it is possible to derive a "cutoff" CE ratio that leads to the same choices as a cost-benefit criterion. However, people who apply and use CE analyses and wish to avoid the valuation of health benefits implicit in such efforts often use an alternative approach based on *league tables*.

The term league table apparently originates from the tables of football team rankings published in European nations. League tables in CE analysis also display rankings. This approach compares the CE ratio of the intervention under study to those of other common medical interventions. By compiling a league table of (incremental) CE ratios of other health interventions, usually culled from the literature, one can demonstrate how the CE ratio of the intervention under study compares with those of the other interventions in the table. If the CE ratio is low, the intervention is termed a good value, while if the CE ratio is high, it is identified as a poor value relative to other accepted interventions. Thus the tabular comparison helps to establish whether the intervention should be used.

3. When does CE analysis lead to optimal decisions?

The league table approach, however, has severe limitations as a guide to medical choices [Birch and Gafni (1994)]. Several problems become apparent to readers of the studies that generated the numbers. For example, the various studies summarized in the table may not use comparable methodology; some of the CE ratios may be incremental, others average; assumptions underlying the cost estimates may differ greatly. Although league tables distinguish between interventions that are relatively good and relatively poor values, that judgment is highly dependent upon the specific alternatives displayed in each table. Unless there is a reason to believe that the interventions appearing in the league table were chosen by a process that maximizes value, we can hardly infer that standing in the league table establishes value in any absolute sense. Finally, even if we could infer whether the intervention was a relatively good or bad value, the league table approach does not establish how much should be spent. This observation leads us back to the question posed at the outset: when we apply the results of CE analysis to allocate health care, do we make optimal decisions? No discussion of the welfare economic foundations and welfare implications of applications of CE analysis is meaningful without consideration of how and why CE analysis is being used. For whom is CE analysis being conducted, and how will its results be used in allocation decisions?

The answers to these questions depend upon the *perspective* of the analysis. The approved practice, under most circumstances, is to adopt a societal perspective, in which we are seeking to make the best decision about health care allocation for a group of people. Often, however, this perspective is taken to mean something more specific: the analysis is intended to aid someone such as a social planner – perhaps the health minister of a country with national health insurance or governmentally provided health care – who must decide which health services to provide or reimburse. The adoption of a societal perspective can give rise to ambiguities. For example, how should the government payer handle heterogeneous preferences, if it recognizes them at all?

The following discussion builds upon the presentation in Garber and Phelps (1997). In that paper, the perspective is that of a "perfect insurer," and CE analysis is treated as a tool to determine which services, in what quantities, the perfect insurer should reimburse. Suppose that there is no specific information to suggest that an individual's risk of various health events differs from the average for the insured population, that utility functions and other characteristics are homogeneous, and that the insurance is actuarially fair. Which services would the optimal policy cover? From this point of view, the usual marginal conditions apply, and CB criteria (i.e., measure benefits and costs accurately and cover those services at quantities that result in maximum net benefit) lead to expected utility maximization. Only those services whose expected benefits equal or exceed their expected costs, which will be included in the premium and copayments, will be covered.

The Garber–Phelps approach has two major characteristics: it uses first-order conditions to derive *cutoff* or *threshold* CE ratios, and it determines when various rules for conducting CE analysis allow the technique to be used to determine optimal health resource allocations. It is possible, for example, that ignoring certain categories of costs, such as earnings lost as a result of early death, would mean that decision rules based on CE analysis would no longer be reliable guides to welfare maximization, or that inappropriately including such costs would also lead to incorrect rankings of alternative health programs.

Garber and Phelps construct the health care allocation problem as a simple von Neumann–Morgenstern utility maximization; essentially, they ask whether the first order conditions can be expressed in a form that leads to a CE criterion. That is, they ask whether it is possible to identify a threshold CE ratio such that acceptance of all interventions whose CE ratio falls below the threshold and rejection of those with higher CE ratios would correspond to the allocation selected by direct utility maximization. In the Garber–Phelps model, the threshold CE ratio for an expenditure on a health intervention in the initial period is simply the ratio between the initial period utility and the marginal utility of income in that period. Fundamental to this approach is an assumption that the effectiveness measure is at least an affine transformation of utility. Embedded in the model is an additional assumption that period-specific income is fixed.

The general model is based on an expected utility function in which first period utility U_0 is a function of initial income Y_0 less expenditures on intervention **a**, whose unit price is w_a, and expenditures on intervention **b**, at unit price w_b. Subsequent period-specific utilities are given by the utility functions $U_i(Y_i)$ weighted by the probability that the individual will be alive in period i, F_i:

$$E_0U = U_0(Y_0 - w_a\mathbf{a} - w_b\mathbf{b}) + \sum_{i=1}^{N} U_i(Y_i)F_i. \tag{4}$$

U_i can be written as $U_i = \vartheta\delta^i k_i$, where $\vartheta = U_0(Y)$. In this formulation, Y is constant over time, and k_i is a period-specific multiplier. Thus the summation term has the form of QALYs, in which the quality adjustment for period i is simply U_i; this corresponds to the common use of the term "*utilities*" to describe the quality adjustments. We denote the summation term by $\mathbf{Q}$.

Interventions **a** and **b** can have effects on the probabilities of survival in the future via F_i and on the utilities via k_i. Both F_i and k_i, and their dependence on **a** and **b**, can have an arbitrary time pattern. Obtaining the first-order conditions for the maximization of utility with respect to expenditures on **a** and **b** is straightforward (note that there can be corner solutions, since optimal expenditures might be zero for either or both interventions). Denote the marginal effect of intervention **a** on future period-specific mortality P_i by $\partial P_i/\partial\mathbf{a} = \varepsilon_i^a$, and let the marginal effect of **a** on period-specific quality adjustments k_i be denoted by $\partial k_i/\partial\mathbf{a} = \psi_i^a$. Using the relationship between conditional mortality and cumulative probability of survival

$$F_i = \prod_{j=1}^{i} P_j, \tag{5}$$

and differentiating expected utility with respect to intervention **a**, we have

$$\frac{\partial E_0U}{\partial\mathbf{a}} = -w_aU_0' + \vartheta\left\{\sum_{i=1}^{N}\delta^i\prod_{j=1}^{i}P_j\left(\psi_i^a + k_i\sum_{k=1}^{i}\frac{\varepsilon_k^a}{P_k}\right)\right\}, \tag{6}$$

which when equated to 0 gives the first order condition

$$w_a = \frac{\vartheta}{U_0'}\frac{\partial Q}{\partial\mathbf{a}}. \tag{7}$$

An analogous relationship results from maximization with respect to intervention **b**:

$$w_a = \frac{\vartheta}{U_0'}\frac{\partial Q}{\partial\mathbf{b}}. \tag{8}$$

The analysis then proceeds to show how the first-order conditions can be translated into CE criteria, in which future unrelated costs of health care are either included or excluded.

First, consider obtaining the optimal cutoff CE ratio when unrelated future costs are ignored. Current medical costs are $C = w_a\mathbf{a} + w_b\mathbf{b}$. Let $\mathbf{z} = \mathrm{d}\mathbf{b}/\mathrm{d}\mathbf{a}$, the marginal rate of substitution between **b** and **a**. Differentiating C with respect to **a** and substituting **z** yields the relationship $\mathrm{d}C/\mathrm{d}\mathbf{a} = w_a + \mathbf{z}w_b$. Then the CE ratio for intervention **a** is

$$\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_a = \frac{\mathrm{d}C/\mathrm{d}\mathbf{a}}{\mathrm{d}Q/\mathrm{d}\mathbf{a}} = \frac{\partial\mathbf{C}/\partial\mathbf{a} + \mathbf{z}w_b}{\partial Q/\partial\mathbf{a} + \mathbf{z}(\partial Q/\partial\mathbf{b})}. \tag{9}$$

Using the first order conditions to solve for the optimal values of $\partial Q/\partial\mathbf{a}$ and $\partial Q/\partial\mathbf{b}$ implies that, at the optimum investment in intervention **a**,

$$\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_a = \frac{w_a + \mathbf{z}w_b}{(w_a + \mathbf{z}w_b)(U_0'/\vartheta)} = \frac{\vartheta}{U_0'}. \tag{10}$$

According to this equation, the ratio of incremental costs to incremental QALYs from further investment in intervention **a** is proportional to the reciprocal of the marginal utility of consumption in the initial period, U_0'. Here, the term *incremental* is completely synonymous with *marginal*, since the CE condition is based on a comparison of an incremental expenditure on **a**, rather than on a comparison to a (discrete) alternative intervention. We can use an analogous procedure to obtain the optimal cutoff CE ratio for intervention **b**, yielding the result that, at optimal investment in **b**,

$$\left(\frac{\mathrm{d}C}{\mathrm{d}Q}\right)_b = \frac{w_a/\mathbf{z} + w_b}{(w_a/\mathbf{z} + w_b)(U_0'/\vartheta)} = \frac{\vartheta}{U_0'}. \tag{11}$$

Thus, when future costs are ignored, the first order conditions imply that a single optimal CE ratio applies to all interventions.

A similar analysis establishes the optimal CE ratio when future costs are included. In this case, the numerator of the CE ratio is the marginal cost of the intervention, including future health care costs. The lifetime costs are

$$C^{\mathrm{tot}} = w_a\mathbf{a} + w_b\mathbf{b} + P_1\delta c_1 + P_1P_2\delta^2c_2 + \cdots, \tag{12}$$

where c_i = total health expenditures in period i. Associated with the use of an intervention are costs of the intervention itself, induced changes in expenditures for the other intervention, along with expenditures that result from living longer:

$$\begin{aligned}\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}\mathbf{a}} = {} & w_a + w_b\frac{\mathrm{d}\mathbf{b}}{\mathrm{d}\mathbf{a}} + \frac{1}{P_1}\left[\frac{\partial P_1}{\partial\mathbf{a}} + \frac{\partial P_1}{\partial\mathbf{b}}\frac{\mathrm{d}\mathbf{b}}{\mathrm{d}\mathbf{a}}\right]\left[\delta P_1c_1 + \delta^2P_1P_2c_2 + \cdots\right] \\ & + \frac{1}{P_2}\left[\frac{\partial P_2}{\partial\mathbf{a}} + \frac{\partial P_2}{\partial\mathbf{b}}\cdot\frac{\mathrm{d}\mathbf{b}}{\mathrm{d}\mathbf{a}}\right]\left[P_1P_2\delta^2c_2 + \cdots\right] + \cdots.\end{aligned} \tag{13}$$

This expression can be rewritten

$$\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}\mathbf{a}} = w_a + w_b\frac{\mathrm{d}\mathbf{b}}{\mathrm{d}\mathbf{a}} + \frac{\partial E}{\partial \mathbf{a}} + z\frac{\partial E}{\partial \mathbf{b}} = \frac{\mathrm{d}C}{\mathrm{d}\mathbf{a}} + \frac{\partial E}{\partial \mathbf{a}} + z\frac{\partial E}{\partial \mathbf{b}}, \tag{14}$$

where $E =$ the net present value of expected health expenditures and as before $\mathbf{z} = \mathrm{d}\mathbf{b}/\mathrm{d}\mathbf{a}$.

By following the procedures used to obtain the optimal CE ratio when future costs are excluded, it is easy to show that

$$\left(\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}Q}\right)_b = \left(\frac{\mathrm{d}C^{\mathrm{tot}}}{\mathrm{d}Q}\right)_a = \frac{\vartheta}{U_0'} + \frac{(1/z)(\partial E/\partial \mathbf{a}) + \partial E/\partial \mathbf{b}}{(1/z)(\partial Q/\partial \mathbf{a}) + \partial Q/\partial \mathbf{b}}. \tag{15}$$

Thus, when unrelated future costs are included, the first order conditions imply a fixed optimal CE ratio that is the same for all interventions. The second ratio on the right-hand side of Equation (15) is a constant when the future costs are unrelated, so the optimal CE ratio when future costs are included is equal to the optimal CE ratio when the future costs are excluded, plus a constant.

This result follows from a number of assumptions. A key one is the optimality of future health care expenditures. If the expenditures are not optimal, it will ordinarily be difficult to apply a CE criterion, since the quality adjustment terms for future years will need to reflect differential utility losses from varying distortions in health care consumption in future years. In addition, this analysis uses a strict definition of "unrelated" future expenditures: conditional on reaching a given age, a person's expenditures on health care do not change with an increase in the quantities of intervention **a** or **b** consumed. Thus the goods under study cannot be close substitutes or complements for other forms of health care (nor can there be changes in the rates of substitution between quality-enhancing and life-prolonging health care). The conditional independence assumption, which is intended to be an accurate representation of the term "unrelated" that often appears in the literature without precise definition, is strict. Even if it can seldom be satisfied exactly, it may be a reasonable approximation for some interventions, such as the treatment of a young accident victim with severe blood loss whose future expected pattern of health may be unaltered by the accident if he or she survives.

This approach does not justify the application of a fixed threshold CE ratio when the first-order conditions cannot be met (e.g., the quantities of **a** and **b** cannot be varied continuously) or when the second-order conditions for a maximum cannot be met. Garber and Phelps argue that the quantities of most health interventions are continuously variable more often than is usually apparent. For example, a screening test might at first seem to be an example of an unambiguously discrete-valued quantity; a woman either has a mammogram or she does not. It is not possible to undergo partial mammographic screening for breast cancer. Yet there are several margins over which the quantity of mammography can be varied, such as the frequency of screening. In addition, the definition of a "positive" test – i.e., one that will lead to further diagnostic evaluation – is often variable (for example, one or more radiologists interpreting a mammogram could estimate the probability that a cancer is present). A more permissive threshold for abnormality results in more true-positive and false-positive test results, usually leading to better health outcomes and higher costs. Variation along such margins can be used to achieve the first-order conditions. As Garber and Phelps note, application of the CE approach in general requires the marginal conditions to hold, because otherwise the use of a fixed CE ratio to be applied across all interventions, as implied by the comparisons in league tables, will be misleading. When the marginal conditions do not hold, optimal health resource allocation will not imply a fixed CE ratio across all interventions.

Restrictions in this model reflect an interpretation of QALYs in utility terms. More flexible utility functions and less restrictive assumptions, such as allowing for variable income and intertemporal reallocation of income and consumption, can change the results, as Meltzer (1997) reported. Extending the Garber–Phelps approach by allowing for borrowing and lending and explicitly distinguishing between health and non-health consumption, he reported that the first-order conditions could no longer be expressed solely in terms of a ratio between marginal costs of health interventions and marginal outcomes. His CE condition implied that "cost-effectiveness analysis must include the total change in future expenditures which results from a medical intervention, regardless of whether those expenditures are medical or non-medical . . . the cost-effectiveness ratio can be viewed as being the sum of a component related to current cost and a component related to future cost." Thus, according to Meltzer, not only "unrelated" future expenditures for health care, but also non-medical consumption expenditures, must be incorporated whenever the intervention under study prolongs life. His results pose a severe challenge for the routine practice of CE analysis, since the utility terms that the quality adjustments need to measure are even further removed from routine measurement of QALYs than under the Garber–Phelps model. Furthermore, the unavailability of accurate health and non-health consumption data has deterred most researchers from implementing any approach that incorporates the present value of non-health consumption as a health cost.

One way to interpret the results of these papers is that the decisions based on CEA can have favorable welfare economic properties, but only if both the costs and outcomes are measured properly. The outcome measure can serve as a basis for determining the first-order conditions only if it is a valid proxy for utility. Common practices in quality of life measurement, however, cast into doubt their ability to proxy overall utility. When developers of instruments for quality of life give respondents any information about what they should assume concerning the socioeconomic status and other factors that might change with a health state, the instructions usually say to consider only health-related aspects. Although rarely are versions of this instruction complete and explicit enough to define "health-related" precisely, their wording often implies that the respondent should ignore financial consequences of a health condition. A treatment that improves an aspect of utility – including utility from consumption expenditures – that is not measured by the effectiveness measure cannot be evaluated properly in this circumstance. But insofar as QALYs or similar outcome measures are used, and are sufficiently broad to serve as a proxy for utility, it becomes much more plausible to represent utility maximization by a CE criterion.

Difficulties with interpreting existing QALY instruments as utility measures should not cast doubt on the theoretical appropriateness of CE analysis. The analysis can have stronger justification as a tool for welfare improvement if a better instrument is used. Furthermore, even CE analysis based on flawed measures of utility can provide a reasonable prioritization of alternative programs to improve health. In many circumstances, the alternative to CE analysis is a decision making process that devotes little attention to either the costs or health consequences of the various policy options. Insofar as it de-emphasizes or ignores considerations such as costs, it would be surprising if such an alternative would consistently prove to be a better guide to improvement of social welfare than even a flawed implementation of CE analysis.

4. Perspective and cost measurement

Despite its prominence as the numerator of the CE ratio, cost typically receives less space and research effort than effectiveness in CE analyses. This disparity may reflect the belief that measuring costs is relatively straightforward or that uncertainty about costs can be addressed adequately in the sensitivity analysis. Typically there are few direct data about the QALYs or life expectancy attributable to the use of a particular health intervention. Even when preference and cost data used for CE calculations are collected as part of a randomized "clinical-economic trial," outcomes must be modeled, as noted previously, because the duration of the trial is too short (typically five years or less) to measure directly the QALYs that result. (Direct measurement of QALYs requires following trial participants until they die.) Cost data, on the other hand, are considered to be relatively explicit and objective.

Estimated costs are usually (but not always) based on prices or, in the case of hospital services, accounting costs. In the US, both accounting- and price-based costs are problematic because both vary greatly. The price of a prescription drug purchased at a retail outlet in New York may differ greatly from the price charged by a hospital pharmacy in Los Angeles, which in turn differs from the price that a managed care organization pays a drug manufacturer. For complex services, such as a major operation, price variation may arise from variation in the definition of the service (not all cardiac valve replacement operations, for example, are the same), and from variation in the prices of factors such as nursing time, surgeon time, and hospital facilities. Although price and accounting cost variation is both large and pervasive in some systems, it is not an insuperable problem for CE analysis. The judicious application of sensitivity analysis can mitigate problems arising from both variation and uncertainty in costs. Furthermore, in most applications, the uncertainty is greatest for costs incurred in the distant future. Such cost estimates require speculation about future health care practices and disease patterns, and thus compound uncertainty about the costs per unit of service. Discounting future costs at an interest rate of 3% or higher, however, means that different methods for measuring costs incurred in the distant future often produce similar present values. Consequently, many CE studies focus on estimation of effectiveness, which often requires indirect inference from results of disparate studies and the use of complex models.

Measurement of costs may nevertheless pose fundamental questions. The most basic is, what is the appropriate measure of cost for use in CE analysis? Should it be marginal cost, average cost, or neither? Many of the leading references on CE analysis say little about specific cost measures. For example, the aforementioned article by Weinstein and Stason (1977) enumerated categories of costs to include in direct medical and health care costs. But the article did not specify whether "costs" are prices in the service market, marginal costs of production, or average costs. In the presence of market imperfections – especially when fixed costs are significant – these alternative measures of cost can differ greatly. In a more detailed discussion of costs, the first edition of the textbook by Drummond et al. (1987) stated that the costs should be "an estimate of the worth of the resources depleted by the programme" (p. 27) and subsequently discussed the various categories of costs (marginal, variable, average, and fixed costs), noting the reasons why different cost measures might be used. Their discussion suggests that the difference in total costs between two alternatives should be used as the measure of costs. Their discussion of how capital costs can be measured, however, stops short of recommending a specific measure to use if fixed or capital costs are large.

The treatment of fixed costs is only one of several controversies surrounding the measurement of costs in both CE and cost-benefit analysis. Experts debate whether only direct costs of the alternatives and of subsequent health care should be included, or whether productivity (indirect) costs (lost earnings or lost value of time) should also be included. They also debate how direct costs should be measured. What if, as is usual in health care, prices do not equal marginal costs? What is the appropriate measure of opportunity cost when markets are imperfect?

4.1. Should the societal perspective be the default?

Although there are not ready answers to all of these questions, they can be best addressed in the context of a specific perspective. Textbooks and review articles routinely emphasize the importance of selecting the perspective of the analysis [US Congress Office of Technology Assessment (1980), Weinstein and Stason (1977)]. Perspective determines whose costs are counted; the perspective of the patient, for example, is usually held to mean that only the costs that the patient bears directly – not the payments of an insurer or government program – matter. Since a typical American with indemnity health insurance bears 20% or less of the price of a covered health service, and in other health care systems the patient's share of costs is often negligible, an intervention that looks very cost-effective when only the patient's out-of-pocket costs are considered may seem like a poor value when the cost measure reflects total costs to the health care system. Opportunity costs, therefore, must be defined with reference to the perspective of the analysis.

The standard recommendation to conduct CE analysis from the societal perspective means that all costs, whether born by patients, insurers, or other parties, are included.

Other perspectives may also be considered, but they are options to be contrasted with the societal perspective, not replacements for it. As in other perspectives, there should not be double-counting of costs (which in turn implies that pure, frictionless transfer payments are not counted as costs), nor in the societal perspective should any relevant costs be omitted. Consider an operation that costs $10,000, for which the insurer pays $8,000 and the patient pays a $2,000 copayment. A CE analysis conducted from the perspective of the patient would assign only a $2,000 cost to the intervention, one conducted from the insurer's perspective would assign $8,000, and one conducted from a societal perspective would assign the full cost of $10,000.

Critics of recommendations to make the societal perspective the default or principal perspective for CE analyses often note that analyses are conducted for a variety of reasons. Consumers and producers of CE analyses can be payers, pharmaceutical companies, providers, and purchasers of health care, so their cost perspectives may be relevant in many important and common situations. These criticisms of the use of the societal perspective are based on an assumption that a payer or government agency, for example, can ignore costs that it does not bear. Yet this assumption is not always realistic. Consider a private insurer; the "*payer's perspective*," as usually conceived, includes reimbursements that the insurer pays but not the out-of-pocket payments of its subscribers. If an insurer does not care about the well-being of its subscribers, so that it can ignore the costs the subscriber bears, then why does it care about maximizing each subscriber's QALYs, which are usually far more difficult to measure? If an insurer sells policies in a competitive market, the value of the policy will depend in part upon the out-of-pocket expenses and time costs that the patient bears. The belief that the insurer ignores costs to the patient overlooks an important fact: insurance programs that account for out-of-pocket expenses and time costs as well as payments by the insurer offer greater benefit to subscribers than do those that ignore such costs. In the face of informational limitations and other forms of market failure, a private insurer may not provide optimal levels and types of insurance coverage, but one that ignores costs borne by the subscriber is unlikely to survive long in the marketplace.

Government programs can also act as payers or as providers (as does Great Britain's National Health Service); the same consideration applies to them. Some government functionaries may consider only the costs that their agencies or programs bear. Implicit in such a strict *government perspective* is an assumption that the health benefits the agency provides are relevant, but monetary benefits and costs, unless directly borne by the agency, are not. Such a point of view, even if widely held by government officials, is at odds with the overt aim of such programs: to serve citizens. The beneficiaries of such programs care about the costs that they bear themselves, in addition to the health improvements that result from the services that they receive. Officials who hold a narrow governmental perspective might recommend extensive centralization of clinical services so that, for example, a diabetic might need to travel for several hours for a routine office visit. Surely the inconvenience and cost to the patient, if regularly ignored, would have repercussions for the official, the agency, and the government. The consequences might not be equally severe or immediate in every society or political system. Nevertheless, government agencies must be concerned about their budgets and the costs and benefits to the populations that they serve. Thus the societal perspective is informative even for payers, government agencies, and other entities that would seem to have an interest in a more limited range of costs.

4.2. The challenge of fixed costs

Implementation of the societal perspective can be difficult, especially when the production of a health intervention requires high fixed costs. The societal perspective usually implies that health services should be used to the point where marginal costs equal the value of the marginal gain in health outcomes. But in the presence of significant fixed costs, price deviates substantially from marginal cost. Large investments for research and development are necessary before many drugs and medical devices can be marketed. Marginal costs of production fail to account for the substantial development investments that are characteristic of pharmaceuticals. Typical recommendations to use marginal costs in CE analysis differ strikingly from typical practice, which uses some measure of the sales price of medications. Price is often many multiples of the marginal cost of producing a drug, at least while the drug is still under patent protection. Many of the same issues arise in joint production and in other situations in which costing is ambiguous.

For the most part, the CE literature gives little guidance on this subject. There is widespread understanding that neither charges nor actual payments for health care are necessarily equal to costs of production, at least as defined in conventional economic terms [Finkler (1982)]. The Panel on Cost-Effectiveness in Cost and Medicine, noting that cost should represent an opportunity cost, went well beyond most of the published CE literature in discussing in comprehensive terms what the alternative measures of cost are, and what measures are theoretically justifiable. The Panel generally urged that long-term marginal costs should be used as the basis for costs, but the specific recommendation depended on the question being asked. They recommended that "fixed costs... should be excluded... costs should not be included for inputs or outputs that are unaffected by changes in the intensity or frequency of an intervention." The panel then made the observation that in the long run there are few fixed costs.

In a discussion of R&D and "first-copy"costs, the report reiterated the recommendation, stating "if the technology has already been developed and the decision addresses the use of the intervention, such as dosage of a drug or frequency of a screening test, then the price should exclude R&D costs. Instead, the relevant costs are the incremental production, distribution, and provision costs." Thus, it suggested that the first-copy or fixed R&D costs should be ignored, implying that the CE analysis should use the marginal cost of the intervention even if the price paid (as for a drug) would often be substantially higher.

This approach might correspond to the outcome that we would seek from a cost-benefit analysis in which we attempted to maximize welfare by adding consumer and producer surplus. The usual teaching (that is, abstracting from the difficult problem of

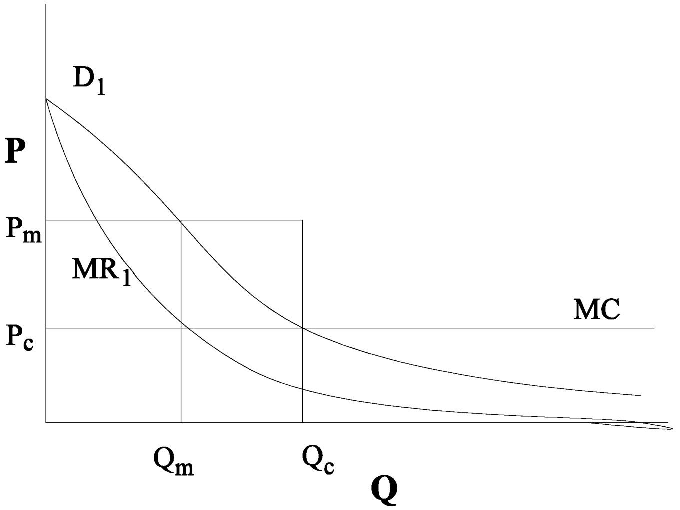


Figure 4. Monopolistic pricing and competitive quantities. The classic monopolist chooses the quantity to set marginal revenue to marginal cost, indicated by Q_{m}, and adopts the price corresponding to that quantity on the demand curve, P_{m}. Presumably implementation of a CE criterion with quantity set according to marginal cost pricing would result in the competitive quantity Q_{c}, but the price would be P_{m} rather than P_{c}. Monopoly revenues would therefore be $P_{\mathrm{m}} * Q_{\mathrm{c}}$ rather than $P_{\mathrm{m}} * Q_{\mathrm{m}}$. If the purchasers are not price-takers the market behavior might correspond more closely to bilateral monopoly, so that the price paid might be less than P_{m}.

determining how to pay the fixed costs) is that the socially optimal level of consumption would be the point at which the marginal benefits equal the marginal costs (see Figure 4), which might be low for a drug.

In a static partial equilibrium analysis that level of consumption would be Pareto optimal, and the effects of changes in price would be purely distributional. As Figure 4 shows, the revenues to a monopolist under an allocation that used marginal costs for the CE criterion but required payment of monopoly prices would lead to larger revenues for the producer than under the conditions of monopolistic supply and competitive demand (price-taking purchasers).

Despite the seeming clarity of their recommendation for excluding fixed costs, the Panel's discussion does not provide unambiguous guidance when fixed costs are substantial. The Panel seemed uncomfortable mandating that only this perspective on costs would be appropriate. Although the Panel did not state this explicitly, if a government agency or insurer announced that it would make coverage or provision decisions based on decision rules that ignored fixed or first-copy costs, they would directly influence research and development decisions for future products and services by assuring high rewards to innovation. In other words, although fixed and first-copy costs for existing technologies have already been borne, investments in fixed costs are endogenous and dependent upon expected revenues, which in turn depend upon the rule for handling such costs in CE analysis.

Recognizing that the authors and readers of CE analysis are rarely concerned with producer's surplus and rents, the Panel's report leaves room for other perspectives:

> ... For perspectives other than societal, the price paid by the decision maker for the good or service is the relevant one, inclusive of whatever return on investment in R&D or rent to patent- or copyright-holder has been incorporated into the price. If a patient or insurance carrier pays a price for zidovudine (AZT) that reflects patent restrictions, for example, the relevant price for a CEA is the one paid, not the opportunity cost of the inputs that went into producing the actual units of AZT consumed...

Since the Panel generally endorsed the societal perspective, what justification can there be for this more limited perspective? Is this perspective appropriate when there are high fixed or first-copy costs?

This more limited perspective is used in most CE analyses of drugs, suggesting that few analysts consider the full societal perspective to be the appropriate one in this context. Few purchasers of health care would be interested in an analysis that evaluated CE of an intervention by assuming a cost much lower than the price at which they could obtain it. That may be why the Panel gave such explicit, and favorable, attention to a perspective that was not societal in the context of high fixed costs. But is the usual practice excessively narrow, ignoring benefits to the producers of interventions?

There is little question about the importance of this issue. New drugs and medical devices are almost always produced by monopolists (albeit sometimes competing with close substitutes), so the disparity between price and marginal cost is large. According to a comprehensive report on pharmaceutical R&D published by the Congressional Office of Technology Assessment in 1993, in the US the cost of bringing to market a drug whose R&D was initiated between 1970 and 1982 was about $194 million [US Congress Office of Technology Assessment (1993)]. This figure is open to debate, and industry sources claim the cost is $250 million or more. Nevertheless, there is no doubt that profits require charging more than marginal cost. Marginal costs – in particular, the costs of manufacturing additional units of a drug – are proprietary information, and are generally unknown.[1] However, because the original producer of a drug is usually believed to have the lowest manufacturing costs (since it is a large-scale producer), the prices of generic compounds after patent expiration give upper limits on the marginal

[1] As part of a study for the Office of Technology Assessment, my colleagues and I attempted to determine the R&D costs and production costs for a very expensive drug (alglucerase) used to treat Gaucher disease, an uncommon genetic disorder. Although we were able to discuss the costs and view internal accounting documents from the company, it was very difficult to ascertain the manufacturing costs and the R&D costs. Production of alglucerase, which was made by chemical modification of an enzyme found in human placentas, was unusually expensive, but nevertheless we estimated that the price of the drug was about twice the marginal cost. The R&D costs born by the company were relatively small, since the drug was discovered by federal scientists and licensed to the company [see Garber et al. (1992)].

costs, and these prices are often small fractions of the prices charged during the period of patent protection. Thus, the disparity between price and marginal cost is likely to be large for most drugs that are under patent protection. Although the same may be true of devices, they have been studied less and production costs may account for a larger share of their average costs.

By recognizing what CE analysis can do best, we can begin to reconcile the contradiction between the usual practice and the usual recommendation of adopting a societal perspective, i.e., one that includes all costs and ignores fixed costs. The technique is not particularly useful for determining the full social optimum, particularly in a dynamic context with large fixed costs, and it is rarely used for that purpose. Instead, *the relevant perspective in most cases is that of consumers and their agents*.

The perspective is essentially that of a perfect insurer, as defined in the Garber and Phelps paper. Mark Pauly has argued that a similar perspective, that of a managed care organization, is often the best one to use in thinking about health care allocation decisions [Pauly (1995)]. This perspective differs from a full societal perspective by ignoring producer surplus. Because the producer surplus is a real component of welfare, government or society should not ignore it. But the practical challenges that must be overcome to maximize the combined surplus by using CE analysis are considerable. For example, if "society" is a province of Canada and the intervention in question is a drug produced by an American company with investors from around the world, Canadians who give greater weight to benefits that accrue to other Canadians will not weigh the company's profits as highly. If the drug or other intervention cannot be obtained at marginal cost, and if health budgets are constrained, can there be any assurance that the attempted application of a CE criterion based on marginal cost will lead to an optimal distribution? A health plan or program that strictly applies the marginal cost concept will treat the costs of two drugs as if they are equal, if the marginal costs of production are similar, even if the price of one is ten times as great as the price of the other.

The attempt to invoke a full societal perspective raises both theoretical and practical difficulties. For example, if buyers purchase pharmaceuticals to the point at which marginal cost and marginal benefits are equal, but pay a monopoly price, monopoly profits should be substantially greater than under the conventional monopoly equilibrium (at which marginal revenue equals marginal cost; see Figure 4). Although the resulting allocation might be Pareto-optimal in a static world, it creates incentives that might cause distortions in investment decisions. The extraordinary profits would induce overinvestment in the development of new pharmaceuticals. Furthermore, as the preceding discussion noted, marginal costs (particularly for drugs still under patent) are usually unknowable, since they constitute proprietary information.

The approach that uses a full societal perspective, with marginal costs as the measure of the costs, implies the need for a nonmarket method of financing. Application of the CE threshold implies that the quantity of a drug purchased will be larger if the CE cost assigned to the intervention is marginal cost rather than the purchase price. To estimate the full optimum, the analyst would have to take into account distortions induced by the method of financing, such as deadweight losses due to income taxation for financing government health care programs. The behavioral change induced by tax incentives can be large, so that the cost of obtaining funds via taxation can greatly exceed the money raised. It is likely that the distortions induced by the modes of financing private health insurance are also large. The distortions introduced by the method of financing present a problem for any attempt to use a CE (or cost-benefit) framework to determine a full social optimum. The marginal cost criterion, with the implied increase in quantity consumed, will exacerbate the problem.

4.3. Distributional considerations

Distributional concerns about CE analysis are raised frequently; such concerns are also prominent in the most vociferous objections to application of CB analysis. Nearly every public program for health care is intended to mitigate inequalities in health, in part by ensuring that the poor have access to effective care. Thus, many discussions of the desirability of CE allocations consider distributional consequences. A strong emphasis on the magnitude of producer's surplus would be incongruous for those nations and groups with deep beliefs about the importance of distributive justice, especially insofar as the owners of companies that produce pharmaceuticals and other health care products are drawn from the upper ranks of the distribution of income and wealth.

4.4. Summary: costs and perspective

Fundamentally, the major issue in defining costs for CE analysis revolves around the definition of opportunity cost. Ordinarily, prices are reasonable proxies for costs. But numerous market imperfections imply that prices are not always good proxies for marginal costs of health care. Because the value of the cost estimate has implications for the adoption and scale of utilization of health interventions when CE analysis is used to aid decision making, these are not merely technical issues. In real-world situations in which the method is likely to be used, the attempt to implement a societal optimum by using nebulous marginal cost figures and purchasing goods and services as if the cost equaled the marginal cost may be unhelpful. Many of the controversies about costs disappear, or at least the problems are mitigated, when analysts present the form of consumer perspective suggested here, in which the premium and out-of-pocket costs of consumers purchasing idealized insurance are the basis for direct cost measurement. Producer benefits also matter, but CE analysis does not offer a comprehensive framework for evaluating them, particularly in a dynamic context. Thus, this perspective is both meaningful and understandable, and is the appropriate perspective for many government agencies, private payers, and providers making decisions about health care.

5. Measuring outcomes

According to the preceding discussion, the welfare economic foundations of CE analysis rest upon the validity of the outcome measure as a representation of utility. This aim was not explicit in the initial development of outcome measures for CE analyses in health care. Whether the purpose of the CE analysis is to maximize utility or to maximize a global measure of health-related quality of life, however, its credibility depends heavily on the comprehensiveness and relevance of the health outcome measure. A highly specific outcome or effectiveness measure like the yield of abnormal test results or the magnitude of the blood pressure response to an antihypertensive drug may be understandable, persuasive, and sensitive to the effects of the intervention under study. But such a measure cannot be used to compare a diverse set of health interventions to be administered to patients with different health conditions. Furthermore, despite occasional claims and implicit assumptions to the contrary, only rarely will such a measure capture all the potential benefits and harms of an intervention. Thus, a comprehensive and general measure of health outcomes is of fundamental importance, whether the analysis is to be justified by appeal to welfare economics or by simple appeal to the inherent plausibility of the health measure.

It is for these reasons that QALYs are most frequently recommended as the outcome measure for CE analysis. More general alternatives, like healthy-year equivalents (HYEs) have attractive theoretical properties [Gafni and Birch (1997), Mehrez and Gafni (1989), Mehrez and Gafni (1993)] but have not gained widespread acceptance, probably because they are perceived as difficult to implement [Johannesson et al. (1993), Gold et al. (1996)].

The measurement of QALYs is the subject of the chapter by Dolan (2000) in this handbook. The following brief discussion emphasizes measurement of the preference weights q_i that appear in Equation (3).

5.1. Steps to measuring QALYs

Three components are needed to calculate an individual's utility at a point in time. First is the definition of the health state in question, which might be a particular disease with specific symptom severity; second is the utility attached to that health state, and third is the probability that the individual will be in that health state. By summing the products of the utilities of each possible health state and their probabilities, it is possible to obtain the expected utility (or QALY contribution) corresponding to the time period in question. This formulation has the advantage of breaking the task of calculating QALYs into manageable components: description of the health state; assessment of utilities toward the health state; and estimating the probability of the health state.

Defining and describing the health state are fundamental to modeling effectiveness. The CE analysis must include each state of health that the intervention might affect, either by preventing or treating illness, or by causing side-effects. Thus, if the intervention under study is surgery for the treatment of coronary artery disease, important health states to model include the presence and severity of angina pectoris, heart attacks, and other symptoms of heart disease or complications of the procedure (or, for that matter, of any alternatives to which it is compared). The scope of available data and analytical tractability limit the number of health states that can be modeled. Many analyses use

Markov modeling and related techniques to describe the progression over time of the probabilities of various health states, and if too many health states are included, there may be few or no transitions between infrequently occurring health states, precluding reliable estimation of some of the parameters of the model.

Dolan's chapter discusses how preference assessment is performed to estimate the utilities or quality weights specific to each health state. A critical issue for preference assessment is whether the respondent – the person whose preferences are being assessed – is asked to place a value on his own current or recent state of health, or is instead asked to place a value on a hypothetical state of health. For example, the preference questions could be directed toward people known to have a particular health state, such as moderately symptomatic coronary artery disease, and they could be asked how their current state of health compares to an ideal state of health. The alternative is to provide a description of a hypothetical state of health and to ask respondents to imagine themselves in that health state and to rate it.

There are several difficulties with rating one's own health state. First, the preferences of people experiencing a state of (usually chronic) ill health may differ systematically from the preferences of the general population. In the face of a disparity, there is no strict consensus about whose preferences should be used. The Panel on Cost-Effectiveness in Health and Medicine argued that when societal (i.e., governmental) resources are used to pay for health care, the preferences should be those of the general population rather than those of individuals with a health condition [Gold et al. (1996)]. Furthermore, it is difficult to study large samples of individuals who have a specific health condition, especially if the condition is uncommon. It is also possible that the disutility associated with a health state may reflect co-existing health conditions or risk factors that predispose to the disease rather than the disease itself. For example, high blood pressure is an asymptomatic condition that increases the risk of heart disease and stroke. People with high blood pressure rate their own health as relatively poor, even when they have not suffered any complications. Because treatment lowers the blood pressure but does not remedy associated health conditions, it does not improve quality of life as greatly as would be predicted from a model in which preferences are obtained from people with the disease and treatment is assumed to restore them to perfect health.

The validity of the alternative approach, *rating hypothetical health states*, is highly dependent on the accuracy and completeness of the description of the hypothetical state(s). The health state description is not critical for a state of health that most respondents have experienced, such as the symptoms of a viral upper respiratory infection or mild low back pain. But for a health state that few respondents have experienced themselves or vicariously through a relative or friend, nearly all the information that the respondent can bring to bear on the question must be provided in the description. This requirement can be an advantage, since it is easier to control the impression that naive respondents have of the health state than the impressions of experienced respondents. But it also means that small and seemingly inconsequential changes in the presentation of the health state can greatly influence the utilities assigned to it. To enhance the reproducibility and validity of ratings of hypothetical states, it is essential to pay close attention to the wording and general design of such elicitations.

5.2. Estimating survival and probabilities of health states

Even for interventions that do not alter the length of life, it is usually necessary to describe patterns of survival since these patterns determine the changes in QALYs that result from use of the interventions. Many treatments, of course, are designed to prevent death, so estimation of survival effects, or the survival probabilities in Equations (2) and (3), is a key component of most CE analyses.

Approaches to measuring survival probabilities vary greatly. Survival estimates nearly always require an element of modeling, since experimental data (from a randomized trial) are usually limited to brief (less than five years) follow-up periods. To estimate the effect on life expectancy, it is necessary to combine such data with observational data about longer-term outcomes in typical practice settings.

The techniques for estimating the pattern of survival associated with an intervention vary. One study of a treatment for heart attacks shows how clinical trial and observational data can be combined to estimate long-term outcomes. Researchers from the GUSTO trial, a study of tissue-type plasminogen activator (t-PA), a drug used to dissolve the blood clots that can cause obstructions in the coronary arteries and precipitate heart attacks, sought to determine the long-term survival benefit by supplementing direct clinical trial data, obtained during an average of 12 months of follow-up, with a model of survival based on an observational database (the Duke Cardiovascular Registry), and a parametric survival function for extrapolating beyond the 14 years of data represented in the observational database. Figure 5 displays the resulting survival curve. Published CE analyses have used a variety of other methods. Some analyses used life table data for either the general population or, where available, for patients who have a specific health condition, and applied a relative risk reduction as estimated in a clinical trial, imposing the assumption that the relative risk reduction is constant across different populations and ages.

By generalizing the methods for estimating survival, one can also estimate probabilities that various states of health will occur in the future, under either the treatment or the intervention. Usually Markov-like modeling offers the most convenient approach to estimating future probabilities of health states. One such approach estimates first the probability that an individual receiving the intervention is alive, say, two years in the future, then uses data from clinical trials or other sources to estimate the probability that, if alive, the patient will be in a symptomatic state of ill-health, and the probability that he or she will be in excellent health. Typically availability of data on rates of adverse events (such as onset or progression of disease, death rates, and morbidity), rather than technical issues (such as the formal structure of the model to depict disease advancement), limits the estimation of probabilities of health states.

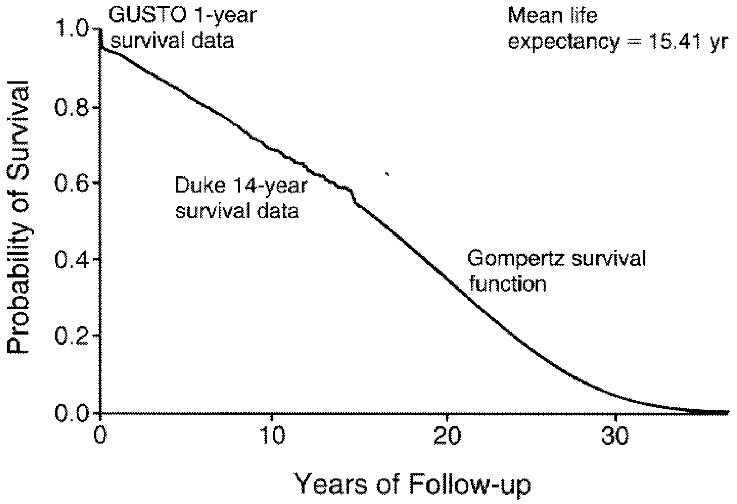


Figure 5. Probability of survival among patients treated with t-PA. A survival function of this type was used to estimate life expectancy for each treatment group. The curve consists of three parts: the survival pattern in the first year after treatment in the GUSTO study, data for an additional 14 years on survivors of myocardial infarction in the Duke Cardiovascular Disease Database, and a Gompertz parametric survival function adjusted to agree with the empirical survival data at the 10-year and 15-year follow-up points.

Source: Mark et al. (1995).

5.3. Preference assessment

The remaining step in calculating QALYs is to assign utilities, or preference weights, to each of the health states. Several reviews describe and compare alternative methods for preference assessment, and Dolan (2000) discusses the topic extensively in this Handbook. Dolan reviews a wide range of issues in assessing preferences and in their interpretation from the point of view of QALY calculation. As his discussion of the methodological issues in assigning utilities to health states implies, preference assessment is sometimes a source of considerable uncertainty in CE analyses. The most reproducible methods of preference assessment, such as the visual analog scale, are not derived from von Neumann–Morgenstern utility theory. Methods that are more firmly grounded in utility theory, such as the standard gamble, are neither perfectly general nor easy for respondents to understand.

Since the validity of CE analysis as a guide to welfare maximization rests upon the validity of QALYs as a measure of utility, the conditions that preference assessment needs to meet are stringent. Usually discussions of quality of life for use in CE analysis emphasize that the measurement should be of *health-related* quality of life. Well known preference-weighted health status indices used to attach utilities to health states – such as the Health Utilities Index of Torrance and colleagues, the Quality of Well-Being scale developed by Kaplan and colleagues, and the Rosser scale – omit mention of non-health consumption and financial status [for an extended discussion of these and other scales, see the book by Patrick and Erickson (1993)]. According to some experts, respondents should be asked to ignore effects of states of ill health on income and other financial repercussions. Yet the plausibility of QALYs as measures of utility depends on the ability to represent fully the changes in well-being that occur with the adoption of an intervention, and often these changes will not be limited to those that are primarily health related. Such concerns may be of little importance if the only financial consequence is loss of earned income, which ordinarily would be incorporated into the numerator of the CE ratio. But if a health state causes alteration of non-health consumption, which is not reflected in the preference assessment procedure (e.g., development of severe arthritis may necessitate changes in clothing, furniture, and use of various non-health services), the adverse effects of the health state will be underestimated.

5.4. Preference heterogeneity and its consequences for CE analysis

Perhaps the greatest practical challenge to the use of QALYs to represent utilities is the variation in preferences that is all but certain to occur in the context of specific health limitations. Just as demand for any good or service varies, so do preferences for states of health. A well-known study of treatment of benign prostatic hyperplasia, which causes a variety of urinary symptoms, demonstrated that variation in attitudes toward specific health limitations can dramatically alter the value of treatment. The most common surgical treatment of prostatic disease is transurethral resection of the prostate, an operation that can be highly effective at relieving the excessive urinary frequency and nocturia (awakening at night to void) and other symptoms that men with prostatic obstruction experience. The operation, however, can cause incontinence, impotence, and other side-effects, some of them permanent. Men who are candidates for surgery vary greatly in their relative preferences for the symptoms of prostatic hyperplasia and the side-effects of the operation, so that the expected quality of life is greater with surgery for some and with nonsurgical management for others [Barry et al. (1988), Fowler et al. (1988)].

Without even considering costs, then, the "best" treatment varies when preferences vary. When CE is used as a criterion for determining the allocation of interventions, preference variation often poses more significant problems. It is possible that every patient who is a candidate for treatment with a particular intervention will gain QALYs from it. But the intervention is much more cost-effective in those patients who experience the greatest disutility from the disease being treated, and who lose little utility from the side-effects of treatment. Other patients who have identical health characteristics may experience little disutility from the disease and more from the treatment. It is very hard for any health care delivery or financing system to distinguish these two types of patients, both of whom would desire the intervention. Although individual clinical decisions can take such heterogeneity into account, even in the physician's office the necessary information, and the ability to use it, may be limited.

5.5. QALY measurement and the application of CE analysis

Technical issues in QALY measurement raise questions about the reliability and validity of QALYs, as usually calculated, as measures of utility. One message from the literature that uses weights based on preferences rather than statistical weights or simple sums to measure quality of life is that comprehensive measures of utility are difficult for study subjects to understand. The reproducibility of such measures, particularly when the underlying preference assessment technique is as complicated as the standard gamble, is often disappointing. The limitations of such measures are partly responsible for the popularity of quality of life measures that are not preference weighted (such as the Rand Corporation's SF-36 scale) or that are not even global measures of quality of life (such as disease-specific quality of life scales). Although these alternative measures offer apparent practical advantages, rarely can they be considered reasonable proxy measures of utility. The major conceptual problem with the preference assessment measures as usually applied is that they do not allow the state of being to be construed broadly enough, a problem that is far worse for disease-specific measures. Measures that are not preference weighted lack the interval scaling properties required for the tradeoff between length and quality of life implicit in QALYs.

The practical problems are particularly great when the benefit from a health intervention is small. Consider, for example, a medicated lotion that relieves the itch of a rash that appears on the arms and back. Even if the lotion completely relieves the rash as soon as it is applied, it will be extremely difficult to assess utilities for the relief of the rash using standard preference assessment techniques. All of the techniques require a tradeoff between a risk of death and symptom relief, but if the symptoms are mild or their duration is brief enough, it is difficult for respondents to estimate the risk of death (or for the time-tradeoff method, the reduction in the length of life) that they would tolerate for an improvement in the symptoms. For this intervention and others that produce small or brief improvements in quality of life, the willingness-to-pay approach used in CB analysis would likely offer a much more suitable approach to valuation.

An ideal measure of health outcomes would be less restrictive than QALYs, abandoning the additive separability embedded in the functional form and the (usually) constant rate of time preference, but preference assessment instruments capable of supporting more general models would impose upon respondents even greater cognitive burdens than current methods. Research on these methods remains active, in some cases reflecting the great interest of governments in applying CE analysis to health care decisions more extensively. As utility measurement improves, claims that the results of CE analysis can be applied to maximize social welfare can be made with greater confidence. Furthermore, although the QALY is not perfectly general as a measure of well-being, it is likely to be a close approximation to more general measures and to represent an acceptable tradeoff between conceptual validity and feasibility. Unlike many competing measures of quality of life, such as the statistically-weighted quality of life indices, QALYs are conceptually appropriate and have the potential to approach the theoretical ideal when preference assessment techniques are developed further.

6. Recommendations

A fundamental but often unstated characteristic of any CE analysis is its purpose. Is that purpose to enable an insurer, a health plan, or a government agency to decide whether to cover a specific intervention? Is it to help a consumer decide which form of treatment to receive? Is it to help a manager make decisions about large investments in health care infrastructure? Is it to help a formulary committee choose which of several drugs should be available in a hospital pharmacy? Or is it to help a decision maker determine the allocation of health care that will achieve a suitably defined social optimum, regardless of who that decision maker is?

Most experts in CE analysis argue that, unless there are compelling reasons to do otherwise, CE analyses should be conducted from the societal perspective. Under this perspective, all costs and all benefits are relevant, but usually analysts assume that the health benefits accrue entirely to the individual receiving care. Exceptions are sometimes made in other circumstances, such as when there are significant externalities. For example, family members may provide care or other people may bear a cost when an individual is injured or ill. Even in the absence of externalities, though, an attempt to use CE analysis to determine a full societal optimum, while laudable, in important circumstances may stretch the technique to the breaking point. Even for a circumscribed measure of optimality like the Kaldor–Hicks criterion (i.e, potential Pareto improvement), such determinations may be difficult for products characterized by economies of scale and by other failures of the assumptions of perfectly competitive markets. How and whether to include the preference of producers in a CE analysis are certain to be controversial, particularly when the profits accrue in a population markedly different from the one that is being treated. Profits are certainly a component of overall welfare, and to remove them from the CE analysis is not the same as saying that they are unimportant. CE analysis does not provide a comprehensive framework for including them.

As common practice dictates, and the abilities of the technique mandate, most CE analyses should be conducted from a consumer-oriented perspective, but not from the one that is generally described as the consumer's or patient's perspective. Rather, the most robust perspective is that of an insurer acting as a perfect agent for its enrollees. Specifically, it assumes that the members of the defined population are behind a "veil of ignorance," having no particular information to distinguish their risk of developing any disease or health condition or desire to utilize services from the average for the defined population. The insurer charges an actuarially fair premium, and has no costs other than the payment of benefits. There are no informational failures of consequence, other than symmetric uncertainty, in the sense that neither the insurer nor any individual has more or less information than others.

Perhaps the most difficult challenge for the implementation of CE analysis is the technique's application in heterogeneous populations. The optimality properties of the CE approach are based upon the application of an individual's specific CE ratio cutoff to decisions about care. For that individual, any intervention whose CE ratio is below the cutoff is welfare-enhancing (i.e., passes a CB criterion), whereas any with a greater CE ratio does not. But for many reasons – income, risk preferences, and various other attitudes and values – CE cutoffs vary greatly across individuals. Many, if not most, CE analyses are used to inform decisions made at a group level yet implicitly apply a single cutoff. Decisions based on a single cutoff cannot claim to have the same optimality properties in a heterogeneous population. The cutoff will be greater than the actual cutoff for some people, and less than the actual cutoff for others. Furthermore, the optimal single cutoff for a heterogeneous population would not necessarily correspond to the average valuation.

The preceding discussion suggests that the welfare implications of the application of CE analysis are clearest when strong conditions are met. The research challenges include better measurement – for example of health outcomes, preferences, and costs – and further investigation into the implications of using CE analysis when ideal conditions do not apply. The measurement of preferences is an area of ongoing research, and it would be helpful to compare the results of analyses that use QALYs with those that use either simpler measures of health outcomes (e.g., life expectancy) or more comprehensive measures (e.g., healthy year equivalents). Further investigation of the theoretical issues would help to clarify the meaning and generalizability of the results of CE analyses. For example, what are the welfare implications of prioritization based on CE ratios when some health services are subsidized but a number of substitutes for them are not? What are the implications of inter-individual variation in rates of time preference? What are the welfare gains from using individual rather than uniform CE cutoffs in heterogeneous populations? Under what circumstances are simple CE analyses accurate guides to welfare maximization?

CE analysis can be a useful aid to decision making in health care. In specific circumstances it can be quite powerful. Yet its grounding in welfare economics has often been implicit, and an explicit examination of how one can use a CE criterion to achieve a potential Pareto improvement demonstrates that the necessary conditions are exacting. Nevertheless, of widely accepted, existing methods for incorporating economic considerations in the prioritization and allocation of health care, CE analysis is probably the most rigorous. Exploration of its welfare economic foundations has the additional advantage of helping to resolve ambiguities in matters such as the measurement of costs, and can help to inform the development of new instruments for measuring quality-of-life effects. CE analysis is not a perfect tool, but in many situations, it may be good enough.

References

Barry, M.J., A.G. Mulley Jr., F.J. Fowler and J.E. Wennberg (1988), "Watchful waiting vs immediate transurethral resection for symptomatic prostatism. The importance of patients' preferences", Journal of the American Medical Association 259:3010–3017.

Birch, S., and A. Gafni (1994), "Cost-effectiveness ratios: in a league of their own", Health Policy 28:133–141.

Briggs, A., and M. Sculpher (1995), “Sensitivity analysis in economic evaluation: a review of published studies”, Health Economics 4:355–371.

Briggs, A., M. Sculpher and M. Buxton (1994), “Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis”, Health Economics 3:95–104.

Dolan, P. (2000), “The measurement of health-related quality of life for use in resource allocation decisions in health care”, in: J.P. Newhouse and A.J. Culyer, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 32.

Drummond, M.F., B. O’Brien, G.L. Stoddart and G.W. Torrance (1997), Methods for the Economic Evaluation of Health Care Programmes (Oxford University Press, New York).

Drummond, M.F., G.L. Stoddart and G.W. Torrance (1987), Methods for the Economic Evaluation of Health Care Programmes (Oxford University Press, Oxford).

Finkler, S.A. (1982), “The distinction between cost and charges”, Annals of Internal Medicine 96:102–109.

Fowler Jr., F.J., J.E. Wennberg, R.P. Timothy, M.J. Barry, A.G. Mulley Jr. and D. Hanley (1988), “Symptom status and quality of life following prostatectomy”, Journal of the American Medical Association 259:3018–3022.

Gafni, A., and S. Birch (1997), “QALYs and HYEs (healthy years equivalent). Spotting the differences”, Journal of Health Economics 16:601–608.

Garber, A.M., A.E. Clarke, D.P. Goldman and M.E. Gluck (1992), Federal and Private Roles in the Development and Provision of Alglucerase Therapy for Gaucher Disease (Congress of the United States, Office of Technology Assessment, Washington, DC).

Garber, A.M., and C.E. Phelps (1997), “Economic foundations of cost-effectiveness analysis”, Journal of Health Economics 16:1–31.

Garber, A.M., and N.A. Solomon (1999), “Cost-effectiveness of alternative test strategies for the diagnosis of coronary artery disease”, Annals of Internal Medicine 130:719–728.

Gold, M.R., J.E. Siegel, L.B. Russell and M.C. Weinstein, eds. (1996), Cost-Effectiveness in Health and Medicine (Oxford University Press, New York).

Hicks, J.R. (1939), “Foundations of welfare economics”, Economic Journal 49:696–712.

Hurley, J. (2000), “An overview of the normative economics of the health sector”, in: J.P. Newhouse and A.J. Culyer, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 2.

Johannesson, M., J.S. Pliskin and M.C. Weinstein (1993), “Are healthy-years equivalents an improvement over quality-adjusted life years?”, Medical Decision Making 13:281–286.

Johannesson, M., and M.C. Weinstein (1993), “On the decision rules of cost-effectiveness analysis”, Journal of Health Economics 12:459–467.

Kaldor, N. (1939), “Welfare propositions of economics and interpersonal comparisons of utility”, Economic Journal 49:549–552.

Karlsson, G., and M. Johannesson (1996), “The decision rules of cost-effectiveness analysis”, PharmacoEconomics 9:113–120.

Keeney, R.L., and H. Raiffa (1993), Decisions with Multiple Objectives (Cambridge University Press, Cambridge).

Kenkel, D.S. (1997), “On valuing morbidity, cost-effectiveness analysis, and being rude”, Journal of Health Economics 16:749–757.

Luce, B.R., W.G. Manning, J.E. Siegel and J. Lipscomb (1996), “Estimating costs in cost-effectiveness analysis”, in: M.R. Gold, J.E. Siegel, L.B. Russell and M.C. Weinstein, eds., Cost-Effectiveness in Health and Medicine (Oxford University Press, New York) 176–213.

Mark, D.B., M.A. Hlatky, R.M. Califf, C.D. Naylor, K.L. Lee, P.W. Armstrong, G. Barbash, H. White, M.L. Simoons and C.L. Nelson (1995), “Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction”, New England Journal of Medicine 332:1418–1424.

Mehrez, A., and A. Gafni (1989), “Quality-adjusted life years, utility theory, and healthy-years equivalents”, Medical Decision Making 9:142–149.

Mehrez, A., and A. Gafni (1993), “Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress”, Medical Decision Making 13:287–292.

Meltzer, D. (1997), “Accounting for future costs in medical cost-effectiveness analysis”, Journal of Health Economics 16:33–64.

Mullahy, J., and W.G. Manning (1994), “Statistical issues in cost-effectiveness analysis”, in: F. Sloan, ed., Valuing Health Care: Costs, Benefits and Effectiveness of Pharmaceuticals and Other Medical Technologies (Cambridge University Press, New York).

O’Brien, B.J., M.F. Drummond, R.J. Labelle and A. Willan (1994), “In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care”, Medical Care 32:150–163.

Patrick, D.L., and P. Erickson (1993), Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation (Oxford University Press, New York).

Pauly, M.V. (1968), “The economics of moral hazard: comment”, American Economic Review 58:531–537.

Pauly, M.V. (1995), “Valuing health care benefits in money terms”, in: F. Sloan, ed., Valuing Health Care: Costs, Benefits, and Effectiveness of Medical Technologies (Cambridge University Press, Cambridge) 99–124.

Phelps, C.E., and A.I. Mushlin (1991), “On the (near) equivalence of cost effectiveness and cost benefit analysis”, International Journal of Technology Assessment in Health Care 7:12–21.

Rawls, J. (1971), A Theory of Justice (Harvard University Press, Cambridge, MA).

Russell, L.L. (1986), Is Prevention Better than Cure? (The Brookings Institution, Washington, DC).

Schelling, T.C. (1968), “The life you save may be your own”, Problems in Public Expenditure Analysis (The Brookings Institution, Washington, DC) 127–159.

Siegel, J.E., M.C. Weinstein and G.W. Torrance (1996), “Reporting cost-effectiveness studies and results”, in: M.R. Gold, J.E. Siegel, L.B. Russell and M.C. Weinstein, eds., Cost-Effectiveness in Health and Medicine (Oxford University Press, New York) 276–303.

US Congress Office of Technology Assessment (1980), The Implications of Cost-Effectiveness Analysis of Medical Technology (US Government Printing Office, Washington, DC).

US Congress Office of Technology Assessment (1993), Pharmaceutical R&D: Costs, Risks and Rewards, OTA-H-522 (US Government Printing Office, Washington, DC).

Wakker, P., and M.P. Klaassen (1995), “Confidence intervals for cost/effectiveness ratios”, Health Economics 4:373–381.

Weinstein, M.C., and W.B. Stason (1977), “Foundations of cost-effectiveness analysis for health and medical practices”, New England Journal of Medicine 296:716–721.

This Page Intentionally Left Blank

Chapter 5

INFORMATION DIFFUSION AND BEST PRACTICE ADOPTION*

CHARLES E. PHELPS

University of Rochester

Contents

*Much of the research supporting this work was carried out with the support of Agency for Health Care Policy and Research Grant Number R01-HS06366. I wish to acknowledge the particular importance of Cathleen Mooney's collaboration in much of that research, especially in the work involving physician practice variations. My thanks for valuable advice also to Joseph Newhouse, Mark Satterthwaite, Tomas Philipson, Marty Gaynor, and Sherry Glied, and other participants in a seminar at the University of Chicago where this and other chapters in this Handbook were presented in preliminary form.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

Incomplete information issues pervade health care markets, with market participants often having relatively little information, and their behavior exhibiting corresponding aberrations from classic market behavior.

Consumers often have relatively little information about prices and quality offered in health care markets, leading to substantial dispersion in prices of apparently identical services. Equilibrium price dispersion increases as the demand elasticity for the product falls. Since health insurance lowers the elasticity of demand, price dispersions should occur more often (and with greater magnitude) in markets such as physician services with relatively complete insurance. Further, many insurance plans blunt incentives for search, compounding the problem.

On the supply side, evidence shows that physicians behave as if they did not share the same information about the productivity of medical care. At the level of geographic regions, numerous studies show the rates at which various medical interventions are used on standardized populations differ hugely – often by an order of magnitude or more from high to low – and these differences in treatment rates do not converge through time as would occur in standard market learning models.

Similarly, individual physicians within a given region also display differences in the propensity to use medical resources. Information from a major study of doctors' "styles" shows large and statistically significant differences in doctors' use of medical resources to treat their patients, even with strong measures of illness severity of the patients included in the models.

Although requiring strong assumptions, one can estimate the welfare losses arising from incomplete information on the provider side of the market. Estimates of the upper bound of these welfare losses place the magnitude of loss in the same range on a per capita basis as the traditionally emphasized welfare losses associated with perverse incentives in health insurance.

The importance of incomplete information leads to discussions of the economic and legal incentives for the production and dissemination of information. Legal incentives to produce such information for medical strategies (treatment protocols) are weak, particularly compared with the incentives in markets for specific products such as prescription drugs. The public good nature of such information and the government role in supporting its production and dissemination form the concluding parts of this chapter.

Keywords

doctor, dissemination, incomplete information, insurance, patient, production of information, property rights, search, variations, welfare

JEL classification: I10, I11, D80, D82, D83

1. Introduction

Information is not free, either in its production, dissemination or acquisition. Yet the "normal" economic model presumes that consumers are fully informed about prices and quality of every good available in the market, and that all producers have complete knowledge of available production technologies. Few people take this model literally, but most economists use it often in analyzing market behavior, in part because the world seems to behave "as if" these assumptions hold in many settings. In other cases, however, the world behaves quite differently, and a more complete understanding of the behavior of market participants (and the outcomes their behavior leads to) requires explicit consideration of the role of production and diffusion of information.

In health care markets, this holds perhaps more than in any other setting.[1] The "footprints" of incomplete information can be found everywhere in health care markets:

- Governments invest massive resources in the production of information about health and medicine, more so than in any other area (perhaps save national defense matters).
- Prices vary substantially within the same geographic area for apparently identical services.
- An individual's chances of receiving a particular medical intervention (e.g., cardiac surgery, carpal tunnel surgery, knee replacement, hospitalization for pneumonia, etc.) can vary by an order of magnitude, depending on where the person lives.
- Individual physicians' resource use to treat an apparently similar groups of patients differs by more than a factor of two within the same market area.
- Governments interfere in market operations in health care more than perhaps any other market, ostensibly to protect consumers against low quality (licensure, drug regulation).

These matters cannot arise simply through the presence of uninformed consumers. Some, such as price variation for identical products, follow directly from models of market equilibrium with incomplete consumer search, as we shall see below. Others, such as the large cross-regional differences in rates of treating patients with particular interventions, almost certainly require major differences in the information set held by health care providers.

One could rightly ask if (and if so, why) these issues are special to the study of health economics. It is obvious that at least some of the problems of incomplete information pervade many markets, and indeed, importantly affect how those markets function. In health care, however, the issues are more pervasive, and some special features of health care markets exacerbate the problem. First and foremost on the consumer search problem, insurance blunts and sometimes completely removes the incentives for consumers

[1] Arrow (1963) provided the essential guidance for this idea in his landmark article, "Uncertainty and the Welfare Economics of Medical Care," in which he set forth many of the propositions we are still just beginning to understand a third of a century later.

to carry out search for lower prices, and other insurance arrangements (e.g., preferred provider arrangements) inhibit search on quality (by limiting the panel of physicians for which the insurance covers services rendered). This does not in general hold even in other markets where insurance is important (e.g., auto repair) because the form of compensation from the insurance contract that so strongly affects search incentives is unique to health insurance.

Second, there are widespread limitations on advertising that restrict the ability of suppliers to transmit information about product quality or price. Benham (1972) used national household survey data to measure the prices of eyeglasses purchased by consumers in states that prohibited advertising by optometrists vs. those where such advertising was permitted. He found that the prohibition of advertising increased eyeglass prices by 25% to more than 100%, depending on the comparisons chosen. Subsequent analysis by Cady (1976) for prescription drugs and by Feldman and Begun (1978) and Kwoka (1984) for optometric examinations shows similar results. Thus, advertising restrictions (more severe in general than in most other markets) inhibit search.

Third (although not unique to health care), western systems of property rights do not in general protect process innovation well, so when doctors learn how to treat patients better, they have little way to reap the benefits of the innovation. Large manufacturing firms and even service delivery "chains" such as McDonalds confront the same problem, but they have internal mechanisms through which they can exploit the gains from process innovation. Small physician offices and, for many processes, even single hospitals don't have such a capability. Thus the incentives to invest in process improvement (better treatment regimen) is minimal. (The issues relating to larger managed care "chains" will be explored in Section 4.2.)

It is also important to understand that the issues about search, while usually couched in terms of price, will (in much of what follows) have symmetric issues in consumer search for information about quality. Indeed, quality search may be more difficult and expensive than price search for most consumers. Effective search requires at least a sampling of the market on the parameter of interest. However, sampling provider quality is expensive, and to the extent that quality must be inferred on the basis of the particular doctor/patient pairing, may require actually changing doctors. While search on price has become less important in markets where it once mattered (e.g., physician markets in the US) due to the major market penetration of "managed care" insurance plans that negotiate the price for the consumer, search on quality remains distinctly important, and indeed, may have become more difficult in some managed care settings. For example, "gatekeeper" arrangements that require a patient to seek care first from a primary care doctor before getting a referral to a specialist almost certainly inhibit search on quality, at least for specialists, since the gatekeeper will seldom be allowed to make more than one speciality referral for a single patient. Of course, one of the alleged advantages of gatekeeper models is that the primary care gatekeeper can assimilate quality information on behalf of patients and make appropriate referrals as a more informed agent. The net effect of these on the level and distribution of quality is unknown at this point.

Proceeding now to the body of this chapter, it divides into three distinct parts. Section 2 analyzes the functioning of markets with incomplete search, building upon work by Wilde and colleagues [Wilde and Schwartz (1979), Schwartz and Wilde (1982), and most importantly, Sadanand and Wilde (1982)], who offer a model that allows direct consideration of the role of health insurance on search in medical care markets. Some further insights from Dionne (1984) extend this discussion. This section includes information about actual price dispersion within single markets for apparently similar products.

The third section analyzes the issue of variable patterns of treatment of patients in settings where neither illness differences nor traditional economic phenomena can explain the variable patterns of behavior. The same problem emerges when one studies either the cross-regional variations in rates of the use of specific medical interventions or the patterns of care rendered by individual physicians with the same community. This section also provides a model to measure the welfare consequences of such variations in medical care use, or at least (in a less direct manner) to focus attention on the areas of medical treatment where better information would have the greatest incremental value.

Sections 2 and 3 interact in the following way: variations in medical practice imply that some patients (at least some – perhaps all) are not getting the proper amount of treatment – either too much or too little – given their medical condition (and their preferences and income). These variations, however, cannot sustain themselves in a market where consumers have perfect information about provider quality. Complete search (or actually, less than complete search) by consumers who could measure quality meaningfully would drive all doctors to the same treatment patterns for a given type of patient (including patient preferences in characterizing patients). In contrast, the data show that the chances of a particular type of patient (but without measuring patient preferences) receiving a particular intervention (e.g., 40 year old female office worker with $30,000 income receiving carpal tunnel surgery) can readily vary by a factor of 4 or 5 or more, depending on where the patient lives and the doctor who advises the patient. Thus, the variations discussed in Section 3 surely exist only because of informational problems such as discussed in Section 2.

Finally, Section 4 analyzes the production and dissemination of information in health care and health sciences, focusing on both the traditional "public good" model of production of basic research and also investigating the legal and economic considerations regarding private production and dissemination of information. These economic and legal considerations provide an improved understanding of why some of the phenomena arise that were discussed in previous sections. This section concludes with some conjectures about market and regulatory interventions that would improve welfare by altering conditions in the market for information in health care, all of which could form the basis for a future research agenda for scholars in this field.

2. Market equilibrium and price variability

2.1. Search and market equilibrium

In many markets, price variability has been observed and well documented, but the assumption almost automatically follows from standard economic thinking that *systematic* differences in price of "the same good" necessarily signal differences in product quality, terms of trade, convenient location of the vendor, or some similar dimension of quality. Indeed, so firmly ingrained is this idea in economists' thinking that they often tautologically assume quality differences must exist to account for observed price differentials. Yet price differentials for apparently identical products are quite common. Pratt, Wise and Zeckhauser (1979) found substantial price dispersion in a variety of markets for standard consumer products. Indeed, even in an environment in which consumers would appear to have very low incremental costs for search, it is easy to find widespread price dispersion.[2]

A number of economic models exist to account for price variability. Stigler (1961) first called attention to this problem with his landmark essay on "The Economics of Information." Since then, a number of other approaches have followed [Salop and Stiglitz (1977), Braverman (1980)], often variations on Stigler's original insight: Information is costly, and efficient search for a lower price will lead to a stopping rule that limits search to something less than a complete sampling of the market. In these models, the extent of search (the optimal stopping rule) is usually driven by a combination of the *a priori* distribution of prices and the cost of search [see Lippman and McCall (1976, 1979), Hey (1979, 1981), Hey and McKenna (1981), McKenna (1986) for further discussion]. Rochaix (1989) adds valuable analysis of health care markets and search.

2.2. Search in health care markets

In health care markets, even these basic approaches fail, because health insurance modifies or sometimes eliminates the incentives for search. Consider first the most simple form of health insurance, in which the consumer is reimbursed for k percent of all expenses. (This has been a very common form of health insurance coverage, e.g., for "major medical insurance" in the US and elsewhere.) This type of insurance has the effect of rotating the consumer's demand curve vertically around the quantity intercept [Phelps (1997, Chapter 4)]. It is well known that this type of insurance both increases

[2] Here is a personal experiment for the reader to conduct: Log onto the World Wide Web and search on several consumer products about which you are knowledgeable to find relevant prices. Select products for which the brand and model are clear and specific. You will likely find significant variability in the price even among firms advertising on the Web, let alone in other media such as catalogs. As an example, I searched recently for Motorola Talkabout Plus hand held personal radios ("walkie talkies" for personal use). Contemporaneous catalog prices ranged from $159 to $139. Prices offered on the Web ranged from a high of $149 to a low of $99. (In all cases, taxes and shipping were extra.)

the quantity demanded and also reduces the consumer's demand elasticity from its uninsured level of η to $C\eta$ where $C = (1 - k)$ is the consumer's coinsurance rate [Phelps and Newhouse (1974)]. The economic desirability of such insurance arises because the post-insurance variance in out of pocket expenditures falls, to a first approximation, from σ^2 to $C^2\sigma^2$, and the risk premium [Pratt (1964)] falls commensurately.[3]

This type of insurance also reduces the consumer's incentive to search. If the returns to search (in terms of finding a lower priced provider) are ρ in a world with no insurance, they become $C\rho$ when the coinsurance rate is C. Since typical major medical insurance policies have parameters like $C = 0.2$, it is easy to understand that the incentives to search fall markedly for consumers with such insurance policies.

Alternative types of insurance contracts produce different effects. An early type of health insurance, now seldom observed, paid $X for specified events (e.g., $20 for the purchase of a medical office visit, $500 for a simple hernia repair surgery, etc.) These types of insurance policies introduce a new type of risk on consumers arising from price variability, but they also obviously completely preserve the economic incentives to search for a lower price. A more common modern insurance policy requires the consumer to pay $M per event (e.g., $5 to $15 for a physician office visit), at which point the insurance policy pays all remaining charges.[4] Endless variations are obviously possible, but the key point to observe is that consumers' incentives to search for lower prices are often blunted or possibly eliminated by modern health insurance arrangements.

Only on rare occasions has there been any actual measurement of the propensity of medical consumers to search. The earliest of these occurred in England and Wales, studying the behavior of patients in the British National Health Service (BNHS). There, price is obviously not a factor since all such care was then free; this study only looked at patient satisfaction. In that setting, 0.7% of patients had changed doctor in the past year for reasons of dissatisfaction [Gray and Cartwright (1953)]. In 1962, a US study found 30% of patients had changed doctors within the past five years, but only 8% for reasons of dissatisfaction, or about 1.6% per year, about double the BNHS rate [Cahal (1962)]. More recently, a study in Utah [Olsen, Kane and Kasteler (1976)] measured not only the rates at which patients changed physicians, but also recorded the primary reason given by the patient for so doing. Well over half of the patients reported that they had changed doctors at some time in the past. Of those who did change, only 9 percent did so because of price, and the propensity to do so was (unsurprisingly) inversely related to income

[3] The risk premium is approximately $-r\sigma^2$, where $r = U''/U'$, the ratio of the second to the first derivatives of the utility function with respect to income. This is in fact the second order Taylor Series expansion of a general utility function, but an approximation that is quite close for many risky distributions.

[4] These types of copayments are very common in modern "HMO" types of insurance plans in the US. These types of policies obliterate any incentives to search, but this is seldom a real economic issue in modern "managed care" settings since the consumer's choice of provider may be limited by contractual agreement, and the insurer likely has negotiated common prices from all eligible providers. In the managed care environment, of course, the insurer has likely negotiated a price with the provider anyway, so search on price is less relevant.

– low socioeconomic status (SES) patients cited cost as the reason for changing doctor at three times the rate of high SES patients. The low rate at which consumers of health care shop for reasons of price is easily explained by the low incentives associated with health insurance coverage, as discussed previously.

2.3. *Incomplete information models*

2.3.1. *A general model of incomplete search*

Several models of incomplete search have been published, mostly exhibiting a common theme: Patients have an *a priori* distribution of prices in the market, a cost of search (possibly varying with opportunity cost of time, external rules affecting the availability of information and advertising, and other related factors). Given that *a priori* distribution of prices, optimal search processes lead to selection of a stopping rule. Consumers search until a reservation price is found, and purchase the good. In many models of search, the assumed costs of search and *a priori* price distributions determine the stopping rule. A nearly equivalent approach, often more tractable mathematically, is simply to assume a fixed number of searches, after which the lowest price found determines the purchase.[5]

One elegant model of this process [Sadanand and Wilde (1982)] using this formulation helps illuminate some key issues in search in health care markets. This model generalizes from previous studies that had assumed discrete demand curves and allows the quantity demanded to depend directly on the lowest price found during search. The Sadanand and Wilde model is summarized next to provide a framework for discussing these issues in more detail.

In this model, the market consists of two types of buyers, A_1 (who randomly select one firm and buy without comparison shopping) and A_n (who search $n \geqslant 2$ times and then buy from the lowest price provider found), occurring in proportions $(1 - \sigma)$ and σ respectively. All firms have the same cost structure, with U-shaped variable cost curves and fixed costs F.

Total average costs are thus $A(q) = [T(q) + F]/q$, which are minimized at output level s. (Thus $A(s) = \min(A(q))$.) The market equilibrium is monopolistically competitive, so expected profits are zero, and the usual monopolistic competition equilibrium occurs where the expected demand curve is just tangent to $A(q)$ along the left-hand portion. Consumer demand follows the usual assumptions: $q = f(p)$, $f'(p) < 0$. Another key parameter is the average number of consumers per firm, α.

The market works in the following way: Firms face an expected demand curve consisting of their proportion of non-shoppers, assumed to distribute themselves uniformly across all the firms in the market (for whom they can charge a higher price) and those

[5] Louis Wilde pointed this out to me in personal communication about work he had done, to be discussed next.

shoppers who, upon shopping, found no lower price. Thus the firm raises profits to non-shoppers by raising prices (but limited by the demand curve eventually) but risks losing sales to shoppers who find a lower price. This demand curve facing any specific firm always slopes downward because non-shoppers will each consume more at a lower price, and shoppers are attracted when they find no lower prices (and hence also add to the demand at lower prices). But profits are not necessarily maximized by lower prices, and this clearly depends upon the extent of shopping in the market. If nobody shops, then each firm faces a scaled-down version of the market demand curve and they price above minimum average cost. If a sufficiently large fraction of the consumers shop, then the only successful strategy is to price at the competitive (minimum average cost) level, since any other price loses too many consumers for the firm to sustain itself. Intermediate degrees of shopping lead to intermediate outcomes. Firms can enter this market whenever the expected demand curve confronting any firm lies above the *AC* curve at some quantity. The condition for this is that $A(\alpha f(p)) \leqslant p$ for some consumer/firm ratio α.

This structure produces three general types of outcomes. When search occurs least often, a distribution of prices arises with no mass point anywhere. At the other extreme, with sufficient amounts of search, the distribution of prices collapses to a single mass point at minimum average cost. An intermediate case arises with intermediate amounts of search, wherein the distribution of prices also contains a mass point at minimum average cost. Two figures [from Sadanand and Wilde (1982)] help understand how this market functions. Suppose the consumer/firm ratio is α^C. Then the demand curve facing the firm begins with its scaled down version of the market demand curve, namely $\alpha^C f(p)$, where $f(p)$ is (recall) the typical consumer's demand curve. If the firm raises its price above p^* (minimum average cost) it only gets $(1-\sigma)$ of those – the non-shoppers, and this is not a successful strategy if that modified demand curve, shown as $(1-\sigma)\alpha^C f(p)$ in Figure 1, lies to the left of the firms average cost curve $A(q)$. Thus, in general, search is sufficient to drive the market to the competitive equilibrium whenever

$$\sigma \geqslant 1 - \bigl[f(A(s))/s\bigr]\bigl[A^{-1}(p)/f(p)\bigr], \quad \forall p \geqslant p^*, \tag{1}$$

where $A^{-1}(p)$ refers to the left hand branch of the *AC* curve. This defines the proportion of shoppers σ necessary to make non-competitive pricing impossible to sustain (or, put slightly differently, it defines the amount of shopping necessary to drive the expected demand curve of non-shoppers below the *AC* curve for all possible outputs of the firm).

Now what happens if there is not sufficient search to produce this outcome (i.e., the condition in Equation (1) is not met)? Figure 2 shows the key conditions. If the expected demand curve confronting each firm cuts the average cost curve at two points, then entry is profitable, since for some output levels, $f(q) > A(q)$. As entry occurs, the expected demand curve from non-shoppers shifts to the left, eventually reaching a tangency in the usual monopolistically competitive equilibrium. The highest price that occurs then is p_U and entry has driven the consumer/firm ratio to α^N. This sets an upper bound on the possible prices in this model, and the firm in this case will produce $A^{-1}(p_U)$ as its output.

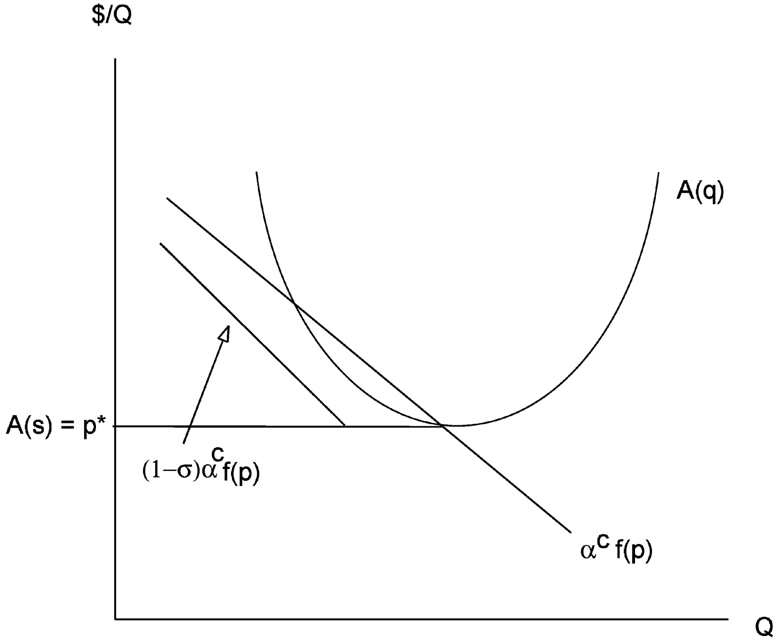


Figure 1. From Sadanand and Wilde (1982).

The other line in Figure 2 shows the outcome for the firm(s) charging the lowest price in the market, which is necessarily $p^* =$ minimum AC. The expected demand curve cutting the AC curve at that point is $[(1-\sigma)f(p^*)+\sigma nf(p^*)]\alpha^N$, consisting of the firm's "share" of non-shoppers plus n times the average share of shoppers (where n is the number of firms searched by shoppers – the lowest price firm gets them all by definition). The value of σ defined by this condition sets the least amount of shopping that still leads to some firms charging p^*.

$$\sigma \leqslant \left[s/\left(\alpha^N f\left(p^*\right)\right)-1\right]/\left[1/(n-1)\right]. \tag{2}$$

If σ falls between the two values defined by Equations (1) and (2), the market supports a distribution of prices between p_U and a lower-bound value p_L that depends on σ and n (the parameters defining the extent of search). Specifically, the firm's expected demand curve is given by

$$\left[(1-\sigma)+\sigma n\right]f(p_L)\alpha^N. \tag{3}$$

In general, as the number of firms searched by each shopper shrinks (n gets smaller) the rightmost of the two demand curves shown in Figure 2 shifts to the left, eventually approaching the leftmost curve that defines the upper bound price p_U. Obviously also,

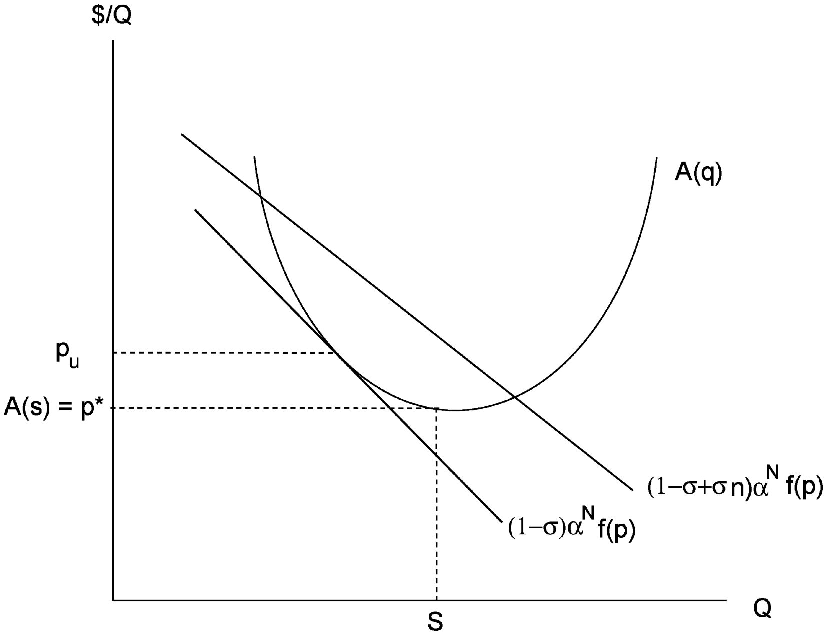


Figure 2. From Sadanand and Wilde (1982).

for a given $n > 1$, as the proportion of consumers in the market (σ) increases, this rightmost demand curve (which defines p_L) shifts back to the right, and p_L approaches p^* as σ increases. But the market need not have complete search, because Equation (1) tells us that once σ reaches a sufficiently high value, the market collapses to competitive pricing at p^* anyway. In the world with limited search, a gap will appear between p^* and p_L and prices will have continuous support between p_L and p_U.[6]

In summary, this model shows how distributions of prices emerge in monopolistically competitive markets, and shows that the way the market functions depends crucially on the proportion of consumers who undertake comparison shopping (σ) and the extent that each of those consumers searches (n). It does not require complete search to approach perfectly competitive pricing, but in general, more search lowers prices for everybody in the market, so those who do search do not capture the full value of their efforts.

[6] As Mark Satterthwaite has pointed out to me, this model of search presumes irrational behavior on the part of individuals if the variance of prices collapses completely. That is, once σ exceed the critical value defined in Equation (1), there can be no returns to search, in which case either n falls to zero or the fraction of searchers itself declines. Thus the model proposed by Sadanand and Wilde requires irrationally large amounts of search to support a complete collapse of prices to the competitive equilibrium. Fortunately, for use in health care markets, this issue can be discarded, since observed prices exhibit wide dispersions, and hence the potential problem of irrational search does not appear to be relevant.

2.3.1.1. Applications in health care markets. One of the unusual features of health care markets, as noted in the introduction, is that health insurance blunts the incentives to search. Using the model outlined above, we can see that there is a precise relationship between the demand elasticity and the number of shoppers necessary to generate a competitive (single price) equilibrium, and we also know [see Phelps and Newhouse (1974)] that reimbursement insurance lowers the demand elasticity of insured consumers. In particular, suppose we specify a constant elasticity demand curve of the form $q = \delta p^{-\eta}$. Sadanand and Wilde show that the necessary condition for a competitive equilibrium can be rewritten as

$$\sigma \geqslant 1 - \left[p/A(s)\right]^{\eta}\left[A^{-1}(p)/s\right], \quad \forall p \geqslant p^{*}. \tag{4}$$

Since health insurance generally drives the demand elasticity towards zero, the requisite proportion of shoppers rises as the insurance coverage becomes more complete. Similarly (and no surprise), when the equilibrium includes a dispersion of prices, the less elastic the demand curves of consumers (more insurance) the more dispersed are the prices. Thus the model of Sadanand and Wilde predicts a direct relationship between the degree of health insurance coverage and the dispersion of prices in the market.

At the same time, health insurance generally provides incentives that limit consumers' propensity to search, as discussed above, decreasing σ and also decreasing n for those shoppers who do engage in search. Thus, both the effect on demand elasticities and the effect on incentives to search drive the market for health services to one with less competitive outcomes (more dispersed prices, less chance of a mass point at the cost-minimizing price). Both of these forces reduce the chances for a competitive equilibrium and increase the dispersion of prices that will emerge.

While not modeled directly by Sadanand and Wilde, it should be obvious that parallel issues arise concerning search on quality. To understand how this works, consider a world where there were two possible qualities in the market (QL_i, $i = 1, 2$), and firms specialize either in higher or lower quality. The uninformed consumer will have two demand curves conditional on quality, but will have to assume some expected value of quality among unknown firms until quality is ascertained. If price is known but quality is not, then searching to find the higher quality firm (for a given price) is similar to searching on price for a known quality (as in the world of Sadanand and Wilde, with homogeneous quality). So we can expect (but the proof awaits further work) that search models in a managed care world will show similar effects of search on the distribution of quality, and also that the incentives of insurance to search on quality are similar to those relating to search on price.

2.3.1.2. Search with unknown product quality. Hey and McKenna (1981) have explored the more complicated and perhaps more pertinent problem of consumer search with both uncertain product quality and price. In their model, both price and quality are unknown; price can be observed before purchase, but quality can only be observed after purchase. They assume (in contrast to Sadanand and Wilde) that product quality is heterogeneous, with the joint distribution of quality and prices known. In this model, both product quality and price depend upon the extent of search.

Further, search rules differ from the case of the single product quality. In the world of only a single quality, endogenous search leads to the well known result that the consumer adopts a reservation price p_r, and search ends once a price lower than p_r is found. In the uncertain quality model, the first step proceeds similarly, with the additional complication that for each price draw from the market, the consumer must infer the expected quality, and then determine if the total bundle affecting utility (price and quality) is satisfactory. If not, the search continues again. Depending on the relationship of quality with price, the optimal strategy may be to "buy expensive" rather than "buy cheap." Indeed, the optimal search strategy varies greatly depending on the relationship of expected quality to price, and how that changes across the price spectrum. As Hey and McKenna show, there can be no simple conclusion drawn from these situations; the problem remains a fruitful area for future investigation.

A further complication exists in models with endogenous price and quality: it may be optimal for producers to signal a high quality by charging a high price. Some markets seem to be characterized by such signaling, while others show a high correlation between price and quality. False quality signals (posting a high price) are likely to emerge when consumers cannot directly measure the quality until they have purchased the product, when purchases are infrequent (or once in a lifetime), and where communication between buyers is unlikely to create adverse word of mouth counter-information. The quintessential example of such a situation would be the traveling salesman of the American west, moving from village to village and selling "snake oil" medicine, the lack of effect of which could only be understood after the salesman had moved on to find other customers (and also escaped retribution of angry buyers). Cases where such signaling would likely prove useless to the deceitful seller would be cases like musical instruments and consumer electronic equipment, where the consumer can readily determine the quality before purchase. Barbers and beauticians in fancy hotels (with few repeat customers) would more likely benefit from deceitful price-signaling of quality than barbers relying on steady customers.

How these issues play out in the world of health care remains a subject yet unanalyzed. Taking the basic ideas, one would expect to find more deceitful quality signaling among doctors who performed once-in-a-lifetime services, and then perhaps even more in areas where personal embarrassment would deter unhappy customers from complaining loudly to their friends. Procedures such as augmentation mammoplasty (breast enlargement) or penile enhancement (commonly advertised daily in the Los Angeles Times sports section) seem like prototypical areas where deceitful price signaling would likely take place.

2.3.1.3. Does insurance insure the costs of search? Dionne (1984) points out another important issue: If the insurance policy covers the costs of search in a way equivalent to the costs of the insured service, then increasing insurance coverage will increase the amount of search under certain patterns of risk aversion (specifically, when relative risk aversion increases with wealth). When search costs are not insured, there are tradeoffs to the consumer balancing risk avoidance issues with search cost issues. If relative risk aversion decreases in wealth, Dionne shows that search tends to diminish, and when search costs themselves are not insured, the result is unambiguous. The tradeoffs are not difficult to comprehend: If medical care is insured (nearly) fully, then the consumer's personal benefits of finding a lower price are small, and when the consumer bears the cost of search, it becomes decreasingly desirable to search as insurance becomes more complete. The offsetting factor, however, is that the acquisition of insurance raises expected utility and the wealth effect arising from that pushes towards more search when risk aversion is increasing in income. When risk aversion is decreasing in income, this wealth effect works in the opposite direction. Unfortunately, the literature on risk premiums is rather sparse even in estimating the magnitude of relative risk aversion, let alone providing an understanding of how that changes with wealth. Garber and Phelps (1997) provide a summary of the relevant literature estimating the magnitude of relative risk aversion in the range of 1 to 4.

As to whether health insurance insures search costs, it depends in part on the nature of the search. In some cases, search for lower prices involves simply making telephone calls, etc., in which case the search will not be insured. In cases where the search involves a direct sampling of the provider's quality, then the insurance effectively treats search costs (or most of them) the same as it treats the purchase of services.

2.3.1.4. Actual price variability. Only sporadic reports appear in the literature showing the distribution of physician firms' prices for known quality of care. The most direct of these comes from Marquis (1985), showing data from the RAND Health Insurance Experiment during the 1970s in two cities (Dayton, OH and Seattle, WA). Her data show the coefficients of variation of fees for general practice and internal medicine (two homogeneous groups of providers in terms of observable credentials). On average, the internal medicine specialists received fees some 10–13% higher than general practice doctors, as befitting their more extensive training. Of particular interest here is that the coefficients of variation in the prices were about 0.15 in Dayton and about 0.23 in Seattle, a wide range of prices. In data such as these, a coefficient of variation of 0.15 implies that the range from high to low will typically be about a factor of two. I know of no other similar price data in a more recent era.

How do these prices contrast with other markets? In one data set available to me through a litigative consulting job, retail prices of gasoline for a single brand were observed repeatedly in a single city in the western US. The homogeneity of the product is nearly perfect in this case, since the brand of the gasoline, the degree of service (self vs. full) and the grade of gasoline (regular vs. premium) were all identified. In these data, coefficients of variation in prices (at any point in time) were an order of magnitude smaller, averaging about 0.015.

In a more wide-ranging study, Pratt, Wise and Zeckhauser (1979) measured the prices of a wide range of consumer commodities, all standardized as to brand. They found coefficients of variation ranging from as low as 0.05 (Raleigh Grand Prix bicycle, mixed concrete) to as much as 0.5 (horoscope, carnations). Thus the variability in physician prices is not unique to this market.

2.3.2. Herd and cascade models

An alternative approach to understanding the behavior of physicians' prescribing patterns may come through studies of herd behavior and information cascades. Although this approach has not been applied directly to behavior in medical markets, it provides a valuable structure for future work. A few examples of this literature are summarized next.

Banerjee (1992) analyzes the behavior of people who observe other economic agents and then (particularly when their own information is weak) follow the choice of the preceding agents on the premise that they may have had special "inside" information. In such a world, Banerjee shows that people end up doing the same thing ("herd" behavior) and that the resulting equilibrium is not efficient. A similar model appears in Bikhedandani, Hirshleifer and Welch (1992), using a model of information cascades. This model, like the herd model of Banerjee, leads to herd-like behavior, but in the information cascade model, the herd's behavior is fragile and can change rapidly. While fashion "fads" may follow this path, most studies of medical practice variations find strong persistence in the observed patterns, so models that commonly lead to fragility and faddish switches in behavior will likely not prove valuable in understanding the physician practice variations phenomenon.

Another approach follows more in the tradition of game theory, analyzing the behavior of an n-person repeated game. Young (1993) studies the evolution of "conventions" (common practices) in such a world, using a model where agents have a sample of information about how other agents have behaved in the past. A similar model is studied by Kandori, Mailath and Rob (1993), focusing on noise in the information signal or mutations in the system.

3. Disagreement about the production function

3.1. The healer's dilemma

This section turns to a completely different problem, namely the role of information in the supply side of the market. This discussion concerns the information healers (doctors, dentists, nurses, etc.) have about the production functions (plural!) that create the final product in the market – "cures" of disease. Investigation of the production function for health providers has been sporadic at best, but almost never focusing on the actual output of value to consumers, the cure of disease or the improvement in health outcomes through such things as pain relief, improved mobility, reduced anxiety, and related outcomes.[7] For ease of discussion, I will call this set of outcomes "cures" hereafter, with

[7] Most studies of the links between inputs and health outcomes use methods of medical decision theory. A few emerging studies use econometric techniques to help understand the relationship between inputs and health outcomes. A recent study by Cutler, McClellan and Newhouse (1999) on the treatment of acute myocardial infarction (AMI) – heart attacks – provides an econometric estimate of the marginal value of interventions in a novel and important way.

an understanding that the set of outcomes under consideration often involves outcomes that do not literally cure disease, but rather relieve the symptoms of the disease in a way that improves patients' utility.

The premise here is that healers confront a staggering array of treatment technologies that they must understand (at least in part) in terms of their costs, side effects, and ability to cure patients in a setting where the technologies' effects may differ greatly depending on unique and (often) difficult-to-observe characteristics of the patient.

The set of diseases that healers must recognize is immense – the relevant codebook for insurance purposes describes literally thousands of diseases, not to mention subclassifications within these categories.[8] The relevant code book for treatments also has thousands of treatments, many of which can potentially affect numerous diseases.[9] For doctors to understand well the complete set of relationships between these treatments and the myriad diseases that their patients may bring to the patient encounter is literally impossible. Indeed, as we shall see in Section 4, it is probably not even possible at the societal level, let alone the individual healer level, to understand these relationships fully.

In slightly more formal notation, consider a utility function for Health (H) and other goods (X), $U(H, X)$, where X can be purchased directly in the market, and Health is produced in by medical care (m) using the production function $H = g(m)$. Consumers confront a stochastic illness ℓ, so $H = H_0 - \ell + g(m)$. Demand for m is then derived, depending on income, prices, the severity of the individual's illness, and the *perceived* efficacy of the treatment in healing (i.e., the shape and slope of $g(m)$).[10] Clearly, if there is disagreement about the shape of $g(m)$, then the eventual demand curves for m will differ across regions. Of course, H, ℓ and m are all vectors of large dimensionality, representing different aspects of health, different illnesses, and different treatments, respectively.

Despite these informational problems, healers *must* act on behalf of patients. They may refer the patient to a more specialized healer, they may try some treatment or another, or they may pursue a course of "watchful waiting" or even tell the patient that there is nothing further to do. Given the impossibility in knowing the specific effects of every possible treatment for every possible disease for every possible patient (remember: the treatment effects may vary greatly by unobservable patient characteristics), it should come as no surprise that doctors disagree about how often (and for which patients) various treatments should be employed. Subsequent parts of this section explore the evidence on the extent to which this disagreement manifests itself in regional differences in patterns of treatment (Subsection 3.2), in the practice patterns of individual

[8] See the International Classification of Diseases, Version 9, or more commonly ICD9, published by the Department of Health and Human Services (1980).

[9] See Current Procedural Terminology, Version 4, or more commonly CPT4, published by the American Medical Association (1990).

[10] This approach was first developed in a deterministic version by Grossman (1972a, 1972b). See Grossman (2000) for further discussion of such models.

physicians (Subsection 3.3). Finally, we seek an understanding of the economic consequences of these disagreements (Subsection 3.4) in terms of welfare loss to patients arising from incomplete information.

3.2. Regional variations

The study of regional variations in medical practice began in England when Sir Allison Glover (1938) presented a paper to the Royal Academy of Medicine describing the frequency with which British schoolchildren of a specific age had received a tonsillectomy. He observed differences of almost an order of magnitude between the towns with the highest and lowest rates of tonsillectomy. Sir Allison, and even his astonished audience that night (their later discussion is also recorded) could not believe that such differences were due to differences in underlying disease patterns. Their discussion – and much modern interpretation of these phenomena – suggested the likelihood that the procedure was overused on average, although one cannot infer from the observed pattern the proper frequency of use (for a defined population). As Section 3.5 discusses, welfare losses occur both because of variability about the mean (even if unbiased) and if there are biases in the average rate of treatment from the "correct" (fully informed) demands.

This approach to studying variations ignores one further facet of the problem: there may not be appropriate mechanisms in place to match actual services provided with patients who have the highest value for the procedure. In general, if the level of service is limited by some external force such as physical capacity or price controls, public mechanisms of rationing will often lead to outcomes where the average value of the services received is below what one would find in a free exchange market with the same quantity sold. This type of problem is not part of the variations issue discussed below, since I assume throughout my discussion that other factors than rationing lead to differences in the amount of service provided.

The typical study of this phenomenon (usually called "regional variations" studies) calculates the rate at which an intervention is used for populations living in a defined geographic area (e.g., a county or state). One must be careful to distinguish between the region of the patient's residence and the region where the procedure is performed, because (particularly for rare treatments) patients are often referred to specialty centers for treatment (by their doctors at home), and failure to account for this would create an odd pattern of apparently very high density of treatment in some communities and none in others. These studies typically also adjust for the age and gender mix of the populations studied, since (at a population level) these variables usually explain a considerable amount of the observed patterns, because disease frequency varies systematically by age and gender for many if not most illnesses. Typically, however, the underlying true prevalence of disease is *not* measured, so it typically remains an assumption that the age and gender adjustment adequately accounts for differences in the underlying burden of illness in the populations analyzed.

Much of the seminal work in this area was conducted by scholars reporting their work in other outlets than commonly referenced by economists. Perhaps the most influential of these was Wennberg and colleagues, who brought the concepts of practice variation into the realm of legitimate scientific inquiry [Wennberg and Gittelsohn (1973, 1982)] and then into the spotlight of public policy inquiry [Wennberg, Freeman and Culp (1987), Wennberg et al. (1989)]. Other key work in this literature comes from Roos and Roos and colleagues from Manitoba, Canada [Roos, Roos and Henteleff (1977), Roos and Roos (1982)].

This literature has historically adopted a policy of reporting the coefficient of variation ($\mathrm{COV} = \sigma/\mu$) of the empirical frequency distribution of treatment rates.[11] The literature reporting these variations – almost entirely in medical journals – is now immense [see Phelps and Mooney (1993) for a summary]. For purposes of illustration of these phenomena, Table 1 shows a single example of regional variations calculated from data showing hospitalization rates for various procedures in 1987 in New York State, using counties as the unit of observation.[12]

In order to support the belief that these and related differences in regional rates of use are due to informational differences, one must eliminate competing explanations. Several natural economic explanations arise, each of which can be eliminated as empirically relevant. We will be able to see that regional variations are not due to income effects, price effects, or substitution among alternative therapies that might produce similar outcomes. It can also be demonstrated that in almost all cases, the variations are "real" in the sense that they are too large and persistent to arise simply from random observations of regions with the same underlying propensity to treat. To understand the first two of these (income and price effects), consider a regression model

$$y = x_1 B_1 + x_2 B_2 + u, \tag{5a}$$

where x_1 and x_2 represent income and price. Taking the variance of this expression and converting to elasticities (η) and coefficients of variation ($\mathrm{COV} = \sigma/\mu$) gives:

$$\mathrm{COV}_y = \left[\mathrm{COV}_1^2 \eta_1^2 + \mathrm{COV}_2^2 \eta_2^2 + \eta_1 \eta_2 \mathrm{COV}_1 \mathrm{COV}_2 + \mathrm{COV}_u^2\right]^{1/2}. \tag{5b}$$

(Here, COV_u is defined to mean σ_u/μ_y, thereby avoiding the obvious complication of computing COV_u scaled by its own expected value of zero.) It is easy to show [see Phelps and Mooney (1993, Appendix 7)] that $|\eta_i|$ provides an upper bound on the rate at which COV_y changes with COV_i.

Data from New York State (using counties as the unit of observation, as appropriate) allow the calculation of the COV for income as 0.2 and the COV for price as approximately 0.1 [see Phelps and Mooney (1993) for details]. The income and price elasticities

[11] As we shall see in Section 3.5, this is fortuitous, since the COV measure enters into a formula to calculate the welfare loss from variations.

[12] These use data from the Statewide Planning and Research Cooperative System (SPARCS) from the New York State Department of Public Health. Other studies using these data are reported in Phelps and Parente (1990), and Phelps and Mooney (1992).

Table 1
Some examples of medical practice variations

Medical			
High Variation		*Low Variation*	
Pediatrics otitis media and URI	0.49	Red blood cell disorders	0.13
Pediatric pneumonia	0.48	Eye disorders	0.13
Acute adjustive reaction	0.45	Gastrointestinal hemorrhage	0.13
Depressive neurosis	0.42	Peripheral vascular disorders	0.12
Atherosclerosis	0.37	Heart failure and shock	0.12
Pediatric gastroenteritis	0.37	Acute myocardial infarction	0.12
Chronic obstructive lung disease	0.35	Respiratory neoplasms	0.10
Concussion	0.34	Kidney and urinary tract disorders	0.10
Circulatory diagnoses, except AMI, with cardiac catheterization	0.32	Special cerebrovascular disorders	0.09
Respiratory signs and symptoms	0.32		
Pediatric bronchitis and asthma	0.30		
Surgical			
High Variation		*Low Variation*	
Dental extractions, restorations	0.61	Ectopic pregnancy	0.14
False labor	0.61	Inguinal and femoral hernia operations	0.13
Extracranial vascular procedures	0.46	Hand operations	0.13
Tubal interruptions	0.38	Major genito-urinary tract operations	0.13
Tonsillectomy and/or adenoidectomy	0.36	Skin grafts	0.12
Carpal tunnel release	0.33	Respiratory system operations	0.12
Vaginal delivery with added procedures	0.32	Cholecystectomy with gall bladder disease	0.12
Pediatric hernia operations	0.32	Vaginal delivery without complications	0.11
		Mastectomy for malignancy	0.11
		Hip procedure, except joint replacement	0.11
		Caesarian section	0.10
		Male reproductive system operations	0.10

of demand are estimated in the RAND Health Insurance Study as approximately 0.1 and -0.2, respectively. Thus income can at most add $0.2 \times 0.1 = 0.02$ to any measured COV, and price differences can at most add $0.2 \times 0.2 = 0.04$ to the observed COV for hospitalizations. Since the observed COVs are generally in the range of 0.2 to 0.5 or larger, it is clear that these standard economic factors (income and price) cannot meaningfully explain these variations.

The issue of substitution requires a different approach. If meaningful substitution exists between two alternative therapies (to produce the same quantity of cures), then there should be negative correlations between the observed rates of use of two treatments (say, T_1 and T_2). Figure 3 shows this phenomenon, showing "iso-treatment contours" with different overall rates of treatment. For two regions with true substitution (points A and B on the same iso-treatment contour) the correlation in observed rates of use will be negative. For two regions on different iso-treatment contours (such as A and C), the

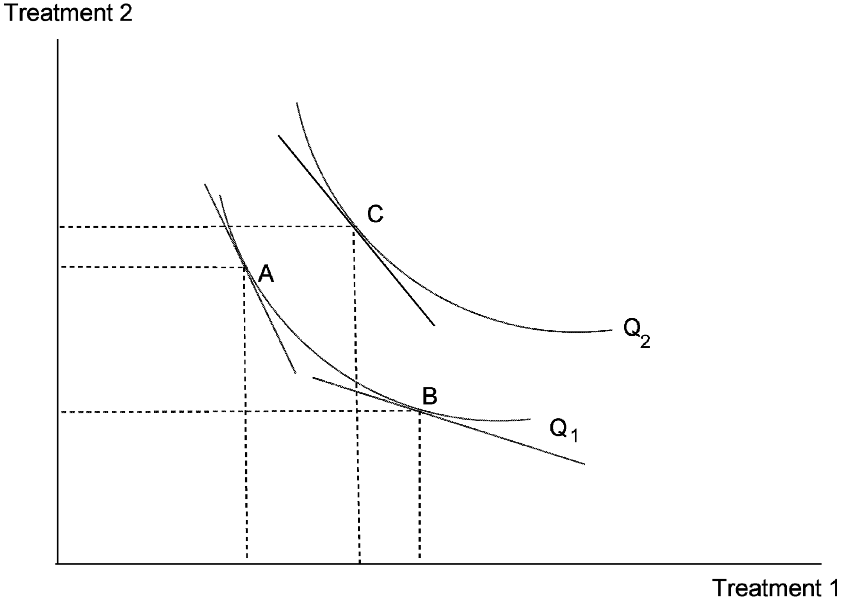


Figure 3. Substitution in production of health.

correlation is likely positive. Note that there can be cross regional differences in the rates of overall treatment and still be negative correlations, but positive correlations in the treatment rates require differences in overall treatment rates.

Table 2 shows the observed correlations from county-level data in New York State for a series of treatments identified by physicians as possible substitutes for treating the same disease within the hospital (the data, remember, reflect hospital admission rates). In almost every case tested, the observed correlations are positive, and usually significant statistically. Only in the case of regular hospital days vs. intensive care unit (ICU) days was substitution observed (significant negative correlations).

Next, we can return to the question of whether the observed variations in treatment rates reflect true differences across regions, or simply arise from random chance. A general test [Diehr et al. (1992)] showed that when using population-weighted coefficients of variation in data such as these NY state data, a simple transformation of the COV has a chi-square distribution with $(n - 1)$ degrees of freedom, where n is the number of counties in the data set. Tests on each of the COVs shown in Table 1 reject the hypothesis that the true COV is zero using this approach. A more powerful approach builds the test up from the regression model such as specified in Equation (5) [Hu (1996)].

Other economic explanations also fail. For example, some observers have posited that these variations arise from differential demand inducement across regions.[13] Such

[13] See McGuire (2000) for a general analysis of demand inducement.

Table 2
Correlation of substitutable procedures/admission rates

Medical back admissions and surgical back admissions for low back problems	0.20
Myelogram for low back problems and CT for low back problems	0.44*
Vaginal hysterectomy and total abdominal hysterectomy for non-malignancy	0.29*
Total abdominal hysterectomy and myomectomy for non-malignancy	0.19
Extracapsular and intracapsular lens extraction	0.33**
Arch-arteriogram and carotid arteriogram for cerebrovascular accident (stroke)	0.49**
Single vessel coronary artery bypass graft (CABG) and single vessel percutaneous transluminal angioplasty (PCTA)	0.50**
Single vessel CABG and multiple vessel CABG	0.70**
Admission for pacemaker insertions and medical admissions for selected arrythmias	0.16
Admission for angina pectoris (chest pain) with and without arteriogram	0.08
Admission for myocardial infarction with and without arteriogram	−0.018
Intensive care unit and non-intensive care unit admissions for myocardial infarction	−0.64**
Intensive care unit and non-intensive care unit Length of Stay for angina or chest pain	−0.37*

$^{*}p < 0.10$; $^{**}p < 0.05$.

an explanation fails to explain why similar patterns of regional variability arise in such countries as Great Britain and Sweden as in the US [McPherson et al. (1981, 1982)], since those countries have comprehensive national health systems where doctors are all on salary, and hence have absolutely no economic incentives to induce demand.

Finally, one must account for the fact that rational demand for treatment shifts directly with the intensity of illness [Phelps (1973, 1997)]. Aggregating individual demands to the regional level (as appropriate for these studies) generally assumes – often without empirical support – that the illness patterns in these regions are similar, once age and gender mix of the populations has been controlled (as is almost universally true in the regional variations literature). Most of the studies of regional variations cannot directly measure illness patterns, so the residual role of illness patterns (holding age and gender mix constant) are typically unknown. The few studies where there are direct measures of illness available [Roos et al. (1977), Roos and Roos (1982), Leape et al. (1990)] show that, if anything, patterns of illness are perversely associated with treatment patterns. Thus, at least on the basis of this sparse evidence, we can reject the belief that regional variations follow patterns of regional illness differences.

A separate issue concerns relationships between patterns of treatment: Treatment patterns are in general unrelated across interventions. Put slightly differently, if Region A is "high" and Region B is low in the rate of use of some specific treatment (say, carpal tunnel surgery), then there is essentially no predictive power as to where Regions A and B will appear with respect to some other treatment (say, knee replacement surgery), even within the domain of the same specialist (here, orthopedic surgery), let alone when one ventures into the domain of other specialists (say, for hospitalization for acute heart attack or pediatric pneumonia). As shown in Phelps and Mooney (1993), the correlations of treatment patterns across interventions is quite low, even within specialty. This in effect eliminates the possibility that the patterns of treatment observed are related to "availability" in some way (either with lower time costs for patients or because of demand inducement that might be occurring).[14]

A still further related issue asks whether the regional patterns of treatment, even if unrelated to physiologic illness patterns, might not relate to patient preferences, since patient preferences for treatment might well differ even if physiologic illness were the same. Here, we must concern ourselves with preference patterns than cannot be explained by income and price, since (see previous discussion) these traditional economic variables offer little to explain regional variations patterns. In this case, while there is no specific evidence that patient preferences drive these patterns, the plausibility that such preference differences drive the treatment differences is extremely low. In order for this to "work" as the explanation, patients must somehow migrate to various geographic regions on the basis of their preferences about being treated for specific diseases, independent of their preferences for other aspects of living in those regions. An alternative is that they acquire disease-specific preferences for treatment from their regional companions after moving to the region, or that some facet of culture in the region controls disease-specific treatment patterns. The herd behavior discussed in Section 2.3.2 may prove useful in studying these issues.

We must remember, however, that any operative patient preferences cannot be simple variations in preferred treatment intensity that appear across all treatments. The

[14] This issue also relates to the consequences of aggregating various interventions to study medical practice variations. The more one aggregates across treatments when doing regional variations studies, the less apparent variability one will find. The reason is simply a consequence of the law of large numbers: when one aggregates a series of uncorrelated treatments in a single region, the overall averages necessarily move towards the common mean. This mistake was propagated most extensively by Stano and Folland (1988), who studied the overall rate of surgical and medical hospitalizations in Michigan and reported that they found no meaningful variations in treatment patterns.

observed treatment patterns for various diseases are poorly correlated across regions, as previously discussed, so there must be very specific preferences in (say) County A by those who wish to be treated aggressively for carpal tunnel injury and at the same time very specific preferences of patients in County B for aggressive treatment of knee injuries. Since the number of potential treatments far exceeds the number of relevant geographic areas, attempts to explain observed treatment patterns on the basis of patient preferences must rely on a very complicated process of preference formation or migration that Occam's razor would dismiss.

As these alternative explanations fail, we are left with the sole remaining plausible hypothesis, that the patterns of regional variation observed for so long over so many countries and health payment schemes are most likely due to disagreement across physician groups about the proper indications for using various treatments, or – in the language of economists – providers disagree about the shape of the production function $g(m)$ that transforms medical care into health. Moreover, the patterns of variation show that the disagreement is geographically clustered.

Models of physician learning suggest that such clustering can be expected, given differential costs of information from different sources. Physicians, recall, confront a hellishly large number of illness and procedures that they must be able to diagnose (in the former case) and use (in the latter case). What are the possible sources of information for physicians about these issues? In every case, the training meaningfully begins in medical school, where the second two years of training include "clinic" work under the tutelage of staff physicians in their hospital of training, and more importantly, from other physicians in training in residency and fellowship programs. When students graduate, they move to residency training programs that last for 3–6 years typically, depending upon their specialty of choice. Some doctors then take on additional subspecialty fellowship training, e.g., in advanced cardiology or neurosurgery. After that training, they move to a medical practice, either in solo practice, with partners (often in a loose business practice arrangement) or as employees of a large group. They will also join a hospital's medical staff (perhaps more than one, but typically concentrating their inpatient practice in a single hospital) for treating seriously ill patients whom they wish to hospitalize. (Note here the careful refusal to use words such as "patients who require hospitalization.")

Now, from what sources to doctors "learn" about the proper indications for a treatment, and how do they modify those beliefs through time? The base learning takes place in medical school and residency training. After that, the young doctor will continue to acquire information, possibly formally, possibly informally, from partners, colleagues on the hospital medical staff, and from reading journal articles and attending "continuing medical education" symposia in locations such as Aspen, CO., Aruba, Cayman (British West Indies), Yosemite National Park, and (during the winter) Palm Springs, CA and Boca Raton, FL.

These sources have differing credibility, and certainly different costs associated with using them for information acquisition. A Bayesian learning process [see Phelps and Mooney (1993)] will likely emphasize local (low cost) sources of information and

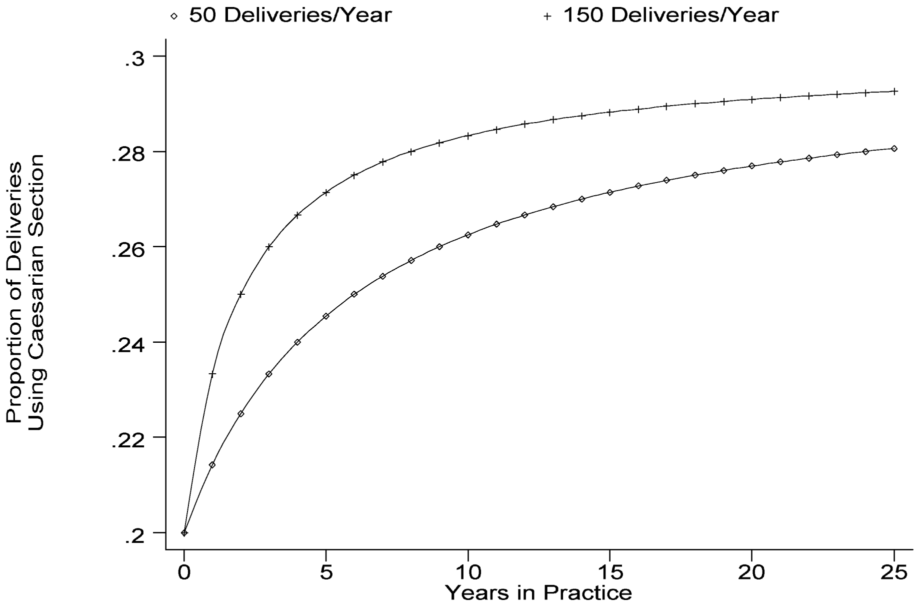


Figure 4. Bayesian learning model.

"stylistic" advice about when to use various medical interventions, most readily from practice partners and hospital medical staff colleagues. To the extent that local styles are already in place, this Bayesian learning process will transmit those styles to new practitioners. This will, if anything, be emphasized by assortative "mating" of new doctors with established doctors and hospitals who share beliefs about medical interventions.

Figure 4 shows a simple example of this process, using the decision of whether or not to perform a Caesarian section instead of a normal vaginal delivery. The indications for this procedure are complicated, with many different circumstances leading to the decision to use the surgical intervention. For convenience, we can summarize the problem in terms of the proportion of a doctor's patients for whom he recommends a C-section. One can formally model the problem by assuming that the physician's prior has a Beta distribution for the proportion of deliveries (θ) where C-section is appropriate, with parameters α and β. The prior distribution, arising from medical school and residency training, has a distribution of

$$h(\theta) = \Gamma(\alpha + \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}/\Gamma(\alpha)\Gamma(\beta) \quad (0 < \theta < 1),$$

with an expected value of $\alpha/(\alpha + \beta)$ representing the fraction of total deliveries observed during training where C-section was used.

Suppose now the young doctor enters practice and begins to observe colleague's behavior in the use of Caesarian sections at a rate y/N. The physician then combines this new information with the prior, in this example treating each "case" adding to both the prior evidence and new evidence with equal stature. Then the expected-loss minimizing posterior distribution has an expected value of

$$w(y) = (\alpha + y)/(\alpha + \beta + N),$$

which obviously can be decomposed to a weighted average of the prior mean and the mean of the rate of accumulated new evidence.

To place this in more concrete terms, suppose the training of the doctor led to a mean C-section rate of 20% with evidence credibility measured by $(\alpha + \beta)$ deliveries. (In Figure 4, $\alpha = 60$ and $\beta = 240$.) Now the doctor begins new practice, observing among his colleagues and other obstetricians in his hospital a C-section rate of 30%. As the years pass by, the observations about the "right" indications for C-sections transmit from the community to the new doctor, and he moves his practice style towards that of the community. As Figure 4 shows, the rate of closure depends on the rate of seeing deliveries (and C-sections) per year, relative to the body of evidence accumulated during the training years. This type of model offers a coherent explanation of how community wide practice styles *persist* once they emerge. It does not explain, however, how they originally emerge.

The problem of how practice styles emerge differently across communities can be explained in part if the doctors from one community tend to come from one medical school (or residency program) while those from another tend to come from another. We do know that doctors tend to train where they intend to practice, but in general, this seems like a weak mechanism to create strong practice patterns. An alternative arises from the likely distribution of outcomes from weakly-designed "local" experiments with procedures – a learning process that has long-dominated the acquisition of new information about the efficacy of various medical interventions. (Section 4.2 discusses this problem in more detail from another perspective.)

Without a well designed randomized controlled trial (which, alas, is the case in most medical interventions) the likely "evidence" to support the use of a new procedure is the outcome of the first handful of cases carried out by an innovating physician in the community, compared with historic outcomes for similar patients using previously available procedures (including "no treatment" in some cases). Suppose the historic procedure had a success rate of 50% and a doctor in the community learns about some new alternative either by reading in journals or attending a training session in some other city (where the trainer will assuredly be a devotee of the new procedure). If 20 new cases are attempted as the "trial" of whether the new procedure works, then the doctor must determine (often with poor understanding of statistical methods) whether the new procedure is better than the old one or not. Assuming an unbiased observation of the outcome (in itself a difficult assumption, since the doctor trying the new procedure will be "invested" in it, and hence likely to observe favorable outcomes wherever possible), the trial will have more or fewer successes than the old procedure depending on the true underlying success rate, the individual doctor's skill at the new procedure, and idiosyncratic patient characteristics.

If this "trial" has fewer than 10 favorable outcomes (assuming a sample size of 20 chosen and a previous success rate of 50%), then the doctor will abandon the new methods in favor of the old, and visa versa. The strength of the outcome will also determine how strongly the advice of that doctor flows to his colleagues. As outcomes randomly differ, it is easy to see how some communities will emerge where the new procedure "worked" and becomes widely adopted, while in other communities, it "fails" and the previous approach to treatment is maintained. This provides one mechanism to generate regional variations in treatments.

3.3. Physician-specific variations

The presence of regional "schools of thought" about various treatments and their proper use raises a separate question: Do individual doctors have "styles" about how aggressively they (individually) use medical intervention for their patients? The approach must necessarily differ from the regional studies described in the previous section. In the study of regional practice variations, the rates of treatment are calculated using geographic regions from which the numerator of the rate is the number of treatments for people within the region and the denominator is the relevant population. An alternative approach uses individual physician practices as the unit of observation, aggregating across all treatments. In this work, the numerator is a measure of treatment rate (e.g., total annual medical spending created by the physician) and the denominator is the total number of patients being treated by each doctor. Since referral patterns would greatly complicate this analysis, it is best carried out using primary care doctors rather than specialists, and it makes most sense to attribute back to the primary care doctors all of the costs of medical care received on behalf of their patients.

This type of work presents two important complications. First, it is normally extremely difficult in the normal US health care market to determine the number of patients under active treatment by any single doctor, even with that doctor's full cooperation. (The problem arises because patients leave the practice to other towns or other doctors, but fail to notify the original physician that they have moved.) The data set used below solves this problem by using data from a health care plan where the primary care provider (PCP) of each patient is clearly identified in the data, and the plan's rules strictly enforce the requirement that all treatment be initiated by the PCP.[15] In the analysis that follows, all treatment received by any patient is "assigned" to that patient's PCP, since it either would have been provided directly by the PCP or else through referrals generated by the PCP.

The second problem arises because the patients in any doctor's practice will typically have different diseases than those in another practice. Thus, if one wishes to aggregate medical care spending across illnesses within a single doctor's practice, it is necessary

[15] During the period of this study, the plan vigorously refused to pay for treatment initiated directly with any other provider than the PCP.

somehow to control for the severity of illness for each patient in each doctor's practice. Fortunately, new severity of illness measures are available that allow such an adjustment, providing a reasonably strong ability to explain individual patient medical care use which, when aggregated to the primary physician level, allows the study of individual practice styles.

The data used in this study consist of the annual medical spending (which also was analyzed by type of care, in results not reported here), derived from the insurance claims paid by the insurance carrier during the relevant calendar year. (Three successive years of data were available for the study.) The insurance plan provided a widely comprehensive scope of benefits (compared with standard US health insurance plans) and had a simple copayment of $5 per office visit as the primary copayment mechanism (virtually identical for all patients in the study). Thus, the data should capture virtually every medical encounter with each patient, and price effects are irrelevant, since the fee paid by each patient was identical for all patients and all physician encounters. (There were similarly small fees for other services, but again, identical across users of the plan.)

The estimation methods are easy to explain. Consider a regression model using individual annual medical spending as the dependent variable, and regressors that control for observed patient characteristics such as age and gender (X), regressors that control for severity of illness (S), and a vector of dummy variables for each doctor in the data set (D). In the actual data discussed below, there are approximately 300,000 patients (the number varies from year to year) and approximately 500 physicians. One can then test to see if the distribution of these indicator variables (D) has any meaningful variability.

The success of this approach hinges closely on the ability of the S vector to control for severity of illness. In previous demand modeling (e.g., earlier regression models using annual survey data of Newhouse and Phelps (1976), or RAND Health Insurance Experiment studies [Newhouse et al. (1993)], severity of illness indicators add at most an R^2 of approximately 0.2. (See also Manning, Newhouse and Ware (1981) for further discussion.) These new indicators of illness severity [Perkins (1991), Starfield et al. (1991), Weiner et al. (1991)] provide incremental R^2 of 0.5 to 0.6 for comparable individual data (in natural logarithms for non zero data).[16]

[16] Both the Perkins (1991) and the Starfield et al. (1991), Weiner et al. (1991) approaches use physician expert opinion to measure illness severity. The Perkins approach generically asks pertinently-trained doctors, disease by disease, to rate "How serious is this disease on a 0–5 scale?" When a patient's medical claim indicates the presence of disease X, the comparable severity of illness is attached to that patient. These indications are made for thousands of specific diseases in the Perkins work. The Weiner approach asks doctors to group diseases into "baskets" that "should have" comparable resource use requirements.
In both cases, since the analysis involved aggregating all demands for a patient over an entire year, one must allow for a patient with multiple diseases. In general, two approaches were followed; one used the maximum severity of disease observed for all of a patient's diseases as the proper indicator of severity (to predict annual spending). The other approach summed the severity indicators of each observed illness. In general, the latter approach yielded higher R^2 measures than the former in the regression models.

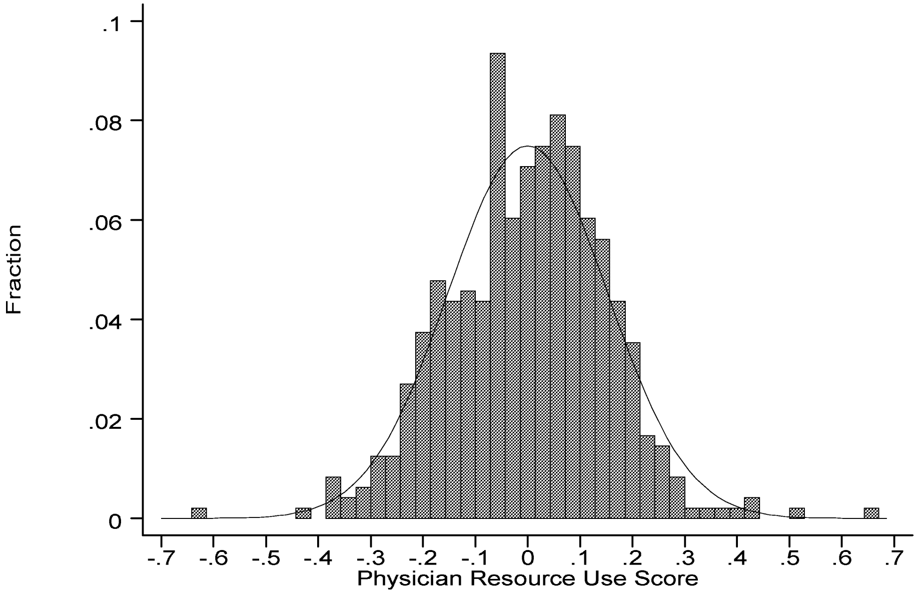


Figure 5.

The residuals from these regressions, when accumulated to the primary care provider, provide an index of the physicians' propensity to prescribe resources for their patients (both direct and through referrals). Figure 5 shows the distribution of physician-specific indicators, defined so that zero indicates a pattern of resource use at the overall mean, -0.1 indicates a 10 percent reduction, $+0.2$ indicates a 20% increase above the mean, etc. Approximately two-thirds of the physicians' indicators are indistinguishable from the mean using a 5% rejection rule, but for the remaining one third of the doctors, the average patient expense, even after controlling well for patients' severity of illness, differs significantly from the mean. Since the distribution of physicians' propensities to use medical resources – their medical "styles" – is rather symmetric, there are approximately as many doctors above the mean as below.

Table 3 shows the average spending by decile of doctors' style indicators. As these data show, doctors in the highest decile use resources to treat their patients at about twice the rate of those in the lowest decile. These data demonstrate that – at the individual physician level in a single community within a single health insurance plan – doctors' beliefs about the efficacy of treatment, and their consequent choices about the use of medical interventions, demonstrate a high degree of variability. In these data, not only are there differences between physicians, but these differences necessarily correlate across treatment types: Some doctors have an "aggressive" treatment style, while others are more conservative. These styles are statistically significant, and economically important. (Note the difference here with regional practice patterns, where there is little

Table 3
Deciles of deviation of per patient expenditure from overall average, by practice

Decile	Number of MDs	Number of patients	Average deviation
1	49	10224	−\$419
2	50	19976	−\$205
3	49	15688	−\$132
4	50	29425	−\$83
5	49	24133	−\$48
6	50	20211	−\$12
7	49	25597	\$46
8	50	20716	\$115
9	50	17658	\$223
10	49	7263	\$594

Note: Overall mean expense approximately \$1000.

if any correlation at the regional level between overall rates of use of various interventions. The links between these two findings remain a topic for subsequent research.)

One final note about this type of work bears discussion. To the extent that patients have systematic preferences about overall intensity of treatment, and to the extent that they can identify physicians who share such predilections, then the analysis of doctors' styles discussed above will overstate the degree to which the results are physician-determined. Assortative "mating" of doctors who are aggressive in treatment with patients who prefer aggressive treatment will produce (qualitatively) the same patterns of per-patient treatment costs as appear in Figure 5. Subsequent analysis using data sets not now available would be necessary to disentangle the effects of doctors' styles vs. those of assortative mating. Either some instruments would have to be available to identify the patients' choice of doctor, or else a randomization of patients to doctors would have to be used to remove possible correlations of patient preferences from doctors' styles.[17] But note that the doctors must have identifiable styles for this to take place, else assortative mating has no benefit to patients. Thus at least some of the patterns of "styles" must indeed be the doctors' styles, but some of it may also be patient preferences aligned with doctors' styles.

3.4. What relationships between regional and individual practice variations exits?

We have two strands of evidence relating to the differential belief systems of doctors about the efficacy of the treatments they use. One approach compares regional average

[17] One instrument recently suggested to me would use prior-year expenditures for each patient, most preferably separately identifying treatments for acute and chronic medical conditions.

use of single specific interventions. Another aggregates many interventions at the individual doctor level. Do these two approaches have any intersection of behavior? One approach to this problem would look at the procedures that exhibit high COV scores in regional variations and ask how much these same procedures account for the variations in individual physician-practice behavior. While this represents a fruitful area for future analysis, preliminary studies to date have found little link between the regional variations literature and the individual physician practice studies [Phelps et al. (1992)].

3.5. Welfare loss from variations

The economic consequences of variations – if due, as discussed above, to incomplete information about the efficacy of medical intervention – can be readily analyzed.[18] Consider Figure 6, which shows demand curves for a single therapy in two regions (1 and 2) and an intermediate "full information" demand curve at the average of the two. (We can relax this assumption momentarily.) The welfare loss from incomplete information about the production function leads to under-use of the therapy in region 1 (X_1) and over-use in region 2 (X_2) relative to the full information demand curve (X_F). The traditional measures of welfare loss from incomplete information are the triangles A in region 1 (foregone cures) and B in region 2 (extra consumption costing more than the incremental value produced).[19] Expanding the number of regions observed would lead to a distribution of rates of use of the therapy similar to the regional variations shown in Table 1.

Continuing with this approach, one can add up all the welfare losses like A and B from incomplete information about the production function, so

$$\mathrm{WL} = \frac{1}{2} \sum_{i=1,N} (X_i - \mu)^2 \Delta P_i, \tag{6a}$$

[18] The approach followed here was first used by Peltzman in his study of the efficacy of the 1962 amendments to the FDA authorization and their effect on consumer well being. See Peltman (1973), Peltzman (1975), and McGuire, Nelson and Spavins (1975).

[19] This analysis bypasses an extensive discussion about the propriety of using triangles from Marshallian demand curves to approximate either compensating or equivalent variation. Harberger (1971) first proposed this approach, but a number of concerns were later expressed, including the importance of income effects and the crucial dependence on the path of price increases when multiple prices change at the same time. Willig (1976) showed that one could bound the error from using such measures in many cases (particularly when only one price changed). McKenzie and Pearce (1982) proposed a "money metric" measure, deriving a Taylor Series expansion from a general utility function for the equivalent variation. This approach has several important advantages, including that it can be expressed to any desired degree of accuracy supported by estimation of relevant demand curves, and the calculation is completely independent of the path of price changes. The welfare triangle shown here approximates the second order Taylor series measure proposed by McKenzie and Pearce except for income effects that can be shown to be quite small empirically in the settings discussed here.

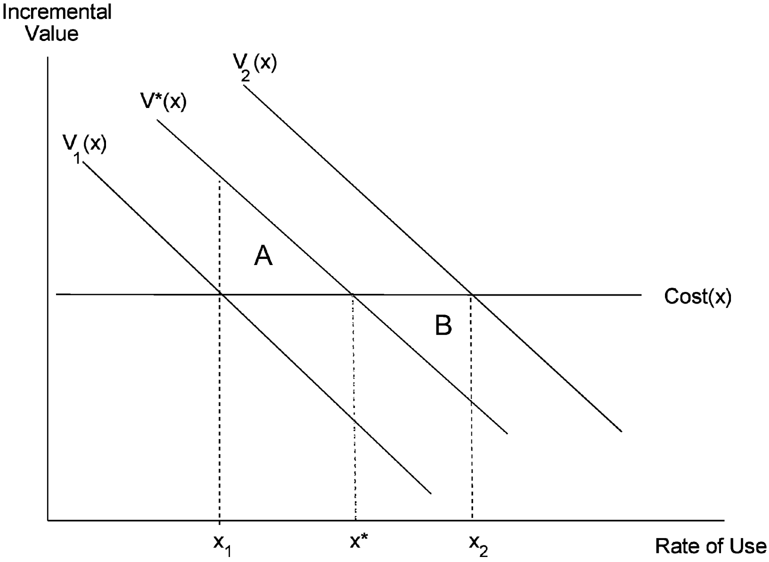


Figure 6. Welfare loss analysis.

which becomes (assuming parallel and straight demand curves for algebraic simplicity)

$$\mathrm{WL} = \frac{1}{2} \sum_{i=1,N} (X_i - \mu)^2 \mathrm{d}P/\mathrm{d}X. \tag{6b}$$

A bit of algebraic manipulation converts this to

$$\mathrm{WL} = \frac{1}{2} \sum_{i=1,N} P_i X_i \mathrm{COV}(X)^2/\eta, \tag{6c}$$

where η is the demand elasticity. This is a simple idea: welfare loss increases directly with total spending on the treatment, directly with the *square* of the coefficient of variation in actual patterns of use, and inversely with the demand elasticity. (The latter phenomenon simply expresses the notion that the welfare triangles like A and B are larger for any given ΔX the steeper are the demand curves.) Table 4 shows the economic losses associated with variability for a variety of procedures carried out in the hospital entirely, calculated from the COV data in Table 1 and extrapolated to national levels from the New York population with which the original spending and variability data were constructed, and converted to $US for year 2000 using the general CPI.

These calculations must be taken with a large dose of caution: In order for society to capture these losses, it would be necessary not only to determine the correct indications for using every medical intervention (including, most desirably, taking patient

Table 4
Annual welfare loss (year 2000 in $US) from medical practice variations when average rate is correct vs. universal adaptation of best practice*

Procedure	Per capita loss ($US)	Aggregate annual loss ($US billion)
Coronary bypass procedures	4.84	1.31
Psychosis	4.63	1.25
Circulatory disorders except AMI with cardiac characterization	2.90	0.78
Chronic obstructive pulmonary disease	2.67	0.72
Angina pectoris	2.07	0.56
Adult gastroenteritis	1.79	0.48
Adult pneumonia	1.78	0.48
Alcohol and drug use	1.68	0.45
Major joint replacement	1.61	0.43
Back and neck procedures	1.45	0.39
Chemotherapy	1.29	0.35
Depressive neurosis	1.26	0.34
Extracranial vascular problems	1.19	0.32
Medical back problems	1.09	0.29
Pediatric pneumonia	1.09	0.29
Cardiac valve procedures	1.06	0.29
Adult bronchitis and asthma	1.01	0.27
Heart failure and shock	1.00	0.27
Acute myocardial infarction (AMI)	0.92	0.25
Pacemaker procedures	0.89	0.24
Respiratory infections/inflammations	0.89	0.24
Infection disease diagnoses	0.88	0.24
Pediatric bronchitis and asthma	0.81	0.22
Cardiac arrythmias	0.81	0.22
Prostatectomy	0.76	0.21
Total, top 25 hospital categories	$40.42	$10.9 billion

Source: Phelps and Mooney (1992), with per capita losses corrected from 1986 to 2000 prices using CPI factor of 1.56 and an aggregate US population of 270 million (usedin final column calculations).

* The welfare loss increases if the current average rate is biased. See Phelps and Parente (1990) and Phelps and Mooney (1992) for details. Welfare loss is overstated to the extent that "best practice" would allow for variation from mean.

preferences into account at the individual level, although we have no reason to believe that such a practice would alter the average rate of use), but it would also be necessary to disseminate the information in a way that was both accessible and credible to every patient and doctor in the country. These are strong requirements, to be sure. It is probably best to think of these calculations as a measure of the magnitude of a problem for which some portion of the welfare loss can be recaptured through study and intervention, and then to compare the potential gains from capturing (say) 10% of these welfare losses (perhaps through improved treatment protocols) with the costs of achieving those gains.

It is also important to remember that the correct "variability" to use in such calculations is the residual after known characteristics of the relevant populations are taken into account, including such obvious factors as age and gender (virtually every study of regional variations uses epidemiologic methods to standardize for age and gender differences across regions), income and price (including insurance coverage), and – much more difficult to obtain in practice – data showing underlying differences in health risks.

The latter issue may never be resolved well, since the relevant health risk will differ procedure by procedure. The risks might be easier to measure if the risk directly relates to an immediate biological or environmental hazard, but most human illnesses arise more from consumption patterns that cumulate for many years. Changing consumption patterns *through time* would then have to be measured at the appropriate regional level (for example, counties in NY state) in order to calculate the potential illness burden. Migration of people across regions would confuse and blur such measures. To consider a simple example, one cannot even meaningfully measure tobacco consumption by county in NY state for a single point in time. State tax levies provide a possible source of data, but the taxes are paid at the wholesale level, with regional distributors supplying many counties. Similarly, it would be virtually impossible to measure dietary cholesterol consumption of populations by county at any single point in time, let alone to measure their cumulative consumption through a number of years, as would be relevant for understanding regional differences in heart disease.

What if the overall average rate of use is incorrect? It is easy to show that the welfare loss increases with the square of the percentage change in the bias of average use, relative to the "correct" rate of use, in addition to the variability around the mean as shown in Table 1. The calculation rests on a phenomenon well known in econometrics, namely that the mean squared error of a biased estimator contains two sources of error – variability about the mean and the squared bias. The calculation in the case of welfare loss for biased rates of treatment is similar, leading to the addition of the following term to the expression found in Equation (5):

$$\text{Additional WL} = \tfrac{1}{2}(\%\text{bias})^2/\eta. \tag{7}$$

The additional welfare losses arise whether the bias is towards overuse or underuse, for similar reasons to the calculation of welfare loss triangles A and B in Figure 6.

In terms of whether we can anticipate biases in the aggregate rate of treatment, the incentives provided by various reimbursement health insurance programs lead towards overuse, while in a capitation-payment system, the incentives lead toward under-use [Woodward and Warren-Boulten (1984)]. An interesting randomized controlled trial of this phenomenon in the realm of pediatric well-care treatment [Hickson, Altmeier and Perrin (1987)] showed the extent to which these predictions are realized. In that study, doctors in a clinic were randomized to a payment scheme (fee for service or flat salary) and patients were randomized to doctor. The standard of "appropriate" care in this study was the recommendations of the American Academy of Pediatrics (AAP) program for well-care visits for healthy children (including routine examinations, vaccinations, etc.). Patients of doctors in the fee for service system received about one more visit annually than those of doctors receiving salary compensation, almost all the extra visits being in the realm of well care (vs. acute illness treatment). Interestingly, doctors on fee for service "over-treated" 22% of their patients (vs. the AAP standard) while at the same time doctors on the salary system did so for only 4% of their patients (vs. the same standard).

Even given the *caveats* about interpretation, these welfare losses are large by any relevant comparison. The most prominent welfare loss discussed in the health economics literature arises from the increased use of medical care due to incentives from insurance arrangements [Arrow (1963), Pauly (1968), Zeckhauser (1970)]. The magnitude of equilibrium welfare losses arising from common health insurance contracts has been estimated by Keeler et al. (1988) using data from the RAND Health Insurance Experiment. On a per capita basis, they estimated that in 1986, an uninsured person would bear approximately $1500 of welfare loss due to risk bearing.[20] A fully insured person would have a welfare loss of $265, all due to "moral hazard." Efficient insurance plans create a net welfare loss of only $50 in their study, minimizing the total losses arising from risk bearing and moral hazard. By contrast, the Phelps and Mooney (1993) estimates of welfare loss from variations is $130 per person, and this only accounts for variations due to hospital admissions, omitting many other sources of welfare loss such as within-hospitalization treatment, out of hospital treatment, etc. Thus, by any meaningful comparison, the welfare loss from medical practice variations is large indeed.

It is worth pointing out that the Arrow/Pauly type of welfare loss has the same problem as the welfare loss calculated from medical practice variations: one cannot meaningfully recapture that loss with currently available mechanisms for insuring against the financial risks of illness. The best mechanisms we know to insure against that risk lead to the welfare losses from excessive use of medical care. If we could somehow magically determine the exact state of illness any individual person had at any moment, then one could conceive of a state-dependent insurance policy that simply transferred income to the individual when a poor health outcome occurred [in the types of insurance plans discussed by Hirshleifer (1966) and Ehrlich and Becker (1972)]. But we do not have such a capability, and hence we do not have such insurance plans. The welfare loss calculations that Arrow (1963) and Pauly's work (1968) motivated simply provide an estimate of the value to society of achieving such a capability. Their insight (and also that provided by Zeckhauser (1970), in building from their work) also provides a basis for thinking about how to minimize that aggregate welfare loss through intelligent choice of insurance parameters.

[20] Their estimate uses a risk aversion parameter that is probably 5 times too large. The implicit relative risk aversion measure in their model is approximately 10, whereas literature estimating this key parameter [see Garber and Phelps (1997)] places the value in the range of 1 to 4, centering on about 2. Thus the welfare loss from risk bearing cited above may be better approximated by something near $300, rather than $1500.

4. Production and dissemination of information

4.1. Property rights to drugs, devices, and ideas

The production of information about medical treatments occurs very differently depending on whether the treatment is manufactured (such as a prescription drug or medical device) or whether it is a "strategy" for treating patients (that may or may not use a proprietary drug or device). Most of the variations discussed in Section 3 arise because of different strategies for treatment being adopted by different health care providers. Thus it is pertinent to consider the economic aspects for producing and disseminating information about these two types of medical interventions.

The case of medical drugs and devices is quite simple – normal patent protection usually applies to such manufactured products. (In the US and elsewhere, there are are additional complications from drug regulation, but these do not alter the basic fact that drugs and medical devices receive patent protection.) Thus, inventors of (say) a new prescription drug that improves treatment of some disease have considerable economic incentive not only to create information about the drug's efficacy, but also to invest considerable resources in disseminating that information. Indeed, drug companies undertake both such activities vigorously. While drug regulation often impinges on the production of information (specifying the amount and nature of research required to demonstrate the drug's safety and efficacy), there is good reason to believe that much of such research would be carried out even without such regulation. In the realm of dissemination of information, we know that drug companies spend immense resources for advertising, not only to physicians, but now increasingly in direct mass advertising to consumers on radio, television, print media, and on the World Wide Web. The most prestigious medical journals in the world often contain more pages devoted to drug advertising than to actual academic manuscripts. Drug companies hire apparently endless hordes of sales agents (called "detail men") to spend time with individual physicians describing the value of using the drugs manufactured by their companies, and one would be hard put to find a physician in the US who did not have multiple memorabilia from such visits – pens, coffee cups, note and prescription pads, and the like.

Drug companies also confront the harsh reality of legal liability if their product harms people in a negligent manner. (Medico-legal liability is one reason why we could expect drug companies to carry out considerable research about a drug's safety and efficacy even if the Food and Drug Administration in the US did not exist.) This provides strong economic incentive to be certain about the drug's characteristics before it is marketed [Danzon (1983)].

In stark contrast, there are only small economic incentives arising from the production of new "strategies" about treatment, and even smaller incentives to invest in the dissemination of such strategies. The reason lies in the failure of most modern laws to define meaningful property rights to a medical treatment strategy. A doctor (or group of doctors) can devise a potentially improved strategy to use in their own practice (and many do) to replace previously used strategies, but the "inventor" has no way to gain the full economic leverage from such an innovation that a drug or device manufacturer has. One cannot patent "strategies," and indeed, the normal practice in most western countries is to publish the strategy in a medical journal, freely available to all who wish to use it. The doctor – most commonly the member of the faculty of a medical school or research institute – gains indirect economic benefit from the "fame" of the publication, and (in US medical schools in particular) the publication of academic manuscripts is a prerequisite for promotion and tenure. But once having published the manuscript, the inventor is very unlikely to expend additional resources to disseminate the information included in the journal article, quite unlike the world of drugs and devices.

The contrast with large manufacturing firms or retail chains is important to understand. In either of those settings, a process improvement (also not patentable) can be exploited either through modification of manufacturing activities within the firm or even in dispersed retail outlets (e.g., McDonalds, Firestone Tires) through employee training, and the benefits captured in lower production costs or higher product quality. The key in the ability to capture the rewards of these innovations centers on having a high volume of repeated activity within the same organization. In the practice of medical care, even the largest medical group will seldom, if ever, achieve these kinds of volume of repeat treatments. (Recall here the very large array of diseases that doctors must treat, drawing on literally thousands of potential treatments.)

The inability for medical doctors to capture gains from process innovation is likely compounded by legal restrictions on medical practice organizational forms. In the US and elsewhere, strong prohibitions against "corporate practice of medicine" were built into nearly every professional practice law. These laws inhibit the development of such things as franchises, organizations that (in concept) could grow sufficiently large to allow them to capture the gains from process innovation internally. Examples in other service industries of this sort of approach include H&R Block's income tax service, McDonald's fast-food chains (which have a highly detailed procedure manual and even a "McDonald's University" to train employees).

Inventors of new strategies also face no liability for creating a bad strategy. Only the doctor who uses the strategy (on a case by case basis) has any potential liability for harm that comes from using a bad strategy. Thus, the inventor of an improved medical strategy not only fails to receive the same economic benefits as would arise from a patentable invention, but also does not confront the incentives for providing a safe product that liability law creates.

4.2. Costs of production of information

If one returned to the list of medical interventions shown in Table 1, or any much wider list, one would find that a surprisingly small fraction of those interventions had ever been tested for effectiveness using even modestly appropriate scientific methodologies. Often, "improvements" in medical practice are adopted on the basis of a study involving only a few patients,[21] commonly using historical rather than concurrent controls, and

[21] One joke summarizes the results of an animal research study in a medical journal as follows: "One third of our subjects were cured, one third died, and the other one escaped."

often with the patients' outcomes evaluated by the doctor conducting the procedure. The opportunities for bias and incorrect conclusions are numerous in this setting – too numerous to detail here.[22]

The problem is essentially one of the difficulty in obtaining adequate power to distinguish the differential effects of one treatment over another. Suppose, for example, that the existing treatment for a disease has a probability of success of 0.7, and we wish to test whether a new treatment has better outcomes. If we use a statistical criterion of (say) $\alpha = 0.05$ (two tailed, since we do now know which treatment is better), then to establish a power of $\beta = 0.8$, we need a sample of approximately 650 cases *for each treatment* to detect a 10% improvement in outcomes (from 0.7 to 0.77 probability of cure).[23] This means that a study to learn about the treatment effects must enroll 1300 patients, randomize them to one of the alternative therapies, and then observe their outcomes for a sufficiently long time to determine ultimate outcomes. (If we are dealing, for example, with many cancer treatments, five years or more are necessary to determine eventual outcome differences.) Unfortunately for the progress of science, most diseases are so rare that this is exceedingly difficult, if not impossible, to accomplish within a few years time.

The problem is of course easier if one seeks only to detect a more substantial improvement in treatment efficacy. If we seek to detect an improvement from a 50% cure rate to a 75% cure rate, then the required sample size falls to 65 cases per arm, a much more manageable problem. The question then turns on the size of the incremental steps in therapeutic efficacy that one can expect, and the nature of the scientific studies necessary to detect relevant changes.

The magnitude of this problem is staggering. The NIH maintains a Web site (www.cancernet.nci.nih.gov/ord/) for "rare diseases" (those with fewer prevalence of less than 200,000 in the entire US population), and has 6,000 such diseases in their data base currently. "Prevalence" measures the number of people who live with the disease. For chronic diseases, the number of new cases annually ("incidence") can be well under ten percent of the prevalence rate. So consider a typical disease in this data base, with prevalence of (say) 75,000 and incidence (new case rate) of 5,000. (This means either that people are cured or die within a 15 year period on average.)

Assume further that these diseases are treated *only* at major medical centers, which number about 500 in the United States (so each would serve a population of approximately 500,000 persons).[24] This means that the average number of cases treated per medical center is about 10. Assembling 1300 cases (see above discussion on power and

[22] Entire courses in graduate programs are organized around this problem, usually titled something such as "clinical epidemiology" or "clinical evaluative sciences." The interested student should venture into medical or public health schools to find such programs of study.

[23] The statistics of these problems are well discussed in Fleiss (1984).

[24] This same calculation illustrates the difficulty for single insurance companies to carry out such a study. Only a few insurance carriers have 500,000 or more individuals enrolled in their plans, so each insurance carrier confronts similar problems to individual medical centers in collecting sufficient numbers of patients to carry out appropriate medical outcomes studies.

sample size) would require the collaboration of 130 such centers for a year, or 130 years for one center, or combinations thereof, *assuming 100% enrollment rate of all eligible patients*. Since there are often biological differences in treatment response by age and gender (or at least the risk of such), these sample criteria must often apply to subsets of the entire population, compounding the enrollment problem. (For example, if we create groups of "young vs. old" and "male vs. female" to study, then the sampling problems increase by a factor of four.) The problem is further magnified if more than one new or potentially improved therapy exists or emerges during an ongoing randomized controlled trial, since subjects enrolled in one study are not eligible (for obvious reasons) to participate in another.[25]

Randomized controlled trials are expensive to carry out because of the size, complexity, and duration necessary to conduct them appropriately, often running in the range of $5–10 million or more. The NIH annual extramural budget of approximately $15 billion is heavily devoted to the production of new basic science knowledge, with about one eighth of the annual budget historically allocated to clinical trials. At that rate (approximately $2 billion a year), at an average cost of $5–10 million, it would take 15–30 years of current-rate NIH funding to support a single clinical trial on each of the 6000 NIH-identified rare diseases. Of course, the production of new scientific knowledge makes it impossible to "keep up" with new potential therapies at this rate, since the underlying rate of technical change leads to new innovations in therapies at a far faster rate. Even with the currently large spending rates on clinical trials by the federal government, there is literally no hope that we can come close to funding valid scientific studies to determine the outcomes of treatments for "all" diseases of potential interest, or even a small fraction thereof.

The private sector cannot be relied upon independently to produce the knowledge necessary to understand treatment outcomes. The issue is not so much the magnitude of spending (indeed, $1 billion a year is a relatively small amount for private sector investment), but rather the public good nature of the problem. As discussed before, property rights to discoveries in the realm of clinical science are weak to non-existent. Further, even large managed care organizations (insurance plans, HMOs, etc.) cannot fully capture the gains from such studies internally. To see this, suppose that a very large HMO decided to fund a series of clinical studies to improve patient outcomes, and advertised that fact to attract patients. The doctors treating patients in that setting would of course have to have access to the new information (improved treatment protocols, etc.) and nothing prevents them from moving to another setting and using the same information. Even if "trade secret" language could be invoked, it would be so alien to the training and culture of physicians that it would be difficult to keep any information intact; the *modus operandi* of physicians is to share information about things that improve patients' outcomes.

The obvious role of the government here is to support, either directly or through subsidies, studies to carry out this work. They do now, through the NIH clinical trials

[25] Their outcome would be under-identified in the econometric sense.

monies and through the Agency for Health Care Research and Quality (AHRQ), but the magnitude of the problem overwhelms the available funding (and the available stock of trained researchers).

How can society determine the proper investment in such efforts? Two approaches both provide direct evidence that the government's current effort represents a considerable under-investment in this field of study. First, the welfare loss calculations related to medical practice variations (see Table 4) show that improvement in the knowledge base in these areas has benefits in terms of reduced variations that far outweigh the costs of carrying out relevant studies in even very rare diseases. To see this, recall that the knowledge produced in a single year has many years of benefit. For purposes of discussion, suppose that the present value of the knowledge is ten times the annual benefit, and that the relevant cost of a study is \$10 million. Thus, any medical activity with an annual welfare loss of \$1 million or more represents a case where the improvement in knowledge has the possibility of paying off. Of course, no study will totally eliminate unwarranted variations in practice patterns, but other gains in welfare appear in addition to the reduction in variance. The most obvious of these arises from the potential for bias in the average rate of treatment, where we know that the incentives for using medical care with traditional health insurance push towards excessive use.

What does it take to produce a welfare loss of \$1 million annually? With a demand elasticity of about -0.15 and a COV of a very modest level of 0.2, the necessary spending rate (using Equation (1)) is about \$8 million per year, or about \$0.03 per person per year in the United States. If the COV rises to 0.4, the spending rate drops to about \$2 million per year, and the requisite annual spending drops below \$1 million for COV in excess of 0.55. Even the rarest of rare diseases cataloged by the NIH are likely to fall into this realm.

Quite separately, a distinctive study by an economist/MD estimated the cost-effectiveness of conducting clinical studies as a way of improving health outcomes [Detsky (1989, 1990)]. He showed that the cost per life year saved from conducting randomized controlled trials ranged from \$2–3 per life year saved to a high (in the set of treatments he investigated) of \$400–700. Many medical interventions have CE ratios of \$20,000 to \$50,000 or more, and recent studies put the willingness to pay for improved life expectancy in the \$25,000 to \$100,000 range [Garber and Phelps (1997)].

These data support strongly the conclusion that our society under-invests in the production and dissemination of new information about the efficacy of various medical interventions. The private sector cannot solve these problems because of the public good nature of this information, and the failure of property rights to create strong incentives to produce such information privately.

References

Arrow, K.J. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53(5):941–973.

American Medical Association (1990), Current Procedural Terminology (CPT), 4th edn (The American Medical Association, Chicago).

Banerjee, A.V. (1992), "A simple model of herd behavior", Quarterly Journal of Economics 107(3):797–817.

Benham, L. (1972), "The effect of advertising on the price of eyeglasses", Journal of Law and Economics 15:337–352.

Bikhchandani, S., D. Hirshleifer and I. Welch (1992), "A theory of fads, fashion, custom, and cultural exchange as informational cascades", Journal of Political Economy 100(5):992–1026.

Braverman, A. (1980), "Consumer search and alternative market equilibria", Review of Economic Studies 47:487–502.

Cady, J. (1976), "An estimate of the price effects of restrictions on drug price advertising", Economic Inquiry 14:493–510.

Cahal, M.F. (1962), "What the public thinks of the family doctor – folklore and facts", GP 25:146–157.

Cutler, D., M. McClellan and J.P. Newhouse (1999), "Quality-adjusted price increases for treating AMI", in: J.E. Triplett, ed., Measuring the Prices of Medical Treatments (Brookings Institution Press, Washington, DC).

Danzon, P.M. (1983), "An economic analysis of the medical malpractice system", Behavioral Sciences and the Law 1(1):39–54.

Department of Health and Human Services (1980), International Classification of Diseases (ICD-9-CM), 2nd edn. (US Government Printing Office (PHS)-80-1260, Washington, DC).

Detsky, A.S. (1989), "Are clinical trials a cost-effective investment?", Journal of the American Medical Association 262(13):1795–1800.

Detsky, A.S. (1990), "Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials", Statistics in Medicine 9:173–183.

Diehr, P., K.C. Cain, W. Kreuter and S. Rosenkranz (1992), "Can small area analysis detect variation in surgery rates? The power of small area variations analysis", Medical Care 30(6):484–502.

Dionne, G. (1984), "Search and insurance", International Economic Review 25(2):357.

Ehrlich, I., and G.S. Becker (1972), "Market insurance, self-insurance, and self-protection", Journal of Political Economy 80(4):623–648.

Feldman, R., and J.W. Begun (1978), "The effect of advertising: lessons from optometry", Journal of Human Resources 13(Suppl.):253–262.

Fleiss, J. (1984), Statistical Methods for Rates and Proportions, 2nd edn. (Wiley and Sons, New York).

Garber, A.M., and C.E. Phelps (1997), "The economic foundations of cost-effectiveness analysis", Journal of Health Economics 16(1):1–31.

Glover, A.F. (1938), "The incidence of tonsillectomy in school children", Proceedings of the Royal Society of Medicine 31:1219–1236.

Gray, P.G., and A. Cartwright (1953), "Choosing and changing doctors", Lancet 2:1308–1309.

Grossman, M. (1972a), "On the concept of health capital and the demand for health", Journal of Political Economy 80(2):223–255.

Grossman, M. (1972b), The Demand for Health: A Theoretical and Empirical Investigation (Columbia University Press, New York).

Grossman, M. (2000), "The human capital model", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 7.

Harberger, A.C. (1971), "Three basic postulates for applied welfare economics: an interpretive essay", Journal of Economic Literature 9(3):22–243.

Hey, J.D. (1979), "A simple generalized stopping rule", Economic Letters 2:115–120.

Hey, J.D. (1981), Economics in Disequilibrium (Martin Robertson, Oxford).

Hey, J.D., and C.J. McKenna (1981), "Consumer search with uncertain product quality", Journal of Political Economy 89(1):54–66.

Hickson, G.B., W.A. Altmeier and J.M. Perrin (1987), "Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study", Pediatrics 80(3):344–350.

Hirshleifer, J. (1966), "Investment under uncertainty: applications of the state-preference approach", Quarterly Journal of Economics 80(1966):252–277.

Hu, T.Y. (1996), Essays on Cardiologists' Behavior: Medical Variations, Impact of PSROs, and Response to PPS, Doctoral dissertation (University of Rochester, Department of Economics).

Kandori, M., G.J. Mailath and R. Rob (1993), "Learning, mutation, and long run equilibria in games", Econometrica 61(1):29–56.

Keeler, E.B., J.L. Buchanan, J.E. Rolph et al. (1988), "The demand for episodes of treatment in the health insurance experiment", Report R-3454-HHS (RAND Corporation, Santa Monica, CA).

Kwoka, J.E. (1984), "Advertising and the price and quality of optometric services", American Economic Review 74:211–216.

Leape, L.L., R.E. Park, D.H. Solomon et al. (1990), "Does inappropriate use explain small area variations in the use of health care services?", Journal of the American Medical Association 263(5):669–672.

Lippman, S.A., and J.J. McCall (1976), "The economics of job search: a survey, Part 1", Economic Inquiry 14(2):155–189.

Lippman, S.A., and J.J. McCall, eds. (1979), Studies in the Economics of Search (North-Holland, Amsterdam).

Manning, W.G., J.P. Newhouse and J.E. Ware (1981), "The status of health in demand estimation: beyond excellent, good, fair, and poor", RAND Report 2696-1-HHS (RAND Corporation, Santa Monica, CA).

Marquis, M.S. (1985), "Cost sharing and the patient's choice of provider", Journal of Health Economics.

McKenna, C.J. (1986), "Theories of individual searching behavior", Bulletin of Economic Research 38(3):189–207.

McKenzie, G.W., and I.F. Pearce (1982), "Welfare analysis: a synthesis", American Economic Review 72(4):669–682.

McGuire, T.G. (2000), "Physician agency", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 9.

McGuire, T., R. Nelson and T. Spavins (1975), "An evaluation of consumer protection legislation: the 1962 drug amendments, comment", Journal of Political Economy 83(3):655–662.

McPherson, K., P.M. Strong, A. Epstein and L. Jones (1981), "Regional variations in the use of common surgical procedures: within and between England and Wales, Canada, and the United States", Social Science in Medicine 15A:273–288.

McPherson, K., J.E. Wennberg, O.B. Hovind and P. Clifford (1982), "Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway", New England Journal of Medicine 307(21):1310–1314.

Newhouse, J.P., and C.E. Phelps (1976), "New estimates of price and income elasticities of demand for medical care services", in: R.N. Rosett, ed., The Role of Health Insurance in the Health Services Sector (National Bureau of Economic Research, New York) 261–313.

Newhouse, J.P., and the RAND Health Insurance Experiment Group (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge, MA).

Olsen, D.M., R.L. Kane and J. Kasteler (1976), "Medical care as a commodity: an exploration of the shopping behavior of patients", Journal of Community Health 2(2):85–91.

Pauly, M.V. (1968), "The economics of moral hazard: comment", American Economic Review 58(3):531–537.

Peltzman, S. (1973), "An evaluation of consumer protection legislation: the 1962 drug amendments", Journal of Political Economy 81(5):1049–1091.

Peltzman, S. (1975), "An evaluation of consumer protection legislation: the 1962 drug amendments, reply", Journal of Political Economy 83(3):663–665.

Perkins, N.A.K. (1991), Case Mix Adjustment and Physician Practice Profiling: Development of a Methodology to Solve Economic and Political Problems in Medicare Part B, Doctoral dissertation (University of Rochester, Department of Political Science).

Phelps, C.E. (1997), Health Economics (Addison-Wesley, Longman, Reading, MA).

Phelps, C.E., and C. Mooney (1992), "Priority setting in medical technology and medical practice assessment: correction and update", Medical Care 31(8):744–751.

Phelps, C.E., and C. Mooney (1993), "Variations in medical practice use: causes and consequences", in: R.J. Arnould, R.F. Rich and W. White, eds., Competitive Approaches to Health Care Reform (The Urban Institute Press, Washington, DC).

Phelps, C.E., and S.T. Parente (1990), "Priority setting in medical technology and medical practice assessment", Medical Care 29(8):703–723.

Phelps, C.E., C. Mooney, A.I. Mushlin and N.A.K. Perkins (1992), Doctors Have Styles, and They Matter! (Department of Community and Preventive Medicine, University of Rochester, Rochester, NY).

Phelps, C.E., and J.P. Newhouse (1974), "Coinsurance and the demand for medical care", Review of Economics and Statistics 56(3):334–342.

Phelps, C.E. (1973), "The demand for health insurance: a theoretical and empirical investigation", Report R-1054-OEO (RAND Corporation, Santa Monica, CA).

Pratt, J.W. (1964), "Risk aversion in the small and in the large", Econometrica 32(1/2):122–136.

Pratt, J.W., D.A. Wise and R. Zeckhauser (1979), "Price differences in almost competitive markets", Quarterly Journal of Economics 93:189–211.

Rochaix, L. (1989), "Information asymmetry and search in the market for physicians' services", Journal of Health Economics 8:53–84.

Roos, N.P., L.L. Roos and P.D. Henteleff (1977), "Elective surgical rates – do high rates mean lower standards? Tonsillectomy and adenoidectomy in Manitoba", New England Journal of Medicine 297:360–365.

Roos, N.P., and L.L. Roos (1982), "Surgical rate variations: do they reflect the health or socioeconomic characteristics of the population?", Medical Care 20(9):945–958.

Sadanand, A., and L.L. Wilde (1982), "A generalized model of pricing for homogeneous goods under imperfect information", Review of Economic Studies 49:229–240.

Salop, S., and J. Stiglitz (1977), "Bargains and ripoffs: a model of monopolistically competitive price dispersion", Review of Economic Studies 44:493–510.

Schwartz, A., and L.L. Wilde (1982), "Competitive equilibria in markets for heterogeneous goods under imperfect information: a theoretical analysis with policy implications", Bell Journal of Economics 13(1):181–193.

Stano, M., and S. Folland (1988), "Variations in the use of physician services by Medicare beneficiaries", Health Care Financing Review 9(3):51–57.

Starfield, B.H., J.P. Weiner, L. Mumford and D.M. Steinwachs (1991), "Ambulatory care groups: a categorization of diagnoses for research and management", Health Services Research 26:53–74.

Stigler, G. (1961), "The economics of information", Journal of Political Economy 69:213–225.

Weiner, J.P., B.H. Starfield, D.M. Steinwachs and L.M. Mumford (1991), "Development and application of a population-oriented measure of ambulatory care case-mix", Medical Care 29:452–472.

Wennberg, J.E., J.L. Freeman and W.J. Culp (1987), "Are hospital services rationed in New Haven or overutilised in Boston?", Lancet 1:1185–1188.

Wennberg, J.E., J.L. Freeman, R.M. Shelton and T.A. Bubolzt (1989), "Hospital use and mortality among Medicare beneficiaries in Boston and New Haven", New England Journal of Medicine 321:1168–73.

Wennberg, J., and A. Gittelsohn (1973), "Small area variations in health care delivery", Science 182:1102–1108.

Wennberg, J., and A. Gittelsohn (1982), "Variations in medical care among small areas", Scientific American 246(4):120–134.

Wilde, L., and A. Schwartz (1979), "Equilibrium comparison shopping", Review of Economic Studies 46:543–553.

Willig, R.D. (1976), "Consumer's surplus without apology", American Economic Review 66:5889–597.

Woodward, R.S., and F. Warren-Boulton (1984), "Considering the effect of financial incentives and professional ethics on 'appropriate' medical care", Journal of Health Economics 3(3):223–237.

Young, P. (1993), "The evolution of conventions", Econometrica 61(1):57–84.

Zeckhauser, R. (1970), "Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives", Journal of Economic Theory 2(1):10–26.

Chapter 6

HEALTH ECONOMETRICS*

ANDREW M. JONES

Department of Economics and Related Studies, University of York

Contents

*The text of this chapter draws on joint work with Vassilios Aletras, Paul Contoyannis, Alan Duncan, Martin Forster, Rob Manning, Nigel Rice, Matt Sutton, and Steven Yen. Also it has benefited from the suggestions made by Will Manning and Edward Norton in their detailed discussions of an earlier version. I am grateful for valuable suggestions and comments from Ignacio Abasolo, Tom Buchmueller, Tony Culyer, Eddy van Doorslaer, Eric French, Marty Gaynor, Antonio Giuffrida, Julie Glanville, Paul Grootendorst, Michael Grossman, Don Kenkel, Sylvia Kuo, Angel López, John Mullahy, Christophe Muller, Joe Newhouse, Owen O'Donnell, Carol Propper, João Santos Silva, Chris Seplaki, Jim Sears, Harri Sintonen, Frank Sloan, Belinda South, Andrew Street, Joe Terza, Volker Ulrich, Dave Vaness, John Wildman, and Amy Wolaver.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

A decade ago, Newhouse (1987) assessed the balance of trade between imports from the econometrics literature into health economics, and exports from health economics to a wider audience. While it is undoubtedly true that imports of concepts and techniques still dominate the balance, the literature reviewed in this chapter shows that the range and volume of applied econometric work in health economics has increased dramatically over the past ten years.

Examples of good practice in health econometrics make extensive use of tests for misspecification and explicit model selection criteria. Robust and distribution-free estimators are of increasing importance, and the chapter gives examples of nonparametric, and semiparametric estimators applied to sample selection, simultaneous equations, count data, and survival models.

Published replications of empirical results remain relatively rare. One way in which this deficit may be remedied is through the appearance of more systematic reviews of econometric studies. The use of experimental data remains an exception and most applied studies continue to rely on observational data from secondary sources. However applied work in health economics is likely to be influenced by the debate concerning the use of data from social experiments.

The chapter illustrates the impressive diversity of applied econometric work over the past decade. Most of the studies reviewed here use individual level data and this has led to the use of a wide range of nonlinear models, including qualitative and limited dependent variables, along with count, survival and frontier models. Because of the widespread use of observational data, particular attention has gone into dealing with problems of self-selection and heterogeneity bias. This is likely to continue in the future, with the emphasis on robust estimators applied to longitudinal and other complex datasets.

JEL classification: C0, I1

1. Introduction

A decade ago, Newhouse (1987) assessed the balance of trade between imports from the econometrics literature into health economics, and exports from health economics to a wider audience. While it is undoubtedly true that imports of concepts and techniques still dominate the balance, the literature reviewed in this chapter shows that the range and volume of applied econometric work in health economics has increased dramatically over the past ten years. What is more, the prevalence of latent variables, unobservable heterogeneity, and nonlinear models make health economics a particularly rich area for applied econometrics.

The chapter is not a systematic review. Instead, it attempts to provide an overview of the econometric methods that have been applied in health economics, and to use a broad range of examples to illustrate their use. The emphasis of the chapter is on the use of individual level data and microeconometric techniques, reflecting the emphasis on microeconomic analysis in health economics generally. The majority of aggregate analyses have used international data and the methodological issues surrounding international comparisons of health care are discussed by Gerdtham and Jonsson (2000) in this Handbook.

The structure of the chapter is organized around the nature of the data to be analyzed and, in particular, the way in which the dependent variable is defined and measured. This puts the emphasis on the specification of models and appropriate methods of estimation. But the emphasis on estimation should not imply a neglect of checks for model misspecification. Although the chapter does not include a separate section devoted to measures of goodness of fit, tests for misspecification, and criteria for comparing and selecting models, examples of the use of diagnostic tests are given throughout the text. The scope of the chapter also limits detailed discussion of the practical problems encountered in working with health data. These include issues such as non-response and attrition, measurement error, the use of proxy variables, missing values and imputation, and problems with self-reported data such as recall and strategic mis-reporting.

2. Identification and estimation

2.1. The evaluation problem

The evaluation problem is whether it is possible to identify causal effects from empirical data. Mullahy and Manning (1996) provide a concise summary of the problem and, while their discussion focuses on clinical trials and cost-effectiveness analysis, the issues are equally relevant for structural econometric models. An understanding of the implications of the evaluation problem for statistical inference will help to provide a motivation for most of the econometric methods discussed in this chapter.

Consider an "outcome" y_{it}, for individual i at time t; for example, an individual's use of primary care services. The problem is to identify the effect of a "treatment", for example, whether the individual has health insurance or not, on the outcome. The causal effect of interest is

$$CE(i,t) = y_{it}^{\mathrm{T}} - y_{it}^{\mathrm{C}}, \tag{1}$$

where T denotes treatment (insurance) and C denotes control (no insurance). The pure causal effect cannot be identified from empirical data because the "counterfactual" can never be observed. The basic problem is that the individual "cannot be in two places at the same time"; that is we cannot observe their use of primary care, at time t, both with and without the influence of insurance.

One response to this problem is to concentrate on the average causal effect

$$ACE(t) = \mathrm{E}\big[y_{it}^{\mathrm{T}} - y_{it}^{\mathrm{C}}\big] \tag{2}$$

and attempt to estimate it with sample data. Here it is helpful to think in terms of estimating a general regression function

$$y = g(x, \mu, \varepsilon), \tag{3}$$

where x is a set of observed covariates, including measures of the treatment, μ represents unobserved covariates, and ε is a random error term reflecting sampling variability. The problem for inference arises if x and μ are correlated and, in particular, if there are unobserved factors that influence whether an individual is selected into the treatment group or how they respond to the treatment. This will lead to biased estimates of the treatment effect.

A randomized experimental design may achieve the desired orthogonality of measured covariates (x) and unobservables (μ); and, in some circumstances, a natural experiment may mimic the features of a controlled experiment [see, e.g., Heckman (1996)]. However, the vast majority of econometric studies rely on observational data gathered in a non-experimental setting. These data are vulnerable to problems of non-random selection and measurement error, which may bias estimates of causal effects.

2.2. Estimation strategies

2.2.1. Estimating treatment effects

In the absence of experimental data attention has to focus on alternative estimation strategies. Mullahy and Manning (1996) identify three common approaches:

(i) Longitudinal data – the availability of panel data, giving repeated measurements for a particular individual, provides the opportunity to control for unobservable individual effects which remain constant over time. The debate over whether to treat these unobservables as fixed or random effects, and methods for estimating both linear and nonlinear panel data models are discussed in Section 6.

(ii) Instrumental variables (IV) – variables (or “instruments”) that are good predictors of the treatment, but are not independently related to the outcome, may be used to purge the bias [see, e.g., McClellan and Newhouse (1997)]. In practice the validity of the IV approach relies on finding appropriate instruments [see, e.g., Bound et al. (1995)]. The use of instrumental variables to deal with heterogeneity and simultaneity bias in both linear and nonlinear models is discussed in Section 5.

(iii) Control function approaches to selection bias – these range from parametric methods such as the Heckit estimator to more recent semiparametric estimators [see, e.g., Vella (1998)]. The use of these techniques in health economics is discussed in Section 4.3.

2.2.2. Model specification and estimation

So far, the discussion has concentrated on the evaluation problem and selection bias. More generally, most econometric work in health economics focuses on the problem of finding an appropriate stochastic model to fit the available data. Estimation of regression functions like Equation (3) typically requires assumptions about the appropriate conditional distribution for the dependent variable and for the functional relationship with one or more covariates. Failure of these assumptions may mean that estimators lose their desired properties and give biased, inconsistent, or inefficient estimates. For this reason attention should be paid to tests for misspecification and robust methods of estimation.

Classical regression analysis assumes that the regression function is linear and that the random error term has a normal distribution

$$y_i = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{4}$$

However, in recent years the econometrics literature has seen an explosion of theoretical developments in nonparametric and semiparametric methods, which relax functional form and distributional assumptions. These are beginning to be used in applied work in health economics. Section 2.3 introduces kernel-based nonparametric estimators, and semiparametric approaches are discussed in Sections 4, 6, and 8.

In health economics empirical analysis is complicated further by the fact that the theoretical models often involve inherently unobservable (latent) concepts such as health endowments, agency and supplier inducement, or quality of life. The problem of latent variables is central to the use of MIMIC models of the demand for health and health status indices (Section 5.1.2); but latent variables are also used to motivate nonlinear models for limited and qualitative dependent variables. The widespread use of individual level survey data means that nonlinear models are common in health economics. Examples include binary responses, such as whether the individual has visited their GP over the previous month (Section 3.1); multinomial responses, such as the choice of provider (Section 3.3); limited dependent variables, such as expenditure on primary care services, which is censored at zero (Section 4); integer counts, such as the number of

GP visits (Section 7); or measures of duration, such as the time elapsed between visits (Section 8).

Maximum likelihood (ML) estimation is widely used in health economics, particularly for nonlinear models involving qualitative or limited dependent variables. ML has desirable properties, such as consistency and asymptotic normality, but these rely on the model being fully and correctly specified. Pseudo (or quasi) maximum likelihood (PML) methods share the properties of ML without having to maintain the assumption that the model is correctly specified [see, e.g., Gourieroux et al. (1984), Gourieroux and Monfort (1993)]. For the class of distributions belonging to the linear exponential family, which includes the binomial, normal, gamma, and Poisson, the PML estimator of the conditional mean is consistent and asymptotically normal. This means that the conditional mean has to be specified but the conditional variance does not. The main use of PML methods in health economics has been in the context of count data regressions (Section 7).

Many of the estimators discussed in this chapter fall within the unifying framework of generalized method of moments (GMM) estimation [see, e.g., Hall (1993)]. This replaces population moment conditions, such as

$$\mathrm{E}\big[f(x,\beta)\big]=0. \tag{5}$$

with their sample analogues

$$m(\beta)=n^{-1}\sum_{i} f(x_i,\beta)=0. \tag{6}$$

The GMM estimator minimizes a quadratic form

$$Q(\beta)=m(\beta)'Wm(\beta), \tag{7}$$

where W is a positive definite matrix, and the optimal W can be selected to give asymptotic efficiency. GMM encompasses many standard estimators. For example, OLS uses the moment conditions $\mathrm{E}[x(y-x\beta)]=0$, instrumental variables with a set of instruments z uses $\mathrm{E}[z(y-x\beta)]=0$, and pseudo maximum likelihood uses $\mathrm{E}[\partial \operatorname{Ln} L/\partial\beta)]=0$. Applications of GMM in health economics are discussed in the context of instrumental variable estimation (Section 5.1.1) and count data models (Section 7.3).

Quantile regression is another semiparametric method which assumes a parametric specification for the qth quantile of the conditional distribution of y,

$$\mathrm{Quantile}_q(y_i \mid x_i)=x_i\beta_q \tag{8}$$

but leaves the error term unspecified [see, e.g., Buchinsky (1998)]. Quantile regression has been applied by Manning et al. (1995) to analyze whether heavy drinkers are more or less responsive to the price of alcohol than other drinkers. They find evidence that the price effect does vary by level of consumption.

Many of the estimators discussed in this chapter rely on the approximation provided by asymptotic theory for their statistical properties. Recent years have seen increasing use of bootstrap methods to deal with cases where the asymptotic theory is intractable or where the asymptotics are known but finite sample properties of an estimator are unknown [see, e.g., Jeong and Maddala (1993)]. The aim of these methods is to reduce bias and to provide more reliable confidence intervals. Bootstrap data are constructed by repeated re-sampling of the estimation data using random sampling with replacement. The bootstrap samples are then used to approximate the sampling distribution of the estimator being used. For example, Nanda (1999) uses bootstrap methods to compute standard errors for two stage instrumental variable estimates in a model of the impact of credit programs on the demand for health care among women in rural Bangladesh.

The growing popularity of bootstrap methods reflects the increased availability of computing power. The same can be said for simulation methods. Monte Carlo simulation techniques can be used to deal with the computational intractability of nonlinear models, such as the multinomial probit, which involve higher order integrals [see, e.g., Hajivassiliou (1993)]. Popular methods include the GHK simulator and Gibbs sampling. These methods can be applied to simulate sample moments, scores, or likelihood functions. Simulation estimates of the multinomial probit are discussed in Section 3.2.4 and estimators for simultaneous equation limited dependent variable models are discussed in Section 5.2.4.

2.2.3. Nonparametric and semiparametric estimators

Most of the estimators discussed in this chapter rely on assumptions about the functional form of the regression equation and the distribution of the error term. However recent developments in the econometrics literature have focused on semiparametric and nonparametric estimation. Many of these are founded on the Rosenblatt–Parzen kernel density estimator. This method uses appropriately weighted local averages to estimate probability density functions of unknown form; in effect using a smoothed histogram to estimate the density. Variants on this basic method of density estimation are also used to estimate distribution functions, regression functionals, and response functions [see, e.g., Ullah (1988), Duncan and Jones (1992)].

The kernel function, $K[\cdot]$, provides the weighting scheme; its bandwidth determines the size of the "window" of observations that are used, and the height of the kernel function gives the weight attached to each observation. This weight varies with the distance between the observation and the point at which the density is being estimated. Consider a random variable with unknown density function $f(x)$. Given a random sample of n observations, the univariate density estimator at a particular point x is

$$f_h(x) = (n \cdot h)^{-1} \sum_i K\big[(x_i - x)/h\big], \tag{9}$$

where $K(\cdot)$ is a kernel function and h is the bandwidth. Usually the kernel will be a positive real function. In addition, kernel functions are often selected to be symmetric and unimodal density functions. In general, the precise shape of the kernel has little impact on the overall appearance of the density. The estimator is easily generalized to multivariate densities by using a multivariate kernel function and a matrix of bandwidths.

A central issue in estimation by local smoothing is the choice of bandwidth. Each bandwidth h is a sequence of numbers such that $h \to 0$ and $nh \to \infty$ as the sample size $n \to \infty$. With a fixed sample, the size of h determines the degree of smoothing and is therefore of crucial importance for the appearance, interpretation, and properties of the final estimate. The choice of bandwidth can be a purely subjective choice, it can be based on some rule of thumb, or the choice can be "automated" by data-driven methods such as cross validation.

One feature of the standard kernel estimator is that the size of bandwidth is independent of the point in the sample space at which the estimator is evaluated. This may mean that excessive weight is given to observations in less dense areas of the sample space. The resulting estimates can produce spurious detail, particularly in the tails of estimated densities. Alternative methods are available to overcome this problem. Such generalizations distinguish themselves from the basic kernel method by adjusting the bandwidth to account for the density of data in particular regions of the sample space, the less dense the data the larger the bandwidth. However it should be borne in mind that the greater robustness of these techniques is bought at extra computational cost. Specific methods include the k-th nearest neighbor and generalized nearest neighbor, variable kernel, and adaptive kernel methods [see, e.g., Duncan and Jones (1992)].

Kernel density estimates form the basis for nonparametric regression analysis. In general the regression functional is

$$\mathrm{E}(y \mid x) = g(x) = \int y f(y \mid x)\, dy = \int y\big(f(y,x)/f(x)\big)\, dy. \tag{10}$$

In nonparametric regression, the regression functional is recovered directly from estimates of the (joint and marginal) density functions. No parametric restrictions are imposed on the form of conditional expectation $g(\cdot)$ or the density function of the implied error term. The Nadaraya–Watson estimator for the bivariate regression model is

$$g_h(x) = \sum y_i W_{hi}(x), \qquad W_{hi}(x) = (n \cdot h)^{-1} K\big[(x_i - x)/h\big]/f_h(x). \tag{11}$$

The nonparametric regression function is therefore a weighted average, with the individual kernel weights $W_{hi}(x)$ dependent on the estimated kernel density of the regressors. Again this is easily generalized for multiple regression.

There appear to have been very few applications of kernel-based nonparametric and semiparametric estimators in health economics. However, as appropriate software becomes more readily available, use of the techniques is likely to increase. Jones (1993) uses data from the 1984 UK Family Expenditure Survey to estimate joint densities and nonparametric regressions for the relationship between household's budget share on tobacco and the logarithm of total non-durable expenditure. Norton (1995) uses kernel estimates to smooth a plot of the fraction of elderly nursing home residents who had "spent-down" at the time of discharge against their time of discharge. Alderson (1997) uses kernel regressions to investigate the shape of the relationship between health related quality of life (HRQoL) and age, without imposing a functional form on the data. She uses data for the EuroQol, EQ-5D, measure of health status collected as part of the ONS Omnibus survey between January and March 1996. The analysis focuses on inequalities in HRQoL and presents separate regressions for males and females and by occupational social class.

Nonparametric estimators can be combined with standard parametric specifications. For example, Dranove (1998) uses a semiparametric approach to investigate economies of scale in a sample of 14 non-revenue generating cost centers in private hospitals in the US. To model the relationship between hospital costs and output he uses the partially linear model, introduced by Robinson (1988)

$$y_i = x_i\beta + g(z_i) + \varepsilon_i, \tag{12}$$

where y is the log of total costs and x contains measures of severity, case-mix, and local wages. Output, z, is measured by the number of discharges and, to allow for a flexible relationship, the form of $g(\cdot)$ is left unspecified.

Estimation of the partially linear model is handled by taking the expectation of (12) conditional on z and then differencing to give

$$y_i - \mathrm{E}(y_i \mid z_i) = \big[x_i - \mathrm{E}(x_i \mid z_i)\big]\beta + \varepsilon_i \tag{13}$$

given the conditional moment conditions $\mathrm{E}(\varepsilon \mid z) = \mathrm{E}(\varepsilon \mid x, z) = 0$. The conditional expectations $\mathrm{E}(y_i \mid z_i)$ and $\mathrm{E}(x_i \mid z_i)$ can be replaced by nonparametric regressions of y and each element of x on z. Then OLS applied to (13) gives $\sqrt{n}$-consistent and asymptotically normal estimates of β, although the asymptotic approximation may perform poorly in finite samples and bootstrap methods are preferable. Finally $g(\cdot)$ can be estimated using a nonparametric regression of $(y - x\hat{\beta})$ on z. In practice, Dranove uses locally weighted least squares to estimate the nonparametric regressions at the first stage and spline regressions for the second stage. Evidence of economies of scale has important implications for the desirability of hospital mergers. Dranove's results suggest that the efficiency gains from mergers would be small and could be easily offset by small price changes resulting from increased market concentration.

One important application of the partially linear model is the sample selection model. Studies by Stern (1996), and Lee et al. (1997) which use kernel based semiparametric estimators of the sample selection model, are discussed in detail in Section 4.3.3.

3. Qualitative dependent variables

3.1. Binary responses

Consider a binary dependent variable, y_i, which indicates whether individual i is a "non-participant" or a "participant". In health economics, binary dependent variables have been used to model an extensive range of phenomena; examples include the use of health care services, purchase of health insurance, and starting or quitting smoking.

If the outcome depends on a set of regressors, x, the conditional expectation of y is

$$\mathrm{E}(y_i \mid x_i) = \mathrm{P}(y_i = 1 \mid x_i) = F(x_i). \tag{14}$$

In order to estimate (14), $F(\cdot)$ could be specified as a linear function, $x_i\beta$; giving the linear probability model. The linear probability model is easy to estimate, using weighted least squares to allow for the implied heteroscedasticity of the non-normal error term, and may be a reasonable approximation if $F(\cdot)$ is approximately linear over the range of sample observations. However the possibility of predicted probabilities outside the range [0, 1] creates a problem of logical inconsistency, which a nonlinear specification of $F(\cdot)$ can avoid.

The most common nonlinear parametric specifications are logit and probit models. These can be given a latent variable interpretation. Let

$$y_i = \begin{cases} 1 & \text{iff } y_i^* > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where

$$y_i^* = x_i\beta + \varepsilon_i$$

and, for a symmetrically distributed error term ε with distribution function $F(\cdot)$

$$\mathrm{P}(y_i = 1 \mid x_i) = \mathrm{P}\big(y_i^* > 0 \mid x_i\big) = \mathrm{P}(\varepsilon_i > -x_i\beta) = F(x_i\beta). \tag{16}$$

Assuming that ε_i has a standard normal distribution gives the probit model, while assuming a standard logistic distribution gives the logit model. These models are usually estimated by maximum likelihood estimation; the log-likelihood for a sample of independent observations is

$$\operatorname{Log} L = \sum_i \big\{(1 - y_i)\log\big(1 - F(x_i\beta)\big) + y_i \log\big(F(x_i\beta)\big)\big\}. \tag{17}$$

Applications of probit, logit, and other models for binary variables are too numerous to list here. One recent example is Buchmueller and Feldstein's (1997) study of the University of California's decision to impose a cap on its contribution to employees' insurance plans in 1994. This natural experiment allows an analysis of how the resulting change in out-of-pocket premiums affected decisions by UC employees to switch insurance plans. The binary dependent variable indicates whether an employee switched plan, and this is modeled by a latent variable representing the net benefit of switching as a function of the change in premium, plan characteristics, and individual demographic characteristics. Plan switching is estimated using probit models on the full sample of 74,478 employees and for separate types of coverage. Simulations of the change in probability of switching associated with changes in the level of premium show large price effects across all of the models.

3.2. *Multinomial and ordered responses*

3.2.1. *Ordered probits and grouped data regression*

The ordered probit model can be used to model a discrete dependent variable that takes ordered multinomial outcomes, e.g., $y = 1, 2, \ldots, m$. A common example is self-assessed health, with categorical outcomes such as excellent, good, fair, poor. The model can be expressed as

$$y_i = j \quad \text{if } \mu_{j-1} < y_i^* \leqslant \mu_j, \ j = 1, \ldots, m, \tag{18}$$

where

$$y_i^* = x_i\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1) \tag{19}$$

and $\mu_0 = -\infty$, $\mu_j \leqslant \mu_{j+1}$, $\mu_m = \infty$. Given the assumption that the error term is normally distributed, the probability of observing a particular value of y is

$$\mathrm{P}_{ij} = \mathrm{P}(y_i = j) = \Phi(\mu_j - x_i\beta) - \Phi(\mu_{j-1} - x_i\beta), \tag{20}$$

where $\Phi(\cdot)$ is the standard normal distribution function. With independent observations, the log-likelihood for the ordered probit model takes the form

$$\operatorname{Log} L = \sum_i \sum_j y_{ij} \log P_{ij}, \tag{21}$$

where y_{ij} is a binary variable that equals 1 if $y_i = j$. This can be maximized to give estimates of β and of the unknown threshold values μ_j. Examples of the use of ordered probit models include Kenkel (1995) who has categorical measures of self-reported health status and of activity limitation from the Health Promotion/Disease Prevention module of the 1985 US National Health Interview Survey, and Chaloupka and Wechsler (1997) who have a categorical measure of average daily cigarette consumption from the 1993 Harvard College Alcohol Study. Levinson and Ullman (1988) apply the ordered probit to a categorical index of the adequacy of prenatal care from a 1994 census of Medicaid births in Wisconsin. An ordered logit specification is used by Theodossiou (1998) for six different measures of mental distress from the 1992 British Household Panel Study, all of which are measured on four point categorical scales. The results show a significant effect of unemployment on the odds of experiencing mental health problems.

Kerkhofs and Lindeboom (1995) develop an ordered probit model for self-reported health, with state-dependent reporting errors. They are concerned with the potential biases that arise in the use of subjective measures of health when responses are influenced by financial incentives and social pressures. In particular they attempt to isolate the impact of employment status on reporting errors. Their model uses three measures of health. A latent variable, H^*, that measures true health; a (categorical) self-reported measure of health, H^{s}; and an objective measure of health based on professional diagnosis, H^{o} (in their case the Hopkins symptom checklist). In order to focus on the possibility of state-dependent reporting errors they assume that H^{o} is a sufficient statistic for the impact of employment status (S) on H^*. They assume that observed self-reported health is given by

$$H^{\mathrm{s}} = j \quad \text{if } \mu_{j-1} < H^* \leqslant \mu_j,\ j = 1, \ldots, m. \tag{22}$$

True health is assumed to depend on $f(H^{\mathrm{o}})$, measured by a set of dummy variables, and demographic characteristics x_1

$$H^* = f(H^{\mathrm{o}}) + x_1\beta + \varepsilon, \quad \varepsilon \sim N(0, 1) \tag{23}$$

and the state-dependent reporting bias is modeled through the threshold values

$$\mu_j = g_j(S, x_2). \tag{24}$$

These depend on employment status and demographic characteristics x_2. Various specifications of $g(\cdot)$ are used to allow for interactions between employment status and demographics. The typical contribution to the likelihood is

$$\begin{aligned} \mathrm{P}(H^{\mathrm{s}} = j) = {} & \Phi\big[g_j(S, x_2) - f(H^{\mathrm{o}}) - x_1\beta\big] \\ & - \Phi\big[g_{j-1}(S, x_2) - f(H^{\mathrm{o}}) - x_1\beta\big]. \end{aligned} \tag{25}$$

The model is estimated with data on heads of household aged 43–63 from the first wave of the Dutch panel survey (CERRA-I). The sample is split by employment status and ordered probit models are estimated with and without the objective measures of health. This gives evidence of state-dependent reporting bias, identified through interactions between employment status and the demographic variables.

Grouped data regression is a variant of the ordered probit model in which the values of the thresholds (μ) are known. Because the μ's are known, the estimates of β are more efficient and it is possible to identify the variance of the error term σ^2. Sutton and Godfrey (1995) use grouped data regression to analyze social and economic influences on drinking by young men. Their analysis uses pooled individual data for males aged 18–24 from the British General Household Survey for 1978–1990. As is often the case with survey measures of alcohol consumption, individuals are assigned to one of seven drinking categories defined by the number of units of alcohol consumed per week, where the range of these intervals is recorded in the survey. They estimate a model in which socio-economic characteristics, along with health-related attitudes and behaviour, are used to predict levels of drinking. A general RESET test for misspecification rejects an OLS specification of the model, but does not reject the grouped data regression. Their results show evidence of a significant interaction between the influence of the price of alcohol and an individual's income. Buchmueller and Zurekas (1998) confront a problem that is common to many health interview surveys, and use grouped data regression to fit a measure of income taken from a categorical scale with varying intervals. Donaldson et al. (1998) suggest the use of grouped data regression to deal with willingness to pay values collected using categorical payment scales.

3.2.2. *The multinomial logit*

Multinomial models apply to discrete dependent variables that can take (unordered) multinomial outcomes, e.g., $y = 1, 2, \dots, m$. In health economics this often applies to the choice of insurance plan or health care provider, but could be used to model the choice of treatment regime for an individual patient. It is helpful to define a set of binary variables to indicate which alternative ($j = 1, \dots, m$) is chosen by each individual ($i = 1, \dots, n$)

$$y_{ij} = \begin{cases} 1 & \text{if } y_i = j, \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

with associated probabilities

$$\mathrm{P}(y_i = j) = \mathrm{P}_{ij}. \tag{27}$$

With independent observations, the log-likelihood for a multinomial model takes the form,

$$\operatorname{Log} L = \sum_i \sum_j y_{ij} \log \mathrm{P}_{ij}. \tag{28}$$

The multinomial logit model uses,

$$\mathrm{P}_{ij} = \exp(x_i\beta_j) \Big/ \sum_k \exp(x_i\beta_k) \tag{29}$$

with a normalization that $\beta_m = 0$. The normalization reflects the fact that only relative probabilities can be identified, with respect to the base alternative (m).

Multinomial models are often motivated by McFadden's random utility model [see, e.g., Dowd et al. (1991)]. Define individual i's utility from choice j as,

$$U_{ij} = V_{ij}(z_i, x_{ij}) + \varepsilon_{ij} \tag{30}$$

or, in linear form

$$U_{ij} = z_i\alpha_j + x_{ij}\beta + \varepsilon_{ij}, \tag{31}$$

where z denotes characteristics that vary across individuals but not across the choices, and x denotes characteristics that vary across the choices. The model assumes that individuals are aware of the unobservable (to the researcher) characteristics ε_{ij}, and the individual is assumed to choose the alternative that gives the maximum utility, so choices are based on net utilities. Typically the ε_{ij} are assumed to be type I extreme value (or Weibull), which has the convenient property that the difference between two Extreme Value I variables has a logistic distribution. The multinomial logit can be derived from the random utility model provided that unmeasured attributes ε_{ij}'s are independent. Then

$$\mathrm{P}_{ij} = \exp(z_i\alpha_j + x_{ij}\beta) \Big/ \sum_k \exp(z_i\alpha_k + x_{ik}\beta) \tag{32}$$

giving a tractable closed form solution. Setting $\beta = 0$ gives the multinomial logit or "characteristics of the chooser" model, while setting $\alpha_j = 0$ gives the conditional logit or "characteristics of the choices" model.

The assumption that the ε_{ij}'s are independent implies the independence of irrelevant alternatives (IIA) property

$$\mathrm{P}_{ij}/\mathrm{P}_{il} = \exp(z_i\alpha_j + x_{ij}\beta)/\exp(z_i\alpha_l + x_{il}\beta). \tag{33}$$

So the odds ratio is unaffected by the existence of alternatives other than j and l (i.e., by changes in the individual's choice set). This implies that if a new alternative is introduced all (absolute) probabilities will be reduced proportionately. Many authors have argued that IIA is too restrictive for many of the applications of multinomial models to health economics. For example, Feldman et al. (1989) argue that, in the case of health insurance plans, the addition of a new plan is more likely to affect the choice of "close substitutes". Much of the recent literature has been concerned with models that relax the IIA assumption such as the nested logit model and the multinomial probit model.

The multinomial logit model can be used in conjunction with two-part models and sample selection models (see Section 4). Haas-Wilson et al. (1988) use data from high option Blue Cross and Blue Shield plans of Federal Employees Benefit Program. The paper makes the case for aggregating health care use by episode of treatment rather than by a fixed period and stresses disaggregation into types of treatment episode; in this case outpatient visits only, outpatient with medication, outpatient with hospitalization, and hospitalization only. A two-part specification with a multinomial logit for types of treatment and OLS for levels of expenditure within episodes is used. The results do not show a significant effect of coinsurance rates on types of episode, but there is a significant effect on levels of expenditure.

Haas-Wilson and Savoca (1990) use a Federal Trade Commission survey of contact lens wearers and their suppliers. A multinomial logit is used to estimate effects of both personal and provider characteristics on the choice of providers between opticians, opthamologists, and optometrists. The choice of provider is estimated jointly with quality of care, using Lee's (1983) generalized selectivity model to estimate regressions for patient outcomes (measured by the "presence of seven potentially pathological eye conditions caused by poorly fitted lenses"). Lee's estimator applies the inverse of the standard normal CDF to the distribution function of the error terms in the multinomial logit. This allows the use of a selectivity model based on bivariate normality. Haas-Wilson and Savoca find evidence of selection bias which leads to an overestimate of quality of care provided by opthamologists. The scope for selection bias arises because outcomes depend partly on patients' behaviour, and differences among patients may be correlated with their choice of provider. The same econometric methods are used by Dowd et al. (1991) who estimate a multinomial logit for choice of insurance plan along with Lee's model for health care utilization, measured by physician contacts and by inpatient days. They do not find evidence of selection bias after controlling for chronic illness and other observed variables.

3.2.3. The nested multinomial logit

Gertler et al. (1987) investigate the impact of user fees on the demand for medical care in urban Peru, using a 1984 Peruvian household survey. They develop a random utility model in which the demand for medical care is modeled as the decision to seek care and, conditional on that, the decision of which provider to use (public clinic, public hospital, or private doctor). The corresponding econometric specification is the nested multinomial logit model, which relaxes the IIA assumption. The empirical model allows them to predict the revenue consequences and welfare effects of increased user fees, and illustrates the trade-off between efficiency and re-distributive goals. Dor et al. (1987) develop the theoretical model used by Gertler et al. (1987) by including access costs in the budget constraint. They apply the nested multinomial logit model to provider choice using 1985 data from the rural Côte d'Ivoire.

A similar approach is adopted by Feldman et al. (1989) who estimate a model using individual data on the demand for health insurance plans among employees of 17 Minneapolis firms. They argue that the existence of "close substitutes" makes the IIA assumption and, hence, the use of a multinomial logit model unrealistic. The assumption is relaxed by using the nested logit specification which drops the IIA assumption between groups of close substitutes. Freedom of choice of doctor is used to distinguish these health plan nests.

The nested logit model generalizes the multinomial/conditional logit as follows. Let $l = 1, \dots, L$ denote "nests" of health plan types. In Feldman et al. there are two nests; IPAs and FFS plans versus HMOs. Within each nest there are $j = 1, \dots, J_l$ plan alternatives. Individual utility is

$$U_{lj} = w_l \delta + x_{lj} \beta + \varepsilon_{lj}, \tag{34}$$

where x_{lj} varies with both the nest and insurance plan, e.g., the premium charged, while w_l varies only with the nest, e.g., freedom to choose a doctor. ε_{lj} is assumed to have a generalized extreme value distribution, which relaxes the assumption that the error terms are independent. Then

$$\mathrm{P}_{lj} = \mathrm{P}_l \cdot \mathrm{P}_{j|l}, \tag{35}$$

where

$$\mathrm{P}_{j|l} = \exp\bigl(x_{ij}\beta/(1-\sigma)\bigr) / \exp(I_l) \tag{36}$$

and

$$I_l = \log\biggl(\sum_k \exp\bigl(x_{ik}\beta/(1-\sigma)\bigr)\biggr) \tag{37}$$

is the "inclusive value", for nest l. β can be estimated up to the scale factor $1/(1-\sigma)$ by using conditional logit within each nest. Then

$$\mathrm{P}_l = \exp\bigl(w_l\delta + (1-\sigma)I_l\bigr) \Big/ \sum_l \exp\bigl(w_l\delta + (1-\sigma)I_l\bigr). \tag{38}$$

This shows that for ease of computation the ML estimation can be done in two steps. First estimate $\beta/(1-\sigma)$ using conditional logit within each nest, then apply conditional logit across the nests to estimate $(1-\sigma)$, including an estimate of the inclusive value.

Feldman et al. (1989) find that Hausman tests, based on the contrast between conditional and nested logit estimates, suggest that the grouping of IPAs and FFS versus HMOs is satisfactory. But they reject the grouping of IPAs and HMOs. Their results show that health plan choices are sensitive to out-of-pocket payments, and they suggest that estimates of the impact of premiums derived from conditional logit models could be misleading.

The use of a nested logit approach implies that choices can be organized into a meaningful nesting or tree structure. This may not be appropriate for some applications. For example, in their study of the choice of provider between government health facilities, mission health facilities, private clinics and self-treatment in the Meru District of Eastern Kenya, Mwabu et al. (1993) argue that there are no a priori grounds for deciding on the correct decision structure for patients. As a result they adopt the simpler multinomial logit specification, using the IIA assumption.

3.2.4. The multinomial probit model

An alternative to the nested logit model is to use a multinomial probit model. Until recently the computational demands of this model have been prohibitive, but the development of simulation based estimators has opened the way for empirical applications. The multinomial probit is used by Börsch-Supan et al. (1992) and by Hoerger et al. (1996) to model choices by elderly disabled people and their families between independent living, living with relatives, and entering a nursing home. Both papers use reduced form equations derived from a random utility framework and the multinomial probit models are estimated using simulated maximum likelihood estimation.

Bolduc et al. (1996) use data from the rural district of Ouidah in Bénin to model the choice of provider between hospital, community health clinic (CHC), private clinic and self-medication. The empirical focus is on the role of user fees (for the CHCs) and precautionary savings (through tontines) to fund health care. They adopt a random utility specification, and compare multinomial logit (ML), independent multinomial probit (IMP), and multinomial probit (MP) specifications. The independent probit model assumes that the ε_{ij} are iid normal. Then the probability that individual i chooses j is

$$\mathrm{P}_{ij} = \int_{-\infty}^{\infty} \prod_{k \neq j} \Phi\big(z_i \alpha_k^* + x_{ik}^* \beta + \varepsilon_{ij}\big) \phi(\varepsilon_{ij})\, d\varepsilon_{ij}, \tag{39}$$

where $\alpha_k^* = \alpha_j - \alpha_k$ and $x_{ik}^* = x_{ij} - x_{ik}$. This specification assumes independence but, unlike the MNL, it does not imply the IIA property.

The multinomial probit model relaxes independence and assumes that the ε_{ij} have a multivariate normal distribution, $N(0, \Omega)$. Then

$$\mathrm{P}_{ij} = \int_{-\infty}^{A_1} \int_{-\infty}^{A_2} \cdots \int_{-\infty}^{A_{j-1}} \phi(u; \Omega)\, du, \tag{40}$$

where $A_k = z\alpha_k^* + x_k^* \beta$. This requires computation of the area under the multivariate normal density $\phi(\cdot)$, such that the utility associated with j is greater than the utility from all the alternatives $k \neq j$. The estimator identifies the α_k^*s, the difference in levels of indirect utility relative to the base alternative (self-medication).

Bolduc et al. estimate this model using simulated maximum likelihood approach using the GHK simulator. They find that an LR test rejects independence in the probit model. Their estimated time and money price elasticities are sensitive to the empirical specification; those for the multinomial probit are "dramatically different" from those for the multinomial logit and independent multinomial probit. In computing these estimates they use hedonic price and travel time equations based on samples of individuals who use each different provider. This is common practice in the literature [see, e.g., Gertler et al. (1987)] but it does raise the issue of potential selection bias.

3.3. Bivariate models

The models discussed in the previous section deal with a single dependent variable that can take multinomial outcomes. The bivariate probit model applies to a pair of binary dependent variables and allows for correlation between the corresponding error terms. It is possible to express the model in terms of latent variables

$$y_{ji}^{*} = x_{ji}\beta_j + \varepsilon_j, \quad j = 1, 2, \quad (\varepsilon_1, \varepsilon_2) \sim N(0, \Omega), \tag{41}$$

where

$$y_{ji} = \begin{cases} 1 & \text{iff } y_{ji}^{*} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

In practice, the health economics literature has made greater use of two variants of the bivariate probit model; the sample selection model and the partial observability probit model. In the model with sample selection y_2 is observed only when $y_1 = 1$. In the partial observability model the researcher observes only $y = y_1 \cdot y_2$.

A variant of the partial observability probit assumes that, if $y_1 = 1$, both y_1 and y_2 are observed, while if $y_1 = 0$, then only $y_1 \cdot y_2$ is observed. The log-likelihood for this case is

$$\begin{aligned} \operatorname{Log} L = {} & \sum_{y_1=0} \log \Phi(-x_1\beta_1) + \sum_{y_1=1,\ y_2=0} \log \Phi(x_1\beta_1, -x_2\beta_2, -\rho) \\ & + \sum_{y_1=1,\ y_2=1} \log \Phi(x_1\beta_1, x_2\beta_2, \rho), \end{aligned} \tag{43}$$

where ρ is the coefficient of correlation between ε_1 and ε_2. In fact, this is identical to the bivariate probit with sample selection; and only the interpretation of the model differs. Examples of the application of these models in health economics are van de Ven and van Praag (1981), Jones (1993) and Kenkel and Terza (1993).

The pioneering use of the sample selection model in health economics is van de Ven and van Praag's (1981) study of the demand for deductibles in private health insurance. They use data on 8,000 respondents from a postal survey of 20,000 policy holders of a large non-profit health insurer in the Netherlands, to model choice between a plan with a deductible and one with complete coverage. The dependent variable is derived from a binary response to a question about their preference for a policy with a deductible. This is modeled as a function of previous use of medical care, self-reported illness days, income, employment and demographics.

Their economic model specifies the expected utility gain from taking a deductible and leads to a basic probit model. However the dataset is prone to selection bias. The survey has a substantial proportion of incomplete responses and these are shown to vary with demographics. van de Ven and van Praag compare a two step estimator with the maximum likelihood estimator of the sample selection model. Incomplete response is predicted by age, gender and family size. Their results show that the two step estimator gives results that are close to ML. They find that health, previous medical consumption and income have significant effects, which implies the potential for adverse selection if individuals can choose between plans with different levels of deductibles.

An example of the partial observability probit model is Kenkel and Terza's (1993) study of the demand for preventive medical care. The motivation for this study is a recognition of the limitations of the neoclassical model of demand for (preventive) medical care, measured by use of diagnostic tests. This stems from the fact that the consumer's (latent) demand is not observed without a visit to doctor, and the actual choice of treatment is influenced by the role of the doctor in mediating patient choice. Together, these mean that a physician visit hurdle comes between latent and observable demand for diagnostic tests.

The use of diagnostic tests is modeled as a partial observability probit based on the latent variables

$$y_2^* = x_2\beta_2 + w_2\alpha_2 + \varepsilon_2 \quad \text{[diagnostic test index]}, \tag{44}$$

$$y_1^* = x_1\beta_1 + \varepsilon_1 \quad \text{[physician visit index]}. \tag{45}$$

Kenkel and Terza's identification strategy relies on the fact that they are modeling sequential decisions. The physician visit is patient initiated, but tests are made after seeing a doctor and are influenced by a set of post-visit influences w_2. Tests for supplier induced demand are based on a sub-set of w_2; those post-visit influences that reflect financial incentives for doctors. Although this is a sequential model, Kenkel and Terza reject a "two-part model", as it rules out positive latent demands for those individuals who do not visit the doctor.

Data from the 1977 National Medical Expenditure Survey are used in separate analyses for men and women and for lab tests and diagnostic tests. The common set of regressors include insurance coverage (private, medicare/caid, none), health (self-assessed and disability days), income, schooling, age, and race. The post-visit variables (w_2) measure outpatient or ER versus office visits, waiting time, and the percentage of the charge paid by private or public insurance. The results show that the correlation between the two error terms is significant for diagnostic tests, but not significant for lab tests. The probability of diagnostic tests increases with private insurance and the fraction of charge paid by private insurance. The results do not support the existence of SID, reflected in the fact that there is no evidence of fewer tests in outpatient/ER compared to office visits, and no effect of waiting time.

4. Limited dependent variables

4.1. Two-part, selectivity, and hurdle models

4.1.1. A taxonomy

Two-part (or multi-part), sample selection, and hurdle models have all been used in the health economics literature to deal with the problem of limited dependent variables. To understand which approach is appropriate for a particular application, it is useful to begin by asking what type of dependent variable is being used. To answer this question it is helpful to introduce some notation. Say that there are two variables of interest: a binary indicator d_i, with associated covariates x_1 and parameters β_1, and a continuous variable y_i, with associated covariates x_2 and parameters β_2, where y_i is coded as $y_i = 0$ if $d_i = 0$.

The first question is whether observations of $y_i = 0$ represent an actual choice of zero. If the answer is no, the problem is one of non-observable response and a sample selection model is potentially appropriate [see, e.g., Heckman (1979)]. For example, this might apply to the case where coinsurance rates (y) are only observed for those who purchase insurance $(d = 1)$, but non-purchase of insurance does not imply that a potential insuree would face a coinsurance rate of zero. If the answer to the question is yes, then zero observations represent a genuine choice of zero.

In the case of "genuine zeros" the second question is whether the choice to consume is influenced by the decision of how much to consume. If the answer is no, a sequential decision model is appropriate. If the answer is yes, a joint decision model is appropriate. When considering joint versus sequential decisions it is important to make the distinction between a chronological sequence of events and sequential choice. For example, the "gate-keeper" role of GPs may mean that an individual has to visit a GP before they can use inpatient care. This limits their opportunity set, but the individual can consider a range of options; do not visit the GP; visit the GP but do not visit consultant; or visit both. Modeling these decisions as a sequential choice suggests a myopic decision rule: visit the GP then decide how to respond to advice. Sequential choices are often used to motivate the two-part model, while joint decisions are associated with generalized Tobit and hurdle models.

The third question to bear in mind is the object of the analysis. Is the object simply prediction of $\mathrm{E}(y \mid x)$, for example, to deal with the problem of imputing missing values due to item non-response in a sample survey? Or is the object to make inferences about β_1 and β_2? The answer to this question will help to determine the appropriate method to adopt.

Defining the dependent variables in this way suggests a taxonomy to distinguish the three approaches. In the sample selection model, knowledge that $y_i = 0$ (as opposed to $d_i = 0$) is uninformative in estimating determinants of the level of y_i. In the two-part model observations for which $y_i = 0$ are uninformative in estimating the determinants of the level of $(y_i \mid y_i > 0)$. In hurdle models, the fact that $y_i = 0$ is used in the estimation of β_2.

It is possible to express the sample selection and hurdle models in terms of latent variables (y^*)

$$y^*_{ji} = x_{ji}\beta_j + \varepsilon_j, \quad j = 1, 2. \tag{46}$$

Then the sample selection model is given by

$$y_i = \begin{cases} y^*_{2i} & \text{iff } y^*_{1i} > 0, \\ \text{unobserved} & \text{otherwise } (= 0 \text{ in generalized Tobit}) \end{cases} \tag{47}$$

and the hurdle model is given by

$$y_i = \begin{cases} y^*_{2i} & \text{iff } y^*_{2i} > 0 \text{ and } y^*_{1i} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

The two-part model is usually estimated by a logit or probit model for the probability of observing a positive value of y, along with OLS on the sub-sample of positive observations.There is no latent variable representation for the two-part model. Instead it is motivated by a conditional mean independence assumption

$$\mathrm{E}(y_i \mid y_i > 0, x_{2i}) = x_{2i}\beta_2. \tag{49}$$

Notice that no assumption is made about the unconditional mean $\mathrm{E}(y \mid x)$, only about the conditional/selected sample. In general, the two-part specification does not assume normality of $(\varepsilon_1, \varepsilon_2)$ and does not require linearity of $\mathrm{E}(y \mid y > 0, x)$.

4.1.2. Two-part versus selectivity models: the debate

The issue of choosing between the two-part model (2PM) and a generalized Tobit or sample selection specification (SSM) to model the demand for medical care has provoked a vigorous, and often heated, debate in the health economics literature. Advocacy of the two-part model is most associated with the empirical strategy adopted for the RAND Health Insurance Experiment (HIE) [see, e.g., Newhouse et al. (1980), Leibowitz et al. (1985), Manning, Newhouse et al. (1987)]. Duan et al. (1983) initiated the subsequent debate by making the case for the two-part model. They argue that the censored data approach requires restrictive distributional assumptions and that, as the censored data are unobservable, these assumptions are not testable. They stress “poor numerical and statistical properties” of the SSM, caused by the existence of multiple local optima in its likelihood function. They also argue that the fact that the residual vector is censored in the SSM poses a problem for standard residual based tests.

Hay and Olsen (1984) criticize the 2PM by claiming that it is also subject to untestable assumptions and they question the existence of any distribution of $(\varepsilon_1, \varepsilon_2)$ that gives a complete normal distribution for $(\varepsilon_2 \mid \varepsilon_1 > -x_1\beta_1)$. To support this argument they show that if ε_1 and ε_2 are not independent, the conditional distribution of

ε_1 is generally a function of $(x_1\beta_1)$. They respond to the argument that the SSM has poor numerical properties by citing an algorithm for finding a global maximum. Also they argue that, even though the 2PM and SSM are non-nested, they can be compared in terms of mean squared forecast error (MSFE). Duan et al. (1984) counter this final point by showing that with the RAND HIE data there is no discernible difference between the 2PM and SSM models according to the MSFE criterion. Also they provide an example designed to show that it is possible to find a distribution of $(\varepsilon_1, \varepsilon_2)$ that contradicts Hay and Olsen's claim.

Maddala (1985) sets out to adjudicate the debate. He stresses the need to understand the nature of the underlying decision process in selecting an empirical model and argues that joint decisions may be more appropriate than the sequential approach implied by the 2PM. He cites van de Ven and van Praag (1981) and argues that decision to use health care will be linked to perceived severity of illness (and hence likely expenditure).

In response to Duan et al. (1984) he points out that semiparametric estimators were available for the SSM and that the normality assumption is testable. Also he considers their "counter-example". Duan et al. (1984) aim to show that there is a joint distribution of $(\varepsilon_1, \varepsilon_2)$ that allows correlation between the two error terms but, for $d = 1$, gives

$$\log(y) = x_2\beta_2 + \varepsilon_2, \varepsilon_2 \sim \mathrm{IN}(0, \sigma^2). \tag{50}$$

They assume that ε_1 is continuous for the whole population, and that ε_2 has a mass point at $\varepsilon_2 = -\infty$ and is continuous over the real line for $d = 1$. They argue that it is possible to construct a joint distribution from these marginals such that ε_1 and ε_2 are correlated. Maddala argues that this is "purely semantic" as the correlation is not estimable. Also, their model is actually specifying conditional distributions for the separate sub-populations, $\varepsilon_1 > x_1\beta_1$ and $\varepsilon_1 \leqslant x_1\beta_1$.

Maddala makes the distinction between sample selection models, in which the criterion function is written in reduced form, and correlation between ε_1 and ε_2 is the only connection between the two equations, and self selection models in which the criterion is written in structural form. He argues that adopting a structural approach "will help in organizing one's thinking properly on why one expects any selectivity bias in the problem". He goes on to argue that "even when decisions are sequential, if there are some common omitted variables the two decisions will be correlated. In this case, it is advisable not to formulate the model in a way that the correlation cannot ever be estimated". Zimmerman Murphy (1987) lists common omitted variables in context of medical care demand; these include insurance status, time costs, marginal valuation of health, time preference, and risk aversion.

Duan et al. (1985) take up Maddala's challenge. They stress that the focus of their own work is on estimating mean medical expenditure and that, in that context, the debate over statistical methods has no relevance for the policy implications of their results. They find that multi-part, ANOVA, and sample selection models all give similar results, and that the debate is "much ado about nothing". Also they argue that "in the specific case of health insurance one does not need an estimate of ρ to estimate mean expenditure" and that many econometrics models are formulated so that "nuisance parameters" are eliminated, these include the Cox partial likelihood, the within-groups estimator for panel data, and zero restrictions in structural models. Maddala (1985) rounds off the exchange by recognizing that the RAND HIE data are special because participants were randomized across insurance plans. But he cautions against use of the 2PM in other contexts.

4.1.3. Monte Carlo evidence

In an attempt to settle the debate over the relative merits of 2PM and SSM specifications, Manning et al. (1987) use Monte Carlo simulations to compare the Heckman two-stage estimator (LIML/Heckit) and the full maximum likelihood estimator (FIML) of the sample selection model with a "naive two-part model" (the true specification omitting the correlation coefficient) and a "data-analytic (testimator) variant" (which adds powers and interactions of x, according to a test criterion).

The debate addresses the case in which valid exclusion restrictions are not available, so $x_1 \equiv x_2$ and identification of the SSM relies on functional form. In discussing this notion of identification by functional form, it is worth making the distinction between the identification of β_2 and of $\mathrm{E}(y \mid x)$. The data-analytic version of the 2PM can be interpreted as giving a Taylor series approximation of the conditional mean function for the SSM which would yield good estimates of $\mathrm{E}(y \mid x)$ but would not identify β_2. As in the earlier work of the RAND HIE researchers, Manning et al. (1987) stress that they are not interested in the coefficients per se, but only in predictions of $\mathrm{E}(y \mid x)$ from

$$\mathrm{E}(y \mid x) = \mathrm{P}(y > 0 \mid x)\mathrm{E}(y \mid y > 0, x). \tag{51}$$

They use the SSM as their theoretical benchmark, but find that 2PM outperforms it on statistical grounds. This leads them to conclude: "based on our experience here and elsewhere, we believe that the data-analytic version of the two part model will be robust – as long as analysts are concerned about the response surface rather than particular coefficients."

A re-assessment of the Monte Carlo evidence in Manning et al. (1987) is provided by Leung and Yu (1996). Leung and Yu argue that their Monte Carlo design creates collinearity problems that bias the results against the SSM and in favor of 2PM. The design problem they identify is that Manning et al. use a model with no exclusion restrictions ($x_1 \equiv x_2$) and simulate $x \sim u(0, 3)$. Leung and Yu argue that this leads to insufficient range of variation in the inverse Mill's ratio. Leung and Yu use $x \sim u(0, 10)$ and find that "collinearity problems vanish and the sample selection model performs much better than the two-part model". Of course, this raises the empirical question of how much variation will be observed with real data. The range used by Leung and Yu is far greater than is likely to be observed in health economics applications.

To understand the collinearity problem, consider the Heckit/LIML estimator of the SSM. This is based on estimating the following regression on the selected sample

$$y = x_2\beta_2 + \lambda(x_1\beta_1) + e_2, \tag{52}$$

where $\lambda(x_1\beta_1) = \mathrm{E}(\varepsilon_2 \mid \varepsilon_1 > -x_1\beta_1)$ and e_2 is a random error term. Assuming joint normality, $\lambda(x_1\beta_1)$ can be estimated by the inverse Mill's ratio, $\phi(x_1\beta_1)/\Phi(x_1\beta_1)$, from a probit regression of d on x_1. With $x_1 \equiv x_2$, identification (of β_2) relies on the nonlinearity of the inverse Mill's ratio $\lambda(\cdot)$, but a plot of $\lambda(\cdot)$ shows that the function is approximately linear for much of its range. This implies that the range of $x_1\hat{\beta}_1$, and hence of x_1, is important and that the degree of censoring is important, as it reduces the range of observed values. Leung and Yu argue that the claim that Heckit will perform poorly when there is a high degree of correlation between $x_1\hat{\beta}_1$ and x_2 is potentially misleading. In their Monte Carlo design, Heckit performs well when x_2 and $x_1\hat{\beta}_1$ are perfectly correlated, as long as the proportion of censored observations is sufficiently small and/or the range of x_1 is sufficiently large (i.e. when the nonlinearity of $\lambda(\cdot)$ comes into play).

Leung and Yu (1996) conclude that the performance of models depends on the empirical context. Collinearity problems can arise if there are few exclusion restrictions, a high degree of censoring, low variability among the regressors (x_1), or a large error variance in the choice equation (i.e. weak instruments). They suggest that applied researchers should always check for collinearity. After looking at a range of measures of collinearity, they favor the condition number. They argue that their Monte Carlo evidence shows that, in the absence of collinearity problems, the t-test on the inverse Mill's ratio can be used to distinguish between the 2PM and SSM. Overall they conclude that "... the merits of the two-part model have been grossly exaggerated in the literature"... "hence the extreme and negative remarks against the sample selection model made by Duan et al. ... are unwarranted and misleading." This conclusion, however, relies on the absence of collinearity problems. These collinearity problems are likely to arise in health data sets, and should be investigated by applied researchers who intend to use the sample selection model.

4.1.4. Empirical evidence

Zimmerman Murphy (1987) estimates sample selection models for physician office visits, hospital outpatient visits, and hospital inpatient days using the 1970 US National Health Survey. She uses the Heckit estimator and finds significant negative coefficients for the inverse Mill's ratio. The results show evidence of the collinearity problem, with the estimates of the selectivity correction becoming less significant the greater the correlation between the inverse Mill's ratio and the other regressors. Hunt-McCool et al. (1994) use a sample of adults from the US National Medical Care Expenditure Survey. Their dependent variables are the quantity of service (office visits, hospital inpatient care) and out-of-pocket expenditure shares. Heckit estimates show positive and significant coefficients on the inverse Mill's ratio.

4.2. Two-part models and retransformation: developments and applications

Applications of the two-part model in health economics have often used logarithmic transformations to deal with dependent variables that are heavily skewed, such as household medical expenditures. This raises the problem of retransforming to the original scale (e.g., dollars rather than log-dollars), in order to make inferences that are relevant for policy [Duan (1983), Duan et al. (1983)]. The retransformation problem has been revisited recently by Manning (1998) and Mullahy (1998), and the material in this section draws heavily on Mullahy's paper: "Much ado about two: reconsidering retransformation and the two-part model in health econometrics". Mullahy focuses on the 2PM applied to "genuine zeros" rather than missing observations. He argues that, due to nonlinearities and retransformations, the estimated parameters from the 2PM are not sufficient for inference about important policy parameters that involve the level of y, such as $\mathrm{E}(y \mid x)$, $\partial \mathrm{E}(y \mid x)/\partial x$, and $\partial \log \mathrm{E}(y \mid x)/\partial \log x$.

In most applications the 2PM is estimated by a probit or logit for $\pi(x) = \mathrm{P}(y > 0 \mid x)$, and least squares on the logarithm of y

$$\begin{aligned} \log(y) &= \log(\mu(x)) + \varepsilon_2, \quad y > 0, \\ &= x\beta_2 + \varepsilon_2. \end{aligned} \tag{53}$$

The problem for inference stems from two issues; the conditioning on $y > 0$, and the need to re-transform from $\log(y)$ to y-space. The identifying assumption for β_2 is the orthogonality condition $\mathrm{E}(\varepsilon_2 \mid y > 0, x) = 0$. Under this assumption the 2PM will give consistent estimates of β_2, but the condition does not identify other parameters such as $\mathrm{E}(y \mid x)$. In general notation, the 2PM implies

$$\begin{aligned} \mathrm{E}(y \mid x) &= \mathrm{P}(y > 0 \mid x) \cdot \mathrm{E}(y \mid y > 0, x) \\ &= \pi(x) \cdot \mu(x) \cdot \mathrm{E}\big(\exp(\varepsilon_2) \mid y > 0, x\big) \\ &= \pi(x) \cdot \mu(x) \cdot \rho(x), \end{aligned} \tag{54}$$

with parametric representations,

$$\mathrm{E}(y \mid x) = \pi(x; \beta_1) \cdot \mu(x; \beta_2) \cdot \rho(x; \gamma). \tag{55}$$

The presence of $\rho(x; \gamma)$ in this expression means that the identification of (β_1, β_2), by the 2PM, is not sufficient to identify $\mathrm{E}(y \mid y > 0, x)$ or $\mathrm{E}(y \mid x)$. Two solutions to this identification problem are:

(1) Assume log-normality of $(y \mid y > 0, x)$ with constant variance σ^2, which implies

$$\mathrm{E}(y \mid y > 0, x) = \exp\big(x\beta_2 + 0.5\sigma^2\big). \tag{56}$$

(2) Instead of assuming a distribution for ε_2, Duan (1983) proposes a nonparametric smearing estimator

$$S = \sum_{i=1}^{n_+} \big[\exp(\varepsilon_{2i})\big]/n_+ \tag{57}$$

the mean of the estimate of $\exp(\varepsilon_{2i})$ over the positive observations (n_+). Duan shows that this is a consistent nonparametric estimator of $\mathrm{E}(\exp(\varepsilon_2))$.

The problem with the smearing estimator is that consistent estimation of β_2 in the 2PM only requires the orthogonality condition $\mathrm{E}(\varepsilon_2 \mid y > 0, x) = 0$. In other words ε_2 could be heteroscedastic, in which case the consistency of the smearing estimation breaks down. Manning (1998) draws attention to the role of heteroscedasticity in the retransformation problem. He notes that many applications in health economics use $\log(y)$ as the dependent variable in order to deal with skewness in the data. But few of these applications make a full correction for heteroscedasticity when estimating the impact of regressors on the level of y. Mullahy (1998) speculates about testing for this problem by running a regression of $\exp(\varepsilon_2)$ on, say, $\exp(x\gamma)$. If $\gamma = 0$ for non-constant elements of x, then Duan's estimator should be adequate. The actual approach adopted by the RAND HIE researchers was to split the sample by discrete x variables and apply separate smearing estimates [see, e.g., Duan et al. (1983)]. Manning (1998) shows how results based on the RAND HIE data are sensitive to corrections for heteroscedasticity.

Given the problems of identifying $\mathrm{E}(y \mid y > 0, x)$ or $\mathrm{E}(y \mid x)$ in the standard 2PM, Mullahy (1998) considers two alternative estimators. First, given that $\mathrm{E}(y \mid y > 0, x)$ must be positive, he suggests using an exponential conditional mean specification; $\mathrm{E}(y \mid y > 0, x) = \exp(x\beta_2)$. Combining this with a logistic specification for $\mathrm{P}(y > 0 \mid x)$, the model gives

$$\begin{aligned} \mathrm{E}(y \mid x) &= \mathrm{P}(y > 0 \mid x) \cdot \mathrm{E}(y \mid y > 0, x) \\ &= \bigl[\exp(x\beta_1)/\bigl(1 + \exp(x\beta_1)\bigr)\bigr] \cdot \exp(x\beta_2) \\ &= \exp\bigl(x(\beta_1 + \beta_2)\bigr)/\bigl(1 + \exp(x\beta_1)\bigr). \end{aligned} \tag{58}$$

The model can be estimated by a two-step estimator (M2PM-2); using logit (or probit) for β_1, and nonlinear least squares (NLLS) for the positive observations. Alternatively it can be estimated in one step (M2PM-1), using the full sample to estimate (58) by NLLS.

Then Mullahy considers the "more primitive assumption", $\mathrm{E}(y \mid x) > 0$. This can be justified by the fact that, for non-negative y, finding $\mathrm{E}(y \mid x) = 0$ means the problem is uninformative (as it implies that y always equals zero). As a consequence he suggests using the exponential conditional mean (ECM) model, $\mathrm{E}(y \mid x) = \exp(x\beta)$, and estimating by NLLS. The advantages of this simple specification are that it is straightforward to use instrumental variables to deal with problems of unobservable heterogeneity in the model, and that the elasticities, $\partial \mathrm{E}(\log y)/\partial \log x$, are simple to compute and interpret. The price of using the simpler specification is that it does not allow separate inferences about $\mathrm{P}(y > 0 \mid x)$ and $\mathrm{E}(y \mid y > 0, x)$. Mullahy notes that the M2PM model reduces to the ECM model when $\beta_1 = 0$, and he proposes a conditional moment test to assess whether a one-part or a two-part specification applies. He also proposes a Wald test based on the contrast between the 2PM and M2PM estimates of β. This can be interpreted as a test of whether $\rho(x)$ is constant. The performance of the competing specifications is illustrated using data on individuals aged 25–64 from the 1992 US National Health Interview Survey. Empirical estimates of models for the number of visits to the doctor over the previous twelve months suggest that the M2PM-2 specification is preferred to the ECM specification which is preferred to OLS and the 2PM.

It is worth noting that the use of exponential conditional mean specifications provides a direct link with the count data regressions discussed in Section 7 of this chapter. The ECM model corresponds to a Poisson regression model, while the M2PM-2 corresponds to the the zero altered Poisson model. These specifications are discussed in greater detail below.

4.3. Selectivity models: developments and applications

4.3.1. Manski bounds

In a recent review, Manski (1993) argues that "the selection problem is, first and foremost, a failure of identification. It is only secondarily a difficulty in sample inference." To illustrate, consider a population characterized by (y, d, x), where d and x are observed but the "outcome" y is only observed if the "treatment" $d = 1$. Interest centers on the unconditional probability

$$\mathrm{P}(y \mid x) = \mathrm{P}(y \mid x, d = 1)\mathrm{P}(d = 1 \mid x) + \mathrm{P}(y \mid d = 0, x)\mathrm{P}(d = 0 \mid x). \tag{59}$$

The selection problem stems from the fact that the term $\mathrm{P}(y \mid d = 0, x)$ cannot be identified from the available data. All that is known is

$$\mathrm{P}(y \mid x) \in \bigl[\mathrm{P}(y \mid x, d = 1)\mathrm{P}(d = 1 \mid x) + \gamma \mathrm{P}(d = 0 \mid x),\ \gamma \in \Gamma\bigr], \tag{60}$$

where Γ is the space of all probability measures on y. To address this problem the statistical literature often assumes independence or ignorable non-response

$$\mathrm{P}(y \mid x) = \mathrm{P}(y \mid d = 0, x) = \mathrm{P}(y \mid d = 1, x). \tag{61}$$

This is a strong assumption which asserts that those individuals who do not receive the treatment would respond in the same way to those who do, conditional on the covariates. But, as Manski points out, "in the absence of prior information this hypothesis is not rejectable"; to see this set $\gamma = \mathrm{P}(y \mid d = 1, x)$. So, in the absence of prior information, the "selection problem is fatal for inference on the mean regression of y on x". Restrictions on $\mathrm{P}(y \mid x)$, $\mathrm{P}(y \mid x,\ d = 0)$, and $\mathrm{P}(d \mid x, y)$ may have identifying power, but restrictions on $\mathrm{P}(y \mid x,\ d = 1)$ and $\mathrm{P}(d \mid x)$ are superfluous as they are already identified by the censored sampling process. These identifying restrictions relate to functional forms, including exclusion restrictions on the regressors that enter the regression equations, and assumptions about the distribution of the error terms. In the econometrics literature, the traditional approach has been the parametric Heckit model, but recent years have seen the development of less restrictive semiparametric estimators which relax some, but not all, of the identifying restrictions. These are discussed in the next section.

The selection problem is fatal for inferences concerning $\mathrm{E}(y \mid x)$ without identifying restrictions, but Manski shows that it is possible to put bounds on other features of the distribution. This leads to nonparametric estimation of the bounds and to estimators for quantile regressions.

4.3.2. Semiparametric estimators

The biostatistics literature has seen the development of the propensity score approach, to deal with the problem of identifying treatment effects when there is self-selection bias in the assignment of patients to treatments. In the econometrics literature, this idea is connected to the development of semiparametric estimators for the sample selection model, some of which have been applied in health economics. These estimators focus on relaxing the distributional assumptions about the error terms in the sample selection model and, in particular, they seek to avoid the assumption of joint normality which is required to identify the Heckit model.

Rosenbaum and Rubin (1983) show that conditioning on the propensity score, which measures the probability of treatment given a set of covariates, can control for confounding by these covariates in estimates of treatment effects. Angrist (1995) provides weak sufficient conditions for conditioning on the propensity score in a general selection problem involving instrumental variables. The main identifying assumption is that the instruments satisfy a simple monotonicity condition, as in Imbens and Angrist (1994). The result implies that, with $\mathrm{P}(d = 1 \mid x)$ fixed, selection bias does not affect IV estimates of slope parameters. This result lies behind Ahn and Powell's (1993) approach to the selection problem, which uses differencing of observations for which nonparametric estimates of $\mathrm{P}(d = 1 \mid x)$ are "close". To illustrate it is worth re-capping a general version of the sample selection model. Assume that the following is observed

$$y = [x_2\beta_2 + \varepsilon_2]\mathbf{1}\big[\varepsilon_1 > -\psi(x_1)\big], \tag{62}$$

where $\mathbf{1}[\cdot]$ is an indicator function, $\psi(x_1)$ is the selection index and d is the observed binary variable, such that $d = \mathbf{1}[\varepsilon_1 > -\psi(x_1)]$. For the selected sample

$$\mathrm{E}[y \mid x, d = 1] = x_2\beta_2 + \mathrm{E}\big[\varepsilon_2 \mid x, \varepsilon_1 > -\psi(x_1)\big]. \tag{63}$$

If the distribution of $(\varepsilon_1, \varepsilon_2)$ is independent of x_1 and x_2, the conditional expectation of ε_2 depends only on $\psi(x_1)$. The propensity score is defined as follows,

$$\mathrm{P}(x_1) = \mathrm{P}(d = 1 \mid x_1) = \mathrm{P}\big[\varepsilon_1 > -\psi(x_1)\big]. \tag{64}$$

When the function is independent of x, it is invertible and it is possible to write $\psi(x_1) = \eta(\mathrm{P}(x_1))$. Then

$$\mathrm{E}[y \mid x, d = 1] = x_2\beta_2 + \tau\big[\mathrm{P}(x_1)\big]. \tag{65}$$

Ahn and Powell (1993) propose an estimator for the general model where the selection term depends on the propensity score. Consider any pair of observations where $\mathrm{P}_i \approx \mathrm{P}_j$. Then, provided the selection function $\tau(\cdot)$ is continuous

$$y_i - y_j \approx [x_{2i} - x_{2j}]\beta_2 + \varepsilon_{ij}. \tag{66}$$

This leads Ahn and Powell to suggest a weighted IV estimator for β_2, using kernel estimates of $(\mathrm{P}_i - \mathrm{P}_j)$ as weights. They show that, under appropriate assumptions, the estimator is $\sqrt{n}$ consistent and asymptotically normal and they provide an estimator for the associated covariance matrix.

The Ahn and Powell approach is particularly flexible because it is based on $\tau[\mathrm{P}(x_1)]$. Many other semiparametric approaches have concentrated on the linear index version of the selectivity model

$$\mathrm{E}[y \mid x, d = 1) = x_2\beta_2 + \lambda(x_1\beta_1). \tag{67}$$

(65) and (67) both have the partially linear form discussed in Section 2.3 and can be estimated using Robinson's (1988) approach.

Stern (1996) provides an example of the semiparametric approach in a study that aims to identify the influence of health, in this case disability, on labor market participation. The paper uses a Heckman style model, using labor market participation to identify the reservation wage (supply) and a selectivity corrected wage equation to identify the offered wage (demand). This proves to be sensitive to distributional assumptions and exclusion restrictions.

Stern's data are a sample of 2,674 individuals from the 1981 US Panel Study on Income Dynamics. Disability is measured by a limit on the amount or kind of work the person can do. Initial estimates are derived from reduced form probits and selectivity corrected reduced form wage equations. He finds that disability is insignificant when controlling for selection but very significant without control (even though the selection term is not significant), a result which seems to highlight the collinearity problems associated with the sample selection model. Structural participation equations, in the form of multiple index binary choice models, were very sensitive to the choice of exclusion restrictions, so Stern turns to semiparametric estimation.

He uses Ichimura and Lee's (1991) estimator for the model

$$y = z_0 + \psi(z_1, z_2) + \varepsilon, \tag{68}$$

where $z_j = x_j\beta_j$. This includes two special cases that are relevant here: first the structural participation model, where $\beta_0 = 0$, z_1 is the demand index, and z_2 is the supply index; and second the Heckman wage equation, where $z_2 = 0$. Ichimura and Lee's approach uses a semiparametric least squares (SLS) estimator and minimizes the criterion

$$(1/n)\sum\big[(y - z_0) - \mathrm{E}(y - z_0 \mid z_1, z_2)\big]^2, \tag{69}$$

where the conditional expectation is given by the nonparametric regression function,

$$\begin{aligned}
&\mathrm{E}\big[(y - z_0 \mid z_1, z_2)\big] \\
&\quad = \frac{1}{n-1}\Bigg[\sum_{j \neq i}(y_j - z_{0j})K\big[(z_{1i} - z_{1j})/h_1,\ (z_{2i} - z_{2j})/h_2\big]\Bigg] \\
&\quad\quad \times \Bigg(\frac{1}{n-1}\sum_{j \neq i} K\big[(z_{1i} - z_{1j})/h_1, (z_{2i} - z_{2j})/h_2\big]\Bigg)^{-1}
\end{aligned} \tag{70}$$

and where $K[\cdot, \cdot]$ is a kernel function and the h's are bandwidths. The Ichimura and Lee estimator is known to be badly behaved in small samples. In Stern's application this shows up in the irregular shape of the estimated supply function. To deal with this he imposes a monotonicity assumption; $\psi_1, \psi_2 \geqslant 0$.

For the multiple index model he reports the correlations for the regressors that are common to both equations. He finds a low degree of correlation and concludes that the "hypothesis that demand and supply are not identified can be rejected" (p. 61). The results suggest that the supply effects of disability are much greater than the demand effects. "Thus effort to improve the handicap accessibility of public transportation or home care programmes for disabled workers (if effective at reducing the supply index) are likely to be more successful than efforts to reduce discrimination among employers or to provide wage subsidies to employers" (p. 68).

Similar semiparametric methods are used by Lee et al. (1997). Like Stern (1996), they adopt a linear index specification and use semiparametric estimators to avoid imposing any assumptions on the distributions of the error terms in their model. Their analysis is concerned with estimating a structural model for anthropometric measures of child health in low income countries. They argue that reduced form estimates of the impact of health interventions, such as improved sanitation, on child health may be prone to selection bias if they are estimated with the sample of surviving children. If the health intervention improves the chances of survival it will lower the average health of the surviving population, as weaker individuals are more likely to survive, and lead to a biased estimate of the effectiveness of the intervention.

They specify a system of structural equations. These consist of a survival equation, based on a binary dependent variable, which includes the influence of water supply and sanitation on child survival, reduced form input demands, measuring calorie intake, and the child health production function, measured by the child's weight. The survival equation is specified as a linear index model with an unknown error distribution, and is estimated by a semiparametric maximum likelihood (SML) procedure. The reduced form input demands, for the surviving children, are estimated as sample selection models by semiparametric least squares (SLS), conditioning on the SML estimates of the survival index. The child health (weight) production function is estimated using the same approach, but the endogenous health inputs are replaced by fitted values from SLS estimates of the reduced forms, giving two-stage semiparametric least squares estimates

(TSLS). The form of the kernel functions and the bandwidths used in the estimation are selected so that the semiparametric estimates are $\sqrt{n}$-consistent and asymptotically normal. Hausman type tests are used to compare the SML estimates of the survival equation with standard probit estimates, to test for the exogeneity of the health inputs, and to test whether there is a problem of sample selection bias.

The models are estimated on two datasets; the 1981–82 Nutrition Survey of Rural Bangladesh and the 1984–85 IFPRI Bukidnon, Philippines Survey. The data are split into sub-samples for children aged 1–6 and 7–14. Tests for normality in the survival equation fail to reject the standard probit model in both of the sub-samples for the Philippines, and for ages 1–6 in Bangladesh. For children aged 7–14 in Bangladesh the estimated effects of maternal schooling and water supply are substantially different, but the estimates for other variables are similar for SML and the probit. For the health production functions they compare a standard simultaneous equations estimator, a simultaneous equations selection model based on joint normality, and the semiparametric estimator. The results do not appear to be sensitive to either the selectivity correction or the normality assumption. Despite this, the authors note that previous reduced form studies may have understated the impact of health interventions, because of the unobservable heterogeneity bias associated with a reduced allocation of resources to child health in households with better facilities.

4.3.3. *Identification by covariance restrictions*

Pitt (1997) develops similar theoretical ideas to Lee et al. (1997), but adopts a different approach to dealing with the selection problem. He argues that fertility selection bias (when parents are influenced by health prospects for potential births) may affect estimates of the determinants of child health and mortality, and that mortality selection bias may influence the analysis of the determinants of child health. This creates an identification problem for standard parametric approaches to the selectivity problem. Pitt argues that in a reduced form specification of child health (mortality) conditional on fertility it is difficult to justify exclusion restrictions, that is to find regressors that influence fertility choices but do not influence child health. In this case identification would have to rely on nonlinearity of the selection correction. For a binary measure of child health (e.g., mortality data) this leads to a bivariate probit with partial observability, and Pitt cites the empirical problems of identifying this model with his data, and in other studies.

Pitt suggests an alternative approach based on identification by covariance restrictions. This provides a strategy for identification "so long as fertility and health outcomes are observed for more than one time period in the life of each woman in the sample". In other words this approach relies on longitudinal data to control for individual effects. Pitt models observed births (F) and deaths (D) in terms of latent variables

$$F_{it}^* = x_{fit}\beta_f + \mu_{fi} + \varepsilon_{fit}, \quad F = 1 \quad \text{if } F^* > 0, \tag{71}$$

$$D_{it}^* = x_{hit}\beta_h + \mu_{hi} + \varepsilon_{hit}, \quad D = 1 \quad \text{if } D^* > 0. \tag{72}$$

Identification relies on there being individual effects that influence fertility (μ_{fi}) which are correlated with the individual effects that influence child health (μ_{hi}). Pitt uses longitudinal data on births to identify these correlated effects. The model adopted is a random effects bivariate probit, which implies that correlation only works through a time invariant effect, i.e. there are no dynamic effects associated with the timing of births.

The model is applied to data from 14 Sub-Saharan Demographic and Health Surveys (DHS). The measure of mortality is deaths before age two. For each country Pitt compares standard probits on the sample of all births, a random effects probit, and a "selection corrected probit", i.e. a random effects bivariate probit with partial observability. There is evidence of correlated effects in all cases. But the random effects only account for a small portion of overall error variance, and there is no marked effect on the derivative of the conditional probability of infant death with respect to parental education.

Pitt also derives trivariate models for continuous anthropometric measures of child health: weight and height. To observe these measures the child must be born and survive and estimation must allow for both sources of selection bias. Estimates from the Zambian DHS show little evidence of selection bias for log(weight) and only limited evidence for log(height).

4.4. Hurdle models: developments and applications

In health survey data, measures of continuous dependent variables such as alcohol and tobacco consumption, or measures of medical care expenditure invariably contain a high proportion of zero observations and appropriate limited dependent variable techniques are required. The special feature of the double hurdle approach is that, unlike the standard Tobit model, the determinants of participation (e.g., whether to start or quit smoking) and the determinants of consumption (e.g., how many cigarettes to smoke) are allowed to differ.

However, a limitation of the standard double hurdle specification is that it is based on the assumption of bivariate normality for the error distribution. Empirical results will be sensitive to misspecification, and ML estimates will be inconsistent if the normality assumption is violated. This may be particularly relevant if the model is applied to a dependent variable that has a highly skewed distribution, as is often the case with survey data on cigarette and alcohol consumption, and for medical care expenditure.

A flexible generalization of the double hurdle model is used by Yen and Jones (1996). The Box–Cox double hurdle model provides a common framework that nests standard versions of the double hurdle model and also includes the generalized Tobit model and 'two-part' dependent variable, as special cases. This allows explicit comparisons of a wide range of limited dependent variable specifications that have been used in the health economics literature. The model for the observed dependent variable (y_i) can be written in terms of two latent variables (y_1^*, y_2^*), where

$$y_{ji}^* = x_{ji}\beta_j + \varepsilon_j, \quad j = 1, 2,$$

$$(\varepsilon_1, \varepsilon_2) \sim N(0, \Omega) \quad \text{and} \quad \Omega = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma^2 \end{bmatrix} \tag{73}$$

and

$$y_{2i}^* = \begin{cases} (y^\lambda - 1)/\lambda & \text{for } \lambda > 0 \\ \log(y_i) & \text{for } \lambda = 0 \\ 0 & \text{otherwise.} \end{cases} \Bigg\} \text{ iff } y_{1i}^* > 0 \text{ and } y_{2i}^* > -1/\lambda, \tag{74}$$

In other words, the conditional distribution of the latent variables is assumed to be bivariate normal. This specification allows participation to depend on both sets of regressors x_{1i} and x_{2i} and permits stochastic dependence between the two error terms. In addition, the use of the Box–Cox transformation relaxes the normality assumption on the conditional distribution of y_i. But this is at the price of making greater demands on the data and care should be taken to check for evidence of over-fitting.

Yen and Jones (1996) show that the log-likelihood function for a sample of independent observations is

$$\begin{aligned}\operatorname{Log} L = {} & \sum_{y=0} \log\bigl[1 - \Phi\bigl(x_1\beta_1, (x_2\beta_2 + 1/\lambda)/\sigma, \rho\bigr)\bigr] \\ & + \sum_{y>0} \log \Phi\bigl[\bigl(x_1\beta_1 + (\rho/\sigma)\bigl\{(y^\lambda - 1)/\lambda - x_2\beta_2\bigr\}\bigr)/\sqrt{(1 - \rho^2)}\bigr] \\ & + \sum_{y>0} (\lambda - 1)\log(y_i) \\ & + \sum_{y>0} \log\bigl[(1/\sigma)\phi\bigl(\bigl\{(y^\lambda - 1)/\lambda - x_2\beta_2\bigr\}/\sigma\bigr)\bigr], \end{aligned} \tag{75}$$

where Φ denotes a univariate or bivariate standard normal CDF, ϕ denotes the univariate standard normal PDF, and $\rho = \sigma_{12}/\sigma$. The general model can be restricted to give various special cases:

(i) $\sigma_{12} = 0$ gives the Box–Cox double hurdle with independent errors.

(ii) $\lambda = 1$ gives the standard double hurdle with dependence. This model is applied to UK data on household tobacco expenditure from the 1984 Family Expenditure Survey (FES) in Jones (1992), and to Spanish Family Expenditure Survey data for 1980–81 in Garcia and Labeaga (1996). The special case in which the error terms are assumed to be independent is applied to FES data on household tobacco expenditure in Atkinson et al. (1984), UK data on individual cigarette consumption from the 1980 General Household Survey (GHS) in Jones (1989), and to US data on wine consumption in Blaylock and Blisard (1993).

(iii) With $\lambda = 0$ the likelihood function corresponds to the generalized Tobit model with $\log(y_i)$ as dependent variable in the regression part of the model. Setting $\sigma_{12} = 0$

gives the special case of the two-part model in which normality is assumed and the equations are linear. Studies of smoking based on the two-part model include Lewit et al. (1981), Wasserman et al. (1991), and Blaylock and Blisard (1992).

Yen and Jones (1996) apply the Box–Cox double hurdle model to data on the number of cigarettes smoked in a sample of current and ex-smokers from the British Health and Lifestyle Survey. The estimated Box–Cox parameter (λ) equals 0.562 which is significantly different from both zero and one at the 0.01 level. Thus, both the standard double hurdle model and generalized Tobit model are rejected.

5. Unobservable heterogeneity and simultaneous equations

5.1. Linear models

5.1.1. Instrumental variables

Problems of unobservable heterogeneity bias and simultaneity have received particular attention in the context of empirical studies of health production. A pioneering paper is Auster et al.'s (1969) analysis of cross sectional data on death rates across the United States in 1960. They specify a Cobb–Douglas model for mortality rates, as a function of medical care and environmental variables. This is estimated by two-stage least squares (2SLS) to allow for the possible endogeneity of medical care, recognizing that aggregate mortality rates may influence the level of spending on medical care at the State level.

Rosenzweig and Schultz (1983) highlight the problem of unobservable heterogeneity bias in a study of child health production and the demand for child health inputs. They consider estimation of a structural health production function

$$y = f(x, z, \mu), \tag{76}$$

where y is a measure of child health, x are goods that affect health such as nutrition, z is medical care, and μ is an unobservable (to the researcher) variable reflecting the child's genetic and environmental endowment. If the child's parents are aware of μ, it may influence the reduced form demands for health inputs; for example, a mother who has a history of complications during previous pregnancies may be more likely to seek early prenatal care. Then the marginal effect of medical care on health is

$$\partial y/\partial z = f_z + f_\mu \partial\mu/\partial z. \tag{77}$$

So, estimates that fail to control for μ will give biased estimates of the effect of medical care on health (f_z). Rosenzweig and Schultz's (1983) proposed solution is to find instruments that predict the use of medical care but do not have an independent effect on health outcomes, and to estimate the model by 2SLS. Data on live births from the US National Natality Followback Surveys for 1967–69 are used and separate models are estimated for birth weight, the length of the gestation period, and the fetal growth rate. Estimates of the impact of the delay before the mother sought medical care change significantly when 2SLS is used rather than OLS.

A similar static health production framework is adopted by Mullahy and Portney (1990) to estimate the impact of smoking and atmospheric pollution on respiratory health. They use individual data from the 1979 US National Health Interview Survey, and estimate models for a binary dependent variable indicating whether the individual experienced days when their activities were limited by respiratory illness, and for the actual number of restricted activity days. Both models are estimated by OLS and by the generalized method of moments (GMM), where the latter uses the price of cigarettes and additional demographic variables to instrument the measure of cigarette smoking and the estimation uses 2SLS with a Huber–White correction for heteroscedasticity. In order to assess the sensitivity of the results to the use of instrumental variables, the models are estimated on different sub-samples and with different instrument sets. The results appear to be robust and show that allowing for unobservable heterogeneity bias increases the estimated impact of smoking relative to the impact of atmospheric pollution.

Mullahy and Sindelar (1996) extend the use of GMM estimation of the linear probability model to a two equation system in which a measure of problem drinking is treated as an endogenous regressor in equations for employment and unemployment (with non-participation in the workforce as the omitted employment status). The study is careful to acknowledge the possibility of IV bias, which arises if the instruments are poor predictors of the endogenous regressor, and it reports F-statistics for the significance of the instruments in the reduced form regressions [see Bound et al. (1995)]. Data from the 1988 Alcohol Supplement of the NHIS are used and the estimates show that problem drinking has a negative effect on employment.

5.1.2. The MIMIC model

In models of the demand for health and of health status indexes, problems of endogeneity are compounded by the fact that the central concept, "health", is inherently unobservable and has to be proxied by indicator variables. Multiple causes-multiple indicators, or MIMIC, models have been widely used to deal with the problem of latent variables. MIMIC models are estimated as LISREL (linear structural relationships) models. Examples of the the use of LISREL models of the demand for health include Erbsland, Ried and Ulrich (1995), Hakkinen (1991), van Doorslaer (1987), van de Ven and van der Gaag (1982), Wagstaff (1986, 1993), Wolfe and van der Gaag (1981). Van de Ven and Hooijmans (1991) and van Vliet and van Praag (1987) concentrate on the derivation of health status indexes from the MIMIC model. Behrman and Wolfe (1987) estimate a structural model of health production functions for maternal and child health in Nicaragua; their latent variables include health inputs such as nutrition, along with community and maternal health endowments.

An illustration of the MIMIC approach is van der Gaag and Wolfe's (1991) study which uses data on adults and children from the 1975 Rochester Community Child

Health Survey. The problem they address is that health has to be proxied by multiple indicators, none of which is a perfect measure of health. To set the scene for their model they show that principal components analysis can be used to reduce the dimensions of the problem; in their case 26 health indicators are reduced to 4 independent factors. They also show that socio-economic factors affect health, and that the estimated effects depend on the particular measure of health that is used. The relationship between socio-economic factors, desired health, and the demand for medical care is explored through a structural model

$$\begin{aligned} H^* &= x\alpha + \varepsilon_1, && (78)\\ D_j &= z\beta_{1j} + H^*\beta_{2j} + \varepsilon_{2j}, \quad j = 1, \ldots, 4, && (79)\end{aligned}$$

where H^* is the (unobserved) desired health status, x and z are socio-economic variables, and the D_js are four observed measures of the demand for medical care. This is combined with the measurement models

$$HP_l = H^*\gamma_l + \varepsilon_{3l}, \quad l = 1, \ldots, L, \tag{80}$$

where the *HP* are proxy measures of health. (78)–(80) are estimated by maximum likelihood as a LISREL model. This assumes joint normality of the error terms and makes use of covariance restrictions to identify the model, so that the unobservable H^* is proxied by a linear combination of the health indicators. Van der Gaag and Wolfe (1991) are careful to point out that the kind of health indicators and measures of health care utilization that commonly arise in survey data are often discrete variables, and the assumptions of the LISREL model may not be appropriate when dealing with discrete variables.

5.2. Nonlinear models

5.2.1. A framework

Blundell and Smith (1993) provide a general framework which is useful to categorize simultaneous equation models involving limited dependent variables. The model consists of an observation mechanism for a limited dependent variable (y_1)

$$y_{1i} = g\big(y_{1i}^*, y_{3i}^*\big) \tag{81}$$

and structural equations

$$\begin{aligned} y_{1i}^* &= \alpha_1 h\big(y_{1i}^*, y_{3i}^*\big) + \gamma_1 y_{2i} + x_{1i}\beta_1 + \varepsilon_{1i}, && (82)\\ y_{2i} &= y_{2i}^* = \alpha_2 h\big(y_{1i}^*, y_{3i}^*\big) + x_{2i}\beta_2 + \varepsilon_{2i}. && (83)\end{aligned}$$

The presence of the additional latent variable y_{3i}^* allows for the possibility of sample selection bias. In models without selection bias $g(y_{1i}^*, y_{3i}^*) = g(y_{1i}^*)$ and $h(y_{1i}^*, y_{3i}^*) =$

$h(y^*_{1i})$. In many of the applications discussed below y_{1i} is a binary variable and y_{2i} is continuous; this is the example that will be pursued here.

Blundell and Smith draw attention to the distinction between what they call Type I and Type II models. In Type I models $h(y^*_{1i}) = y^*_{1i}$, and the identification condition for the model is $\alpha_1 = 0$. This implies that a Type I specification is appropriate if the structural model is based on a simultaneous equations involving the latent variables

$$y^*_{1i} = \gamma_1 y_{2i} + x_{1i}\beta_1 + \varepsilon_{1i}, \tag{84}$$

$$y_{2i} = \alpha_2 y^*_1 + x_{2i}\beta_2 + \varepsilon_{2i}. \tag{85}$$

An example of the use of a Type I specification in health economics is Hamilton et al.'s (1997) study of the impact of unemployment on mental health. In their model y^*_1 is a latent index of employability and y_{2i} is mental health, measured by the Psychiatric Symptom Index. However if their structural model had predicted that an individual's actual employment status influenced their mental health, then a Type II specification would be more appropriate.

In a Type II model $h(y^*_{1i}) = y_{1i}$. This is appropriate if the outcome y_{2i} depends on the actual realization y_{1i}

$$y^*_{1i} = \alpha_1 y_{1i} + \gamma_1 y_{2i} + x_{1i}\beta_1 + \varepsilon_{1i}, \tag{86}$$

$$y_{2i} = \alpha_2 y_{1i} + x_{2i}\beta_2 + \varepsilon_{2i}. \tag{87}$$

Type II specifications raise the problem of coherency conditions. These reflect the logical consistency of the model and are required for the model to have unique reduced form solutions. For example in the model (86) and (87) the restriction, $\alpha_1 + \alpha_2\gamma_1 = 0$, ensures that the probabilities $\mathrm{P}(y_{1i} = 0)$ and $\mathrm{P}(y_{1i} = 1)$ sum to one.

Estimation of the LDV equations in Type I and Type II models requires different approaches. The Type I specification gives two equations to estimate: one, (84), with the LDV as dependent variable and one, (85), with the continuous dependent variable. Various estimators are available for the LDV Equation (84). Of these, two have been favored in the health economics literature. The two-step or IV estimator replaces the actual values of y_{2i} with fitted values from OLS estimates of the reduced form. The use of predicted values means that the covariance matrix of the estimates should be adjusted to allow for the additional sampling variability [see, e.g., Maddala (1983)]. The conditional maximum likelihood approach (CML), developed by Smith and Blundell (1986) for the Tobit model and by Rivers and Vuong (1988) for the probit, adds the OLS residuals to the equation. The t-statistic for the residuals provides a simple test for the exogeneity of y_2. Blundell and Smith (1993) propose a CML estimator for the Type II LDV Equation (86).

5.2.2. Applications

In two related papers, Kenkel (1990, 1991) estimates models for health related behaviour in which continuous measures of health knowledge are treated as endogenous regressors, due to unobservable heterogeneity bias, and replaced by fitted values. Kenkel (1990) uses a survey of 5,336 household from a 1975–76 survey carried out by the Centre for Health Administration Studies and the National Opinion Research Center (CHAS-NORC) of the University of Chicago and looks at the relationship between a general index of health knowledge and physician visits. The probability of a physician visit is modeled using the two-stage probit estimator, replacing the actual values of health knowledge with fitted values from an OLS reduced form. The number of visits is estimated using a simultaneous equation version of the sample selection model, using fitted values along with the Inverse Mill's Ratio from a reduced form probit equation. Kenkel (1991) uses data from the 1985 US National Health Interview Survey for three measures of health related behaviour, smoking, drinking and exercise. These are all left censored variables and Tobit models are used. In all cases the Smith-Blundell test rejects the exogeneity of health knowledge. Kenkel discusses the goodness of fit of the OLS reduced forms. He argues that the results are reasonable for the measures of knowledge about the health effects of smoking and exercise ($R^2 \approx 0.12$–0.19), but rather poor for alcohol ($R^2 = 0.02$).

Bollen et al. (1995) use data from the Tunisian Demographic and Health Survey to illustrate the practical relevance of Monte Carlo experiments on the performance of different estimators for simultaneous equation probit models reported by Guilkey et al. (1992). Their model involves a binary measure of contraceptive use and a measure of the family's desired number of children, which for the purposes of the analysis is treated as continuous and susceptible to unobservable heterogeneity bias. Like Guilkey et al., they apply a range of estimators: these are the standard probit model, the two-step probit estimator, the conditional ML estimator (CML), FIML, GMM, and a LISREL specification. In the case of the two-step and CML estimators, they rely on Monte Carlo evidence from Guilkey et al. to justify using the standard estimates of the covariance matrix, rather than adjusting the standard errors to allow for the fact that predicted values are being used rather than actual values. Overidentification tests are used to assess the validity of the instruments. These are implemented by comparing the log-likelihood for the model with fitted values and with an unrestricted version in which instruments are added to the equation directly. In their empirical application the exogeneity of the desired number of children cannot be rejected and the simple one-step probit model is favored. The empirical results are used to reinforce the message from Guilkey et al.'s Monte Carlo evidence that the performance of the two-step estimators relative to the simple probit model depends on the goodness of fit in the reduced form equations and on the degree of identification, reflected by the number of regressors (x) that are common to both equations.

5.2.3. Switching regressions

The models discussed above include the case of an endogenous binary variable which, in effect, shifts the intercept of the regression function under different regimes. The switching regression model extends this to deal with the case where the whole regression function, slope coefficients as well as the intercept, switches under different regimes. Examples from the health economics literature include O'Donnell's (1993) study of disability and labor supply, in which the income function depends on an individual's labor market status; and Gaynor's (1989) model of nonprice competition within group practices, in which the regression equation for the efficient price locus switches between regimes when demand is constrained or unconstrained.

O'Donnell (1993) uses data from the UK OPCS Disability Survey to investigate the influence of disability benefits on labor market participation by disabled people. The nature of the tax-benefit system means that individuals face a non-convex budget constraint and labor market participation is modeled using a fixed hours specification. A linear utility model leads to a structural labor market participation index

$$d_i^* = \alpha(y_{1i} - y_{2i}) + x_i\beta + \varepsilon_i. \tag{88}$$

This gives the net utility from working, and depends on the gap between income in work (y_{1i}) and income out of work (y_{2i}), along with socio-economic characteristics x_i. The problem with estimating (88) is that, for a particular individual, only one level of income can be observed. In order to measure the income gap, incomes have to be predicted using reduced form functions,

$$y_{1i} = z_{1i}\alpha_1 + \varepsilon_{1i}, \quad \varepsilon_1 \sim N(0, \sigma_1^2), \tag{89}$$

$$y_{2i} = z_{2i}\alpha_2 + \varepsilon_{2i}, \quad \varepsilon_2 \sim N(0, \sigma_2^2). \tag{90}$$

Because labor market participation is a choice that depends on the levels of y_{1i} and y_{0i}, this gives a switching regression model with endogenous switching. To estimate the model, (89) and (90) are substituted into (88) to give a reduced form participation equation. This is estimated as a probit model, and the inverse Mill's ratio is added to the income equations to obtain selectivity corrected estimates. The structural model is identified by exclusion restrictions on β. As long as the model is over-identified, the predicted values of y_{1i} and y_{0i} from the income equations can then be used to obtain consistent, but inefficient, estimates of the parameters of the structural participation equation. In his empirical results, O'Donnell finds that the income gap has a significant effect on labor market participation, although the magnitude of the effect is sensitive to the functional form adopted. He uses the estimates to simulate the impact of the introduction of Disability Working Allowance on employment.

Hay (1991) estimates a variant of the switching regression model with a multinomial logit participation criterion. This allows him to estimate a model for physician's incomes in which their choice of specialty (between GP, internal medicine, and other specialties) may involve self-selection and be influenced by income differentials. Using US data from the Seventh Periodic Survey of Physicians for 1970, he finds the estimated effect of income on choice of specialty changes sign in estimates that take account of selectivity bias.

5.2.4. Simulation estimators

The generic simultaneous equation model outlined by Blundell and Smith is relatively tractable in that it only involves one limited dependent variable and one continuous dependent variable. This allows the use of two-step instrumental variables estimators. Estimating systems of multiple LDV equations entails joint ML estimation and, due to the numerical integration involved, computation soon becomes intractable. Simulation estimators provide an alternative [see, e.g., Hajivassiliou (1993)]. Examples of these techniques are only beginning to appear in the health economics literature. For example, Pudney and Shields (1997) use simulated maximum likelihood (SML) estimation to deal with unobservable heterogeneity in a model of pay and promotion in the internal labor market for British NHS nurses. They use information from a one-off postal survey conducted for the Department of Health to model the speed of promotion, and the focus of the analysis is on an ordered probit equation the six nursing grades. Pudney and Shields allow for the possibility that unobservable heterogeneity bias affects five variables which reflect the individual's employment and training history, such as career breaks and in-service training. These variables are themselves measured as binary or ordered probits.

Hamilton (1999) uses a Bayesian panel data Tobit model of Medicare expenditures for recent US retirees that allows for deaths over the course of the panel. A Tobit model is used because the individual data on monthly medical expenditures from the New Beneficiary Survey contains a high proportion of zeros. This is combined with a probit equation for mortality to give a simultaneous equations LDV model. Hamilton argues that estimation can be easily handled by Bayesian Markov Chain Monte Carlo (MCMC) methods. The estimates are generated by an iterative algorithm based on the idea of data augmentation, where numerical simulations are used to generate the latent variables corresponding to the LDVs and then standard linear systems estimators are used (SURE). The model is implemented using a multivariate t distribution rather than normality to allow for heavy tails in the distribution of medical expenditure. The results suggest that survival effects are important, with a higher probability of mortality associated with higher medical expenditure in the last year of life.

6. Longitudinal and hierarchical data

6.1. Multilevel models

Multilevel models are used to analyze data that fall naturally into hierarchical structures consisting of multiple macro units, and multiple micro units within each macro unit. Emphasis is placed on defining and exploring variations at each level of the hierarchy after conditioning on the set of explanatory variables of interest. To illustrate the basic structure of a multilevel model, consider a simple linear model consisting of two levels which may represent patients ($i = 1, \dots, n$) nested within hospitals ($j = 1, \dots, m$). y_{ij}

represents the outcome of interest which is related to a vector of explanatory variables x in the following manner:

$$y_{ij} = x_{ij}\beta + \mu_j + \varepsilon_{ij}. \tag{91}$$

Assume that the random error term for patient i in hospital j, ε_{ij}, has zero mean and constant variance σ_ε^2. The effects of hospitals are estimated through μ_j which is assumed random and again has a mean of zero and constant variance σ_μ^2. Finally assume that patient and hospital effects are uncorrelated, $\mathrm{cov}(\varepsilon_{ij}, \mu_j) = 0$.

For the ith patient within the jth hospital, the conditional variance is $\mathrm{var}(y_{ij} \mid x_{ij}\beta) = \sigma_\mu^2 + \sigma_\varepsilon^2$ and hence, the overall variance is partitioned into components for both hospitals and patients. The partitioning of the variance in this manner leads to the intra-group correlation coefficient, $\rho = \sigma_\mu^2(\sigma_\mu^2 + \sigma_\varepsilon^2)^{-1}$, which measures the strength of nesting within the data hierarchy and is fundamental to the estimation procedures for multilevel models. In the presence of a non-zero intra-group correlation, estimation usually proceeds through the use of generalized least squares (GLS). Various estimation routines have been developed for the analysis of hierarchical data structures and these are reviewed in Rice and Jones (1997).

An alternative to the use of GLS is the Generalized Estimating Equations (GEE) approach; however this is principally concerned with estimates of fixed part parameters (for explanatory variables) rather than exploring the random part. GEE is typically used for clustered data, where there are a large number of clusters. These kinds of data are common in evaluations of prevention programmes which randomize clusters of individuals, rather than specific individuals, to treatment programmes. Norton et al. (1996) use GEE estimates for linear and logistic regressions in an evaluation of Drug Abuse Resistance Education (DARE) using individual data based on a random sample of schools.

It is conceivable that the relationship between an explanatory variable and the response is not the same across all hospitals. Certain hospitals may have the effect of increasing the average response (for example, length of stay) of younger patients while decreasing the stays of older patients. The exploration of different ‘higher level effects’ can be obtained by the inclusion of random coefficients [see, e.g., Gatsonis et al. (1995)]. Then, the slope effect associated with an explanatory variable (x_{ij}) can be represented by

$$y_{ij} = x_{ij}\beta + x_{ij}\gamma_j + \mu_j + \varepsilon_{ij}. \tag{92}$$

In (92) there are three random terms, two of which are random at the hospital level, γ_j and μ_j. This highlights the use of random coefficients by allowing regression coefficients to vary across level 2 units. However more complex variance structures can be introduced at any level of the hierarchy, including level 1, and this may lead to interesting interpretations and better model specification.

The models discussed so far represent the most basic form of a multilevel model where a continuous response is linearly related to a set of explanatory variables and the structure of the hierarchy is simple. In terms of the contributions to health and health economics research, more complex multilevel models may have the most to offer. For example, interest may be focused on the efficiency of both clinicians and provider units when assessing performance. In such a situation the hierarchy consists of patients within clinicians within provider units, and a multilevel model containing three levels is required. Alternatively, data may consist of a series of repeated measurements on patients attending different hospitals. This structure can be modeled using three levels; observations within patients, within hospitals. In reality, clinicians may operate in more than one hospital. In such situations the hierarchy is termed cross-classified. This occurs when individuals within a lower level cluster are grouped into a different higher level unit than peers from the same cluster.

Many health applications are not suited to a simple model with a linear link function and further extensions to incorporate generalized linear models, including link functions for logit, probit, Poisson, negative binomial, duration (survival) and multinomial models may be specified. The range of applications of multilevel models in health economics is discussed in Rice and Jones (1997).

An example of a linear multilevel model is the analysis of intertemporal preferences for future health by Cairns and van der Pol (1997). Survey respondents were asked to identify what future level of benefit make them indifferent between a specified benefit to be received one year in the future and the more distant delayed benefit. Each respondent was asked to provide estimates of their chosen future level of benefit for two different periods of delay. From the sample data collected implied discount rates for each respondent were calculated and regressed against the set of explanatory variables. The results compare an OLS specification and a multilevel specification of a hyperbolic discounting model. First, it appears that the OLS standard errors are underestimated, and hence the significance of the coefficients are exaggerated. Second, the partitioning of the variance between that observed across responses within respondents and that across respondents themselves allows the intra-class correlation to be estimated. The vast majority of variation (98%) exists across individuals. This suggests that respondents vary greatly in their time preferences, but in comparison appear to be reasonably consistent in applying discount rates to different periods of delay. An advantage of applying a multilevel specification to these data is that the heterogeneity across individuals is modeled whilst preserving degrees of freedom. Due to the lack of multiple responses elicited from individuals, a fixed effects specification would be prohibitive in this application.

Scott and Shiell (1997) apply multilevel analysis with a binary logit link function. Their study analyses the impact of a change in the reimbursement of Australian GPs in 1990. This involved a move from a system based on the length of consultation, to one based on fee descriptors reflecting the content of the consultation, for those GPs on the vocational register. Data are taken from the 1990–91 Australian Morbidity and Treatment Survey. Their working dataset consists of 4,185 consultations for upper respiratory tract infections and sprain/strains, nested within 412 GPs, within 25 types of local area. Three binary dependent variables are investigated measuring prescribing, therapeutic treatments, and counselling. The multilevel model can be expressed in terms of the log-odds ratio for patient i being treated by GP j

$$\log\left[\pi_{ij}/(1-\pi_{ij})\right]=x_{ij}\beta+z_j\gamma+\mu_j+\varepsilon_{ij}, \tag{93}$$

where the x's are measured patient characteristics and the z's are measured GP characteristics. Estimation is based on software which linearizes the model and uses a quasi-likelihood procedure. The results do not show a significant effect of the change in reimbursement, proxied by membership of the vocational register, on counselling or treatment, but they do show that prescribing is reduced.

6.2. Random versus fixed effects

The literature on Panel data techniques places emphasis on the relative merits of treating higher level units as random or fixed effects. In model (91), the individual effects (μ_j) are specified as random effects, but they could be specified as fixed effects, to be estimated together with β. The choice of specification requires careful consideration and may be determined by the data generating process and the type of inference sought. If individual effects are not of intrinsic importance in themselves, and are assumed to be random draws from a population of individuals and inferences concerning population effects and their characteristics are sought, then a random specification may be more suitable. However, if inferences are to be confined to the effects in the sample only, and the effects themselves are of substantive interest, then a fixed effects specification may be more appropriate.

Another important consideration is whether the explanatory variables are correlated with the effects. In such circumstances, random or fixed effects approaches may lead to very different estimates, and again careful consideration of the model specification is warranted. The situation can be extended to the multilevel model depicted in (91). When μ_j and x_{ij} are correlated, and group sizes are relatively small, the iterative generalized least squares estimator for the parameters β will be inconsistent. Treating the effects μ_j as fixed and applying a least squares dummy variable (LSDV) or within-groups/covariance (CV) estimator leads to consistent estimates. However, when group sizes are large, the two estimators can be shown to be equivalent [see Blundell and Windmeijer (1997)]. As the random effects estimator is both consistent and efficient when the effects are uncorrelated but inconsistent when they are correlated, and the fixed effects estimator is consistent regardless of the correlation, the two specifications can be compared by using a Hausman test.

In the situation where an explanatory variable is correlated with the higher level effects, and the sole concern of the analyst is the consistent estimation of the parameters associated with the explanatory variables or the mean effect of the higher levels, a fixed effects specification is likely to be preferable. However, in the multilevel framework, intrinsic interest lies in the estimation and interpretation of higher level variances, after conditioning on the set of explanatory variables.

6.3. Individual effects in panel data

6.3.1. Linear models

Applied work in health economics frequently has to deal with both the existence of unobservable individual effects that are correlated with relevant explanatory variables, and with the need to use nonlinear models to deal with qualitative and limited dependent variables. The combined effect of these two problems creates difficulties for the analysis of longitudinal data, particularly if the model includes dynamic effects such as lagged adjustment or addiction.

To understand these problems, first consider the standard linear panel data regression model, in which there are repeated measurements ($t = 1, \dots, T$) for a sample of n individuals ($i = 1, \dots, n$)

$$y_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}. \tag{94}$$

Failure to account for the correlation between the unobservable individual effects (μ) and the regressors (x) will lead to inconsistent estimates of the βs. Adding a dummy variable for each individual will solve the problem, but the least squares dummy variable approach (LSDV) may be prohibitive if there are a large number of cross section observations. The fixed effects can be swept from the equation by transforming variables into deviations from their within-group means. Applying least squares to the transformed equation gives the covariance or within-groups estimator of β (CV). Similarly, the model could be estimated in first differences to eliminate the time-invariant fixed effects. It should be clear that identification of β rests on there being sufficient variation within groups. In practice, fixed effects may only work well when there are many observations and much variation within groups.

One disadvantage of using mean deviations or first differences, is that parameters associated with any time invariant regressors, such as gender or years of schooling, are swept from the equation along with the fixed effects. Kerkhofs and Lindeboom (1997) describe a simple two-step procedure for retrieving these parameters; in which estimates of the fixed effects from the differenced equation are regressed on the time invariant variables. This is applied to a model of the impact of labor market status on self-assessed health.

The within-groups estimator breaks-down in dynamic models such as

$$y_{it} = \alpha y_{it-1} + \mu_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{iid}. \tag{95}$$

This is because the group mean, $y_{it-1} = (1/T)\sum_t y_{it-1}$, is a function of ε_{it} and ε_{it-1}. An alternative is to use the differenced equation

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta\varepsilon_{it} \tag{96}$$

in which case both y_{it-2} and Δy_{it-2} are valid instruments for Δy_{it-1} as long as the error term (ε_{it}) does not exhibit autocorrelation.

First differences are used by Bishai (1996) to deal with individual and family fixed effects in a model of child health. He develops a model of child health production which emphasizes the interaction between a caregiver's education and the amount of time they actually spend caring for the child. The aim is to get around the confounding of, effectively time invariant, levels of education with unobservable (maternal) health endowments. This is done by comparing the productivity of child care time given by members of the family with different levels of education. The model is estimated using the 1978 Intrafamily Food Distribution and Feeding Practices Survey from Bangladesh and the estimator used is the lagged instruments fixed effects estimator (LIFE) of Rosenzweig and Wolpin (1995). This uses differencing to remove the fixed effects, and then estimates the model by 2SLS, using lagged values of childcare time, family resource allocation, and child health as instruments to deal with the potential endogeneity of health inputs and the measures of health.

6.3.2. The conditional logit estimator

Now consider a nonlinear model, for example, a binary choice model based on the latent variable specification

$$y_{it}^* = x_{it}\beta + \mu_i + \varepsilon_{it}, \quad \text{where } y_{it} = 1 \text{ if } y_{it}^* > 0,\ 0 \text{ otherwise.} \tag{97}$$

Then, assuming that the distribution of ε_{it} is symmetric with distribution function $F(\cdot)$,

$$\mathrm{P}(y_{it} = 1) = \mathrm{P}(\varepsilon_{it} > -x_{it}\beta - \mu_i) = F(x_{it}\beta + \mu_i). \tag{98}$$

This illustrates the "problem of incidental parameters": as $n \to \infty$ the number of parameters to be estimated (β, μ_i) also grows. In linear models β and μ are asymptotically independent, which means that taking mean deviations or differencing allows the derivation of estimators for β that do not depend on μ. In general this is not possible in nonlinear models and the inconsistency of estimates of μ carries over into estimates of β.

An exception to this general rule is the conditional logit estimator. The conditional logit estimator uses the fact that $\sum_t y_{it}$ is a sufficient statistic for μ_i [see, e.g., Chamberlain (1980)]. This means that conditioning on $\sum_t y_{it}$ allows a consistent estimator for β to be derived. For example, with $T = 2$, $\sum_t y_{it} = 0$ is uninformative as it implies that $y_{i1} = 0$ and $y_{i2} = 0$. Similarly $\sum_t y_{it} = 2$ is uninformative as it implies that $y_{i1} = 1$ and $y_{i2} = 1$. But there are two ways in which $\sum_t y_{it} = 1$ can occur; either $y_{i1} = 1$ and $y_{i2} = 0$, or $y_{i1} = 0$ and $y_{i2} = 1$. Therefore analysis is confined to those individuals whose status changes over the two periods. Using the logistic function

$$\mathrm{P}(y_{it} = 1) = F(x_{it}\beta + \mu_i) = \exp(x_{it}\beta + \mu_i)/\big(1 + \exp(x_{it}\beta + \mu_i)\big) \tag{99}$$

it is possible to show that

$$\mathrm{P}\big[(0,1)\mid(0,1)\text{ or }(1,0)\big]=\exp\big((x_{i2}-x_{i1})\beta\big)/\big(1+\exp\big((x_{i2}-x_{i1})\beta\big)\big). \tag{100}$$

In other words, the standard logit model can be applied to differenced data and the individual effect is swept out.

Bjorklund (1985) uses the conditional logit model to analyze the impact of the occurrence and duration of unemployment on mental health using data from the Swedish Level of Living Survey. This includes longitudinal data which allows him to focus on individuals whose mental health status changed during the course of the survey. Bjorklund's estimates compare the conditional logit with cross section models applied to the full sample. He finds that the cross section estimates cannot, on the whole, be rejected when compared to the panel data estimates.

6.3.3. Parameterizing the individual effect

Another approach to dealing with individual effects that are correlated with the regressors is to specify $\mathrm{E}(\mu\mid x)$ directly. For example, in dealing with a random effects probit model Chamberlain (1980, 1984) suggests using

$$\mu_i=x_i\alpha+u_i,\quad u_i\sim\text{iid }N(0,\sigma^2), \tag{101}$$

where $x_i=(x_{i1},\ldots,x_{iT})$. Then, by substituting (101) into (97), the distribution of y_{it} conditional on x but marginal to μ_i has the probit form

$$\mathrm{P}(y_{it}=1)=\Phi\big[(1+\sigma^2)^{-1/2}(x_{it}\beta+x_i\alpha)\big]. \tag{102}$$

The model could be estimated directly by maximum likelihood (ML), but Chamberlain suggests a minimum distance estimator. This takes the estimates from reduced form probits on x_i, for each cross section, and imposes the restrictions implied by (102) to retrieve the parameters of interest (β,σ).

Labeaga (1993, 1999) develops the Chamberlain approach to deal with situations that combine a dynamic model and limited dependent variables. In Labeaga (1993) he uses panel data from the Spanish Permanent Survey of Consumption, dating from the second quarter of 1977 to the fourth quarter of 1983. Data on real household expenditure on tobacco is used to estimate the Becker and Murphy (1988) rational addiction model; a model that includes past and future consumption as endogenous regressors. The data contain around 40 per cent of zero observations and a limited dependent variable approach is required. The problems of endogeneity and censoring are dealt with separately, by using a GMM estimator on the sample of positive observations to deal with endogeneity and using reduced form T-Tobit models to deal with the limited dependent variable problem.

In Labeaga (1999) the two problems are dealt with simultaneously. To illustrate, consider a structural model for the latent variable of interest (say the demand for cigarettes)

$$y_{it}^* = \alpha y_{it-1}^* + x_{it}\beta + x_{it-1}\gamma + z_i\eta + \mu_i + \varepsilon_{it}. \tag{103}$$

This allows for dynamics in the latent variable (y^*) and the time varying regressors (x) as well as time invariant regressors (z). The observed dependent variable (y) is related to the latent variable by the observation rule

$$y_{it}^* = g(y_{it}), \tag{104}$$

where $g(\cdot)$ represents any of the common LDV specifications, such as probit, Tobit, etc.

This specification raises two problems: the inconsistency of ML in nonlinear models with fixed effects and a fixed T, and the correlation between the fixed effect and y_{it-1}^*. Labeaga's solution to this problem combines Chamberlain's approach to correlated individual effects with the within-groups estimator. Assume

$$\mu_i = w_i\alpha + u_i \quad \text{and} \quad \mathrm{E}\big(y_{i0}^* \mid w_i\big) = w_i\theta, \tag{105}$$

where $w_i = [x_{i1}, \ldots, x_{iT}, z_i$, and nonlinear terms in x_i and $z_i]$. The second assumption addresses the problem of the initial condition for the value of y^*. Using these assumptions it is possible to derive T reduced form equations, one for each cross section of data

$$y_{it}^* = w_i\pi_t + e_{it}. \tag{106}$$

Each of these can be estimated using the appropriate LDV model, implied by $g(\cdot)$, and specification tests can be carried out on these reduced form models. Once reduced form estimates of π_t have been obtained for each of the cross sections, they could be used in a minimum distance estimator. However Labeaga suggests applying the within-groups estimator to equation (103) using the reduced form fitted values of the latent variables (y_{it}^* and y_{it-1}^*). This gives consistent estimates of (α, β, γ), although they are less efficient than the minimum distance estimator. This approach can also deal with continuous endogenous explanatory variables (y_2) by using predictions from the OLS reduced form

$$\mathrm{E}(y_{2it} \mid w_i) = w_i\pi_2 \tag{107}$$

in the within-groups estimation.

Labeaga's (1993, 1999) results confirm the existence of addiction effects on the demand for cigarettes, even after controlling for unobservable individual heterogeneity. They show evidence of a significant, but inelastic, own-price effect.

Löpez (1998) makes use of Labeaga's approach to estimate the demand for medical care using the Spanish Continuous Family Expenditure Survey. The dependent variable measures expenditure on non-refundable visits to medical practitioners, for which 60 per cent of households make at least one purchase during the 8 quarters that they are measured. This leads López to use an infrequency of purchase specification for the LDV model $g(\cdot)$. He adopts the model of Blundell and Meghir (1987) which allows a separate hurdle for non-participation (identified as no purchases during 8 quarters) and which makes use of the identifying condition that $\mathrm{E}(y^*) = \mathrm{E}(y)$. In specifying the demand for medical care López combines the logarithmic version of the Grossman model with the partial adjustment model used by Wagstaff (1993). The estimates, for the impact of age, education, and the log(wage), show that controlling for censoring and unobservable individual effects does influence the results. This is to be expected, as unobservable heterogeneity is likely to be a particular problem in the use of expenditure survey data which do not contain any direct measures of morbidity.

The work of Dustmann and Windmeijer (1996) brings together many of the ideas discussed so far in this section. They develop a model of the demand for health care based on a variant of the Grossman model in which the demand for health capital is derived solely from the utility of increased longevity. Given the optimal path for health, they assume that there are transitory random shocks to the individual's health. If these fall below a threshold, the individual visits their GP. The model implies that the demand for medical care will depend on the ratio of the initial values of the individual's marginal utilities of wealth and of health; in other words the model contains an unobservable individual effect. The model is estimated with the first four waves of the German Socio-Economic Panel for 1984–87, using a sample of males who are measured throughout the period and who report visits to a GP. Poisson and negbin2 models are estimated for the number of visits and logit models are estimated for contact probabilities. The specifications of the Poisson and negbin2 models are discussed in more detail below in Section 7.

Dustmann and Windmeijer compare three strategies for dealing with the individual effects. The first is to use a random effects specification. In the negbin2 model the GEE approach is used to allow for the clustering of the data. For the logit model, a nonparametric approach is adopted. This approximates the distribution of unobservable heterogeneity using a finite set of mass points, μ_s, with associated probabilities, p_s [Heckman and Singer (1984)]. The likelihood function for this model is

$$L = \prod_i \sum_s p_s \left[\prod_t (\lambda_{its})^{yit} (1 - \lambda_{its})^{(1-yit)} \right], \tag{108}$$

where

$$\lambda_{its} = \exp(x_{it}\beta + \mu_s) / \bigl(1 + \exp(x_{it}\beta + \mu_s)\bigr) \tag{109}$$

and μ_s and p_s are parameters to be estimated. This finite density estimator has been used in other health economics applications, using both count data and survival data, and these are discussed in Sections 7 and 8.

The second strategy is to parameterize the individual effects. They adopt Mundlak's (1978) approach and parameterize the individual effects as a function of the group means for the time varying regressors (they report that they found very similar results with Chamberlain's approach of using all leads and lags of the variables).

The third strategy is to use conditional likelihood estimates of the logit and Poisson models. The log-likelihood for the conditional Poisson is similar to the logit model and takes the form,

$$\operatorname{Log} L = \sum_i \sum_t \Gamma(y_{it} + 1) - \sum_i \sum_t y_{it} \log\left[\sum_s \exp\bigl(-(x_{it} - x_{is})\beta\bigr)\right], \tag{110}$$

where $\Gamma(\cdot)$ is the gamma function ($\Gamma(q) = \int_0^\infty p^{q-1} e^{-p}\, dp$). Overall they find that the second and third strategies, which control for correlated effects, give similar estimates but that they differ dramatically from the random effects specifications. With the fixed effect estimators, the estimated impact of current income is reduced and becomes insignificant. This is consistent with their theoretical model which predicts that permanent rather than transitory income will affect the demand for health, and that the ratio of marginal utilities of wealth and health is a function of lifetime income.

6.3.4. *A semiparametric approach: the pantob estimator*

The Ministry of Health in British Columbia gives enhanced insurance coverage for prescription drugs to residents aged 65 and over. Grootendorst (1997) uses the "natural experiment" of someone turning 65 to investigate whether the effect of insurance on prescription drug use is permanent or transitory, and whether changes are concentrated among those on low incomes. He uses longitudinal claims data for around 18,000 elderly people for 1985–92. This dataset does not include measures of health status and it has to be treated as an "individual specific fixed endowment subject to a common rate of decay", which is modeled as a fixed effect, μ_i, and an (observable) age effect.

The measure of prescription drug utilization is censored at the deductible limit and Grootendorst uses Honoré's (1992) panel Tobit estimator (pantob). This estimator deals with censoring and fixed effects, and allows for a non-normal error term. It requires that the latent variable (y^*), after controlling for covariates, is independently and identically distributed for each individual over time. For the case of $T = 2$

$$y_{it}^* = x_{it}\beta + \mu_i + \varepsilon_{it}, \quad t = 1, 2. \tag{111}$$

If ε_{i1} and ε_{i2} are i.i.d. then the distribution of (y_1^*, y_2^*) is symmetric around a 45° line through $(x_{i1}\beta, x_{i2}\beta)$. This symmetry gives a pair of orthogonality conditions which imply objective functions that can be used to derive estimators of β. Honoré shows that the estimators are consistent and asymptotically normal for T fixed and $n \to \infty$. Grootendorst's results suggest that there is no permanent effect on drug use, except for low income males. There is little evidence of a transitory effect and it appears that insurance coverage only makes a minor contribution to the growth in utilization.

7. Count data regressions

7.1. Count data

Count data regression is appropriate when the dependent variable is a non-negative integer-valued count, $y = 0, 1, 2, \ldots$. Typically these models are applied when the distribution of the dependent variable is skewed to the left, and contains a large proportion of zeros and a long right hand tail. The most common examples in health economics are measures of health care utilization, such as numbers of GP visits or the number of prescriptions dispensed over a given period.

Cameron and Trivedi (1986) use a range of measures of health care utilization from the 1977–78 Australian Health Survey (AHS), and this dataset has become a test-bed for many of the recent methodological innovations in the area. Cameron et al. (1988) use a sample of single person households from the AHS and their dependent variables include the number of hospital admissions and the number of days in hospital over the previous year, along with the number of prescribed and the number of non-prescribed medicines taken. Cameron and Trivedi (1993) use the same set of models to illustrate conditional moment tests for independence of the different count variables. Cameron and Windmeijer (1996) use the same data and models as Cameron and Trivedi (1986) to compare a range of models of goodness of fit for count data regressions, favoring those based on deviance residuals. Cameron and Johansson (1997) use the count of visits to (non-doctor) health professionals to illustrate a new estimator based on squared polynomial expansions of the Poisson model. Mullahy (1997b) uses the measure of number of consultations in the previous two weeks to explore the role of unobservable heterogeneity in accounting for excess zeros in count data.

Other applications to health care utilization include Cauley (1987), who estimates Poisson regressions for the number of outpatient visits during a year, using a random sample of individuals from the Southern California region of Kaiser Permante Medical Care Programs. Arinen et al. (1996) compare simple and two-part versions of the Poisson and negbin models for dental visits by young adults in Finland. Grootendorst (1995) uses self-reported utilization of medicines by individuals aged 55–75 in the 1990 Ontario Health Survey. Pohlmeier and Ulrich (1995) use cross-section data from the 1985 wave of the German Socio-Economic Panel to estimate hurdle models for the demand for ambulatory care, measured by the number of physician visits during the year. The same dataset is used by Geil et al. (1997), who exploit the unbalanced panel data for 1984–89 and 1992–94 to estimate models for the number of hospital trips each year. Primoff Vistnes and Hamilton (1995) use data from the 1987 US National Medical Expenditure Survey to estimate a negbin model of mothers' demand for pediatric ambulatory care. Deb and Trivedi (1997) use the same survey to estimate models for six different measures of health care utilization by the elderly; these include office visits and hospital outpatient visits to both physicians and non-physicians, along with emergency room visits, and inpatient stays. Coulson et al. (1995) use information on the number of prescriptions filled or re-filled over two weeks, among a sample of Medicare enrolled Pennsylvanians. Häkkinen et al. (1996) use information from telephone surveys on physician visits, over the previous six months, to analyze the impact of recession on the use of physician services in Finland. Gurmu et al. (1997) use data from Santa Barbara and Ventura counties taken from the 1986 Medicaid Consumer Survey to estimate models for the number of doctor and health center visits over a four month (120 day) period, presenting separate results for Medicaid eligible recipients and AFDC beneficiaries. Windmeijer and Santos Silva (1997) and Santos Silva and Windmeijer (1997) use the British Health and Lifestyle Survey to estimate models for the number of GP visits over the past month. Gerdtham (1997) uses measures of the number of physician visits and weeks of care over the past year from the Swedish Level of Living Survey for 1991.

Despite this emphasis on measures of health care utilization, count data models have proved useful in other areas. Mullahy (1997a) uses data on cigarette smoking from the 1979 US National Health Interview Survey, and on birthweight from the 1988 Child Health Supplement of the NHIS. Kenkel and Terza (1993) use count data on the number of drinks consumed over the two weeks, from the 1990 US National Health Interview Survey.

7.2. *The basic model: counts and durations*

To understand the nature of count data models consider the following simple example. Assume that the probability of an event (e.g., a GP visit), during a brief period of time (dt), is constant and proportional to its duration. So the probability equals λdt, where λ is known as the intensity of the process. Now consider the count of events from zero up to time t, say (y, t). These are random variables, and the discrete density function must satisfy

$$f(y, t + dt) = f(y - 1, t)\lambda\, dt + f(y, t)(1 - \lambda\, dt). \tag{112}$$

Letting $dt \to 0$ gives a differential equation which solves to give

$$f(y, t) = \mathrm{e}^{-\lambda t}\big[(\lambda t)^y / y!\big], \tag{113}$$

which is the joint density of y and t. This yields two additional distributions; the first for the count of events (y) over a fixed interval of time $(t = 1)$; and the second for the time (t) until the first occurrence of the event $(y = 1)$, or the "time until failure". This illustrates the point that the count data models discussed in this section are, in general, dual to the duration models discussed in Section 8. This duality applies to particular parametric models; if the count is Poisson the duration is exponential, if it is negative binomial the duration is Weibull. For example, in order to model episodes of mental health services in the RAND HIE, Keeler et al. (1988) use a Weibull model of the time elapsed before the first visit, rather than a negbin model for the number of visits. If suitable data are available, duration models are usually more precise than count data models as they can exploit the continuous variation in durations, and are not confined to an integer scale. For example, more information can be extracted by modeling the age of starting smoking rather than simply estimating a binary choice model for whether someone has started [e.g., Douglas and Hariharan (1994)].

Setting $t = 1$ in (113), gives the starting point for count data regression; the Poisson process

$$\mathrm{P}(y_i) = \mathrm{e}^{-\lambda i} \lambda^{yi} / y_i!. \tag{114}$$

This gives the probability of observing a count of y_i events, during a fixed interval. In order to condition the outcome (y) on a set of regressors (x), it is usually assumed that

$$\lambda_i = \mathrm{E}(y_i \mid x_i) = \exp(x_i \beta). \tag{115}$$

An important feature of the Poisson model is the equidispersion property; that $\mathrm{E}(y_i \mid x_i) = \mathrm{Var}(y_i \mid x_i) = \lambda_i$. Experience shows that this property is often violated in empirical data. In particular, the overwhelming majority of the empirical studies of health care utilization cited above show evidence of overdispersion $(\mathrm{E}(y_i \mid x_i) < \mathrm{Var}(y_i \mid x_i))$. With overdispersion, the Poisson model will tend to under-predict the actual frequency of zeros, and of values in the right hand tail of the distribution. The need for tests and remedies for overdispersion provide the motivation for many of the methodological developments discussed below.

There are two basic approaches that have been used to estimate count data regressions. Maximum likelihood estimation (ML) uses the fully specified probability distribution and maximizes the log-likelihood

$$\mathrm{Log}\, L = \sum_i \log\bigl[\mathrm{P}(y_i)\bigr]. \tag{116}$$

For the Poisson model, the ML estimator solves the first order conditions

$$x'(y - \lambda) = x'\bigl(y - \exp(x\beta)\bigr) = 0. \tag{117}$$

If the conditional mean specification is correct but there is under- or overdispersion, then the ML estimates of the standard errors will be biased. However the theory of pseudo-maximum likelihood (PML) estimation ensures that the estimates of β are consistent, and the standard errors can be adjusted by using an appropriate estimator of the covariance matrix [see, e.g., Gourieroux et al. (1984), Mullahy (1998), Windmeijer and Santos Silva (1997)].

The first-order moment condition (117) implies an alternative formulation of the Poisson model, as a nonlinear regression equation

$$\mathrm{E}(y_i \mid x_i) = \exp(x_i \beta). \tag{118}$$

This is the exponential conditional model (ECM) discussed in Section 4. An alternative approach to estimation, suggested by (118), is to use moment-based estimators, such as nonlinear least squares (NLLS) or generalized method of moments (GMM). For example the GMM estimator minimizes

$$(y-\lambda)'xW^{-1}x'(y-\lambda), \tag{119}$$

where W is a positive definite weighting matrix. As this approach only uses the first moment rather than the full probability distribution, it is more robust than ML. In fact the exponential conditional model encompasses other parametric specifications, such as the geometric and the negative binomial, both of which have the same conditional expectation.

7.3. *Overdispersion and excess zeros*

Many applied studies find that the frequency of zeros in count data is greater than the Poisson model would predict. One source of excess zeros in count data is overdispersion. Mullahy (1997b) emphasizes that the presence of excess zeros "is a strict implication of unobserved heterogeneity". In other words, the existence of unobservable heterogeneity may be sufficient to explain excess zeros, without recourse to alternative specifications such as zero inflated or hurdle models. He concentrates on the case where heterogeneity is modelled as a mixture; $\exp(x_i\beta+\mu_i)=\exp(x_i\beta)\eta_i$, with $\mathrm{E}(\eta_i)=1$. This includes the negbin model as a special case. Mullahy demonstrates that $\mathrm{P}(y_i=0)$ is greater for mixing models than for the Poisson model (where $\eta_i=1$ for all i). A similar result applies to the probability of events in the upper tail of the distribution. The intuition behind these results is that the additional dispersion, associated with mixing, spreads the distribution out to the tails. In this sense, the phenomenon of excess zeros is no more than a symptom of overdispersion.

The negative binomial specification allows for overdispersion by specifying, $\exp(x_i\beta+\mu_i)=[\exp(x_i\beta)]\eta_i$ where η_i is a gamma distributed error term (see, e.g., Cameron and Trivedi (1986)]. Then

$$\mathrm{P}(y_i)=\bigl\{\Gamma(y_i+\psi_i)/\Gamma(\psi_i)\Gamma(y_i+1)\bigr\}\bigl(\psi_i/(\lambda_i+\psi_i)\bigr)^{\psi_i}\bigl(\lambda_i/(\lambda_i+\psi_i)\bigr)^{y_i}, \tag{120}$$

where $\Gamma(\cdot)$ is the gamma function. Letting the "precision parameter" $\psi=(1/a)\lambda^k$, for $a>0$, gives

$$\mathrm{E}(y)=\lambda \quad \text{and} \quad \mathrm{Var}(y)=\lambda+a\lambda^{2-k}. \tag{121}$$

This leads to two special cases: setting $k=1$ gives the negbin 1 model with the variance proportional to the mean, $(1+a)\lambda$; and setting $k=0$ gives the negbin 2 model where the variance is a quadratic function of the mean, $\lambda + a\lambda^2$. Setting $a = 0$ gives the Poisson model, and this nesting can be tested using a conventional t-test. In general, a does not have to be constant and can be specified as a function of the regressors. The negative binomial has been applied extensively in studies of health care utilization; examples include Arinen et al. (1996), Cameron and Trivedi (1986), Cameron et al. (1988), Cameron and Windmeijer (1996), Cameron and Johansson (1997), Geil et al. (1997), Gerdtham (1997), Grootendorst (1995), Häkkinen et al. (1996), Pohlmeier and Ulrich (1995).

Although overdispersion can account for excess zeros, it may be that there is something special about zero observations *per se*, and an excess of zero counts may not be associated with increased dispersion throughout the distribution. This may reflect the role of the participation decision in the underlying economic model. Many studies of health care utilization have emphasized the principal-agent relationship between doctor and patient and stressed the distinction between patient initiated decisions, such as the first contact with a GP, and decisions that are influenced by the doctor, such as repeat visits, prescriptions, and referrals [see, e.g., Pohlmeier and Ulrich (1995)]. There are two approaches which place particular emphasis on the role of zeros; zero inflated models and hurdle, or two-part, models.

The "zero inflated" or "with zeros" model is a mixing specification which adds extra weight to the probability of observing a zero [see, e.g., Mullahy (1986)]. This can be interpreted as a splitting mechanism which divides individuals into non-users, with probability $q(x_{1i}\beta_1)$, and potential-users, with probability $1 - q(x_{1i}\beta_1)$. So the probability function for the zero inflated Poisson model, $\mathrm{P}^{\mathrm{ZIP}}(y \mid x)$ is related to the standard Poisson model, $\mathrm{P}^{\mathrm{P}}(y \mid x)$, as follows

$$\mathrm{P}^{\mathrm{ZIP}}(y \mid x) = 1(y = 0)q + (1 - q)\mathrm{P}^{\mathrm{P}}(y \mid x). \tag{122}$$

Zero inflated Poisson and negbin models can be estimated by maximum likelihood. However researchers often report problems in getting the estimates to converge when the full set of regressors are included in the splitting mechanism [see, e.g., Grootendorst (1995), Gerdtham (1997)].

In the count data literature, unlike the limited dependent variable literature discussed in Section 4, hurdle and two-part specifications are often treated as synonymous. The hurdle model assumes the participation decision and the positive count are generated by separate probability processes $P_1(\cdot)$ and $P_2(\cdot)$. The log-likelihood for the hurdle model is

$$\begin{aligned}\operatorname{Log} L = {} & \sum_{y=0} \log\bigl[1 - P_1(y > 0 \mid x)\bigr] \\ & + \sum_{y>0} \bigl\{ \log\bigl[P_1(y > 0 \mid x)\bigr] + \log\bigl[P_2(y \mid x, y > 0)\bigr]\bigr\}\end{aligned}$$

$$\begin{aligned}
&= \left\{\sum_{y=0}\log\bigl[1-P_1(y>0\mid x)\bigr]+\sum_{y>0}\log\bigl[P_1(y>0\mid x)\bigr]\right\}\\
&\quad+\left\{\sum_{y>0}\log\bigl[P_2(y\mid x, y>0)\bigr]\right\}\\
&= \operatorname{Log}L_1+\operatorname{Log}L_2. \qquad (123)
\end{aligned}$$

This shows that the two parts of the model can be estimated separately; with a binary process ($\operatorname{Log}L_1$) and the truncated at zero count model ($\operatorname{Log}L_2$). Mullahy (1986) introduces the hurdle specification for Poisson and exponential models, while Pohlmeier and Ulrich (1995) extend it by using a negbin 1 specification for both stages. Grootendorst (1995) applies the two-part model with a probit for the first stage and a negbin 2 model for the second, while Häkkinen et al. (1996) and Gerdtham (1997) use a logit for the first stage and a negbin 2 model for the second stage.

Grootendorst (1995) provides an empirical comparison of two-part and zero inflated specifications. The study uses data from the 1990 Ontario Health Survey to analyze the impact of copayments on the utilization of prescription drugs by the elderly, exploiting the fact that Ontario residents become eligible for zero copayments under the Ontario Drug Benefit Program on their 65th birthday. Zero inflated and two-part models are not parsimonious, often doubling the number of parameters to be estimated. As always, more complicated models may be prone to over-fitting, and to allow for this Grootendorst uses within-sample forecasting accuracy to evaluate their performance. The models are estimated on a random sample of 70 per cent of the observations. The estimated models are used to compute predictions for the remaining 30 per cent (the forecast sample). Models are then compared on the basis of the mean squared error for the forecast sample. In addition to the split-sample analysis, Voung's non-nested test is computed. The two-part models outperform the other specifications on all of the criteria. Having established this, Grootendorst goes on to show evidence of heteroscedastity in both the probit and negbin components of the model and parameterizes the heterogeneity, but the comparison of models is not repeated.

Pohlmeier and Ulrich (1995) are careful to point out that a limitation of the hurdle model is that it implies that the measure of repeat visits to the doctor relates to a single spell of illness, an issue that may be especially problematic with their annual data. This issue is explored by Santos Silva and Windmeijer (1997) who propose some alternative two stage count models that allow for multiple spells of illness. The observed total number of visits (y) is modeled as

$$y=\sum_{j=1}^{S}(1+R_j), \qquad (124)$$

where S is the number of illness spells, and R_j is the number of referrals, or repeat visits, during the jth spell. It should be clear that this definition of an illness spell implies that the individual will always make at least one visit to the doctor when they are ill. This perspective leads Santos Silva and Windmeijer to question the need for a truncated model in the second stage of hurdle models. However this seems to be an empirical issue; and allowing individuals to have zero visits during an illness spell may be relevant in studies of unmet need.

(124) implies that y has a stopped sum distribution. Santos Silva and Windmeijer consider two special cases. When S is Poisson and the R_j are independent identical Poisson variates, y has a Thomas distribution. When S is Poisson and $(1 + R_j)$ are logarithmic, y has a negative binomial distribution. The stopped sum specification allows S and R to be parameterized separately, as functions of variables that influence the first visit and that influence referrals. In the light of this, Santos Silva and Windmeijer argue that failure to recognize that the negbin model may reflect a two stage decision process and hence to parameterize the dispersion, may bias comparisons in favor of hurdle models.

Given assumptions about the distributions of S and R_j, the model could be estimated by ML. But pseudo-ML results do not apply, and misspecification of the stopped sum distribution can lead to inconsistent estimates. Instead, Santos Silva and Windmeijer rely on a first-order moment condition and derive a GMM estimator. They use the hypothesis of a single spell ($S = 1$) to generate testable overidentifying restrictions. The estimator is applied to data on the number of GP visits over the past month from the British Health and Lifestyle Survey. They find that the overidentification test does not reject the hypothesis that the observations are generated by a single spell of illness, suggesting that the hurdle specification may be adequate. The implication is that data collected over longer periods, such as a year, may be prone to the problem of multiple spells, and that, where possible, information should be collected for separate illness spells or episodes of care.

It is often argued that the zero inflated model illustrates the fact that excess zeros can arise even when there is no unobservable heterogeneity [see, e.g., Grootendorst (1995), Mullahy (1997b)]. For example, Grootendorst (1995) argues that comparing the negbin 2 model with a zero inflated negbin 2 allows the analyst to discriminate between unobservable heterogeneity and the splitting mechanism. However a recent paper by Deb and Trivedi (1997) puts a different perspective on the issue. They interpret the zero inflated model as a restrictive special case of a general mixture model with unobservable heterogeneity.

Deb and Trivedi deal with unobservable heterogeneity by using a finite mixture approach. The intuition is that observed counts are sampled from a mixture of different populations. They argue that zero inflated models are a special case of the mixture model, in which the zero counts alone are sampled from a mixture of two populations (non-users and potential users). Their model is implemented using a finite density estimator, where each population, j, is represented by a probability mass point, p_j [see, Heckman and Singer (1984)]. The C-point finite mixture negbin model takes the form

$$\mathrm{P}(y_i \mid \cdot) = \sum_{j=1}^{C} p_j \cdot P_j(y_i \mid \cdot), \quad \sum_{j=1}^{C} p_j = 1, \quad 0 \leqslant p_j \leqslant 1, \tag{125}$$

where each of the $P_j(y_i \mid \cdot)$ is a separate negbin model, and the p_js are estimated along with the other parameters of the model.

The model is applied to the demand for medical care among individuals aged 66 and over, in the 1987 US National Medical Care Expenditure Survey. Demand is measured by six different measures of utilization for a one year period, and the finite mixture model is compared to hurdle and zero inflated specifications. The finite mixture models are estimated by maximum likelihood, using two and three points of support. The models are compared on the basis of likelihood ratio (LR) and information criterion tests (IC), along with measures of goodness of fit. The negbin 1 models with two points of support are preferred on the basis of these statistical criteria. Deb and Trivedi interpret the points of support as two latent populations of "healthy" and "ill" individuals, reflecting unobserved frailty. Perhaps it is not surprising that a model of health care utilization among the elderly over a full year which splits the population in this way proves more applicable than the zero inflated and hurdle models, which split individuals into sub-populations of users and non-users.

While Deb and Trivedi apply a finite density estimator to the distribution of unobservable heterogeneity, Gurmu (1997) adopts a semiparametric approach, using a Laguerre series approximation of the unknown density function. This is applied to hurdle models because, unlike the standard model, misspecification of the density leads to inconsistent estimates of the conditional mean in hurdle models. The Laguerre polynomials are complex, but they do have closed form solutions and the model can be estimated by maximum likelihood. The model nests the Poisson hurdle model and the negbin hurdle with a binary logit for the first stage. In order to balance goodness of fit and parsimony, the number of terms in the Laguerre polynomials is selected according to the Akaike information criterion ($\mathrm{AIC} = -\{2 \operatorname{Log} L + 2 \cdot (\text{number of free parameters})\}/n$). Estimates of models for the number of doctor and health center visits from the 1986 Medicaid Consumer Survey suggest that the semiparametric estimator dominates the Poisson and negbin hurdle models, in terms of the maximized log-likelihood and the AIC. According to the AIC, a first order polynomial is preferred for the first stage, in other words, a logit model is adequate. While a second order polynomial is preferred at the second stage, giving a specification with greater flexibility than the standard negbin model.

Cameron and Johansson (1997) propose a new estimator that uses squared polynomial expansions around a Poisson baseline. This differs from Gurmu's (1997) approach in that the expansion is around the count density itself, rather than around the density of unobservable heterogeneity. This affects the mean as well as the dispersion. The model is estimated by maximum likelihood using a fast simulated annealing algorithm to deal with the problem of multiple local optima. Cameron and Johansson argue that their estimator is particularly suited for underdispersed data, which are rare in health applications. However for overdispersed data it provides an alternative to the negbin model. They apply the estimator to (non-doctor) health professional visits in the 1977–78 Australian Health Survey and find that their preferred specification, based on a 5th order polynomial, outperforms a negbin 2 model.

7.4. Unobservable heterogeneity and simultaneity biases

Count data models typically assume that unobservable heterogeneity is uncorrelated with the regressors (the same is true of the duration models discussed in Section 8). Mullahy (1997a) argues that this assumption may not hold in many applications, particularly when the unobservable heterogeneity (μ) represents unmeasured omitted regressors. He cites the example of health care utilization, where μ may reflect an individual's propensity for illness, in which case regressors measuring an individual's insurance coverage may be prone to self selection bias. Similarly, Dustmann and Windmeijer's (1996) model suggests that health care utilization will depend on correlated individual effects reflecting the ratio of the initial values of the individual's marginal utilities of wealth and of health. The problem may not be confined to individual characteristics; Pohlmeier and Ulrich (1995) argue that unobservable heterogeneity may reflect supply side factors that are not recorded in individual survey data. These variables may well be correlated with individual characteristics that influence their choice of provider as well as their rate of utilization of health care. The presence of correlated unobservable heterogeneity means that the standard estimators (ML, PML, NLLS) are inconsistent estimators of β. Mullahy (1997a) proposes the use of nonlinear instrumental variables, estimated by the generalized method of moments (GMM), as a fairly general solution to this problem.

The standard nonlinear instrumental variables estimator deals with the case in which unobservables are additively separable

$$y_i = \exp(x_i\beta) + \mu_i + \varepsilon_i, \tag{126}$$

where ε is a random error that is independent of x. But if μ is to be regarded as an omitted variable it may seem more natural to treat measured and unmeasured regressors "symmetrically" [see, e.g., Mullahy (1997a), and Terza (1998)]. This implies that a multiplicative specification should be used, including μ in the linear index

$$y_i = \exp(x_i\beta + \mu_i) + \varepsilon_i = \exp(x_i\beta)\eta_i + \varepsilon_i. \tag{127}$$

While this specification may seem more natural, it raises problems for the use of nonlinear IV estimators. In this context, the assumptions that define a set of valid instruments, z, are

$$\mathrm{E}(y \mid x, \eta, z) = \mathrm{E}(y \mid x, \eta), \tag{128}$$

$$\mathrm{E}(\eta \mid z) = 1. \tag{129}$$

Now consider the "standard" residual

$$u_i = y_i - \exp(x_i\beta), \tag{130}$$

where, from (127),

$$u_i = \exp(x_i\beta)(\eta_i - 1) + \varepsilon_i. \tag{131}$$

The problem is that this expression involves the product of functions of x and η. So, in general, $\mathrm{E}(u \mid z) \neq 0$, even if (128) and (129) hold. This means that nonlinear IV will be an inconsistent estimator of β. Mullahy's (1997a) solution to this problem is to transform the model so that the transformed residuals (u^{T}) do satisfy the standard conditions for the consistency of IV. Let

$$u_i^{\mathrm{T}} = u_i/\lambda_i = u_i/\exp(x_i\beta) = \exp(-x_i\beta)y_i - 1 = \eta_i + \exp(-x_i\beta)\varepsilon_i. \tag{132}$$

The transformed residual is additively separable in η_i, and Mullahy shows that $\mathrm{E}(u^{\mathrm{T}} \mid z) = 0$. He then derives an optimal GMM estimator using the transformed residuals to define the moment conditions.

The choice between multiplicative and additive specifications is taken up by Windmeijer and Santos Silva (1997) in the context of simultaneous equations models for count data. They emphasize that, in general, a particular set of instruments, z, will not be orthogonal to both u_i and u_i^{T}. They appear to be skeptical of the claim that a multiplicative specification is more natural, and argue that the choice is an empirical issue. This can be settled using tests for the overidentifying restrictions in cases where there are more instruments than endogenous regressors.

Windmeijer and Santos Silva use data from the 1991 British Health and Lifestyle Survey to investigate simultaneous equations models for GP visits, in which self-assessed health is treated as a binary endogenous regressor. They adopt the Blundell and Smith (1993) framework, discussed in Section 5, and compare type I and type II specifications. In the type II model, recorded health status is assumed to influence GP visits. In the type I model it is the latent health index that influences the number of visits. The coherency conditions for the type II model imply that the model is only logically consistent when it is specified as a recursive system. In other words, the type II specification can only be coherent when the endogeneity of self-assessed health stems from unobservable heterogeneity bias rather than classical simultaneous equations bias. Additive and multiplicative specifications of the type II model are estimated by GMM (alternative estimators for the multiplicative model are discussed by Terza (1998)); the type I specification is estimated using a two-step approach. The tests of the overidentifying restrictions favor the additive specification, although Hausman tests do not reject the exogeneity of self-assessed health.

8. Duration analysis

8.1. Survival and duration data

Statistical models of "time until failure" tend to be labeled survival analysis in the epidemiology and biostatistics literature, while the labor economics literature uses the label duration analysis. In health economics, the techniques have been applied to a range of datasets. The most obvious application of survival analysis is to individual lifespan and mortality rates, usually in the context of models of individual health production. For example, Behrman et al. (1990) use the Dorn survey of mortality among US veterans. While Behrman et al. (1991) analyze racial inequality in age specific death rates for males from the US Retirement History Survey (RHS). The RHS is also used by Butler et al. (1989) in a competing risks model for transitions into re-employment or death. Forster and Jones (1997a) use data on mortality from the British Health and Lifestyle Survey (HALS) to estimate a model of the demand for longevity.

However the techniques are not confined to studies of mortality rates. Keeler et al. (1988) model the time elapsed before the first use of mental health services among participants in the RAND HIE. Philipson (1996) uses the child health supplement of the 1991 US National Health Interview Survey (NHIS) to analyze the time elapsed before children have their first MMR vaccination. Douglas and Hariharan (1994) use the 1978 and 1979 smoking supplements of the NHIS to estimate a model for the age of starting smoking, while Forster and Jones (1997b) use the HALS dataset to analyze the number of years that someone smokes and the decision to quit, and Douglas (1998) uses the 1987 NHIS to model both age of starting and years of smoking. Morris et al. (1994) use data from a social experiment involving 36 for-profit nursing homes in San Diego to analyze length of stay by Medicaid recipients. Norton (1995) analyses the time to "spend-down" in nursing homes, modeling the time elapsed before an individual's personal assets are exhausted and they become eligible for Medicaid. Siddiqui (1997) uses the German Socio-Economic Panel to model the impact of chronic illness and disability on the probability of early retirement using a discrete time hazard rate model. Lindeboom et al. (1995) use a semi-Markov model for sickness, work, and job exit to explain sickness absenteeism among public school teachers in the Netherlands. Bhattacharya et al. (1996) use information on around 440,000 patients from the Japanese Ministry of Health and Welfare's 1990 Patient Survey to estimate a Cox proportional hazards model for the time elapsed between outpatient visits. The delay before adopting a new technology is used by Escarce (1996), to analyze the diffusion of laparoscopic cholecystectomy in a 1992 survey of US surgeons. Hamilton et al. (1996) and Hamilton and Hamilton (1997) use a competing risks specification for post-surgery length of stay and inpatient mortality to estimate the impact of waiting time on surgical outcomes and the volume-outcome relationship. Burgess and Propper (1998) use a discrete time logistic hazard model for the time to first marriage or cohabitation in a study of the impact of early health related behaviours on later life chances based on the US National Longitudinal Survey of Youth for 1979–92.

8.2. Methods

8.2.1. Semiparametric models

The key concept in duration models is the hazard function, defined as the rate of failure at a point in time, given survival to that time. Nonparametric, semiparametric and parametric duration models make assumptions of varying degrees of strength about the hazard function underlying the data generating process. The most commonly used semiparametric duration model is the proportional hazards model of Cox (1972). Applications of this approach in health economics include Behrman et al. (1990, 1991), Bhattacharya et al. (1996), Forster and Jones (1997a, 1997b), and Philipson (1996).

In the Cox model, the hazard function at time t for individual i, $h_i(t, x_i)$, is defined as the product of a baseline hazard function, $h_o(t)$, and a proportionality factor $\exp(x_i\beta)$

$$h_i(t, x_i) = h_o(t) \cdot \exp(x_i\beta), \tag{133}$$

where x_i is a vector of covariates and β is a parameter vector. The covariates may be time invariant, or the model can be extended to allow for time-varying covariates. For example Philipson (1996) sets out to estimate the "prevalence elasticity" of the demand for MMR vaccinations, and treats regional measles caseloads as a time-varying covariate.

Cox's method is described as being semiparametric because it does not specify the baseline hazard function $h_o(t)$. Estimation uses the partial log-likelihood function

$$\operatorname{Log} L = \sum_i \delta_i \left\{ x_i\beta - \log\left(\sum_{l \in R_i} \exp(x_l\beta) \right) \right\}, \tag{134}$$

where δ_i is a dummy variable equal to 1 if the observation exits the process of interest (for example, the age at death of an individual) and 0 if the observation is censored (for example, if an individual is still alive at the end of the data collection period). $l \in R_i$ are those observations in the risk set, R_i, at the time of exit of individual i. R_i includes those observations still alive and uncensored at the time of exit of individual i and whose entry time to the survey is less than or equal to the exit time of the individual (this controls for left truncation). By conditioning on the risk set the baseline hazard $h_o(t)$ is factored out of the partial likelihood function, in the same way that fixed effects are dealt with in the conditional logit model. The sampling distribution of the β that maximizes the partial likelihood is asymptotically normal, and the standard results of maximum likelihood estimation apply. In proportional hazards models, the estimates of the parameter vector β measure the effect of a unit change in the covariates of the model on the log of the proportionate shift in the baseline hazard function.

The partial likelihood approach relies on the proportionality of the hazard function to factor out the baseline hazard, and the estimates are sensitive to violations of proportionality. A related model, used in Forster and Jones (1997a), is the stratified proportional hazards model

$$h_{iv}(t) = h_{ov}(t) \cdot \exp(x_{iv}\beta), \tag{135}$$

where $h_{iv}(t)$ is the hazard function for individual i in stratum v, β is the common shift parameter vector, x_{iv} is the vector of explanatory variables for individual i in stratum v, and $h_{ov}(t)$. is the baseline hazard function in stratum v. This model can be used when misspecification tests suggests that non-proportional hazards exist for one or more covariates.

8.2.2. *Parametric models*

Partial likelihood methods discard information on actual failure times and use only their rank order. This reduces the efficiency of the estimates. An alternative is to adopt a parametric approach. Parametric models assume a functional form for the baseline hazard function. Many applied studies compare a variety of different functional forms in order to assess the best empirical specification. Behrman et al. (1990) use the Weibull, log-normal, log-logistic, and generalized gamma. Behrman et al. (1991) use the Weibull and log-logistic. Morris et al. (1994) use the exponential, Weibull, log-normal and generalized gamma. Norton (1995) compares the Weibull, log-normal, log-logistic and generalized gamma. Escarce (1996) uses the Weibull model with and without unobserved gamma heterogeneity.

Specifying the baseline hazard function as $h_o(t) = hpt^{p-1}$ gives the Weibull proportional hazards model

$$h_i(t) = hpt^{p-1} \cdot \exp(x_i\beta), \tag{136}$$

where p is known as the shape parameter. In the Weibull model, the shape of the baseline hazard function, pt^{p-1}, is shifted by the proportionality factor $h \cdot \exp(x_i\beta)$. The hazard is monotonically increasing for $p > 1$, showing increasing duration dependence, and monotonically decreasing for $p < 1$, showing decreasing duration dependence. The hazard function, $h(t) = f(t)/S(t)$, can be used to derive the probability density function, $f(t)$, and the survival function, $S(t)$, for the Weibull model, and the likelihood function with right censoring is

$$L = \prod_i \{f_i(t)/S_i(t)\}^{\delta_i} \cdot S_i(t). \tag{137}$$

Standard maximum likelihood estimation can be used to obtain estimates of the parameters p and β.

The Weibull model may also be estimated in what is called the accelerated time to failure format, which expresses the log of time as a function of the dependent variables and the shape parameter. Taking logs of both sides of (136) and simplifying gives

$$\log(t_i) = (1/p)\bigl\{-\log(h) - x_i\beta + \log\bigl(-\log(S_i(t))\bigr)\bigr\}, \tag{138}$$

where $\log(-\log(S_i(t)))$ has an extreme value distribution. In the accelerated failure time version of the Weibull model, the parameters $-\beta/p$ measure the effect of a one unit change in a covariate on the log of failure time. The Weibull model (and its special case the exponential model, when $p = 1$) is the only parametric model that can be expressed in both the proportional hazards and accelerated time to failure format. But a variety of functional forms are available for the latter. These include non-monotonic hazard functions such as the log-logistic and the generalized gamma.

In their analysis of US data on the age of starting smoking, Douglas and Hariharan (1994) argue that the standard survival analysis may not be appropriate and that a split-population model should be used. The standard survival analysis would treat individuals who had not started smoking by the time of the survey as incomplete spells, and it is assumed that all of these individuals will eventually "fail". The split-population specification allows for the possibility that some people will remain confirmed non-smokers. It augments the standard model by adding a probability, modeled as a probit, that an individual will never fail. A log-logistic specification is used for the hazard function; this is non-monotonic and captures the peak in starting smoking during the mid-teens. This approach is extended in Douglas (1998), who uses an ordered probit to split the sample between those who never start, those who start and eventually quit and those who start and never quit. A log-logistic hazard is used to model starting and a Weibull to model quitting

8.2.3. Unobservable heterogeneity

The existence of unobservable heterogeneity will bias estimates of duration dependence. To illustrate, imagine that survival data is sampled from two groups, a "frail" group and a "healthy" group, both of which have constant hazard rates. As time goes by the sample will contain a higher proportion of those with the lower hazard rate; as those with the higher hazard will have died. This will lead to a spurious estimate of negative duration dependence.

Kiefer (1988) shows how unobservable heterogeneity can be incorporated by adding a general heterogeneity effect μ and specifying

$$f(t) = \int f(t \mid \mu)p(\mu)\,d\mu. \tag{139}$$

The unknown distribution $p(\mu)$ can be modeled parametrically using mixture distributions. Alternatively a non-parametric approach can be adopted which gives μ a discrete distribution characterized by the mass-points

$$P(\mu = \mu_i) = p_i, \quad i = 1, \ldots, I, \tag{140}$$

where the parameters $(\mu_1, \ldots, \mu_n, p_1, \ldots, p_n)$ are estimated as part of the maximum likelihood estimation. This is the basis for the finite support density estimator of Heckman and Singer (1984).

Behrman et al. (1990, 1991) provide comprehensive treatments of unobservable heterogeneity in their studies of mortality risks; using parametric, semiparametric, and nonparametric estimators. They adopt two special cases of the Box–Cox conditional hazard used by Heckman and Singer (1984), and they consider two ways in which unobservable frailty (μ) can affect the hazard

$$h\big(t \mid x(t), \mu(t)\big) = \exp\big(x(t)\beta + \gamma(t^k - 1)/k + \mu(t)\big) \tag{141}$$

and

$$h\big(t \mid x(t), \mu(t)\big) = \mu(t) \cdot \exp\big(x(t)\beta + \gamma(t^k - 1)/k\big). \tag{142}$$

Their parametric approach uses a normal distribution for $f(\mu)$ in the additive specification and an inverse Gaussian distribution in the multiplicative specification. Both versions of the hazard function can be expressed in the form, $h(t) = h_o(t)\exp(x(t)\beta)$, and their semiparametric estimator uses the Cox partial likelihood approach to factor the baseline hazard out of the likelihood function. The nonparametric Heckman and Singer approach can be applied by using a finite support density estimator for $f(\mu)$.

In addition to these well known approaches, Behrman et al. (1991) apply a maximum penalized likelihood estimator (MPLE). The rationale for this approach is that it avoids over-parameterizing the heterogeneity, and it avoids the computational problems associated with the finite density estimator, particularly when there is a high degree of censoring and the distribution of heterogeneity has a long tail. In general, the penalized log-likelihood takes the form

$$\operatorname{Log} L_\alpha(f) = \sum_i \log\big(f(x_i)\big) - \alpha R(f). \tag{143}$$

The penalty term, $\alpha R(f)$, takes account of the "roughness" or local variability in the joint density of the data. The smoothing parameter α, which controls the balance between smoothness and goodness of fit, is typically chosen by cross-validation.

Behrman et al. (1990) evaluate the performance of their models using the maximized value of the likelihood function as a measure of goodness of fit and they test for unobserved heterogeneity using Lancaster's IM test, based on Cox–Snell residuals. They find evidence of heterogeneity but conclude that "modeling of unobserved heterogeneity directly in a proportional hazard setting may not be as important as allowing the covariates to affect the hazard in the highly nonlinear way that the gamma accelerated failure-time model allows". Behrman et al. (1991) find that the "introduction of nonparametric or parametric heterogeneity yields a small improvement in fit, similar parameter estimates, and changed significance levels".

8.3. Competing risks and multiple spells

So far the focus has been on duration models with a single destination, such as an individual's death. However, the techniques can be extended to allow for multiple destinations; or competing risks. For example, Butler, Anderson and Burkhauser (1989) use a competing risks specification with transitions out of retirement either back into employment or due to death. In their study of sickness absence among Dutch teachers, Lindeboom et al. (1995) use a three state Markov model that allows transitions from spells of work into sickness absence or exit from the job, and from spells of sickness back into work or exit from the job. Their model uses a partial likelihood approach to allow for school specific fixed effects.

Hamilton and Hamilton's (1997) study of the surgical volume-outcome relationships for patients undergoing surgery for hip fractures in Quebec between 1991–93 provides an example that combines competing risks, unobservable heterogeneity, and fixed effects. They use longitudinal data from the MED-ÉCHO database of hospital discharge abstracts. This allows them to attribute differences in the quality of providers to hospital specific fixed effects, modeled by dummy variables, and to analyze the within-hospital volume-outcome relationship; thereby discriminating between the "practice makes perfect effect" and "selective referral effect" (that hospitals with good outcomes will get more referrals).

Their competing risks specification allows for a correlation between the two outcomes; post-surgery length of stay and inpatient mortality. This is important as, ceteris paribus, a death in hospital is more likely for a patient with a longer length of stay. With two exhaustive and mutually exclusive destinations for discharges, alive (a) or dead (d), the probability of exit to state r, after a length of stay m, for patient i, in hospital h, at period t, is assumed to be

$$f_r(m_{iht} \mid x_{iht}) = \lambda_r(m_{iht} \mid x_{iht}) \prod_{j \in a,d} \exp\left[-\int_0^{miht} \lambda_j(u \mid x_{iht})\, du\right], \quad r = a, d. \tag{144}$$

The first term on the right hand side, $\lambda_r(m_{iht} \mid x_{iht})$, is the transition intensity, the equivalent of the hazard rate in single destination models. The second term is the survivor function, giving the probability of surviving to m without death or discharge. A proportional hazards specification is used

$$\lambda_r(\cdot) = \exp(x_{iht}\beta_r + \theta_{hr} + \pi_r\mu)\lambda_{or}(m_{iht}), \quad r = a, d \text{ and } \pi_a = 1, \tag{145}$$

where θ_{hr} is the hospital fixed effect, and a log-logistic baseline hazard is used

$$\lambda_{or}(m) = (\rho_r\alpha_r m^{\alpha r-1})/(1 + \rho_r m^{\alpha r}), \quad \alpha_r > 0,\ \rho_r > 0. \tag{146}$$

Unobserved frailty is modeled as the scalar random variable μ, and its distribution is estimated using the Heckman–Singer approach. The likelihood takes the form

$$L = \prod_i \sum_k p_k \cdot f_a(m_{iht} \mid x_{iht}, \theta_h, \mu_k)^{\delta ia} \cdot f_d(m_{iht} \mid x_{iht}, \theta_h, \mu_k)^{\delta id}, \qquad \sum_k p_k = 1, \tag{147}$$

where the points of support (μ_k) and associated probabilities (p_k) are estimated along with the other parameters. Hamilton and Hamilton (1997) use three mass points, which they interpret as a distribution made up of three types of patients.

The results of the study show that when hospital fixed effects are added to the model the coefficient on volume, measured by the logarithm of live discharges, declines substantially and is insignificant. Volume does not have a significant effect on inpatient deaths with or without hospital fixed effects, although cruder models without unobservable heterogeneity and with fewer controls for comorbidities do show a significant effect.

9. Stochastic frontiers

9.1. Cost function studies

A recent systematic review by Aletras (1996) identifies approximately 100 studies which provide evidence on the existence of economies of scale and scope in hospitals. Many of these are econometric studies which use regression analysis to explore the average cost of hospital treatment. Other methods include data envelopment analysis (DEA), market survival methods, and before-and-after studies. These attempts to estimate empirical production functions and cost functions for hospitals and other health care organizations face some common methodological problems.

It would be desirable to define a hospital's output in terms of health outcomes, measured as health gains, but typically these kinds of data are not available and measures of throughput have to be used (e.g., admissions, discharges, number of procedures performed). Output is multi-dimensional, and it is important to control for case-mix, by including variables for the proportion of patients in each specialty, the number of discharges, or the average length of stay by specialty or case-mix grouping. However these case-mix adjustments may miss intra-category variations in severity, and inter-hospital variation in case-mix. Measures of the quantity of output may neglect differences in quality across hospitals, which may bias estimates of economies of scale. Similar arguments apply to the neglect of differences in the quality of inputs. Also, in econometric studies, the level of output is usually assumed to be exogenous, reflecting the demand for health care from patients or purchasers. The possibility of an incomplete agency relationship between purchasers and providers may lead to simultaneous equations bias.

9.2. *Frontier models*

9.2.1. *Cross section estimators*

Rather than discussing hospital cost studies in general, this section concentrates on the econometric techniques that have been used to analyze the efficiency of health care organizations, and in particular the use of stochastic frontier models. This builds on earlier surveys by Wagstaff (1989a, 1989b) and Aletras (1996). The section does not cover data envelopment analysis (DEA) which is a nonparametric approach that uses linear programming methods to identify efficiency scores. Typically, applications of DEA in health economics do not allow for a random error term and are likely to be sensitive to the influence of outliers.

Feldstein's (1967) pioneering econometric analysis of hospitals costs in the British NHS uses the following empirical specification

$$y_i = \beta_o + \sum_j \beta_j x_{ij} + u_i, \tag{148}$$

where y_i is the average cost per case and x_{ij} is the proportion of hospital i's patients in the j-th case-mix category. In this model the residuals are distributed symmetrically around the cost function and it cannot be interpreted as a frontier. This is relaxed by deterministic cost frontier (DCF) models, which assume $u_i \geqslant 0$ for all i. In this case the error term moves hospitals above the (deterministic) cost frontier. One estimator for this model is corrected OLS, which simply adjusts the OLS estimates of the intercept β_o and the residuals by adding $\min(u_i)$ to the intercept and subtracting it from the residuals. The drawbacks of this method are that it treats the most efficient hospital as 100 per cent efficient, and that the whole of the error term is assumed to reflect inefficiency. This ignores random "noise" due to measurement errors and unobservable heterogeneity.

To relax these assumptions stochastic cost frontiers (SCF) are based on the two-error model

$$y_i = \beta_o + \sum_j \beta_j x_{ij} + u_i + \varepsilon_i, \quad u_i \geqslant 0, \tag{149}$$

where it is assumed that u_i measures inefficiency and ε_i is a random error term. The identifying assumption in this model is that there is zero skewness in the distribution of the random error term; this allows statistical evidence of skewness in the residuals to be given an economic interpretation as inefficiency. To estimate parametric versions of this model by maximum likelihood it is necessary to make assumptions about the distributions of u and ε. For example, Aigner et al. (1977) assume that ε is normal and u is half-normal. Other common assumptions are that u is truncated normal, exponential, or gamma distributed.

Vitaliano and Toren (1994) apply stochastic frontiers to estimate cost inefficiency in New York nursing homes, using the 1987 and 1990 waves of a panel dataset. After experimenting with truncated normal and exponential distributions, they choose to estimate the model using a half normal inefficiency term. They use Jondrow et al.'s method to decompose the estimated error term; this computes an estimate of inefficiency conditional on the estimated residual, $\mathrm{E}(u_i \mid u_i + \varepsilon_i)$. Their results suggest a mean inefficiency of 29 per cent.

Stochastic frontiers are applied to a multiproduct hospital cost function by Zuckerman et al. (1994). They use data on 1,600 US hospitals from the AHA Annual Survey, Medicare hospital cost reports, and MEDPAR data system to estimate translog cost functions that include measures of illness severity, output quality, and patient outcomes. The SCF models are estimated by ML using a half-normal distribution for inefficiency, these suggest a mean inefficiency of 13.6 per cent. The authors are concerned about possible endogeneity of the output measures, and find that Hausman–Wu tests reject exogeneity in non-frontier specifications. However, they are not able to find estimates that converge when instrumental variables are used in the frontier models.

The use of stochastic frontiers is not confined to estimates of hospital cost functions. Gaynor and Pauly (1990) use production frontiers to investigate the effects of different compensation arrangements on productive efficiency in medical group practices. They compare "traditional" production functions, which only include inputs, with "behavioral" functions, which include variables that reflect incentives. Data on 6,353 physicians within 957 group practices, from a survey carried out by Mathematica Policy Research in 1978, are used to estimate stochastic frontiers using normal and truncated normal error components. The potential endogeneity of variables that measure the firm's compensation structure is dealt with using instrumental variables. The results suggest that incentives do influence productivity, with larger groups reducing productivity and greater average experience within a group increasing productivity.

Most cross-section frontier models are estimated by maximum likelihood, imposing specific parametric distributions on both u and ε. Kopp and Mullahy (1990, 1993) propose semiparametric estimators which relax the distributional assumptions about ε, simply requiring that it is symmetrically distributed. Given the symmetry assumption, they are able to derive restrictions for the higher order moments of the composite error term. In Kopp and Mullahy (1990) these moment conditions are used to motivate a GMM estimator, and in Kopp and Mullahy (1993) they are used to motivate a COLS or corrected moment (CM) estimator. These estimators do not seem to have been applied to health data as yet.

9.2.2. Panel data estimators

The fact that cross-section frontier models rely on skewness to identify inefficiency is often criticized [see, e.g., Dor (1994), Newhouse (1994), Skinner (1994), Wagstaff (1989b)]. The danger is that skewness in the distribution of the random error term could be mistakenly attributed to inefficiency. Section 9.1 has described the problems of controlling for multi-dimensional output, case-mix, quality of output, and quality of inputs. All of these are potential omitted variables which may lead to skewness in the residuals and be labeled as inefficiency. An alternative is to use panel data estimators. On the assumption that inefficiency remains constant over time, the stochastic frontier model takes on a form similar to the standard panel data regression (see Equation (94))

$$y_{it} = \beta_o + \sum_j \beta_j x_{ijt} + u_i + \varepsilon_{it}, \quad u_i \geqslant 0. \tag{150}$$

This model can be estimated using fixed or random effects estimators, and the results are subject to the strengths and weaknesses of these estimators, as discussed in Section 6. In particular, the fixed effects models raises the problem of separately identifying inefficiency and the effects of time invariant regressors, while the random effects specification is biased if the inefficiency is correlated with the regressors. Park et al. (1998) show that the within-groups (CV) estimator is the efficient semiparametric estimator for (150), when no particular structure is placed on the dependence between the regressors (x) and random effect (u_i). Also they derive efficient semiparametric estimators for the cases in which the dependence is restricted to a sub-set of the regressors, and when the dependence is through the mean of the sub-set of regressors.

Wagstaff (1989b) uses data on 49 Spanish public hospitals to compare cross section and panel data estimators. Cross section estimates based on the half-normal model suggest that mean cost inefficiency is only 10 per cent, and it is not possible to reject the null hypothesis that there is no skewness. However estimates of the fixed effects specification suggest that around one third of the variation in costs can be attributed to inefficiency. Also the stochastic frontier leads to quite different efficiency rankings than the fixed effects and deterministic cost frontier models. This ambiguity leads Wagstaff to recommend that a range of methods are compared to assess the sensitivity of the efficiency estimates to model specification.

Koop et al. (1997) develop a Bayesian fixed effects estimator, using the prior that the inefficiency effects will be one-sided and independent. They also develop a random effects estimator that allows the inefficiency to depend on time invariant hospital characteristics. These estimators are applied to a panel of 382 US non-teaching hospitals for 1987–91. Estimates of a translog cost function show that for-profit hospitals are less efficient, although these results are based on highly aggregated measures of output and may neglect differences in quality.

The assumption that inefficiency remains constant over time can be relaxed. For example, Battese and Coelli (1992) propose a panel data estimator model in which firm specific inefficiency takes the form

$$u_{it} = \exp\{-\eta(t - T)\}u_i \geqslant 0. \tag{151}$$

This allows inefficiency to change over time, but on the assumption that the rate of change, η, is common to all firms. The model is estimated by ML, on the assumption that that ε is normal and u is truncated or half-normal. Battese and Coelli (1995) propose an alternative specification in which

$$u_{it} = z_{it}\delta + \omega_{it} \geqslant 0. \tag{152}$$

The z_{it} variables are determinants of cost inefficiency and the distribution of u_{it} is assumed to be truncated normal. Linna (1998) applies both of these models, along with nonparametric data envelopment analysis (DEA), to Finnish panel data covering 43 acute hospitals for 1988–94. He finds that the nonparametric and parametric methods compare well with respect to individual efficiency scores, and measures of time-varying inefficiency and technological change. The correlation between the different measures is greater than in previous studies based on cross sectional data.

10. Conclusion

In documenting the influence of econometrics on the development of health economics, Newhouse (1987) grouped imports from econometrics under four headings: specification tests, robust estimators, replication, and experimentation. Ten years on, the first two of these remain dominant themes in applied work. Examples of good practice in health econometrics make extensive use of tests for misspecification and explicit model selection criteria. Robust and distribution-free estimators are of increasing importance, and this chapter has given examples of nonparametric, and semiparametric estimators applied to sample selection, simultaneous equations, count data, and survival models. As the use of these techniques widens, it will be interesting to see whether they have an impact on the economic and policy relevance of the results produced. Even if the impact proves to be small, researchers will be able to place more confidence in earlier results that were generated by less robust methods.

Published replications of empirical results remain relatively rare, perhaps reflecting the incentives surrounding academic publication in economics. One way in which this deficit may be remedied is through the appearance of more systematic reviews of econometric studies, such as the work of Aletras (1996). This chapter has shown that certain datasets are widely used, allowing results to be compared across studies, and many of the studies reviewed here are careful to compare new techniques with established methods. The use of experimental data remains an exception and most applied studies continue to rely on observational data from secondary sources. However applied work in health economics is likely to be influenced by the debate concerning the use of instrumental variables to analyze social experiments [see, e.g., Angrist et al. (1996), Heckman (1997)].

This chapter has illustrated the impressive diversity of applied econometric work over the past decade. It has emphasized the range of models and estimators that have been applied, but that should not imply a neglect of the need for sound economic theory and careful data collection and analysis in producing worthwhile econometric research. Most of the studies reviewed here use individual level data and this has led to the use of a wide range of nonlinear models, including qualitative and limited dependent variables, along with count, survival and frontier models. Because of the widespread use of observational data, particular attention has gone into dealing with problems of self-selection and heterogeneity bias. This is likely to continue in the future, with the emphasis on robust estimators applied to longitudinal and other complex datasets.

References

Ahn, H., and J.L. Powell (1993), "Semiparametric estimation of censored selection models with a nonparametric selection mechanism", Journal of Econometrics, Annals 58:3–30.

Aigner, D.J., C.A.K. Lovell and P. Schmidt (1977), "Formulation and estimation of stochastic production function models", Journal of Econometrics 6:21–37.

Alderson, C. (1997), "Exporing a modified 'fair innings' approach to addressing social class inequalities in lifetime health", unpublished M.Sc. dissertation (University of York).

Aletras, V. (1996), "Concentration and choice in the provision of hospital services", Technical Appendix 2, NHS Centre for Reviews and Dissemination (University of York).

Angrist, J.D. (1995), "Conditioning on the probability of selection to control selection bias", NBER Technical Working Paper #181.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of causal effects using instrumental variables", Journal of the American Statistical Association 91:444–455.

Arinen, S.-S., H. Sintonen and G. Rosenqvist (1996), "Dental utilisation by young adults before and after subsidisation reform", Discussion Paper 149 (Centre for Health Economics University of York).

Atkinson, A.B., J. Gomulka and N.H. Stern (1984), "Household expenditure on tobacco 1970–1980: evidence from the family expenditure survey", Discussion Paper no. 60 (London School of Economics).

Auster, R., I. Leveson and D. Sarachek (1969), "The production of health an exploratory study", Journal of Human Resources 15:411–436.

Battese, G.E., and T.J. Coelli (1992), "Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India", Journal of Productivity Analysis 3:153–169.

Battese, G.E., and T.J. Coelli (1995), "A model for technical inefficiency effects in a stochastic frontier production function for panel data", Empirical Economics 20:325–332.

Becker, G.S., and K.M. Murphy (1988), "A theory of rational addiction", Journal of Political Economy 96:675–700.

Behrman, J.R., R.C. Sickles and P. Taubman (1990), "Age-specific death rates with tobacco smoking and occupational activity: sensitivity to sample length, functional form, and unobserved frailty", Demography 27:267–284.

Behrman, J.R., R. Sickles, P. Taubman and A. Yazbeck (1991), “Black-white mortality inequalities”, Journal of Econometrics 50:183–203.

Behrman, J.R., and B.L. Wolfe (1987), “How does mother’s schooling affect family health, nutrition, medical care usage, and household sanitation?”, Journal of Econometrics 36:185–204.

Bhattacharya, J., W.B. Vogt, A. Yoshikawa and T. Nakahara (1996), “The utilization of outpatient medical services in Japan”, Journal of Human Resources 31:450–476.

Bishai, D.M. (1996), “Quality time: how parents’ schooling affects child health through its interaction with childcare time in Bangladesh”, Health Economics 5:383–407.

Björklund, A. (1985), “Unemployment and mental health: some evidence from panel data”, Journal of Human Resources 20:469–483.

Blaylock, J.R., and W.N. Blisard (1992), “Self-evaluated health status and smoking behaviour”, Applied Economics 24:429–435.

Blaylock, J.R., and W.N. Blisard (1993), “Wine consumption by US men”, Applied Economics 25:645–651.

Blundell, R.W., and C. Meghir (1987), “Bivariate alternatives to the Tobit Model”, Journal of Econometrics 34:179–200.

Blundell, R.W., and R.J. Smith (1993), “Simultaneous microeconometric models with censored or qualitative dependent variables”, in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier, Amsterdam) 117–143.

Blundell, R.W., and F.A.G. Windmeijer (1997), “Correlated cluster effects and simultaneity in multilevel models”, Health Economics 6:439–443.

Bolduc, D., G. Lacroix and C. Muller (1996), “The choice of medical providers in rural Bénin: a comparison of discrete choice models”, Journal of Health Economics 15:477–498.

Bollen, K.A., D.K. Guilkey and T.A. Mroz (1995), “Binary outcomes and endogenous explanatory variables: tests and solutions with an application to the demand for contraceptive use in Tunisia”, Demography 32:111–131.

Börsch-Supan, A., V. Hajivassiliou, L.J. Kotlikoff and J.N. Morris (1992), “Health, children, and elderly living arrangements: a multiperiod-multinomial probit model with unobserved heterogeneity and autocorrelated errors”, in: D.A. Wise, ed., Topics in the Economics of Aging (University of Chicago Press, Chicago).

Bound, J., D. Jaeger and R. Baker (1995), “Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak”, Journal of the American Statistical Association 90:443–450.

Buchinsky, M. (1998), “Recent advances in quantile regression models”, Journal of Human Resources 33:88–126.

Buchmueller, T.C., and P.J. Feldstein (1997), “The effect of price on switching among health plans”, Journal of Health Economics 16:129–260.

Buchmueller, T.C., and S.H. Zurekas (1998), “Drug use, drug abuse and labour market outcomes”, Health Economics 7:229–245.

Burgess, S.M., and C. Propper (1998), “Early health related behaviours and their impact on later life chances: evidence from the US”, Health Economics 7:381–399.

Butler, J.S., K.H. Anderson and R.V. Burkhauser (1989), “Work and health after retirement: a competing risks model with semiparametric unobserved heterogeneity”, The Review of Economics and Statistics 71:46–53.

Cairns, J.A., and M. van der Pol (1997), “Saving future lives: a comparison of three discounting models”, Health Economics 6:341–350.

Cameron, A.C., and P. Johansson (1997), “Count data regression using series expansions: with applications”, Journal of Applied Econometrics 12:203–223.

Cameron, A.C., and F.A.G. Windmeijer (1996), “R-squared measures for count data regression models with applications to health care utilization”, Journal of Business and Economic Statistics 14:209–220.

Cameron, A.C., and P.K. Trivedi (1986), “Econometric models based on count data: comparisons and applications of some estimators and tests”, Journal of Applied Econometrics 1:29–53.

Cameron, A.C., P.K. Trivedi, F. Milne and J. Piggott (1988), “A microeconometric model of demand for health care and health insurance in Australia”, Review of Economic Studies 55:85–106.

Cameron, A.C., and P.K. Trivedi (1993), "Tests of independence in parametric models with applications and illustrations", Journal of Business and Economic Statistics 11:29–43.

Cauley, S.D. (1987), "The time price of medical care", Review of Economics and Statistics 69:59–66.

Chaloupka, F.J., and H. Wechsler (1997), "Price, tobacco control policies and smoking among young adults", Journal of Health Economics 16:359–373.

Chamberlain, G. (1980), "Analysis of covariance with qualitative data", Review of Economic Studies 47:225–238.

Chamberlain, G. (1984), "Panel data", in: Z. Griliches and M. Intrilligator, eds., Handbook of Econometrics (North-Holland, Amsterdam) 1247–1318.

Coulson, N.E., J.V. Terza, C.A. Neslusan and B.C. Stuart (1995), "Estimating the moral-hazard effect of supplemental medical insurance in the demand for prescription drugs by the elderly", AEA Papers and Proceedings 85:122–126.

Cox, D. (1972), "Regression models with life tables", Journal of the Royal Statistical Society 74:187–220.

Deb, P., and P.K. Trivedi (1997), "Demand for medical care by the elderly: a finite mixture approach", Journal of Applied Econometrics 12:313–336.

Donaldson, C., A.M. Jones, T.J. Mapp and J.A. Olsen (1998), "Limited dependent variables in willingness to pay studies: applications in health care", Applied Economics 30:667–677.

Dor, A. (1994), "Non-minimum cost functions and the stochastic frontier: On applications to health care providers", Journal of Health Economics 13:329–334.

Dor, A., P. Gertler and J. van der Gaag (1987), "Non-price rationing and the choice of medical care providers in rural Cote D'Ivoire", Journal of Health Economics 6:291–304.

Douglas, S. (1998), "The duration of the smoking habit", Economic Inquiry 36:49–64.

Douglas, S., and G. Hariharan (1994), "The hazard of starting smoking: estimates from a split population duration model", Journal of Health Economics 13:213–230.

Dowd, B., R. Feldman, S. Cassou and M. Finch (1991), "Health plan choice and the utilization of health care services", Review of Economics and Statistics 73:85–93.

Dranove, D. (1998), "Economies of scale in non-revenue producing cost centers: implications for hospital mergers", Journal of Health Economics 17:69–83.

Duan, N. (1983), "Smearing estimate: a nonparametric retransformation method", Journal of the American Statistical Association 78:605–610.

Duan, N., W.G. Manning, C.N. Morris and J.P. Newhouse (1983), "A comparison of alternative models for the demand for medical care", Journal of Business and Economic Statistics 1:115–126.

Duan, N., W.G. Manning, C.N. Morris and J.P. Newhouse (1984), "Choosing between the sample-selection and multi-part model", Journal of Business and Economic Statistics 2:283–289.

Duan, N., W.G. Manning, C.N. Morris and J.P. Newhouse (1985), "Comments on selectivity bias", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 19–24.

Duncan, A.S., and A.M. Jones (1992), "NP-REG: an interactive package for kernel density estimation and nonparametric regression", Working Paper W92/7 (Institute for Fiscal Studies).

Dustmann, C., and F.A.G. Windmeijer (1996), "Health, wealth and individual effects – a panel data analysis", presented at Fifth European Workshop on Econometrics and Health Economics, Barcelona.

Ellis, R.P., D.K. McInnes and E.H. Stephenson (1994), "Inpatient and outpatient health care demand in Cairo, Egypt", Health Economics 3:183–200.

Erbsland, M., W. Ried and V. Ulrich (1995), "Health, health care, and the environment. Econometric evidence from German micro data", Health Economics 4:169–182.

Escarcé, J.J. (1996), "Externalities in hospitals and physicina adoption of a new surgical technology: an exploratory analysis", Journal of Health Economics 15:715–734.

Feldman, R., M. Finch, B. Dowd and S. Cassou (1989), "The demand for employment-based health insurance plans", Journal of Human Resources 24:115–142.

Feldstein, M.S. (1967), Economic Analysis for Health Service Efficiency: Econometric Studies of the British National Health Service (North-Holland, Amsterdam).

Forster, M., and A.M. Jones (1997a), "Inequalities in optimal life-span: a theoretical and empirical investigation", mimeo (University of York).

Forster, M., and A.M. Jones (1997b), "The optimal time path of consumption of an unhealthy good: a theoretical and empirical investigation of smoking durations", mimeo (University of York).

van der Gaag, J., and B.L. Wolfe (1991), "Estimating demand for medical care: health as a critical factor for adults and children", in: G. Duru and J.H.P. Paelinck, eds., Econometrics of Health Care (Kluwer, Amsterdam) 31–58.

Garcia, J., and J.M. Labeaga (1996), "A cross-section model with zeros: an application to the demand for tobacco", Oxford Bulletin of Economics and Statistics 58:489–506.

Gatsonis, C.A., A.M. Epstein, J.P. Newhouse, S.-L. Normand and B.J. McNeil (1995), "Variations in the utilization of coronary angiography for elderly patients with an acute myocardial infraction: an analysis using hierarchical logistic regression", Medical Care 33:625–642.

Gaynor, M. (1989), "Competition within the firm: theory plus some evidence from medical group practice", RAND Journal of Economics 20:59–76.

Gaynor, M., and M.V. Pauly (1990), "Compensation and productive efficiency in partnerships: evidence from medical group practice", Journal of Political Economy 98:544–573.

Geil, P., A. Million, R. Rotte and K.F. Zimmermann (1997), "Economic incentives and hospitalization in Germany", Journal of Applied Econometrics 12:295–311.

Gerdtham, U.-G. (1997), "Equity in health care utilization: further tests based on hurdle models and Swedish micro data", Health Economics 6:303–319.

Gerdtham, U.-G., and B. Jonsson (2000), "International comparisons of healthcare expenditure: theory, data and regression analysis", in: J.P. Newhouse and A.J. Culyer, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 1.

Gertler, P., L. Locay and W. Sanderson (1987), "Are user fees regressive? The welfare implications of health care financing proposals in Peru", Journal of Econometrics 36:67–88.

Gourieroux, C.A., A. Monfort and A. Trognon (1984), "Pseudo maximum likelihood methods: applications to Poisson models", Econometrica 52:701–720.

Gourieroux, C.A., and A. Monfort (1993), "Pseudo-likelihood methods", in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier, Amsterdam) 335–362.

Grootendorst, P.V. (1995), "A comparison of alternative models of prescription drug utilization", Health Economics 4:183–198.

Grootendorst, P.V. (1997), "Health care policy evaluation using longitudinal insurance claims data: an application of the panel Tobit estimator", Health Economics 6:365–382.

Guilkey, D.K., T.A. Mroz and L. Taylor (1992), "Estimation and testing in simultaneous equations models with discrete outcomes using cross section data", unpublished manuscript.

Gurmu, S. (1997), "Semi-parametric estimation of hurdle regression models with an application to medicaid utilization", Journal of Applied Econometrics 12:225–242.

Haas-Wilson, D., A. Cheadle and R. Scheffler (1988), "Demand for mental health services: an episode of treatment approach", Southern Economic Journal 55:219–232.

Haas-Wilson, D., and E. Savoca (1990), "Quality and provider choice: a multinomial logit-least squares model with selectivity", Health Services Research 2:791–809.

Hajivassiliou, V.A. (1993), "Simulation estimation methods for limited dependent variable models", in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier, Amsterdam) 519–543.

Hakkinen, U. (1991), "The production of health and the demand for health care in Finland", Social Science and Medicine 33:225–237.

Hakkinen, U., G. Rosenquist and S. Aro (1996), "Economic depression and the use of physician services in Finland", Health Economics 5:421–434.

Hall, A. (1993), "Some aspects of generalized method of moments estimation", in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier, Amsterdam) 393–417.

Hamilton, B. (1999), "The impact of HMOs on Medicare costs: Bayesian MCMC estimation of a robust panel data Tobit model with survival", Health Economics 8:403–414.

Hamilton, B.H., and V.H. Hamilton (1997), "Estimating surgical volume-outcome relationships applying survival models: accounting for frailty and hospital fixed effects", Health Economics 6:383–395.

Hamilton, B.H., V.H. Hamilton and N.E. Mayo (1996), "What are the costs of queueing for hip fracture surgery in Canada?", Journal of Health Economics 15:161–185.

Hamilton, V.H., P. Merrigan and E. Dufresne (1997), "Down and out: estimating the relationship between mental health and unemployment", Health Economics 6:397–406.

Hay, J.W. (1991), "Physicians' specialty choice and specialty income", in: G. Duru and J.H.P. Paelinck, eds., Econometrics of Health Care (Kluwer, Amsterdam) 95–113.

Hay, J., and R.J. Olsen (1984), "Let them eat cake: a note on comparing alternative models of the demand for health care", Journal of Business and Economic Statistics 2:279–282.

Heckman, J.J. (1979), "Sample selection bias as a specification error", Econometrica 47:153–161.

Heckman, J.J. (1996), "Randomization as an instrumental variable", Review of Economics and Statistics 78:336–341.

Heckman, J.J. (1997), "Instrumental variables. A study of implicit behavioral assumptions used in making program evaluations", Journal of Human Resources 32:441–461.

Heckman, J.J., and B. Singer (1984), "A method of minimizing the distributional impact in econometric models for duration data", Econometrica 52:271–230.

Hoerger, T.J., G.A. Picone and F.A. Sloan (1996), "Public subsidies, private provision of care and living arrangements of the elderly", Review of Economics and Statistics 78:428–439.

Honoré, B.E. (1992), "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects", Econometrica 60:533–565.

Hunt-McCool, J., B.F. Kiker and Y.C. Ng (1994), "Estimates of the demand for medical care under different functional forms", Journal of Applied Econometrics 9:201–218.

Ichimura, H., and L.F. Lee (1991), "Semiparametric estimation of multiple index models: single equation estimation", in: W.A. Barnett, J. Powell and G. Tauchen, eds., Nonparametric and Semiparametric Methods in Econometrics and Statistics (Cambridge University Press, New York).

Imbens, G.W., and J.D. Angrist (1994), "Identification of local average treatment effects", Econometrica 62:467–475.

Jeong, J., and G.S. Maddala (1993), "A perspective on application of bootstrap methods in econometrics", in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier, Amsterdam) 519–543.

Jones, A.M. (1989), "A double-hurdle model of cigarette consumption", Journal of Applied Econometrics 4:23–39.

Jones, A.M. (1993), "Starters, quitters and smokers: choice or addiction", prepared for the Inaugural Labelle Lectureship, CHEPA, McMaster University.

Keeler, E.B., W.G. Manning and R.B. Wells (1988), "The demand for episodes of mental health services", Journal of Health Economics 7:369–392.

Kenkel, D.S. (1990), "Consumer health information and the demand for medical care", The Review of Economics and Statistics 72:587–595.

Kenkel, D.S. (1991), "Health behaviour, health knowledge and schooling", Journal of Political Economy 99:287–305.

Kenkel, D.S. (1995), "Should you eat breakfast? Estimates from health production functions", Health Economics 4:15–29.

Kenkel, D.S., and J.V. Terza (1993), "A partial observability probit model of medical demand", mimeo (Pennsylvania State University).

Kerkhofs, M., and M. Lindeboom (1995), "Subjective health measures and state dependent reporting errors", Health Economics 4:221–235.

Kerkhofs, M., and M. Lindeboom (1997), "Age related health dynamics and changes in labour market status", Health Economics 6:407–423.

Kiefer, N. (1988), “Economic duration data and hazard functions”, Journal of Economic Literature 26:646–679.

Koop, G., J. Osiewalski and M.F.J. Steel (1997), “Bayesian efficiency analysis through individual effects: hospital cost frontiers”, Journal of Econometrics 76:77–105.

Kopp, R.J., and J. Mullahy (1990), “Moment-based estimation and testing of stochastic frontier models”, Journal of Econometrics 46:165–183.

Kopp, R.J., and J. Mullahy (1993), “Least squares estimation of econometric frontier models: consistent estimation and inference”, Scandinavian Journal of Economics 95:125–132.

Labeaga, J.M. (1993), “Individual behaviour and tobacco consumption: a panel data approach”, Health Economics 2:103–112.

Labeaga, J.M. (1999), “A dynamic panel data model with limited dependent variables: an application to the demand for tobacco”, Journal of Econometrics 93:49–72.

Lee, L.-F. (1983), “Generalized econometric models with selectivity”, Econometrica 51:507–512.

Lee, L.-F., M.R. Rosenzweig and M.M. Pitt (1997), “The effects of improved nutrition, sanitation, and water quality on child health in high mortality populations”, Journal of Econometrics 77:209–235.

Leibowitz, A., W.G. Manning and J.P. Newhouse (1985), “The demand for prescription drugs as a function of cost-sharing”, Social Science and Medicine 21:1063–1069.

Leung, S.F., and S. Yu (1996), “On the choice between sample selection and two-part models”, Journal of Econometrics 72:197–229.

Levinson, A., and F. Ullman (1998), “Medicaid managed care and infant health”, Journal of Health Economics 17:351–368.

Lewit, E.M., D. Coate and M. Grossman (1981), “The effects of government regulation on teenage smoking”, Journal of Law and Economics 24:545–570.

Lindeboom, M., M. Kerkhofs and L. Aarts (1995), “Sickness absenteeism of primary school teachers in the Netherlands”, mimeo (Leiden University).

Linna, M. (1998), “Measuring the hospital cost efficiency with panel data models”, Health Economics 7:415–427.

López, A. (1998), “Unobserved heterogeneity and censoring in the demand for health care”, Health Economics 7:429–437.

McClellan, M., and J.P. Newhouse (1997), “The marginal cost-effectiveness of medical technology: a panel instrumental variables approach”, Journal of Econometrics 77:39–64.

Maddala, G.S. (1983), Limited-Dependent and Qualitative Variables in Econometrics (Cambridge University Press, Cambridge).

Maddala, G.S. (1985), “A survey of the literature on selectivity bias as it pertains to health care markets”, in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 3–17.

Manning, W.G. (1998), “The logged dependent variable, heteroscedasticity, and the retransformation problem”, Journal of Health Economics 17:283–295.

Manning, W.G., L. Blumberg and L.H. Moulton (1995), “The demand for alcohol: the differential response to price”, Journal of Health Economics 14:123–148.

Manning, W.G., N. Duan and W.H. Rogers (1987), “Monte Carlo evidence on the choice between sample selection and two-part models”, Journal of Econometrics 35:59–82.

Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler, A. Leibowitz and M.S. Marquis (1987), “Health insurance and the demand for medical care: evidence from a randomized experiment”, American Economic Review 77:251–277.

Manski, C.F. (1993), “The selection problem in econometrics and statistics”, in: G.S. Maddala, C.R. Rao and H.D. Vinod, eds., Handbook of Statistics, Vol. 11 (Elsevier, Amsterdam) 73–84.

Morris, C.N., E.C. Norton and X.H. Zhou (1994), “Parametric duration analysis of nursing home usage”, in: N. Lange et al., eds., Case Studies in Biometry (John Wiley & Sons, New York) 231–248.

Mullahy, J. (1986), “Specification and testing of some modified count data models”, Journal of Econometrics 33:341–365.

Mullahy, J. (1997a), "Instrumental variable estimation of count data models. Applications to models of cigarette smoking behaviour", Review of Economics and Statistics 79:586–593.

Mullahy, J. (1997b), "Heterogeneity, excess zeros, and the structure of count data models", Journal of Applied Econometrics 12:337–350.

Mullahy, J. (1998), "Much ado about two: reconsidering retransformation and the two-part model in health econometrics", Journal of Health Economics 17:247–281.

Mullahy, J., and W. Manning (1996), "Statistical issues in cost-effectiveness analysis", in: F.A. Sloan, ed., Valuing Health Care (Cambridge University Press, Cambridge) 149–184.

Mullahy, J., and P.R. Portney (1990), "Air pollution, cigarette smoking, and the production of respiratory health", Journal of Health Economics 9:193–205.

Mullahy, J., and J. Sindelar (1996), "Employment, unemployment, and problem drinking", Journal of Health Economics 15:409–434.

Mundlak, Y. (1978), "On the pooling of time series and cross section data", Econometrica 46:69–85.

Mwabu, G., M. Ainsworth and A. Nyamete (1993), "Quality of medical care and choice of medical treatment in Kenya. An empirical analysis", Journal of Human Resources 28:838–862.

Nanda, P. (1999), "The impact of women's participation in credit programs on the demand for quality health care in rural Bangladesh", Health Economics 8:415–428.

Newhouse, J.P. (1987), "Health economics and econometrics", American Economic Review 77:269–274.

Newhouse, J.P. (1994), "Frontier estimation: how useful a tool for health economics", Journal of Health Economics 13:317–322.

Newhouse, J.P., C.E. Phelps and M.S.M. Marquis (1980), "On having your cake and eating it too, Econometric problems in estimating the demand for health services", Journal of Econometrics 13:365–390.

Norton, E.C. (1995), "Elderly assets, Medicaid policy, and spend-down in nursing homes", Review of Income and Wealth 41:309–329.

Norton, E.C., G.S. Bieler, S.T. Ennett and G.A. Zarkin (1996), "Analysis of prevention program effectiveness with clustered data using generalized estimating equations", Journal of Consulting and Clinical Psychology 64:919–926.

O'Donnell, O. (1993), "Income transfers and the labour market participation of disabled individuals in the UK", Health Economics 2:139–148.

Park, B.U., R.C. Schmidt and L. Simar (1998), "Stochastic panel frontiers: a semiparametric approach", Journal of Econometrics 84:273–301.

Philipson, T. (1996), "Private vaccination and public health: an empirical examination for US measles", Journal of Human Resources 31:611–630.

Pitt, M.M. (1997), "Estimating the determinants of child health when fertility and mortality are selective", Journal of Human Resources 32:129–158.

Pohlmeier, W., and V. Ulrich (1995), "An econometric model of the two-part decisionmaking process in the demand for health care", Journal of Human Resources 30:339–360.

Primoff Vistnes, J., and V. Hamilton (1995), "The time and monetary costs of outpatient care for children", American Economic Review Papers and Proceedings 85:117–121.

Pudney, S., and M. Shields (1997), "Gender and racial inequality in pay and promotion in the internal labour market for NHS nurses", Discussion Paper in Public Sector Economics no. 97/4 (University of Leicester).

Rice, N., and A.M. Jones (1997), "Multilevel models and health economics", Health Economics 6:561–575.

Rivers, D., and Q.H. Vuong (1988), "Limited information estimators and exogeneity tests for simultaneous probit models", Journal of Econometrics 39:347–366.

Robinson, P. (1988), "Root-N-consistent semiparametric regression", Econometrica 56:931–954.

Rosenbaum, P.R., and D.B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", Biometrika 70:41–55.

Rosenzweig, M.R., and T.P. Schultz (1983), "Estimating a household production function: heterogeneity, the demand for health inputs, and their effects on birth weight", Journal of Political Economy 91:723–746.

Rosenzweig, M.R., and K.T. Wolpin (1995), "Sisters, siblings, and mothers: the effect of teenage childbearing and birth outcomes in a dynamic family context", Econometrica 63:303–326.

Santos Silva, J.M.C., and F.A.G. Windmeijer (1997), "Stopped-sum models for health care demand", presented at Sixth European Workshop on Econometrics and Health Economics, Lisbon.

Scott, A., and A. Shiell (1997), "Do fee descriptors influence treatment choices in general practice? A multilevel discrete choice model", Journal of Health Economics 16:323–342.

Siddiqui, S. (1997), "The impact of health on retirement behaviour: empirical evidence from West Germany", Health Economics 6:425–438.

Skinner, J. (1994), "What do stochastic cost frontiers tell us about inefficiency?", Journal of Health Economics 13:323–328.

Smith, R.J., and R.W. Blundell (1986), "An exogeneity test for a simultaneous equation Tobit model with an application to labor supply", Econometrica 54:679–685.

Stern, S. (1996), "Semiparametric estimates of the supply and demand effects of disability on labor force participation", Journal of Econometrics 71:49–70.

Sutton, M., and C. Godfrey (1995), "A grouped data regression approach to estimating economic and social influences on individual drinking behaviour", Health Economics 4:237–247.

Terza, J.V. (1998), "Estimating count data models with endogenous switching: sample selection and endogenous treatment effects", Journal of Econometrics 84:93–127.

Theodossiou, I. (1998), "The effects of low-pay and unemployment on psychological well-being: a logistic regression approach", Journal of Health Economics 17:85–104.

Ullah, A. (1988), "Non-parametric estimation of econometric functionals", Canadian Journal of Economics 21:625–658.

van Doorslaer, E.K.A. (1987), Health, Knowledge and the Demand for Medical Care (van Gorcum, Assen/Maasricht).

van der Gaag, J., and B. Wolfe (1991), "Estimating demand for medical care: health as a critical factor for adults and children", in: G. Duru and J.H.P. Paelinck, eds., Econometrics of Health Care (Kluwer, Amsterdam) 31–58.

van de Ven, W.P.M.M., and J. van der Gaag (1982), "Health as an unobservable: a MIMIC model for health care demand", Journal of Health Economics 1:157–183.

van de Ven, W.P.M.M., and E.M. Hooijmans (1991), "The MIMIC health status index", in: G. Duru and J.H.P. Paelinck, eds., Econometrics of Health Care (Kluwer, Amsterdam) 19–29.

van de Ven, W.P.M.M., and B.M.S. van Praag (1981), "The demand for deductibles in private health insurance", Journal of Econometrics 17:229–252.

van Vliet, R.C.J.A., and B.M.S. van Praag (1987), "Health status estimation on the basis of MIMIC health care models", Journal of Health Economics 6:27–42.

Vella, F. (1998), "Estimating models with sample selection bias", Journal of Human Resources 33:127–169.

Vitaliano, D.F., and M. Toren (1994), "Cost and efficiency in nursing homes: a stochastic frontier approach", Journal of Health Economics 13:281–300.

Wagstaff, A. (1986), "The demand for health, Some new empirical evidence", Journal of Health Economics 5:195–233.

Wagstaff, A. (1989a), "Econometric studies in health economics", Journal of Health Economics 8:1–51.

Wagstaff, A. (1989b), "Estimating efficiency in the hospital sector: a comparison of three statistical cost frontiers", Applied Economics 21:659–672.

Wagstaff, A. (1993), "The demand for health: an empirical reformulation of the Grossman model", Health Economics 2:189–198.

Wasserman, J., W.G. Manning, J.P. Newhouse and J.D. Winkler (1991), "The effects of excise taxes and regulations on cigarette smoking", Journal of Health Economics 10:43–64.

Windmeijer, F.A.G., and J.M.C. Santos Silva (1997), "Endogeneity in count data models: an application to demand for health care", Journal of Applied Econometrics 12:281–294.

Wolfe, B., and J. van der Gaag (1981), "A new health status index for children", in: J. van der Gaag and M. Perlman, eds., Health, Economics, and Health Economics (North-Holland, Amsterdam) 283–304.

Yen, S.T., and A.M. Jones (1996), "Individual cigarette consumption and addiction: a flexible limited dependent variable approach", Health Economics 5:105–117.

Zimmerman Murphy, M. (1987), "The importance of sample selection bias in the estimation of medical care demand equations", Eastern Economic Journal 13:19–29.

Zuckerman, S., J. Hadley and L. Iezzoni (1994), "Measuring hospital efficiency with frontier cost functions", Journal of Health Economics 13:255–280.

PART 2

DEMAND AND REIMBURSEMENT FOR MEDICAL SERVICES

This Page Intentionally Left Blank

Chapter 7

THE HUMAN CAPITAL MODEL*

MICHAEL GROSSMAN

City University of New York Graduate School and National Bureau of Economic Research

Contents

*I am indebted to Robert Kaestner, Sara Markowitz, Tomas Philipson, and Walter Ried for helpful comments.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This chapter contains a detailed treatment of the human capital model of the demand for health which was originally developed in 1972. Theoretical predictions are discussed, and theoretical extensions of the model are reviewed. Empirical research that tests the predictions of the model or studies causality between years of formal schooling completed and good health is surveyed. The model views health as a durable capital stock that yields an output of healthy time. Individuals inherit an initial amount of this stock that depreciates with age and can be increased by investment. The household production function model of consumer behavior is employed to account for the gap between health as an output and medical care as one of many inputs into its production. In this framework the "shadow price" of health depends on many variables besides the price of medical care. It is shown that the shadow price rises with age if the rate of depreciation on the stock of health rises over the life cycle and falls with education (years of formal schooling completed) if more educated people are more efficient producers of health. An important result is that, under certain conditions, an increase in the shadow price may simultaneously reduce the quantity of health demanded and increase the quantities of health inputs demanded.

JEL classification: I10

1. Introduction

Almost three decades have elapsed since I published my National Bureau of Economic Research monograph [Grossman (1972b)] and Journal of Political Economy paper [Grossman (1972a)] dealing with a theoretical and empirical investigation of the demand for the commodity "good health."[1] My work was motivated by the fundamental difference between health as an output and medical care as one of a number of inputs into the production of health and by the equally important difference between health capital and other forms of human capital. According to traditional demand theory, each consumer has a utility or preference function that allows him or her to rank alternative combinations of goods and services purchased in the market. Consumers are assumed to select that combination that maximizes their utility function subject to an income or resource constraint: namely, outlays on goods and services cannot exceed income. While this theory provides a satisfactory explanation of the demand for many goods and services, students of medical economics have long realized that what consumers demand when they purchase medical services are not these services per se but rather better health. Indeed, as early as 1789, Bentham included relief of pain as one of fifteen "simple pleasures" which exhausted the list of basic arguments in one's utility function [Bentham (1931)]. The distinction between health as an output or an object of choice and medical care as an input had not, however, been exploited in the theoretical and empirical literature prior to 1972.

My approach to the demand for health has been labeled as the human capital model in much of the literature on health economics because it draws heavily on human capital theory [Becker (1964, 1967), Ben-Porath (1967), Mincer (1974)]. According to human capital theory, increases in a person's stock of knowledge or human capital raise his productivity in the market sector of the economy, where he produces money earnings, and in the nonmarket or household sector, where he produces commodities that enter his utility function. To realize potential gains in productivity, individuals have an incentive to invest in formal schooling and on-the-job training. The costs of these investments include direct outlays on market goods and the opportunity cost of the time that must be withdrawn from competing uses. This framework was used by Becker (1967) and by Ben-Porath (1967) to develop models that determine the optimal quantity of investment in human capital at any age. In addition, these models show how the optimal quantity varies over the life cycle of an individual and among individuals of the same age.

Although Mushkin (1962), Becker (1964), and Fuchs (1966) had pointed out that health capital is one component of the stock of human capital, I was the first person to construct a model of the demand for health capital itself. If increases in the stock of health simply increased wage rates, my undertaking would not have been necessary, for

[1] My monograph was the final publication in the Occasional Paper Series of the National Bureau of Economic Research. It is somewhat ironic that the publication of a study dealing with the demand for health marked the death of the series.

one could simply have applied Becker's and Ben-Porath's models to study the decision to invest in health. I argued, however, that health capital differs from other forms of human capital. In particular, I argued that a person's stock of knowledge affects his market and nonmarket productivity, while his stock of health determines the total amount of time he can spend producing money earnings and commodities.

My approach uses the household production function model of consumer behavior [Becker (1965), Lancaster (1966), Michael and Becker (1973)] to account for the gap between health as an output and medical care as one of many inputs into its production. This model draws a sharp distinction between fundamental objects of choice – called commodities – that enter the utility function and market goods and services. These commodities are Bentham's (1931) pleasures that exhaust the basic arguments in the utility function. Consumers produce commodities with inputs of market goods and services and their own time. For example, they use sporting equipment and their own time to produce recreation, traveling time and transportation services to produce visits, and part of their Sundays and church services to produce "peace of mind." The concept of a household production function is perfectly analogous to a firm production function. Each relates a specific output or a vector of outputs to a set of inputs. Since goods and services are inputs into the production of commodities, the demand for these goods and services is a derived demand for a factor of production. That is, the demand for medical care and other health inputs is derived from the basic demand for health.

There is an important link between the household production theory of consumer behavior and the theory of investment in human capital. Consumers as investors in their human capital *produce* these investments with inputs of their own time, books, teachers' services, and computers. Thus, some of the outputs of household production directly enter the utility function, while other outputs determine earnings or wealth in a life cycle context. Health, on the other hand, does both.

In my model, health – defined broadly to include longevity and illness-free days in a given year – is both demanded and produced by consumers. Health is a choice variable because it is a source of utility (satisfaction) and because it determines income or wealth levels. That is, health is demanded by consumers for two reasons. As a consumption commodity, it directly enters their preference functions, or, put differently, sick days are a source of disutility. As an investment commodity, it determines the total amount of time available for market and nonmarket activities. In other words, an increase in the stock of health reduces the amount of time lost from these activities, and the monetary value of this reduction is an index of the return to an investment in health.

Since health capital is one component of human capital, a person inherits an initial stock of health that depreciates with age – at an increasing rate at least after some stage in the life cycle – and can be increased by investment. Death occurs when the stock falls below a certain level, and one of the novel features of the model is that individuals "choose" their length of life. Gross investments are produced by household production functions that relate an output of health to such choice variables or health inputs as medical care utilization, diet, exercise, cigarette smoking, and alcohol consumption. In addition, the production function is affected by the efficiency or productivity of a given consumer as reflected by his or her personal characteristics. Efficiency is defined as the amount of health obtained from a given amount of health inputs. For example, years of formal schooling completed plays a large role in this context.

Since the most fundamental law in economics is the law of the downward sloping demand function, the quantity of health demanded should be negatively correlated with its "shadow price." I stress that the shadow price of health depends on many other variables besides the price of medical care. Shifts in these variables alter the optimal amount of health and also alter the derived demand for gross investment and for health inputs. I show that the shadow price of health rises with age if the rate of depreciation on the stock of health rises over the life cycle and falls with education (years of formal schooling completed) if more educated people are more efficient producers of health. I emphasize the result that, under certain conditions, an increase in the shadow price may simultaneously reduce the quantity of health demanded and increase the quantities of health inputs demanded.

The task in this paper is to outline my 1972 model of the demand for health, to discuss the theoretical predictions it contains, to review theoretical extensions of the model, and to survey empirical research that tests the predictions made by the model or studies causality between years of formal schooling completed and good health. I outline my model in Section 2 of this paper. I include a new interpretation of the condition for death, which is motivated in part by analyses by Ehrlich and Chuma (1990) and by Ried (1996, 1998). I also address a fundamental criticism of my framework raised by Ehrlich and Chuma involving an indeterminacy problem with regard to optimal investment in health. I summarize my pure investment model in Section 3, my pure consumption model in Section 4, and my empirical testing of the model in Section 5. While I emphasize my own contributions in these three sections, I do treat closely related developments that followed my 1972 publications. I keep derivations to a minimum because these can be found in Grossman (1972a, 1972b).[2] In Section 6 I focus on theoretical and empirical extensions and criticisms, other than those raised by Ehrlich and Chuma and by Ried.

I conclude in Section 7 with a discussion of studies that investigate alternative explanations of the positive relationship between years of formal schooling completed and alternative measures of adult health. While not all this literature is grounded in demand for health models, it is natural to address it in a paper of this nature because it essentially deals with complementary relationships between the two most important components of the stock of human capital. Currently, we still lack comprehensive theoretical models in which the stocks of health and knowledge are determined simultaneously. I am somewhat disappointed that my 1982 plea for the development of these models has gone unanswered [Grossman (1982)]. The rich empirical literature treating interactions between schooling and health underscores the potential payoffs to this undertaking.

[2] Grossman (1972b) is out of print but available in most libraries.

2. Basic model

2.1. Assumptions

Let the intertemporal utility function of a typical consumer be

$$U = U(\phi_t H_t, Z_t), \quad t = 0, 1, \dots, n, \tag{1}$$

where H_t is the stock of health at age t or in time period t, ϕ_t is the service flow per unit stock, $h_t = \phi_t H_t$ is total consumption of "health services," and Z_t is consumption of another commodity. The stock of health in the initial period (H_0) is given, but the stock of health at any other age is endogenous. The length of life as of the planning date (n) also is endogenous. In particular, death takes place when $H_t \leqslant H_{\min}$. Therefore, length of life is determined by the quantities of health capital that maximize utility subject to production and resource constraints.

By definition, net investment in the stock of health equals gross investment minus depreciation:

$$H_{t+1} - H_t = I_t - \delta_t H_t, \tag{2}$$

where I_t is gross investment and δ_t is the rate of depreciation during the tth period $(0 < \delta_t < 1)$. The rates of depreciation are exogenous but depend on age. Consumers produce gross investment in health and the other commodities in the utility function according to a set of household production functions:

$$I_t = I_t(M_t, TH_t; E), \tag{3}$$

$$Z_t = Z_t(X_t, T_t; E). \tag{4}$$

In these equations M_t is a vector of inputs (goods) purchased in the market that contribute to gross investment in health, X_t is a similar vector of goods inputs that contribute to the production of Z_t, TH_t and T_t are time inputs, and E is the consumer's stock of knowledge or human capital exclusive of health capital. This latter stock is assumed to be exogenous or predetermined. The semicolon before it highlights the difference between this variable and the endogenous goods and time inputs. In effect, I am examining the consumer's behavior after he has acquired the optimal stock of this capital.[3] Following Michael (1972, 1973) and Michael and Becker (1973), I assume that an increase in knowledge capital raises the efficiency of the production process in the nonmarket or household sector, just as an increase in technology raises the efficiency of the production process in the market sector. I also assume that all production functions are linear homogeneous in the endogenous market goods and own time inputs.

[3] Equations (3) and (4) assume that E does not vary over the life cycle. In Grossman (1972b, pp. 28–30), I consider the impacts of exogenous variations in this stock with age.

In much of my modeling, I treat the vectors of goods inputs, M_t and X_t, as scalars and associate the market goods input in the gross investment production function with medical care. Clearly this is an oversimplification because many other market goods and services influence health. Examples include housing, diet, recreation, cigarette smoking, and excessive alcohol use. The latter two inputs have negative marginal products in the production of health. They are purchased because they are inputs into the production of other commodities such as "smoking pleasure" that yield positive utility. In completing the model I will rule out this and other types of joint production, although I consider joint production in some detail in Grossman (1972b, pp. 74–83). I also will associate the market goods input in the health production function with medical care, although the reader should keep in mind that the model would retain its structure if the primary health input purchased in the market was something other than medical care. This is important because of evidence that medical care may be an unimportant determinant of health in developed countries [see Auster et al. (1969)] and because Zweifel and Breyer (1997) use the lack of a positive relationship between correlates of good health and medical care in micro data to criticize my approach.

Both market goods and own time are scarce resources. The goods budget constraint equates the present value of outlays on goods to the present value of earnings income over the life cycle plus initial assets (discounted property income):

$$\sum_{t=0}^{n} \frac{P_t M_t + Q_t X_t}{(1+r)^t} = \sum_{t=0}^{n} \frac{W_t TW_t}{(1+r)^t} + A_0. \tag{5}$$

Here P_t and Q_t are the prices of M_t and X_t, W_t is the hourly wage rate, TW_t is hours of work, A_0 is initial assets, and r is the market rate of interest. The time constraint requires that Ω, the total amount of time available in any period, must be exhausted by all possible uses:

$$TW_t + TH_t + T_t + TL_t = \Omega, \tag{6}$$

where TL_t is time lost from market and nonmarket activities due to illness and injury.

Equation (6) modifies the time budget constraint in Becker's (1965) allocation of time model. If sick time were not added to market and nonmarket time, total time would not be exhausted by all possible uses. I assume that sick time is inversely related to the stock of health; that is $\partial TL_t / \partial H_t < 0$. If Ω is measured in hours ($\Omega = 8{,}760$ hours or 365 days times 24 hours per day if the year is the relevant period) and if ϕ_t is defined as the flow of healthy time per unit of H_t, h_t equals the total number of healthy hours in a given year. Then one can write

$$TL_t = \Omega - h_t. \tag{7}$$

From now on, I assume that the variable h$_t$ in the utility function coincides with healthy hours.[4]

By substituting for hours of work (TW_t) from Equation (6) into Equation (5), one obtains the single "full wealth" constraint:

$$\sum_{t=0}^{n} \frac{P_t M_t + Q_t X_t + W_t(TL_t + TH_t + T_t)}{(1+r)^t} = \sum_{t=0}^{n} \frac{W_t \Omega}{(1+r)^t} + A_0. \tag{8}$$

Full wealth, which is given by the right-hand side of Equation (8), equals initial assets plus the discounted value of the earnings an individual would obtain if he spent all of his time at work. Part of this wealth is spent on market goods, part of it is spent on nonmarket production, and part of it is lost due to illness. The equilibrium quantities of H_t and Z_t can now be found by maximizing the utility function given by Equation (1) subject to the constraints given by Equations (2), (3), and (8). Since the inherited stock of health and the rates of depreciation are given, the optimal quantities of gross investment determine the optimal quantities of health capital.

2.2. Equilibrium conditions

First-order optimality conditions for gross investment in period $t-1$ are[5]

$$\begin{aligned} \frac{\pi_{t-1}}{(1+r)^{t-1}} = {} & \frac{W_t G_t}{(1+r)^t} + \frac{(1-\delta_t) W_{t+1} G_{t+1}}{(1+r)^{t+1}} + \cdots \\ & + \frac{(1-\delta_t)\cdots(1-\delta_{n-1}) W_n G_n}{(1+r)^n} \\ & + \frac{Uh_t}{\lambda} G_t + \cdots (1-\delta_t)\cdots(1-\delta_{n-1}) \frac{Uh_n}{\lambda} G_n, \end{aligned} \tag{9}$$

$$\pi_{t-1} = \frac{P_{t-1}}{\partial I_{t-1}/\partial M_{t-1}} = \frac{W_{t-1}}{\partial I_{t-1}/\partial TH_{t-1}}. \tag{10}$$

The new symbols in these equations are: $Uh_t = \partial U/\partial h_t$, the marginal utility of healthy time; λ, the marginal utility of wealth; $G_t = \partial h_t/\partial H_t = -(\partial TL_t/\partial H_t)$, the marginal

[4] Clearly this is a simplification. No distinction is made between the quality and the quantity of healthy time. If the stock of health yielded other services besides healthy time, ϕ_t would be a vector of service flows. These services might or might not be perfect substitutes in the utility function.

[5] An increase in gross investment in period $t-1$ increases the stock of health in all future periods. These increases are equal to

$$\frac{\partial H_t}{\partial I_{t-1}} = 1, \quad \frac{\partial H_{t+1}}{\partial I_{t-1}} = (1-\delta_t), \ldots, \frac{\partial H_n}{\partial I_{t-1}} = (1-\delta_t)(1-\delta_{t+1})\cdots(1-\delta_{n-1}).$$

product of the stock of health in the production of healthy time; and π_{t-1}, the marginal cost of gross investment in health in period $t-1$.

Equation (9) states that the present value of the marginal cost of gross investment in health in period $t-1$ must equal the present value of marginal benefits. Discounted marginal benefits at age t equal

$$G_t\left[\frac{W_t}{(1+r)^t}+\frac{Uh_t}{\lambda}\right],$$

where G_t is the marginal product of health capital – the increase in the amount of healthy time caused by a one-unit increase in the stock of health. Two monetary magnitudes are necessary to convert this marginal product into value terms because consumers desire health for two reasons. The discounted wage rate measures the monetary value of a one-unit increase in the total amount of time available for market and nonmarket activities, and the term Uh_t/λ measures the discounted monetary value of the increase in utility due to a one-unit increase in healthy time. Thus, the sum of these two terms measures the discounted marginal value to consumers of the output produced by health capital.

Condition (9) holds for any capital asset, not just for health capital. The marginal cost as of the current period, obtained by multiplying both sides of the equation by $(1+r)^{t-1}$, must be equated to the discounted flows of marginal benefits in the future. This is true for the asset of health capital by labeling the marginal costs and benefits of this particular asset in the appropriate manner. As I will show presently, most of the effects of variations in exogenous variables can be traced out as shifting the marginal costs and marginal benefits of the asset.

While Equation (9) determines the optimal amount of gross investment in period $t-1$, Equation (10) shows the condition for minimizing the cost of producing a given quantity of gross investment. Total cost is minimized when the increase in gross investment from spending an additional dollar on medical care equals the increase in total cost from spending an additional dollar on time. Since the gross investment production function is homogeneous of degree one in the two endogenous inputs and since the prices of medical care and time are independent of the level of these inputs, the average cost of gross investment is constant and equal to the marginal cost.

To examine the forces that affect the demand for health and gross investment, it is useful to convert Equation (9) into an equation that determines the optimal stock of health in period t. If gross investment in period t is positive, a condition similar to (9) holds for its optimal value. From these two first-order conditions

$$G_t\left[W_t+\left(\frac{Uh_t}{\lambda}\right)(1+r)^t\right]=\pi_{t-1}\big(r-\tilde{\pi}_{t-1}+\delta_t\big), \tag{11}$$

where $\tilde{\pi}_{t-1}$ is the percentage rate of change in marginal cost between period $t-1$ and period t.[6] Equation (11) implies that the undiscounted value of the marginal product of the optimal stock of health capital at any age must equal the supply price of capital, $\pi_{t-1}(r - \tilde{\pi}_{t-1} + \delta_t)$. The latter contains interest, depreciation, and capital gains components and may be interpreted as the rental price or user cost of health capital.

Equation (11) fully determines the optimal quantity at time t of a capital good that can be bought and sold in a perfect market. The stock of health capital, like the stock of knowledge capital, cannot be sold because it is imbedded in the investor. This means that gross investment cannot be nonnegative. Although sales of capital are ruled out, provided gross investment is positive, there exists a user cost of capital that in equilibrium must equal the value of the marginal product of the stock. In Grossman (1972a, p. 230); (1972b, pp. 6–7), I provide an intuitive interpretation of this result by showing that exchanges over time in the stock of health by an individual substitute for exchanges in the capital market.

2.3. *Optimal length of life*[7]

So far I have essentially reproduced the analysis of equilibrium conditions in my 1972 National Bureau of Economic Research monograph and Journal of Political Economy article. A perceptive reader may have noted that an explicit condition determining length of life is absent. The discounted marginal benefits of an investment in health in period 0 are summed from periods 1 through n, so that the consumer is alive in period n and dead in period $n+1$.[8] This means that H_{n+1} is equal to or less than $H_{\min}$, the death stock, while H_n and H_t $(t < n)$ exceed $H_{\min}$. But how do we know that the optimal quantities of the stock of health guarantee this outcome? Put differently, length of life is supposed to be an endogenous variable in the model, yet discounted income and expenditure flows in the full wealth constraint and discounted marginal benefits in the first-order conditions appear to be summed over a fixed n.

I was bothered by the above while I was developing my model. As of the date of its publication, I was not convinced that length of life was in fact being determined by the model. There is a footnote in my Journal of Political Economy article [Grossman (1972a, p. 228, footnote 7)] and in my National Bureau of Economic Research monograph [Grossman (1972b, p. 4, footnote 9)] in which I impose the constraints that $H_{n+1} \leqslant H_{\min}$ and $H_n > H_{\min}$.[9] Surely, it is wrong to impose these constraints in a maximization problem in which length of life is endogenous.

[6] Equation (11) assumes $\delta_t \tilde{\pi}_{t-1} \cong 0$.

[7] First-time readers of this chapter can skip Sections 2.3 and 2.4. The material in the remaining sections does not depend on them.

[8] Since the initial period is period 0, a consumer who is alive in year n and dead in year $n+1$ lives for $n+1$ years.

[9] Actually, I assert that I am assuming $H_n \leqslant H_{\min}$. That is incorrect because $H_n > H_{\min}$ if the consumer is alive in period n. The corrected footnote should read: "The constraints are imposed that $H_{n+1} \leqslant H_{\min}$ and $H_n > H_{\min}$."

My publications on the demand for health were outgrowths of my 1970 Columbia University Ph.D. dissertation. While I was writing my dissertation, my friend and fellow Ph.D. candidate, Gilbert R. Ghez, pointed out that the determination of optimal length of life could be viewed as an iterative process. I learned a great deal from him, and I often spent a long time working through the implications of his comments.[10] It has taken me almost thirty years to work through his comment on the iterative determination of length of life. I abandoned this effort many years ago but returned to it when I read Ried's (1996, 1998) reformulation of the selection of the optimal stock of health and length of life as a discrete time optimal control problem. Ried (1998, p. 389) writes: "Since [the problem] is a free terminal time problem, one may suspect that a condition for the optimal length of the planning horizon is missing in the set of necessary conditions .... However, unlike the analogous continuous time problem, the discrete time version fails to provide such an equation. Rather, the optimal final period ... has to be determined through the analysis of a sequence of fixed terminal time problems with the terminal time varying over a plausible domain." This is the same observation that Ghez made. I offer a proof below. I do not rely on Ried's solution. Instead, I offer a much simpler proof which has a very different implication than the one offered by Ried.

A few preliminaries are in order. First, I assume that the rate of depreciation on the stock of health (δ_t) rises with age. As we shall see in more detail later on, this implies that the optimal stock falls with age. Second, I assume that optimal gross investment in health is positive except in the very last year of life. Third, I define V_t as $W_t + (Uh_t/\lambda)(1+r)^t$. Hence, V_t is the undiscounted marginal value of the output produced by health capital in period t. Finally, since the output produced by health capital has a finite upper limit of 8,760 hours in a year, I assume that the marginal product of the stock of health (G_t) diminishes as the stock increases ($\partial G_t/\partial H_t < 0$).

Consider the maximization problem outlined in Section 2.1 except that the planning horizon is exogenous. That is, an individual is alive in period n and dead in period $n+1$. Write the first-order conditions for the optimal stocks of health compactly as

$$V_t G_t = \pi_{t-1}\big(r - \tilde{\pi}_{t-1} + \delta_t\big), \quad t < n, \tag{12}$$

$$V_n G_n = \pi_{n-1}(r+1). \tag{13}$$

Note that Equation (13) follows from the condition for optimal gross investment in period $n-1$. An investment in that period yields returns in one period only (period n) since the individual dies after period n. Put differently, the person behaves as if the rate of depreciation on the stock of health is equal to 1 in period n.

[10] Readers seeking a definitive and path-breaking treatment of the allocation of goods and time over the life cycle, should consult Ghez's pioneering monograph on this topic with Becker [Ghez and Becker (1975)]. In the late 1970s and 1980s, there was a tremendous growth in the literature on life-cycle labor supply and consumption demand using the concept of demand functions that hold the marginal utility of wealth constant. All this literature can be traced to Ghez's treatment of the topic.

I also will make use of the first-order conditions for gross investment in health in periods 0 and n:

$$\pi_0 = \frac{V_1 G_1}{(1+r)} + \frac{d_2 V_2 G_2}{(1+r)^2} + \cdots + \frac{d_n V_n G_n}{(1+r)^n}, \tag{14}$$

$$I_n = 0. \tag{15}$$

In Equation (14), d_t is the increase in the stock of health in period t caused by an increase in gross investment in period 0:

$$d_1 = 1, d_t(t > 1) = \prod_{j=1}^{t-1}(1 - \delta_j).$$

Obviously, gross investment in period n is 0 because the individual will not be alive in period $n + 1$ to collect the returns.

In order for death to take place in period $n + 1$, $H_{n+1} \leqslant H_{\min}$. Since $I_n = 0$,

$$H_{n+1} = (1 - \delta_n)H_n. \tag{16}$$

Hence, for the solution (death after period n) to be fully consistent

$$H_{n+1} = (1 - \delta_n)H_n \leqslant H_{\min}. \tag{17}$$

Suppose that condition (17) is violated. That is, suppose maximization for a fixed number of periods equal to n results in a stock in period $n + 1$ that exceeds the death stock. Then lifetime utility should be re-maximized under the assumption that the individual will be alive in period $n + 1$ but dead in period $n + 2$. As a first approximation, the set of first-order conditions for H_t $(t < n)$ defined by Equation (12) still must hold so that the stock in each of these periods is not affected when the horizon is lengthened by 1 period.[11] But the condition for the stock in period n becomes

$$V_n^* G_n^* = \pi_{n-1}\big(r - \tilde{\pi}_{n-1} + \delta_n\big), \tag{18}$$

where asterisks are used because the stock of health in period n when the horizon is $n + 1$ is not equal to the stock when the horizon is n (see below). Moreover,

$$V_{n+1}^* G_{n+1}^* = \pi_n(r + 1), \tag{19}$$

[11] This is a first approximation because it assumes that that λ does not change when the horizon is extended by one period. Consider the standard intertemporally separable lifetime utility function in which the current period utility function, $\psi(h_t, Z_t)$, is strictly concave. With full wealth constant, an increase in the horizon causes λ to rise. But full wealth increases by $W_{n+1}\Omega/(1 + r)^{n+1}$, which causes λ to decline. I assume that these two effects exactly offset each other. This assumption is not necessary in the pure investment model described in Sections 2.5 and 3 because V_t does not depend on λ in that model.

$$I_{n+1} = 0, \tag{20}$$

$$H_{n+2} = (1 - \delta_{n+1}) H^*_{n+1}. \tag{21}$$

If the stock defined by Equation (21) is less than or equal to $H_{\min}$, death takes place in period $n + 2$. If H_{n+2} is greater than $H_{\min}$, the consumer re-maximizes lifetime utility under the assumption that death takes place in period $n + 3$ (the horizon ends in period $n + 2$).

I have just described an iterative process for the selection of optimal length of life. In words, the process amounts to maximizing lifetime utility for a fixed horizon, checking to see whether the stock in the period after the horizon ends (the terminal stock) is less than or equal to the death stock ($H_{\min}$), and adding one period to the horizon and re-maximizing the utility function if the terminal stock exceeds the death stock.[12] I want to make several comments on this process and its implications. Compare the condition for the optimal stock of health in period n when the horizon lasts through period n [Equation (13)] with the condition for the optimal stock in the same period when the horizon lasts through period $n + 1$ [Equation (18)]. The supply price of health capital is smaller in the latter case because $\delta_n < 1$.[13] Hence, the undiscounted value of the marginal product of health capital in period n when the horizon is $n + 1 (V^*_n G^*_n)$ must be smaller than the undiscounted value of the marginal product of health capital in period n when the horizon is n $(V_n G_n)$. In turn, due to diminishing marginal productivity, the stock of health in period n must rise when the horizon is extended by one period $(H^*_n > H_n)$.[14]

[12] If $H_n \leqslant H_{\min}$, the utility function is re-maximized after shortening the horizon by 1 period.

[13] It may appear that the supply price of capital given by the right-hand side of Equation (18) is smaller than the one given by the right-hand side of Equation (13) as long as $1 - \delta_n > -\tilde{\pi}_{n-1}$. But Equation (18) is based on the assumption $\delta_n \tilde{\pi}_{n-1} \cong 0$. The exact form of (18) is

$$V^*_n G^*_n = \pi_{n-1}(r + 1) - (1 - \delta_n)\pi_n.$$

The difference between the right-hand side of this equation and the right-hand side of Equation (13) is $-(1 - \delta_n)\pi_n < 0$.

[14] While G^*_n is smaller than G_n, it is not clear whether V^*_n is smaller than V_n. The last term can be written

$$V_n = W_n + \frac{Uh_n}{U_n} m_t,$$

where U_n is the marginal utility of Z_n and m_n is the marginal cost of producing Z_n. The wage rate in period n (W_n) and m_n are not affected when the length of the horizon is increased from n to $n + 1$. In equilibrium,

$$\frac{Uh_n}{U_n} = \left[\frac{(\pi_{n-1}(1 + r))/G_n - W_n}{m_n}\right]$$

and

$$\frac{Uh^*_n}{U^*_n} = \left[\frac{(\pi_{n-1}(1 + r) - \pi_n(1 - \delta_n))/G^*_n - W_n}{m_n}\right].$$

When the individual lives for $n+1$ years, the first-order condition for gross investment in period 0 is

$$\pi_0 = \frac{V_1 G_1}{(1+r)} + \frac{d_2 V_2 G_2}{(1+r)^2} + \cdots + \frac{d_n V_n^* G_n^*}{(1+r)^n} + \frac{d_n(1-\delta_n)V_{n+1}^* G_{n+1}^*}{(1+r)^{n+1}}. \tag{22}$$

Note that the discounted marginal benefits of an investment in period 0 are the same whether the person dies in period $n+1$ or in period $n+2$ [compare the right-hand sides of Equations (14) and (22)] since the marginal cost of an investment in period 0 does not depend on the length of the horizon. This may seem strange because one term is added to discounted marginal benefits of an investment in period 0 or in any other period when the horizon is extended by one period – the discounted marginal benefit in period $n+1$. This term, however, is exactly offset by the reduction in the discounted marginal benefit in period n. The same offset occurs in the discounted marginal benefits of investments in every other period except for periods $n-1$ and n.

A proof of the last proposition is as follows. The first $n-1$ terms on the right-hand sides of Equations (14) and (22) are the same. From Equations (13), (18), and (19),

$$\begin{aligned} V_n^* G_n^* &= \frac{V_n G_n (r - \tilde{\pi}_{n-1} + \delta_n)}{(1+r)}, \\ V_{n+1}^* G_{n+1}^* &= V_n G_n \big(1 + \tilde{\pi}_{n-1}\big). \end{aligned}$$

Hence, the sum of the last two terms on the right-hand side of Equation (22) equals the last term on the right-hand side of Equation (14):[15]

$$\frac{d_n V_n^* G_n^*}{(1+r)^n} + \frac{d_n(1-\delta_n)V_{n+1}^* G_{n+1}^*}{(1+r)^{n+1}} = \frac{d_n V_n G_n}{(1+r)^n}. \tag{23}$$

Using the last result, one can fully describe the algorithm for the selection of optimal length of life. Maximize the lifetime utility function for a fixed horizon. Check to see

Suppose that

$$G_n^* = G_n \left[\frac{\pi_{n-1}(r+1) - \pi_n(1-\delta_n)}{\pi_{n-1}(r+1)} \right].$$

Then $Uh_n^*/U_n^* = Uh_n/U_n$ and $V_n^* = V_n$. In this case there is no incentive to substitute healthy time for the other commodity in the utility function in period n when the horizon is increased by one period. In other cases this type of substitution will occur. If it does occur, I assume that the lifetime utility function is separable over time so that the marginal rate of substitution between h and Z in periods other than period n is not affected. Note the distinction between H_{n+1} and H_{n+1}^*. The former stock is the one associated with an n period horizon and $I_n = 0$. The latter stock is the one associated with an $n+1$ horizon and $I_n > 0$. Clearly, $H_{n+1}^* > H_{n+1}$.

[15] In deriving Equation (23) I use the approximation that $\delta_n \tilde{\pi}_{n-1} \cong 0$. If the exact form of Equation (18) is employed (see footnote 13), the approximation is not necessary.

whether the terminal stock is less than or equal to the death stock. If the terminal stock exceeds the death stock, add one period to the horizon and redo the maximization. The resulting values of the stock of health must be the same in every period except for periods n and $n+1$. The stock of health must be larger in these two periods when the horizon equals $n+1$ than when the horizon equals n. The stock in period t depends on gross investment in period $t-1$, with gross investment in previous periods held constant. Therefore, gross investment is larger in periods $n-1$ and n but the same in every other period when the horizon is increased by one year. A rise in the rate of depreciation with age guarantees finite life since for some j[16]

$$H_{n+j} = (1-\delta_{n+j-1})H_{n+j-1} \leqslant H_{\min}.$$

I have just addressed a major criticism of my model made by Ehrlich and Chuma (1990). They argue that my analysis does not determine length of life because it "... does not develop the required terminal (transversality) conditions needed to assure the consistency of any solutions for the life cycle path of health capital and longevity" [Ehrlich and Chuma (1990, p. 762)]. I have just shown that length of life is determined as the outcome of an iterative process in which lifetime utility functions with alternative horizons are maximized. Since the continuous time optimal control techniques employed by Ehrlich and Chuma are not my fields of expertise, I invite the reader to study their paper and make up his or her own mind on this issue.

As I indicated at the beginning of this subsection, Ried (1996, 1998) offers the same general description of the selection of length of life as an iterative process. He proposes

[16] If $W_{n+1} = W_n$ and $\pi_n = \pi_{n-1}$, then H^*_{n+1} (the optimal stock when the horizon is $n+1$) $= H_n$ (the optimal stock when the horizon is n). Hence,

$$\begin{aligned} H_{n+2} &= (1-\delta_{n+1})H_n, \\ H_{n+1} &= (1-\delta_n)H_n. \end{aligned}$$

Since $\delta_{n+1} > \delta_n$, H_{n+2} could be smaller than or equal to $H_{\min}$, while at the same time H_n could exceed $H_{\min}$. Note that one addition to the algorithm described is required. Return to the case when maximization for a fixed number of periods equal to n results in a stock in period $n+1$ that is smaller than or equal to the death stock. The consumer should behave as if the rate of depreciation on the stock of health is equal to 1 in period n and consult Equation (13) to determine I_{n-1} and H_n. Suppose instead that he behaves as if the rate of depreciation is the actual rate in period n ($\delta_n < 1$). This is the same rate used by the consumer who dies in period $n+2$. Denote I^*_{n-1} as the quantity of gross investment that results from using Equation (18) to select the optimal stock in period n. Under the alternative decision rule, the stock in period $n+1$ could exceed the death stock. The difference in the stock in period $n+1$ that results from these two alternative decision rules is

$$H^*_{n+1} - H_{n+1} = I^*_{n-1} + \delta_{n-1}\Big(I_{n-1} - I^*_{n-1}\Big),$$

where I assume that I_n (gross investment in period n when death takes place in period $n+2$) equals I_{n-1} (gross investment in period $n-1$ when death takes place in period $n+1$). This difference falls as I^*_{n-1} falls. In turn, I^*_{n-1} falls as δ_n rises with rates of depreciation in all other periods held constant.

a solution using extremely complicated discrete time optimal control techniques. Again, I leave the reader to evaluate Ried's solution. But I do want to challenge his conclusion that "... sufficiently small perturbations of the trajectories of the exogenous variables will not alter the length of the individual's planning horizon .... [T]he uniqueness assumption [about length of life] ensures that the planning horizon may be treated as fixed in comparative dynamic analysis .... Given a fixed length of the individual's life, it is obvious that the mortality aspect is entirely left out of the picture. Thus, the impact of parametric changes upon individual health is confined to the quality of life which implies the analysis to deal [sic] with a pure morbidity effect" (p. 389).

In my view it is somewhat unsatisfactory to begin with a model in which length of life is endogenous but to end up with a result in which length of life does not depend on any of the exogenous variables in the model. This certainly is not an implication of my analysis of the determination of optimal length of life. In general, differences in such exogenous variables as the rate of depreciation, initial assets, and the marginal cost of investing in health across consumers of the same age will lead to differences in the optimal length of life.[17]

To be concrete, consider two consumers: a and b. Person a faces a higher rate of depreciation in each period than person b. The two consumers are the same in all other respects. Suppose that it is optimal for person a to live for n years (to die in year $n+1$). Ried argues that person b also lives for n years because both he and person a use Equation (13) to determine the optimal stock of health in period n. That equation is independent of the rate of depreciation in period n. Hence, the stock of health in period n is the same for each consumer. For person a, we have

$$H_{n+1}^{\mathrm{a}} = \left(1-\delta_n^{\mathrm{a}}\right) H_n^{\mathrm{a}} \leqslant H_{\min},$$

where the superscript a denotes values of variables for person a. But for person b,

$$H_{n+1}^{\mathrm{b}} = \left(1-\delta_n^{\mathrm{b}}\right) H_n^{\mathrm{a}}.$$

Since $\delta_n^{\mathrm{b}} < \delta_n^{\mathrm{a}}$, there is no guarantee that $H_{n+1}^{\mathrm{b}} \leqslant H_{\min}$. If $H_{n+1}^{\mathrm{b}} > H_{\min}$, person b will be alive in period $n+1$. He will then use Equation (18), rather than Equation (13), to pick his optimal stock in period n. In this case person b will have a larger optimal stock in period n than person a and will have a longer length of life.

Along the same lines parametric differences in the marginal cost of investment in health (differences in the marginal cost across people of the same age), differences in initial assets, and parametric differences in wage rates cause length of life to vary among individuals. In general, any variable that raises the optimal stock of health in each period

[17] Technically, I am dealing with parametric differences in exogenous variables (differences in exogenous variables across consumers) as opposed to evolutionary differences in exogenous variables (differences in exogenous variables across time for the same consumer). This distinction goes back to Ghez and Becker (1975) and is explored in detail by MaCurdy (1981).

of life also tends to prolong length of life.[18] Thus, if health is not an inferior commodity, an increase in initial assets or a reduction in the marginal cost of investing in health induces a longer optimal life. Persons with higher wage rates have more wealth; taken by itself, this prolongs life. But the relative price of health (the price of h_t relative to the price of Z_t) may rise as the wage rate rises. If this occurs and the resulting substitution effect outweighs the wealth effect, length of life may fall.

According to Ried, death occurs if $H_{n+1} < H_{\min}$ rather than if $H_{n+1} \leqslant H_{\min}$. The latter condition is the one that I employ, but that does not seem to account for the difference between my analysis and his analysis. Ried's only justification of his result is in the context of a dynamic model of labor supply. He assumes that a non-negativity constraint is binding in some period and concludes that marginal changes in any exogenous variable will fail to bring about positive supply.

Ried's conclusion does not appear to be correct. To see this in the most simple manner, consider a static model of the supply of labor, and suppose that the marginal rate of substitution between leisure time and consumption evaluated at zero hours of work is greater than or equal to the market wage rate at the initial wage. Hence, no hours are supplied to the market. Now suppose that the wage rate rises. If the marginal rate of substitution at zero hours equaled the old wage, hours of work will rise above zero. If the marginal rate of substitution at zero hours exceeded the old wage, hours could still rise above zero if the marginal rate of substitution at zero hours is smaller than the new wage. By the same reasoning, while not every parametric reduction in the rate of depreciation on the stock of health will increase optimal length of life (see footnote 18), some reductions surely will do so. I stand by my statement that it is somewhat unsatisfactory to begin with a model in which length of life is endogenous and end up with a result in which length of life does not depend on any of the exogenous variables in the model.

2.4. "Bang-bang" equilibrium

Ehrlich and Chuma (1990) assert that my "key assumption that health investment is produced through a constant-returns-to-scale ... technology introduces a type of indeterminacy ('bang-bang') problem with respect to optimal investment and health maintenance

[18] Consider two people who face the same rate of depreciation in each period. Person b has higher initial assets than person a and picks a larger stock in each period. Suppose that person a dies after period n. Hence,

$$H^{\mathrm{a}}_{n+1} = (1-\delta_n)H^{\mathrm{a}}_n \leqslant H_{\min}.$$

Suppose that person b, like person a, invests nothing in period n. Then

$$H^{\mathrm{b}}_{n+1} = (1-\delta_n)H^{\mathrm{b}}_n.$$

Clearly, $H^{\mathrm{b}}_{n+1} > H^{\mathrm{a}}_{n+1}$ since $H^{\mathrm{b}}_n > H^{\mathrm{a}}_n$. But both people could die in period $n+1$ if $H^{\mathrm{a}}_{n+1} < H_{\min}$. This is a necessary, but not a sufficient, condition. For this reason, I use the term "tends" in the text. This ambiguity is removed if the condition for death is defined by $H_{n+1} = H_{\min}$. That definition is, however, unsatisfactory because the rate of depreciation in period n does not guarantee that it is satisfied. That is, death takes place in $n+1$ if $\delta_n \geqslant (H_n - H_{\min})/H_n$.

choices. … [This limitation precludes] a systematic resolution of the choice of *both* (their italics) optimal health paths and longevity. … Later contributions to the literature spawned by Grossman … suffer in various degrees from these shortcomings. … Under the linear production process assumed by Grossman, the marginal cost of investment would be constant, and no interior equilibrium for investment would generally exist" (pp. 762, 764, 768).[19]

Ried (1998) addresses this criticism by noting that an infinite rate of investment is not consistent with equilibrium. Because Ehrlich and Chuma's criticism appears to be so damaging and Ried's treatment of it is brief and not convincing, I want to deal with it before proceeding to examine responses of health, gross investment, and health inputs to evolutionary (life-cycle) and parametric variations in key exogenous variables. Ehrlich and Chuma's point is as follows. Suppose that the rate of depreciation on the stock of health is equal to zero at every age, suppose that the marginal cost of gross investment in health does not depend on the amount of investment, and suppose that none of the other exogenous variables in the model is a function of age.[20] Then the stock of health is constant over time (net investment is zero). Any discrepancy between the initial stock and the optimal stock is erased in the initial period. In a continuous time model, this means an infinite rate of investment to close the gap followed by no investment after that. If the rate of depreciation is positive and constant, the discrepancy between the initial and optimal stock is still eliminated in the initial period. After that, gross investment is positive, constant, and equal to total depreciation; while net investment is zero.

To avoid the "bang-bang" equilibrium (an infinite rate of investment to eliminate the discrepancy between the initial and the desired stock followed by no investment if the rate of depreciation is zero), Ehrlich and Chuma assume that the production function of gross investment in health exhibits diminishing returns to scale. Thus, the marginal cost of gross or net investment is a positive function of the amount of investment. Given this, there is an incentive to reach the desired stock gradually rather than instantaneously since the cost of gradual adjustment is smaller than the cost of instantaneous adjustment.

The introduction of diminishing returns to scale greatly complicates the model because the marginal cost of gross investment and its percentage rate of change over time become endogenous variables that depend on the quantity of investment and its rate of change. In Section 6, I show that the structural demand function for the stock of health at age t in a model with costs of adjustment is one in which H_t depends on the stock at age $t+1$ and the stock at age $t-1$. The solution of this second-order difference equation results in a reduced form demand function in which the stock at age t depends on all past and future values of the exogenous variables. This makes theoretical and econometric analysis very difficult.

[19] For an earlier criticism of my model along the same lines, see Usher (1975).

[20] In addition, I assume that the market rate of interest is equal to the rate of time preference for the present. See Section 4 for a definition of time preference.

Are the modifications introduced by Ehrlich and Chuma really necessary? In my view the answer is no. The focus of my theoretical and empirical work and that of others who have adopted my framework [Cropper (1977), Muurinen (1982), Wagstaff (1986)] certainly is not on discrepancies between the inherited or initial stock and the desired stock. I am willing to assume that consumers reach their desired stocks instantaneously in order to get sharp predictions that are subject to empirical testing. Gross investment is positive (but net investment is zero) if the rate of depreciation is positive but constant in my model. In the Ehrlich–Chuma model, net investment can be positive in this situation. In both models consumers choose an infinite life. In both models life is finite and the stock of health varies over the life cycle if the rate of depreciation is a positive function of age. In my model positive net investment during certain stages of the life cycle is not ruled out. For example, the rate of depreciation might be negatively correlated with age at early stages of the life cycle. The stock of health would be rising and net investment would be positive during this stage of the life cycle.

More fundamentally, Ehrlich and Chuma introduce rising marginal cost of investment to remove an indeterminacy that really does not exist. In Figure 1 of their paper (p. 768), they plot the marginal cost of an investment in health as of age t and the discounted marginal benefits of this investment as functions of the quantity of investment. The discounted marginal benefit function is independent of the rate of investment. Therefore, no interior equilibrium exists for investment unless the marginal cost function slopes upward. This is the basis of their claim that my model does not determine optimal investment because marginal cost does not depend on investment.

Why, however, is the discounted marginal benefit function independent of the amount of investment? In a personal communication, Ehrlich informed me that this is because the marginal product of the stock of health at age t does not depend on the amount of investment at age t. Surely that is correct. But an increase in I_t raises the stock of health in all future periods. Since the marginal product of health capital diminishes as the stock rises, discounted marginal benefits must fall. Hence, the discounted marginal benefit function slopes downward, and an interior equilibrium for gross investment in period t clearly is possible even if the marginal cost of gross investment is constant.

Since discounted marginal benefits are positive when gross investment is zero, the discounted marginal benefit function intersects the vertical axis.[21] Thus corner solutions for gross investment are not ruled out in my model. One such solution occurs if the rate

[21] This follows because

$$\frac{\partial h_{t+j}}{\partial I_t} = \frac{\partial h_{t+j}}{\partial H_{t+j}} \frac{\partial H_{t+j}}{\partial I_t},$$

or

$$\frac{\partial h_{t+j}}{\partial I_t} = G_{t+j}(1-\delta_{t+1})(1-\delta_{t+2})\cdots(1-\delta_{t+j-1}) = G_{t+j}d_{t+j}.$$

Clearly, d_{t+j} is positive and finite when I_t equals zero. Moreover G_{t+j} is positive and finite when I_t equals zero as long as H_{t+j} is positive and finite.

of depreciation on the stock of health equals zero in every period. Given positive rates of depreciation, corner solutions still are possible in periods other than the last period of life because the marginal cost of gross investment could exceed discounted marginal benefits for all positive quantities of investment. I explicitly rule out corner solutions when depreciation rates are positive, and Ried and other persons who have used my model also rule them out. Corner solutions are possible in the Ehrlich–Chuma model if the marginal cost function of gross investment intersects the vertical axis. Ehrlich and Chuma rule out corner solutions by assuming that the marginal cost function passes through the origin.

To summarize, unlike Ried, I conclude that exogenous variations in the marginal cost and marginal benefit of an investment in health cause optimal length of life to vary. Unlike Ehrlich and Chuma, I conclude that my 1972 model provides a simple but logically consistent framework for studying optimal health paths and longevity. At the same time I want to recognize the value of Ried's emphasis on the determination of optimal length of life as the outcome of an iterative process in a discrete time model. I also want to recognize the value of Ehrlich and Chuma's model in cases when there are good reasons to assume that the marginal cost of investment in health is not constant.

2.5. Special cases

Equation (11) determines the optimal stock of health in any period other than the last period of life. A slightly different form of that equation emerges if both sides are divided by the marginal cost of gross investment:

$$\gamma_t + a_t = r - \tilde{\pi}_{t-1} + \delta_t. \tag{24}$$

Here $\gamma_t \equiv W_t G_t/\pi_{t-1}$ defines the marginal monetary return on an investment in health and $a_t \equiv [(Uh_t/\lambda)(1+r)^t G_t]/\pi_{t-1}$ defines the psychic rate of return.[22] In equilibrium, the total rate of return to an investment in health must equal the user cost of capital in terms of the price of gross investment. The latter variable is defined as the sum of the real-own rate of interest and the rate of depreciation and may be termed the opportunity cost of health capital.

In Sections 3 and 4, Equation (24) is used to study the responses of the stock of health, gross investment in health, and health inputs to variations in exogenous variables. Instead of doing this in the context of the general model developed so far, I deal with two special cases: a pure investment model and a pure consumption model. In the former model the psychic rate of return is zero, while in the latter the monetary rate of return is zero. There are two reasons for taking this approach. One involves an appeal

[22] The corresponding condition for the optimal stock in the last period of life, period n, is

$$\gamma_n + a_n = r + 1.$$

to simplicity. It is difficult to obtain sharp predictions concerning the effects of changes in exogenous variables in a mixed model in which the stock of health yields both investment and consumption benefits. The second is that most treatments of investments in knowledge or human capital other than health capital assume that monetary returns are large relative to psychic returns. Indeed, Lazear (1977) estimates that the psychic returns from attending school are negative. Clearly, it is unreasonable to assume that health is a source of disutility, and most discussions of investments in infant, child, and adolescent health [see Currie (2000)] stress the consumption benefits of these investments. Nevertheless, I stress the pure investment model because it generates powerful predictions from simple analyses and because the consumption aspects of the demand for health can be incorporated into empirical estimation without much loss in generality.

3. Pure investment model

If healthy time did not enter the utility function directly or if the marginal utility of healthy time were equal to zero, health would be solely an investment commodity. The optimal amount of H_t $(t < n)$ could then be found by equating the marginal monetary rate of return on an investment in health to the opportunity cost of capital:

$$\frac{W_t G_t}{\pi_{t-1}} \equiv \gamma_t = r - \tilde{\pi}_{t-1} + \delta_t. \tag{25}$$

Similarly, the optimal stock of health in the last period of life would be determined by

$$\frac{W_n G_n}{\pi_{n-1}} \equiv \gamma_n = r + 1. \tag{26}$$

Figure 1 illustrates the determination of the optimal stock of health capital at age t. The demand curve MEC shows the relationship between the stock of health and the rate of return on an investment or the marginal efficiency of health capital. The supply curve S shows the relationship between the stock of health and the cost of capital. Since the real-own rate of interest $(r - \tilde{\pi}_{t-1})$ and the rate of depreciation are independent of the stock, the supply curve is infinitely elastic. Provided the MEC schedule slopes downward, the equilibrium stock is given by H_t^*, where the supply and demand functions intersect.

The wage rate and the marginal cost of gross investment do not depend on the stock of health. Therefore, the MEC schedule would be negatively inclined if and only if the marginal product of health capital (G_t) diminishes as the stock increases. I have already assumed diminishing marginal productivity in Section 2 and have justified this assumption because the output produced by health capital has a finite upper limit of 8,760 hours in a year. Figure 2 shows a plausible relationship between the stock of health and the amount of healthy time. This relationship may be termed a production function

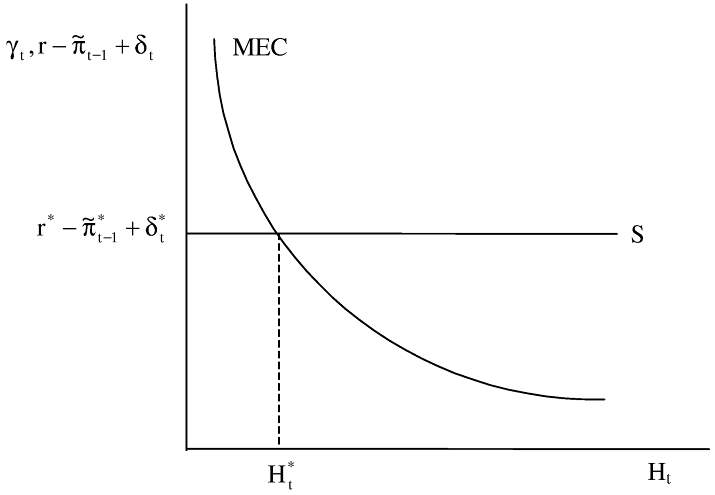

$\gamma_t, r - \tilde{\pi}_{t-1} + \delta_t$
MEC
$r^* - \tilde{\pi}^*_{t-1} + \delta^*_t$
S
H^*_t
H_t

Figure 1.

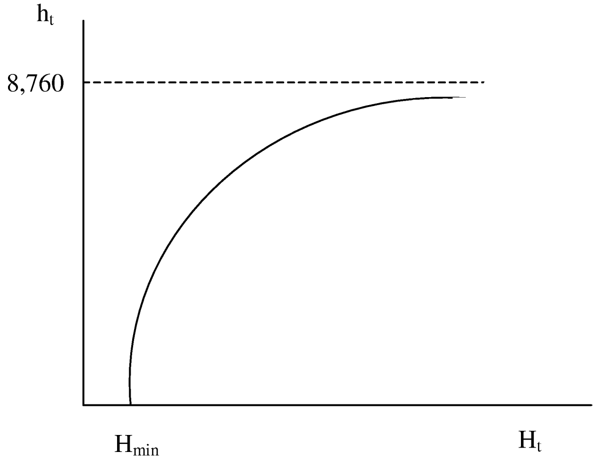


Figure 2.

of healthy time. The slope of the curve in the figure at any point gives the marginal product of health capital. The amount of healthy time equals zero at the death stock, H_{min}. Beyond that stock, healthy time increases at a decreasing rate and eventually approaches its upper asymptote as the stock becomes large.

Equations (25) and (26) and Figure 1 enable one to study the responses of the stock of health and gross investment to variations in exogenous variables. As indicated in

Section 2, two types of variations are examined: evolutionary (differences across time for the same consumer) and parametric (differences across consumers of the same age). In particular I consider evolutionary increases in the rate of depreciation on the stock of health with age and parametric variations in the rate of depreciation, the wage rate, and the stock of knowledge or human capital exclusive of health capital (E).

3.1. Depreciation rate effects

Consider the effect of an increase in the rate of depreciation on the stock of health (δ_t) with age. I have already shown in Section 2 that this factor causes the stock of health to fall with age and produces finite life. Graphically, the supply function in Figure 1 shifts upward over time or with age, and the optimal stock in each period is lower than in the previous period.

To quantify the magnitude of percentage rate of decrease in the stock of health over the life cycle, assume that the wage rate and the marginal cost of gross investment in health do not depend on age so that $\tilde{\pi}_{t-1} = 0$. Differentiate Equation (25) with respect to age to obtain[23]

$$\widetilde{H}_t = -s_t \varepsilon_t \tilde{\delta}_t. \tag{27}$$

In this equation, the tilde notation denotes a percentage time or age derivative

$$\widetilde{H}_t = \frac{dH_t}{dt}\frac{1}{H_t}, \text{ etc.,}$$

and the new symbols are $s_t = \delta_t/(r + \delta_t)$, the share of the depreciation rate in the cost of health capital; and

$$\varepsilon_t = -\frac{\partial \ln H_t}{\partial \ln(r + \delta_t)} = -\frac{\partial \ln H_t}{\partial \ln \gamma},$$

the elasticity of the MEC schedule.

Provided the rate of depreciation rises over the life cycle, the stock of health falls with age. The life cycle profile of gross investment does not, however, simply mirror that of health capital. The reason is that a rise in the rate of depreciation not only reduces the amount of health capital demanded by consumers but also reduces the amount of capital supplied to them by a given amount of gross investment. If the change in supply exceeds the change in demand, individuals have an incentive to close the gap by increasing gross

[23] As Ghez and Becker (1975) point out, none of the variables in a discrete time model are differentiable functions of time. Equation (27) and other equations involving time or age derivatives are approximations that hold exactly in a continuous time model.

investment. On the other hand, if the change in demand exceeds the change in supply, gross investment falls over the life cycle.

To begin to see why gross investment does not necessarily fall over the life cycle, first consider the behavior of one component of gross investment, total depreciation ($D_t = \delta_t H_t$), as the rate of depreciation rises over the life cycle. Assume that the percentage rate of increase in the rate of depreciation with age ($\tilde{\delta}_t$) and the elasticity of the MEC schedule (ε_t) are constant. Then

$$\widetilde{D}_t = \tilde{\delta}(1 - s_t\varepsilon) \gtreqless 0 \quad \text{as } \varepsilon \lesseqgtr \frac{1}{s_t}.$$

From the last equation, total depreciation increases with age as long as the elasticity of the MEC schedule is less than the reciprocal of the share of the depreciation rate in the cost of health capital. A sufficient condition for this to occur is that ε is smaller than one.

If ε_t and $\tilde{\delta}_t$ are constant, the percentage change in gross investment with age is given by

$$\widetilde{I}_t = \frac{\tilde{\delta}(1 - s_t\varepsilon)(\delta_t - s_t\varepsilon\tilde{\delta}) + s_t^2\varepsilon\tilde{\delta}^2}{(\delta_t - s_t\varepsilon\tilde{\delta})}. \tag{28}$$

Since health capital cannot be sold, gross investment cannot be negative. Therefore, $\delta_t \geqslant -\widetilde{H}_t$ or $\delta_t \geqslant -s_t\varepsilon\tilde{\delta}$. Provided gross investment is positive, the term $\delta_t - s_t\varepsilon\tilde{\delta}$ in the numerator and denominator of Equation (28) must be positive. Thus, a sufficient condition for gross investment to be positively correlated with the depreciation rate is $\varepsilon < 1/s_t$. Clearly, $\widetilde{I}_t$ is positive if $\varepsilon < 1$.

The important conclusion is reached that, if the elasticity of the MEC schedule is less than one, gross investment and the depreciation rate are positively correlated over the life cycle, while gross investment and the stock of health are negatively correlated. In fact, the relationship between the amount of healthy time and the stock of health suggests that ε is smaller than one. A general equation for the healthy time production function illustrated by Figure 2 is

$$h_t = 8{,}760 - BH_t^{-C}, \tag{29}$$

where B and C are positive constants. The corresponding MEC schedule is

$$\ln\gamma_t = \ln BC - (C + 1)\ln H_t + \ln W + \ln\pi. \tag{30}$$

The elasticity of this schedule is $\varepsilon = 1/(1 + C) < 1$ since $C > 0$.

Observe that with the depreciation rate held constant, increases in gross investment increase the stock of health and the amount of healthy time. But the preceding discussion indicates that because the depreciation rate rises with age it is likely that unhealthy (old) people will make larger gross investments in health than healthy (young) people. This means that sick time (TL_t) will be positively correlated with the market good or medical care input (M_t) and with the own time input (TH_t) in the gross investment production function over the life cycle.

The framework used to analyze life cycle variations in depreciation rates can easily be used to examine the impacts of variations in these rates among persons of the same age. Assume, for example, a uniform percentage shift in δ_t across persons so that the depreciation rate function can be written as $\delta_t = \delta_0 \exp(\tilde{\delta} t)$ where δ_0 differs among consumers. It is clear that such a shift has the same kind of effects as an increase in δ_t with age. That is, persons of a given age who face relatively high depreciation rates would simultaneously reduce their demand for health but increase their demand for gross investment if $\varepsilon < 1$.

3.2. Market and nonmarket efficiency

Persons who face the same cost of health capital would demand the same amount of health only if the determinants of the rate of return on an investment were held constant. Changes in the value of the marginal product of health capital and the marginal cost of gross investment shift the MEC schedule and, therefore, alter the quantity of health capital demanded even if the cost of capital does not change. The consumer's wage rate and his or her stock of knowledge or human capital other than health capital are the two key shifters of the MEC schedule.[24]

Since the value of the marginal product of health capital equals WG, an increase in the wage rate (W) raises the monetary equivalent of the marginal product of a given stock. Put differently, the higher a person's wage rate the greater is the value to him of an increase in healthy time. A consumer's wage rate measures his market efficiency or the rate at which he can convert hours of work into money earnings. Hence, the wage is positively correlated with the benefits of a reduction in lost time from the production of money earnings due to illness. Moreover, a high wage induces an individual to substitute market goods for his own time in the production of commodities. This substitution continues until in equilibrium the monetary value of the marginal product of consumption time equals the wage rate. Thus, the benefits from a reduction in time lost from nonmarket production are also positively correlated with the wage.

If an upward shift in the wage rate had no effect on the marginal cost of gross investment, a 1 percent increase in the wage would increase the rate of return (γ) associated with a fixed stock of capital by 1 percent. In fact this is not the case because own time is an input in the gross investment production function. If K is the fraction of the total

[24] In this section I deal with uniform shifts in variables that influence the rate of return across persons of the same age. I also assume that these variables do not vary over the life cycle. Finally, I deal with the *partial* effect of an increase in the wage rate or an increase in knowledge capital, measured by the number of years of formal schooling completed, on the demand for health and health inputs. For a more general treatment, see Grossman (1972b, pp. 28–30).

cost of gross investment accounted for by time, a 1 percent rise in W would increase marginal cost (π) by K percent. After one nets out the correlation between W and π, the percentage growth in γ would equal $1 - K$, which exceeds zero as long as gross investment is not produced entirely by time. Hence, the quantity of health capital demanded rises as the wage rate rises as shown in the formula for the wage elasticity of capital:

$$e_{HW} = (1 - K)\varepsilon. \tag{31}$$

Although the wage rate and the demand for health or gross investment are positively related, W has no effect on the amount of gross investment supplied by a given input of medical care. Therefore, the demand for medical care rises with the wage. If medical care and own time are employed in fixed proportions in the gross investment production function, the wage elasticity of M equals the wage elasticity of H. On the other hand, given a positive elasticity of substitution in production (σ_p) between M and TH, M increases more rapidly than H because consumers have an incentive to substitute medical care for their relatively more expensive own time. This substitution is reflected in the formula for the wage elasticity of medical care:

$$e_{MW} = K\sigma_p + (1 - K)\varepsilon. \tag{32}$$

The preceding analysis can be modified to accommodate situations in which the money price of medical care is zero for all practical purposes because it is fully financed by health insurance or by the government, and care is rationed by waiting and travel time. Suppose that q hours are required to obtain one unit of medical care, so that the price of care is Wq. In addition, suppose that there are three endogenous inputs in the gross investment production function: M, TH, and a market good (X) whose acquisition does not require time. Interpret K as the share of the cost of gross investment accounted for by M and TH. Then Equation (31) still holds, and an increase in W causes H to increase. Equation (32) becomes

$$e_{MW} = (1 - K)(\varepsilon - \sigma_{MX}), \tag{33}$$

where σ_{MX} is the partial elasticity of substitution in production between M and the third input, X. If these two inputs are net substitutes in production, σ_{MX} is positive. Then

$$e_{MW} \gtreqless 0 \quad \text{as } \varepsilon \gtreqless \sigma_{MX}.$$

In this modified model the wage elasticity of medical care could be negative or zero. This case is relevant in interpreting some of the empirical evidence to be discussed later.

As indicated in Section 2, I follow Michael (1972, 1973) and Michael and Becker (1973) by assuming that an increase in knowledge capital or human capital other than health capital (E) raises the efficiency of the production process in the nonmarket or household sector, just as an increase in technology raises the efficiency of the production process in the market sector. I focus on education or years of formal schooling completed as the most important determinant of the stock of human capital. The gross investment production function and the production function of the commodity Z are linear homogeneous in their endogenous inputs [see Equations (3) and (4)]. Therefore, an increase in the exogenous or predetermined stock of human capital can raise output only if it raises the marginal products of the endogenous inputs.

Suppose that a one unit increase in E raises the marginal products of M and TH in the gross investment production function by the same percentage (ρ_H). This is the Hicks- or factor-neutrality assumption applied to an increase in technology in the nonmarket sector. Given factor-neutrality, there is no incentive to substitute medical care for own time as the stock of human capital rises.

Because an increase in E raises the marginal products of the health inputs, it reduces the quantity of these inputs required to produce a given amount of gross investment. Hence, with no change in input prices, the marginal or average cost of gross investment falls. In fact, if a circumflex over a variable denotes a percentage change per unit change in E, one easily shows

$$\hat{\pi} = -\rho_H. \tag{34}$$

With the wage rate held constant, an increase in E would raise the marginal efficiency of a given stock of health. This causes the MEC schedule in Figure 1 to shift upward and raises the optimal stock of health.

The percentage increase in the amount of health capital demanded for a one unit increase in E is given by

$$\widehat{H} = \rho_H \varepsilon. \tag{35}$$

Since ρ_H indicates the percentage increase in gross investment supplied by a one unit increase in E, shifts in this variable would not alter the demand for medical care or own time if ρ_H equaled $\widehat{H}$. For example, a person with 10 years of formal schooling might demand 3 percent more health than a person with 9 years of formal schooling. If the medical care and own time inputs were held constant, the former individual's one extra year of schooling might supply him with 3 percent more health. Given this condition, both persons would demand the same amounts of M and TH. As this example illustrates, any effect of a change in E on the demand for medical care or time reflects a positive or negative differential between $\widehat{H}$ and ρ_H:

$$\widehat{M} = \widehat{TH} = \rho_H(\varepsilon - 1). \tag{36}$$

Equation (36) suggests that the more educated would demand more health but less medical care if the elasticity of the MEC schedule were less than one. These patterns are opposite to those that would be expected in comparing the health and medical care utilization of older and younger consumers.

4. Pure consumption model

If the cost of health capital were large relative to the monetary rate of return on an investment in health and if $\tilde{\pi}_{t-1} = 0$, all t, then Equation (11) or (24) could be approximated by

$$\frac{Uh_t G_t}{\lambda} = \frac{UH_t}{\lambda} = \frac{\pi(r+\delta_t)}{(1+r)^t}. \tag{37}$$

Equation (37) indicates that the monetary equivalent of the marginal utility of health capital must equal the discounted user cost of H_t.[25] It can be used to highlight the differences between the age, wage, or schooling effect in a pure consumption model and the corresponding effect in a pure investment model. In the following analysis, I assume that the marginal rate of substitution between H_t and H_{t+1} depends only on H_t and H_{t+1} and that the marginal rate of substitution between H_t and Z_t depends only on H_t and Z_t. I also assume that one plus the market rate of interest is equal to one plus rate of time preference for the present (the ratio of the marginal utility of H_t to the marginal utility of H_{t+1} when these two stocks are equal minus one). Some of these assumptions are relaxed, and a more detailed analysis is presented in Grossman (1972b, Chapter III).

With regard to age-related depreciation rate effects, the elasticity of substitution in consumption between H_t and H_{t+1} replaces the elasticity of the MEC schedule in Equations (27) and (28). The quantity of health capital demanded still falls over the life cycle in response to an increase in the rate of depreciation. Gross investment and health inputs rise with age if the elasticity of substitution between present and future health is less than one.

Since health enters the utility function, health is positively related to wealth in the consumption model provided it is a superior good. That is, an increase in wealth with no change in the wage rate or the marginal cost of gross investment causes the quantity of health capital demanded to rise. This effect is absent from the investment model because

[25] Solving Equation (37) for the monetary equivalent of the marginal utility of healthy time, one obtains

$$\frac{Uh_t}{\lambda} = \frac{\pi(r+\delta_t)/G_t}{(1+r)^t}.$$

Given diminishing marginal productivity of health capital, the undiscounted price of a healthy hour, $\pi(r+\delta_t)/G_t$, would be positively correlated with H or h even if π were constant. Therefore, the consumption demand curve would be influenced by scale effects. To emphasize the main issues at stake in the consumption model, I ignore these effects essentially by assuming that G_t is constant. The analysis would not be greatly altered if they were introduced.

the marginal efficiency of health capital and the market rate of interest do not depend on wealth.[26] Parametric wage variations across persons of the same age induce wealth effects on the demand for health. Suppose that we abstract from these effects by holding the level of utility or real wealth constant. Then the wage elasticity of health is given by

$$e_{HW} = -(1-\theta)(K - K_Z)\sigma_{HZ}, \tag{38}$$

where θ is the share of health in wealth, K_Z is the share of total cost of Z accounted for by time, and σ_{HZ} is the positive elasticity of substitution in consumption between H and Z.[27] Hence,

$$e_{HW} \lesseqgtr 0 \quad \text{as } K \gtreqless K_Z.$$

The sign of the wage elasticity is ambiguous because an increase in the wage rate raises the marginal cost of gross investment in health and the marginal cost of Z. If time costs were relatively more important in the production of health than in the production of Z, the relative price of health would rise with the wage rate, which would reduce the quantity of health demanded. The reverse would occur if Z were more time intensive than health. The ambiguity of the wage effect here is in sharp contrast to the situation in the investment model. In that model, the wage rate would be positively correlated with health as long as K were less than one.

Instead of examining a wage effect that holds utility constant, Wagstaff (1986) and Zweifel and Breyer (1997) examine a wage effect that holds the marginal utility of wealth constant. This analysis is feasible only if the current period utility function, $\Psi(H, Z)$, is strictly concave:

$$\Psi_{HH} < 0,\ \Psi_{ZZ} < 0,\ \Psi_{HH}\Psi_{ZZ} - \Psi_{HZ}^2 > 0.$$

With the marginal utility of wealth (λ) held constant, the actual change in health caused by a one percent increase in the wage rate is given by

$$\frac{\partial H}{\partial \ln W} = \frac{K\Psi_H\Psi_{ZZ} - K_Z\Psi_Z\Psi_{HZ}}{\Psi_{HH}\Psi_{ZZ} - \Psi_{HZ}^2}. \tag{39}$$

[26] For a different conclusion, see Ehrlich and Chuma (1990). They argue that wealth effects are present in the investment model given rising marginal cost of investment in health and endogenous length of life.

[27] I assume that the elasticity of substitution in consumption between H_t and Z_t is the same in every period and that the elasticity of substitution between H_t and Z_j $(t \neq j)$ does not depend on t or j. The corresponding equation for the wage elasticity of medical care is

$$e_{MW} = K\sigma_p - (1-\theta)(K - K_Z)\sigma_{HZ}.$$

Equation (39) is negative if $\Psi_{HZ} \geqslant 0$. The sign of the wage effect is, however, ambiguous if $\Psi_{HZ} < 0$.[28]

The human capital parameter in the consumption demand function for health is

$$\widehat{H} = \rho\eta_H + (\rho_H - \rho_Z)(1 - \theta)\sigma_{HZ}, \tag{40}$$

where ρ_Z is the percentage increase in the marginal product of the Z commodity's goods or time input caused by a one unit increase in E (the negative of the percentage reduction in the marginal or average cost of Z), η_H is the wealth elasticity of demand for health, and $\rho = \theta\rho_H + (1-\theta)\rho_Z$ is the percentage increase in real wealth as E rises with money full wealth and the wage rate held constant. The first term on the right-hand side of Equation (40) reflects the wealth effect and the second term reflects the substitution effect. If E's productivity effect in the gross investment production function is the same as in the Z production function, then $\rho_H = \rho_Z$ and $\widehat{H}$ reflects the wealth effect alone. In this case, a shift in human capital, measured by years of formal schooling completed or education is "commodity-neutral," to use the term coined by Michael (1972, 1973). If $\rho_H > \rho_Z$, E is "biased" toward health, its relative price falls, and the wealth and substitution effects both operate in the same direction. Consequently, an increase in E definitely increases the demand for health. If $\rho_H < \rho_Z$, E is biased away from health, its relative price rises, and the wealth and substitution effects operate in opposite directions.

The human capital parameter in the consumption demand curve for medical care is

$$\widehat{M} = \rho(\eta_H - 1) + (\rho_H - \rho_Z)\big[(1 - \theta)\sigma_{HZ} - 1\big]. \tag{41}$$

If shifts in E are commodity-neutral, medical care and education are negatively correlated unless $\eta_H \geqslant 1$. If on the other hand, there is a bias in favor of health, these two variables will still tend to be negatively correlated unless the wealth and price elasticities both exceed one.[29]

The preceding discussion reveals that the analysis of variations in nonmarket productivity in the consumption model differs in two important respects from the corresponding analysis in the investment model. In the first place, wealth effects are not relevant in the investment model, as has already been indicated. Of course, health would have a positive wealth elasticity in the investment model if wealthier people faced lower rates of interest. But the analysis of shifts in education assumes money wealth is fixed. Thus, one could not rationalize the positive relationship between education and health in terms of an association between wealth and the interest rate.

[28] I assume that $\Psi_Z\Psi_{HZ} > \Psi_H\Psi_{ZZ}$ so that the pure wealth effect is positive and a reduction in λ raises health. When Ψ_{HZ} is negative, this condition does not guarantee that Equation (39) is negative because K_Z could exceed K.

[29] The term $(1 - \theta)\sigma_{HZ}$ in Equation (40) or Equation (41) is the compensated or utility-constant price elasticity of demand for health.

In the second place, if the investment framework were utilized, then whether or not a shift in human capital is commodity-neutral would be irrelevant in assessing its impact on the demand for health. As long as the rate of interest were independent of education, H and E would be positively correlated.[30] Put differently, if individuals could always receive, say, a 5 percent real rate of return on savings deposited in a savings account, then a shift in education would create a gap between the cost of capital and the marginal efficiency of a given stock.

Muurinen (1982) and Van Doorslaer (1987) assume that an increase in education lowers the rate of depreciation on the stock of health rather than raising productivity in the gross investment production function. This is a less general assumption than the one that I have made since it rules out schooling effects in the production of nondurable household commodities. In the pure investment model, predictions are very similar whether schooling raises productivity or lowers the rate of depreciation.[31] In the pure consumption model the assumption made by Muurinen and Van Doorslaer is difficult to distinguish from the alternative assumption that ρ_Z is zero. Interactions between schooling and the lagged stock of health in the demand function for current health arise given costs of adjustment in the Muurinen–Van Doorslaer model. These are discussed in Section 6.

5. Empirical testing

In Grossman (1972b, Chapter IV), I present an empirical formulation of the pure investment model, including a detailed outline of the structure and reduced form of that model. I stress the estimation of the investment model rather than the consumption model because the former model generates powerful predictions from simple analysis and more innocuous assumptions. For example, if one uses the investment model, he or she does not have to know whether health is relatively time-intensive to predict the effect of an increase in the wage rate on the demand for health. Also, he or she does not have to know whether education is commodity-neutral to assess the sign of the correlation between health and schooling. Moreover, the responsiveness of the quantity of health demanded to changes in its shadow price and the behavior of gross investment depend essentially on a single parameter – the elasticity of the MEC schedule. In the consumption model, on the other hand, three parameters are relevant – the elasticity of substitution in consumption between present and future health, the wealth elasticity of demand for health, and the elasticity of substitution in consumption between health and the Z commodity. Finally, while good health may be a source of utility, it clearly is a

[30] I relax the assumption that all persons face the same market rate of interest in Section 7.

[31] If an increase in schooling lowers the rate of depreciation at every age by the same percentage, it is equivalent to a uniform percentage shift in this rate considered, a matter briefly in Section 3.3 and in more detail in Grossman (1972b, pp. 19, 90). The only difference between this model and the productivity model is that a value of the elasticity of the MEC schedule smaller than one is a sufficient, but not a necessary, condition for medical care to fall as schooling rises.

source of earnings. The following formulation is oriented toward the investment model, yet I also offer two tests to distinguish the investment model from the consumption model.

5.1. Structure and reduced form

With the production function of healthy time given by Equation (29), I make use of three basic structural equations (intercepts are suppressed):

$$\ln H_t = \varepsilon \ln W_t - \varepsilon \ln \pi_t - \varepsilon \ln \delta_t, \tag{42}$$

$$\ln \delta_t = \ln \delta_0 + \tilde{\delta} t, \tag{43}$$

$$\ln I_t \equiv \ln H_t + \ln\bigl(1 + \widetilde{H}_t/\delta_t\bigr) = \rho_H E + (1-K)\ln M_t + K \ln TH_t. \tag{44}$$

Equation (42) is the demand function for the stock of health and is obtained by solving Equation (30) for $\ln H_t$. The equation contains the assumption that the real-own rate of interest is equal to zero. Equation (43) is the depreciation rate function. Equation (44) contains the identity that gross investment equals net investment plus depreciation and assumes that the gross investment production function is a member of the Cobb–Douglas class.

These three equations and the least-cost equilibrium condition that the ratio of the marginal product of medical care to the marginal product of time must equal the ratio of the price of medical care to the wage rate generate the following reduced form demand curves for health and medical care:

$$\ln H_t = (1-K)\varepsilon \ln W_t - (1-K)\varepsilon \ln P_t + \rho_H \varepsilon E - \tilde{\delta}\varepsilon t - \varepsilon \ln \delta_0, \tag{45}$$

$$\begin{aligned} \ln M_t = {} & \bigl[(1-K)\varepsilon + K\bigr]\ln W_t - \bigl[(1-K)\varepsilon + K\bigr]\ln P_t + \rho_H(\varepsilon - 1)E \\ & + \tilde{\delta}(1-\varepsilon)t + (1-\varepsilon)\ln \delta_0 + \ln\bigl(1 + \widetilde{H}_t/\delta_t\bigr). \end{aligned} \tag{46}$$

If the absolute value of the rate of net disinvestment ($\widetilde{H}_t$) were small relative to the rate of depreciation, the last term on the right-hand side of Equation (46) could be ignored.[32] Then Equations (45) and (46) would express the two main endogenous variables in the system as functions of four variables that are treated as exogenous within the context of this model – the wage rate, the price of medical care, the stock of human capital, and

[32] Recall that $\delta_t > -\widetilde{H}_t$ since gross investment must be positive. Given that the real rate of interest is zero

$$\frac{-\widetilde{H}_t}{\delta_t} = \varepsilon \delta_t^2 \frac{d\delta_t}{dt}.$$

Since ε is likely to be smaller than one and the square of the rate of depreciation is small for modest rates of depreciation, δ_t is likely to be much larger than $-\widetilde{H}_t$.

age – and one variable that is unobserved – the rate of depreciation in the initial period. With age subscripts suppressed the estimating equations become

$$\ln H = B_w \ln W + B_P \ln P + B_E E + B_t t + u_1, \tag{47}$$

$$\ln M = B_{WM} \ln W + B_{PM} \ln P + B_{EM} E + B_{tM} t + u_2, \tag{48}$$

where $B_W = (1 - K)\varepsilon$, et cetera, $u_1 = -\varepsilon \ln \delta_0$ and $u_2 = (1 - \varepsilon) \ln \delta_0$. The investment model predicts $B_W > 0$, $B_P < 0$, $B_E > 0$, $B_t < 0$, $B_{WM} > 0$, and $B_{PM} < 0$. In addition, if $\varepsilon < 1$, $B_{EM} < 0$ and $B_{tM} > 0$.

The variables u_1 and u_2 represent disturbance terms in the reduced form equations. These terms are present because depreciation rates vary among people of the same age, and such variations cannot be measured empirically. Provided $\ln \delta_0$ were not correlated with the independent variables in (47) and (48), u_1 and u_2 would not be correlated with these variables. Therefore, the equations could be estimated by ordinary least squares.

The assumption that the real-own rate of interest equals zero can be justified by noting that wage rates rise with age, at least during most stages of the life cycle. If the wage is growing at a constant percentage rate of $\widetilde{W}$, then $\tilde{\pi}_t = K\widetilde{W}$, all t. So the assumption implies $r = K\widetilde{W}$. By eliminating the real rate of interest and postulating that $-\widetilde{H}_t$ is small relative to δ_t, $\ln H$ and $\ln M$ are made linear functions of age. If these assumptions are dropped, the age effect becomes nonlinear.

Since the gross investment production function is a member of the Cobb–Douglas class, the elasticity of substitution in production between medical care and own time (σ_p) is equal to one, and the share of medical care in the total cost of gross investment or the elasticity of gross investment with respect to medical care, $(1 - K)$, is constant. If σ_p were not equal to one, the term K in the wage and price elasticities of demand for medical care would be multiplied by this value rather than by one. The wage and price parameters would not be constant if σ_p were constant but not equal to one, because K would depend on W and P. The linear age, price, and wage effects in Equations (47) and (48) are first-order approximations to the true effects.

I have indicated that years of formal schooling completed is the most important determinant of the stock of human capital and employ schooling as a proxy for this stock in the empirical analysis described in Section 5.2. In reality the amount of human capital acquired by attending school also depends on such variables as the mental ability of the student and the quality of the school that he or she attends. If these omitted variables are positively correlated with schooling and uncorrelated with the other regressors in the demand function for health, the schooling coefficient is biased upwards. These biases are more difficult to sign if, for example, mental ability and school quality are correlated with the wage rate.[33]

There are two empirical procedures for assessing whether the investment model gives a more adequate representation of people's behavior than the consumption model. In

[33] See Section 7 for a detailed analysis of biases in the regression coefficient of schooling due to the omission of other variables.

the first place, the wage would have a positive effect on the demand for health in the investment model as long as K were less than one. On the other hand, it would have a positive effect in the consumption model only if health were relatively goods-intensive $(K < K_Z)$. So, if the computed wage elasticity turns out to be positive, then the larger its value the more likely it is that the investment model is preferable to the consumption model. Of course, provided the production of health were relatively time intensive, the wage elasticity would be negative in the consumption model. In this case, a positive and statistically significant estimate of B_W would lead to a rejection of the consumption model.

In the second place, health has a zero wealth elasticity in the investment model but a positive wealth elasticity in the consumption model provided it is a superior good. This suggests that wealth should be added to the set of regressors in the demand functions for health and medical care. Computed wealth elasticities that do not differ significantly from zero would tend to support the investment model.[34]

In addition to estimating demand functions for health and medical care, one could also fit the gross investment function given by Equation (44). This would facilitate a direct test of the hypothesis that the more educated are more efficient producers of health. The production function contains two unobserved variables: gross investment and the own time input. Since, however, $-\widetilde{H}_t$ has been assumed to be small relative to δ_t, one could fit[35]

$$\ln H = \alpha \ln M + \rho_H E - \tilde{\delta} t - \ln \delta_0. \tag{49}$$

The difficulty with the above procedure is that it requires a good estimate of the production function. Unfortunately, Equation (49) cannot be fitted by ordinary least squares (OLS) because $\ln M$ and $\ln \delta_0$, the disturbance term, are bound to be correlated. From the demand function for medical care

$$\text{Covariance}(\ln M, \ln \delta_0) = (1 - \varepsilon)\text{Variance}(\ln \delta_0).$$

Given $\varepsilon < 1$, $\ln M$ and $\ln \delta_0$ would be positively correlated. Since an increase in the rate of depreciation lowers the quantity of health capital, the coefficient of medical care

[34] Health would have a positive wealth elasticity in the investment model for people who are not in the labor force. For such individuals, an increase in wealth would raise the ratio of market goods to consumption time, the marginal product of consumption time, and its shadow price. Hence, the monetary rate of return on an investment in health would rise. Since my empirical work is limited to members of the labor force, a pure increase in wealth would not change the shadow price of their time.

[35] In my monograph, I argue that a one percent increase in medical care would be accompanied by a one percent increase in own time if factor prices do not vary as more and more health is produced. Therefore, the regression coefficient α in Equation (49) would reflect the sum of the output elasticities of medical care and own time. Given constant returns to scale, the true value of α should be unity [Grossman (1972b, p. 43)]. This analysis becomes much more complicated once joint production is introduced [see Grossman (1972b), Chapter VI and the brief discussion of joint production below].

would be biased *downward*. The same bias exists if there are unmeasured determinants of efficiency in the production of gross investments in health.

The biases inherent in ordinary least squares estimates of health production functions were first emphasized by Auster et al. (1969). They have been considered in much more detail in the context of infant health by Rosenzweig and Schultz (1983, 1988, 1991), Corman et al. (1987), Grossman and Joyce (1990), and Joyce (1994). Consistent estimates of the production function can be obtained by two-stage least squares (TSLS). In the present context, wealth, the wage rates, and the price of medical care serve as instruments for medical care. The usefulness of this procedure rests, however, on the validity of the overidentification restrictions and the degree to which the instruments explain a significant percentage of the variation in medical care [Bound et al. (1995), Staiger and Stock (1997)]. The TSLS technique is especially problematic when the partial effects of several health inputs are desired and when measures of some of these inputs are absent. In this situation the overidentification restrictions may not hold because wealth and input prices are likely to be correlated with the missing inputs.

In my monograph on the demand for health, I argued that "a production function taken by itself tells nothing about producer or consumer behavior, although it does have implications for behavior, which operate on the demand curves for health and medical care. Thus, they serve to rationalize the forces at work in the reduced form and give the variables that enter the equations economic significance. Because the reduced form parameters can be used to explain consumer choices and because they can be obtained by conventional statistical techniques, their interpretation should be pushed as far as possible. Only then should one resort to a direct estimate of the production function" [Grossman (1972b, p. 44)]. The reader should keep this position in mind in evaluating my discussion of the criticism of my model raised by Zweifel and Breyer (1997) in Section 6.

5.2. Data and results

I fitted the equations formulated in Section 5.1 to a nationally representative 1963 United States survey conducted by the National Opinion Research Center and the Center for Health Administration Studies of the University of Chicago. I measured the stock of health by individuals' self-evaluation of their health status. I measured healthy time, the output produced by health capital, either by the complement of the number of restricted-activity days due to illness or injury or the number of work-loss days due to illness or injury. I measured medical care by personal medical expenditures on doctors, dentists, hospital care, prescribed and nonprescribed drugs, nonmedical practitioners, and medical appliances. I had no data on the actual quantities of specific types of services, for example the number of physician visits. Similarly, I had no data on the prices of these services. Thus, I was forced to assume that the price of medical care (P) in the reduced form demand functions either does not vary among consumers or is not correlated with the other regressors in the demand functions. Neither assumption is likely to be correct in light of the well known moral hazard effect of private health insurance.[36] The main independent variables in the regressions were the age of the individual, the number of years of formal schooling he or she completed, his or her weekly wage rate, and family income (a proxy for wealth).

The most important regression results in the demand functions are as follows. Education and the wage rate have positive and statistically significant coefficients in the health demand function, regardless of the particular measure of health employed. An increase in age simultaneously reduces health and increases medical expenditures. Both effects are significant. The signs of the age, wage, and schooling coefficients in the health demand function and the sign of the age coefficient in the medical care demand function are consistent with the predictions contained in the pure investment model.

In the demand function for medical care the wage coefficient is negative but not significant, while the schooling coefficient is positive but not significant. The sign of the wage coefficient is not consistent with the pure investment model, and the sign of the schooling coefficient is not consistent with the version of the investment model in which the elasticity of the MEC schedule is less than one. In Grossman (1972b, Appendix D), I show that random measurement error in the wage rate and a positive correlation between the wage and unmeasured determinants of nonmarket efficiency create biases that may explain these results. Other explanations are possible. For example, the wage elasticity of medical care is not necessarily positive in the investment model if waiting and travel time are required to obtain this care [see Equation (33)]. Schooling is likely to be positively correlated with the generosity of health insurance coverage leading to an upward bias in its estimated effect.

When the production function is estimated by ordinary least squares, the elasticities of the three measures of health with respect to medical care are all *negative*. Presumably, this reflects the strong positive relation between medical care and the depreciation rate. Estimation of the production function by two-stage least squares reverses the sign of the medical care elasticity in most cases. The results, however, are sensitive to whether or not family income is included in the production function as a proxy for missing inputs.

The most surprising finding is that healthy time has a negative family income elasticity. If the consumption aspects of health were at all relevant, a literal interpretation of this result is that health is an inferior commodity. That explanation is, however, not consistent with the positive and significant income elasticity of demand for medical care. I offer an alternative explanation based on joint production. Such health inputs as cigarettes, alcohol, and rich food have negative marginal products. If their income elasticities exceeded the income elasticities of the beneficial health inputs, the marginal cost of gross investment in health would be positively correlated with income. This explanation can account for the positive income elasticity of demand for medical care. Given

[36] See Zweifel and Manning (2000) for a review of the literature dealing with the effect of health insurance on the demand for medical care.

its assumptions, higher income persons simultaneously reduce their demand for health and increase their demand for medical care if the elasticity of the MEC schedule is less than one.

I emphasized in Section 2 that parametric changes in variables that increase healthy time also prolong length of life. Therefore, I also examine variations in age-adjusted mortality rates across states of the United States in 1960. I find a close agreement between mortality and sick time regression coefficients. Increases in schooling or the wage rate lower mortality, while increases in family income raise it.

6. Extensions

In this section I deal with criticisms and empirical and theoretical extensions of my framework. I begin with empirical testing with cross-sectional data by Wagstaff (1986), Erbsland et al. (1995), and Stratmann (1999) in Section 6.1. I pay particular attention to Wagstaff's study because it serves as the basis of a criticism of my approach by Zweifel and Breyer (1997), which I also address in Section 6.1. I turn to empirical extensions with longitudinal data by Van Doorslaer (1987) and Wagstaff (1993) in Section 6.2. These studies introduce costs of adjustment, although in a rather ad hoc manner. I consider theoretical developments by Cropper (1977), Muurinen (1982), Dardanoni and Wagstaff (1987, 1990), Selden (1993), Chang (1996), and Liljas (1998) in Section 6.3. With the exception of Muurinen's work, these developments all pertain to uncertainty.

6.1. Empirical extensions with cross-sectional data

Wagstaff (1986) uses the 1976 Danish Welfare Survey to estimate a multiple indicator version of the structure and reduced form of my demand for health model. He performs a principal components analysis of nineteen measures of non-chronic health problems to obtain four health indicators that reflect physical mobility, mental health, respiratory health, and presence of pain. He then uses these four variables as indicators of the unobserved stock of health. His estimation technique is the so-called MIMIC (multiple indicators-multiple causes) model developed by Jöreskog (1973) and Goldberger (1974) and employs the maximum likelihood procedure contained in Jöreskog and Sörbom (1981). His contribution is unique because it accounts for the multidimensional nature of good health both at the conceptual level and at the empirical level.

Aside from the MIMIC methodology, there are two principal differences between my work and Wagstaff's work. First, the structural equation that I obtain is the production function. On the other hand, the structural equation that he obtains is a conditional output demand function. This expresses the quantity demanded of a health input, such as medical care, as a function of health output, input prices, and exogenous variables in the production function such as schooling and age. In the context of the structure that I specified in Section 5.1, the conditional output demand function is obtained by solving Equation (44) for medical care as a function of health, the own time input, schooling, age, and the rate of depreciation in the initial period and then using the cost-minimization condition to replace the own time input with the wage rate and the price of medical care. Since an increase in the quantity of health demanded increases the demand for health inputs, the coefficient of health in the conditional demand function is positive.

Second, Wagstaff utilizes a Frisch (1964) demand function for health in discussing and attempting to estimate the pure consumption model. This is a demand function in which the marginal utility of lifetime wealth is held constant when the effects of variables that alter the marginal cost of investment in health are evaluated. I utilized it briefly in treating wage effects in the pure consumption model in Section 4 but did not stress it either theoretically or empirically. The marginal utility of lifetime wealth is not observed but can be replaced by initial assets and the sum of lifetime wage rates. Since the data are cross-sectional, initial assets and wage rates over the life cycle are not observed. Wagstaff predicts the missing measures by regressing current assets and the current wage on age, the square of age, and age-invariant socioeconomic characteristics.

Three health inputs are contained in the data: the number of physician visits during the eight months prior to the survey, the number of weeks spent in a hospital during the same period, and the number of complaints for which physician-prescribed or self-prescribed medicines were being taken at the time of the interview. To keep my discussion of the results manageable, I will focus on the reduced form and conditional demand functions for physician visits and on the demand function for health. The reader should keep in mind that the latent variable health obtained from the MIMIC procedure is a *positive* correlate of good health. Good health is the dependent variable in the reduced form demand function for health and one of the right-hand side variables in the conditional demand function for physician visits.

Wagstaff estimates his model with and without initial assets and the sum of lifetime wage rates. He terms the former a pure investment model and the latter a pure consumption model. Before discussing the results, one conceptual issue should be noted. Wagstaff indicates that medical inputs in Denmark are heavily subsidized and that almost all of the total cost of gross investment is accounted for by the cost of the own time input. He then argues that the wage coefficient should equal zero in the pure investment demand function since K, the share of the total cost of gross investment accounted for by time, is equal to one. He also argues that the coefficient of the wage in the demand function for medical care should equal one [see the relevant coefficients in Equations (45) and (46)].

Neither of the preceding propositions is necessarily correct. In Section 3 I developed a model in which the price of medical care is zero, but travel and waiting time (q hours to make one physician visit) are required to obtain medical care as well as to produce health. I also assumed three endogenous inputs in the health production function: M, TH, and a market good whose acquisition does not require time. I then showed that the wage elasticity of health is positive, while the wage elasticity of medical care is indeterminate in sign [see Equations (32) and (33), both of which hold q constant]. If there are only two inputs and no time required to obtain medical care, the wage elasticities of health and medical care are zero. The latter elasticity is zero because the marginal product of medical care would be driven to zero if its price is zero for a given wage rate. An increase in the wage rate induces no further substitution in production. With travel and waiting time, the marginal product of care is positive, but the price of medical care relative to the price of the own time input (Wq/W) does not depend on W.

An additional complication is that Wagstaff includes a proxy for Wq – the respondent's wage multiplied by the time required to travel to his or her physician – in the demand function for medical care. He asserts that the coefficient of the logarithm of this variable should equal the coefficient of the logarithm of the wage in the demand function for medical care in a model in which the price of medical care is not zero. This is not correct because the logarithm of W is held constant. Hence increases in Wq are due solely to increases in q. As q rises, H falls and M falls because the price of M relative to the price of $TH[(P + Wq)/W]$ rises.[37]

In Wagstaff's estimate of the reduced form of the pure investment model, the wage rate, years of formal schooling completed, and age all have the correct signs and all three variables are significant in the demand function for health. In the demand function for physician visits the schooling variable has a negative and significant coefficient. This finding differs from mine and is in accord with the predictions of the pure investment model. The age coefficient is positive and significant at the 10 percent level on a one-tailed test but not at the 5 percent level.[38] The wage coefficient, however, is negative and not significant. The last finding is consistent with the three-input model outlined above and is not necessarily evidence against the investment model.

The time cost variable has the correct negative sign in the demand function for physician visits, but it is not significant. Wagstaff, however, includes the number of physicians per capita in the respondent's county of residence in the same equation. This variable has a positive effect on visits, is likely to be negatively related to travel time, and may capture part of the travel time effect.

Wagstaff concludes the discussion of the results of estimating the reduced form of the investment model as follows: "Broadly speaking ... the coefficients are similar to those reported by Grossman and are consistent with the model's structural parameters being of the expected sign. One would seem justified, therefore, in using the ... data for exploring the implications of using structural equation methods in this context" (p. 214). When this is done, the coefficient of good health in the conditional demand function for physician visits has the wrong sign. It is negative and very significant.

This last finding is used by Zweifel and Breyer (1997) to dismiss my model of the demand for health. They write: "Unfortunately, empirical evidence consistently fails to

[37] The complete specification involves regressing H on W and q and regressing M on W and q, where it is understood that all variables are in logarithms. In the three input model in which the money price of medical care is zero, the coefficients of q are negative in both equations. The coefficient of W is positive in the health equation and ambiguous in sign in the medical care equation.

[38] A one-tailed test is appropriate since the alternative hypothesis is that the age coefficient is positive. The age coefficient is positive and significant at all conventional levels in the demand functions for hospital stays and medicines.

confirm this crucial prediction [that the partial correlation between good health and medical care should be positive]. When health status is introduced as a latent variable through the use of simultaneous indicators, all components of medical care distinguished exhibit a very definite and highly significant *negative* (their italics) partial relationship with health .... The notion that expenditure on medical care constitutes a demand derived from an underlying demand for health cannot be upheld because health status and demand for medical care are negatively rather than positively related" (pp. 60, 62).

Note, however, that biases arise if the conditional demand function is estimated with health treated as exogenous for the same reason that biases arise if the production function is estimated by ordinary least squares. In particular, the depreciation rate in the initial period (the disturbance term in the equation) is positively correlated with medical care and negatively correlated with health. Hence, the coefficient of health is biased downward in the conditional medical care demand function, and this coefficient could well be negative. The conditional demand function is much more difficult to estimate by two-stage least squares than the production function because no exogenous variables are omitted from it in the investment model. Input prices cannot be used to identify this equation because they are relevant regressors in it. Only wealth and the prices of inputs used to produce commodities other than health are omitted from the conditional demand function in the consumption model or in a mixed investment-consumption model. Measures of the latter variables typically are not available.

In his multiple indicator model, Wagstaff does not treat the latent health variable as endogenous when he obtains the conditional demand function. He is careful to point out that the bias that I have just outlined can explain his result. He also argues that absence of measures of some inputs may account for his finding. He states: "The identification of medical care with market inputs in the health investment production function might be argued to be a source of potential error. If non-medical inputs are important inputs in the production of health – as clearly they are – one might argue that the results stem from a failure to estimate a *system* (his italics) of structural demand equations for health inputs" (p. 226). Although I am biased, in my view these considerations go a long way toward refuting the Zweifel–Breyer critique.

As I indicated above, Wagstaff estimates a reduced form demand function for health in the context of what he terms a pure consumption model as well as in the context of a pure investment model. He does this by including initial assets and the lifetime wage variable in the demand functions as proxies for the marginal utility of wealth. Strictly speaking, however, this is not a pure consumption model. It simply accommodates the consumption motive as well as the investment motive for demanding health. Wagstaff proposes but does not stress one test to distinguish the investment model from the consumption model. If the marginal utility of health does not depend on the quantity of the Z-commodity in the current period utility function, the wage effect should be negative in the demand function for health. Empirically, the current wage coefficient remains positive and significant when initial assets and the lifetime wage are introduced as regressors in this equation. As I noted in Section 4, this result also is consistent with a pure consumption model in which the marginal utility of H is negatively related to Z.

The initial assets and lifetime wage coefficients are highly significant. As Wagstaff indicates, these variables are highly correlated with schooling, the current wage, and other variables in the demand function for health since both are predicted from these variables. These intercorrelations are so high that the schooling coefficient becomes negative and significant in the consumption demand function. Wagstaff stresses that these results must be interpreted with caution.

Zweifel and Breyer (1997) have a confusing and incorrect discussion of theoretical and empirical results on wage effects. They claim that their discussion, which forms part of the critique of my model, is based on Wagstaff's study. In the demand function for health, they indicate that the lifetime wage effect is negative in the pure consumption model and positive in the pure investment model. If the current wage is held constant, both statements are wrong. There is no lifetime wage effect in the investment model and a positive effect in the consumption model provided that health is a superior commodity.[39] If their statements pertain to the current wage effect, the sign is positive in the investment model and indeterminate in the consumption model whether utility or the marginal utility of wealth is held constant.

Zweifel and Breyer's (1997) discussion of schooling effects can be characterized in the same manner as their discussion of wage effects. They claim that the consumption model predicts a positive schooling effect in the demand function for health and a negative effect in the demand function for medical care. This is not entirely consistent with the analysis in Section 4. They use Wagstaff's (1986) result that schooling has a positive coefficient in the conditional demand function for physician visits as evidence against my approach. But as Wagstaff and I have stressed, estimates of that equation are badly biased because health is not treated as endogenous.

A final criticism made by Zweifel and Breyer is that the wage rate does not adequately measure the monetary value of an increase in healthy time due to informal sick leave arrangements and private and social insurance that fund earnings losses due to illness. They do not reconcile this point with the positive effects of the wage rate on various health measures in my study and in Wagstaff's study. In addition, sick leave and insurance plans typically finance less than 100 percent of the loss in earnings. More importantly, they ignore my argument that "... 'the inconvenience costs of illness' are positively correlated with the wage rate .... The complexity of a particular job and the amount of responsibility it entails certainly are positively related to the wage. Thus, when an individual with a high wage becomes ill, tasks that only he can perform accumulate. These increase the intensity of his work load and give him an incentive to avoid illness by demanding more health capital" [Grossman (1972b, pp. 69–70)].

Erbsland et al. (1995) provide another example of the application of the MIMIC procedure to the estimation of a demand for health model. Their database is the 1986 West

[39] Joint production could account for a negative lifetime wage effect, but Zweifel and Breyer do not consider this phenomenon.

German Socio-economic Panel. The degree of handicap, self-rated health, the duration of sick time, and the number of chronic conditions, all as reported by the individual, serve as four indicators of the unobserved stock of health. In the reduced form demand function for health, schooling has a positive and significant coefficient, while age has a negative and significant coefficient. In the reduced form demand function for visits to general practitioners, the age effect is positive and significant, while the schooling effect is negative and significant. These results are consistent with predictions made by the investment model. The latent variable health, which is treated as exogenous, has a negative and very significant coefficient in the conditional demand function for physician visits. This is the same finding reported by Wagstaff (1986).

In my 1972 study [Grossman (1972b)], I showed that the sign of the correlation between medical care and health can be reversed if medical care is treated as endogenous in the estimation of health production functions. Stratmann (1999) gives much more recent evidence in support of the same proposition. Using the 1989 US National Health Interview Survey, he estimates production functions in which the number of work-loss days due to illness in the past two weeks serves as the health measure and a dichotomous indicator for a doctor visit in the past two weeks serves as the measure of medical care. In a partial attempt to control for reverse causality from poor health to more medical care, he obtains separate production functions for persons with influenza, persons with impairments, and persons with chronic asthma.

In single equation tobit models, persons who had a doctor visit had significantly more work-loss than persons who did not have a visit for each of the three conditions. In simultaneous equations probit-tobit models in which the probability of a doctor visit is endogenous, persons who had a doctor visit had significantly less work-loss. The tobit coefficient in the simultaneous equations model implies that the marginal effect of a doctor visit is a 2.7 day reduction in work loss in the case of influenza.[40] The corresponding reductions for impairments and chronic asthma are 2.9 days and 6.9 days, respectively.

6.2. Empirical extensions with longitudinal data

Van Doorslaer (1987) and Wagstaff (1993) fit dynamic demand for health models to longitudinal data. These efforts potentially are very useful because they allow one to take account of the effects of unmeasured variables such as the rate of depreciation and of reverse causality from health at early stages in the life cycle to the amount of formal schooling completed (see Section 7 for more details). In addition, one can relax the assumption that there are no costs of adjustment, so that the lagged stock of health becomes a relevant determinant of the current stock of health.

[40] Stratmann identifies his model with instruments reflecting the price of visiting a physician which differs according to the type of health insurance carried by individuals. The specific measures are Medicaid coverage, private insurance coverage, membership in a Health Maintenance Organization, and whether or not the employer paid the health insurance premium.

Van Doorslaer (1987) employs the 1984 Netherlands Health Interview Survey. While this is a cross-sectional survey, respondents were asked to evaluate their health in 1979 as well as in 1984. Both measures are ten-point scales, where the lowest category is very poor health and the highest category is very good health.

Van Doorslaer uses the identity that the current stock of health equals the undepreciated component of the past stock plus gross investment:

$$H_t = (1 - \delta_{t-1})H_{t-1} + I_{t-1}. \tag{50}$$

He assumes that gross investment is a function of personal background variables (schooling, age, income, and gender). Thus, he regresses health in 1984 on these variables and on health in 1979. To test Muurinen's (1982) hypothesis that schooling lowers the rate of depreciation (see Section 4.2), he allows for an interaction between this variable and health in 1979 in some of the estimated models.

Van Doorslaer's main finding is that schooling has a positive and significant coefficient in the regression explaining health in 1984, with health in 1979 held constant. The regressions in which schooling, past health, and an interaction between the two are entered as regressors are plagued by multicollinearity. They do not allow one to distinguish Muurinen's hypothesis from the hypothesis that schooling raises efficiency in the production of health.

Wagstaff (1993) uses the Danish Health Study, which followed respondents over a period of 12 months beginning in October 1982. As in his 1986 study, a MIMIC model is estimated. Three health measures are used as indicators of the unobserved stock of health capital in 1982 (past stock) and 1983 (current stock). These are a dichotomous indicator of the presence of a health limitation, physician-assessed health of the respondent as reported by the respondent, and self-assessed health.[41] Both of the assessment variables have five-point scales. Unlike his 1986 study, Wagstaff also treats gross investment in health as a latent variable. There are six health care utilization indicators of gross investment: the number of consultations with a general practitioner over the year, the number of consultations with a specialist over the year, the number of days as an inpatient in a hospital over the year, the number of sessions with a physiotherapist over the year, the number of hospital outpatient visits over the year, and the number of hospital emergency room visits during the year.

Wagstaff explicitly assumes partial adjustment instead of instantaneous adjustment. He also assumes that the reduced form demand for health equation is linear rather than log-linear. He argues that this makes it compatible with the linear nature of the net investment identity (net investment equals gross investment minus depreciation). The

[41] The health limitation variable for 1982 is based on responses during October, November, and December of that year. The corresponding variable for 1983 is based on responses during the months of January through September. It is not clear which month is used for the self- and physician-rated health measures. I assume that the 1982 measures come from the October 1982 survey and the 1983 measure comes from the September 1983 survey.

desired stock in period t is a linear function of age, schooling, family income, and gender. A fraction (μ) of the gap between the actual and the desired stock is closed each period. Hence the lagged stock enters the reduced form demand function with a coefficient equal to $1-\mu$. Solving Equation (50) for gross investment in period $t-1$ and replacing H_t by its demand function, Wagstaff obtains a demand function for I_{t-1} that depends on the same variables as those in the demand function for H_t. The coefficient of each sociodemographic variable in the demand function for H_t is the same as the corresponding coefficient in the demand function for I_{t-1}. The coefficient on the lagged stock in the latter demand function equals $-(\mu-\delta_{t-1})$. By estimating the model with cross-equation constraints, μ and δ_{t-1} are identified.

Wagstaff emphasizes that the same variables enter his conditional demand function for I_{t-1} in his cost-of-adjustment model as those that enter the conditional demand function for I_{t-1} in my instantaneous adjustment model. The interpretation of the parameters, however, differs. In my case, the contemporaneous health stock has a positive coefficient, whereas in his case the coefficient is negative if μ exceeds δ_{t-1}. In my case, the coefficient of schooling, for example, is equal to the negative of the schooling coefficient in the production function. In his case, it equals the coefficient of schooling in the demand function for H_t.

To allow for the possibility that the rate of depreciation varies with age, Wagstaff fits the model separately for adults under the age of forty-one and for adults greater than or equal to this age. For each age group, schooling has a positive and significant effect on current health with past health held constant. The coefficient of H_{t-1} in the demand function for I_{t-1} is negative, suggesting costs of adjustment and also suggesting that μ exceeds δ_{t-1}. The implied value of the rate of depreciation is, however, larger in the sample of younger adults than in the sample of older adults. Moreover, in the latter sample, the estimated rate of depreciation is *negative*. These implausible findings may be traced to the inordinate demands on the data attributed to the MIMIC methodology with two latent variables and cross-equation constraints.

Some conceptual issues can be raised in evaluating the two studies just discussed. Wagstaff (1993) estimates input demand functions which include availability measures as proxies for travel and waiting time (for example, the per capita number of general practitioners in the individual's district in the demand function for the number of consultations with general practitioners). Yet he excludes these variables from the demand functions for H_t and I_{t-1}. This is not justified. In addition, Wagstaff implies that the gross investment production function is linear in its inputs, which violates the cost-minimization conditions.

More fundamentally, both Van Doorslaer (1987) and Wagstaff (1993) provide ad hoc cost-of-adjustment models. I now show that a rigorous development of such a model contains somewhat different demand functions than the ones that they estimate. To simplify, I assume that the pure investment model is valid, ignore complications with cost-of-adjustment models studied by Ehrlich and Chuma (1990), and fix the wage rate at

\$1. I also make use of the exact form of the first-order condition for H_t in a discrete time model (see footnote 13):

$$G_t = (1+r)\pi_{t-1} - (1-\delta_t)\pi_t. \tag{51}$$

Note that π_{t-1} is the marginal cost of gross investment in health. Since marginal cost rises as the quantity of investment rises in a model with costs of adjustment, the marginal cost of investment exceeds the average cost of investment. Also to simplify and to keep the system linear, I assume that G_t is a linear function of H_t and that π_{t-1} is a linear function of I_{t-1} and P_{t-1} (the price of the single market input used in the gross investment production function):

$$G_t = \varphi - \alpha H_t, \tag{52}$$

$$\pi_{t-1} = P_{t-1} + I_{t-1}. \tag{53}$$

Given this model, the optimal stock of health in period t is

$$H_t = \frac{\varphi}{\alpha} - \frac{(1+r)}{\alpha}P_{t-1} - \frac{(1+r)}{\alpha}I_{t-1} + \frac{(1-\delta_t)}{\alpha}P_t + \frac{(1-\delta_t)}{\alpha}I_t. \tag{54}$$

Since $I_{t-1} = H_t - (1-\delta_{t-1})H_{t-1}$ and $I_t(1-\delta_t) = (1-\delta_t)H_{t+1} - (1-\delta_t)^2 H_t$,

$$\begin{aligned} H_t = {} & \frac{\varphi}{D} - \frac{(1+r)}{D}P_{t-1} + \frac{(1+r)(1-\delta_{t-1})}{D}H_{t-1} \\ & + \frac{(1-\delta_t)}{D}H_{t+1} + \frac{(1-\delta_t)}{D}P_t, \end{aligned} \tag{55}$$

where $D = \alpha + (1+r) + (1-\delta_t)^2$. Alternatively, substitute Equation (55) into the definition of I_{t-1} to obtain

$$\begin{aligned} I_{t-1} = {} & \frac{\varphi}{D} - \frac{(1+r)}{D}P_{t-1} - \frac{(1-\delta_{t-1})[\alpha + (1-\delta_t)^2]}{D}H_{t-1} \\ & + \frac{(1-\delta_t)}{D}H_{t+1} + \frac{(1-\delta_t)}{D}P_t. \end{aligned} \tag{56}$$

Equation (55) is the demand for health function obtained by Van Doorslaer (1987) and Wagstaff (1993). Their estimates are biased because the stock of health in period t depends on the stock of health in period $t+1$ as well as on the stock of health in period $t-1$ in a model with costs of adjustment. Equation (56) is the gross investment demand function obtained by Wagstaff. His estimate is biased because gross investment in period $t-1$ depends on the stock of health in period $t+1$ as well as on the stock of health in period $t-1$. Note that this equation and Equation (55) also depend on measured and unmeasured determinants of market and nonmarket efficiency in periods $t-1$ and t.

The second-order difference equations given by (55) and (56) can be solved to express H_t or I_{t-1} as functions of current, past, and future values of all the exogenous variables. Similarly, H_{t-1} and H_{t+1} depend on this set of exogenous variables. Since one of the members of this set is the disturbance term in (55) or (56), the lagged and future stocks are correlated with the regression disturbance. Consequently, biases arise if either equation is estimated by ordinary least squares. Consistent estimates can be obtained by fitting the equations by two-stage least squares with past and future values of the exogenous variables serving as instruments for the one-period lead and the one-period lag of the stock.[42] Note that consistent estimates cannot be obtained by the application of ordinary least squares to a first-difference model or to a fixed-effects model. Lagged and future health variables do not drop out of these models and are correlated with the time-varying component of the disturbance term.

I conclude that cost-of-adjustment models require at least three data points (three observations on each individual) to be estimated. Calculated parameters of this model are biased if they are obtained by ordinary least squares. There is an added complication that arises even if all the necessary data are available because the procedure that I have outlined assumes that individuals have perfect information about the future values of the exogenous variables. This may or may not be the case.[43] While the two studies that I have reviewed are provocative, they do not contain enough information to compare instantaneous-adjustment models to cost-of-adjustment models.

6.3. *Theoretical extensions*

Muurinen (1982) examines comparative static age, schooling, and wealth effects in the context of a mixed investment-consumption model with perfect certainty. This approach is more general than mine because it incorporates both the investment motive and the consumption motive for demanding health. In deriving formulas for the effects of increases in age and schooling on the optimal quantities of health capital and medical care, Muurinen assumes that the undiscounted monetary value of the marginal utility of healthy time in period t, given by $(Uh_t/\lambda)(1+r)^t$, is constant for all t. If m_t is the marginal cost of the Z commodity and U_t is its marginal utility, the undiscounted monetary value of the marginal utility of healthy time also is given by $(Uh_t/U_t)m_t$. Hence, Muurinen is assuming that the marginal rate of substitution between healthy time and the Z commodity is constant or that the two commodities are perfect substitutes. Clearly, this is a very restrictive assumption.[44]

[42] The minimum requirements for instruments are measures of P_{t+1} and P_{t-2}. The one-period lead of the price has no impact on H_t with H_{t+1} held constant. Similarly, the two-period lag of the price has no impact on H_t with H_{t-1} held constant.

[43] The same issue arises in estimating the rational addiction model of consumer behavior. For a detailed discussion, see Becker et al. (1994).

[44] Ried (1998) also develops a framework for examining the impacts of changes in exogenous variables in the context of a mixed investment-consumption model. He uses Frisch (1964) demand curves to decompose

In my formal development of the demand for health, I ruled out uncertainty. Surely that is not realistic. I briefly indicated that one could introduce this phenomenon by assuming that a given consumer faces a probability distribution of depreciation rates in every period. I speculated, but did not prove, that consumers might have incentives to hold an excess stock of health in relatively desirable "states of the world" (outcomes with relatively low depreciation rates) in order to reduce the loss associated with an unfavorable outcome. In these relatively desirable states, the marginal monetary return on an investment in health might be smaller than the opportunity cost of capital in a pure investment model [Grossman (1972b, pp. 19–21)].

Beginning with Cropper (1977), a number of persons formally have introduced uncertainty into my pure investment model. Cropper assumes that illness occurs in a given period if the stock of health falls below a critical sickness level, which is random. Income is zero in the illness state. An increase in the stock of health lowers the probability of this state. Cropper further assumes that savings are not possible (all income takes the form of earnings) and that consumers are risk-neutral in the sense that their objective is to maximize the expected discounted value of lifetime wealth.[45]

In my view, Cropper's main result is that consumers with higher incomes or wealth levels will maintain higher stocks of health than poorer persons. While this may appear to be a different result than that contained in my pure investment model with perfect certainty, it is not for two reasons. First, an increase in the stock of health lowers the probability of illness but has no impact on earnings in non-illness states. Hence the marginal benefit of an increase in the stock is given by the reduction in the probability of illness multiplied by the difference between income and gross investment outlays. With these outlays held constant, an increase in income raises the marginal benefit and the marginal rate of return on an investment.[46] Therefore, this wealth or income effect is analogous to the wage effect in my pure investment model with perfect certainty. Second, consider a pure investment model with perfect certainty, positive initial assets but no possibility to save or borrow in financial markets. In this model, investment in health is the only mechanism to increase future consumption. An increase in initial assets will increase the optimal stock of health provided future consumption has a positive wealth elasticity.

Later treatments of uncertainty in the context of demand for health models have assumed risk-averse behavior, so that an expected utility function that exhibits diminishing marginal utility of present and future consumption is maximized. Dardanoni and Wagstaff (1987), Selden (1993), and Chang (1996) all employ two-period models in

the total effect into an effect that holds the marginal utility of wealth constant and an effect attributable to a change in the marginal utility of wealth. He obtains few, if any, unambiguous predictions. I leave it to the reader to evaluate this contribution.

[45] Cropper begins with a model in which consumers are risk-averse, but the rate of depreciation on the stock of health does not depend on age. She introduces risk-neutrality when she allows the rate of depreciation to depend on age.

[46] Cropper assumes that gross investment in health is produced only with medical care, but the above result holds as long as the share of the own time input in the total cost of gross investment is less than one.

which the current period utility function depends only on current consumption. Uncertainty in the second period arises because the earnings-generating function in that period contains a random variable. This function is $Y_2 = Wh_2(H_2, R) = F(H_2, R)$, where Y_2 is earnings in period two, h_2 is the amount of healthy time in that period, H_2 is the stock of health, and R is the random term. Clearly, $F_1 > 0$ and $F_{11} < 0$, where F_1 and F_{11} are the first and second derivatives of H_2 in the earnings function. The second derivative is negative because of my assumption that the marginal product of the stock of health in the production of healthy time falls as the stock rises. An increase in R raises earnings ($F_2 > 0$). In addition to income or earnings from health, income is available from savings at a fixed rate of return.

Given uncertainty, risk-averse individuals make larger investments in health than they would in its absence. Indeed, the expected marginal rate of return is smaller than the rate with perfect certainty. This essentially confirms a result that I anticipated in the brief discussion of uncertainty in my monograph.

The main impact of the introduction of uncertainty is that the quantities of health capital and gross investment depend on initial assets, with the wage rate held constant. The direction of these effects, however, is ambiguous because it depends on the way in which risk is specified. Dardanoni and Wagstaff (1987) adopt a multiplicative specification in which the earnings function is $Y_2 = RH_2$. They show that an increase in initial assets raises the optimal quantities of health and medical care if the utility function exhibits decreasing absolute risk aversion.[47] Selden (1993) adopts a linear specification in which the earnings function is $Y_2 = F(H_2 + R)$ and $\partial^2 Y_2/\partial(H_2 + R)^2 < 0$. He reaches the opposite conclusion: health and medical care fall as assets rise given declining absolute risk aversion.

Chang (1996) generalizes the specification of risk. He shows that the sign of the asset effect depends on the sign of the second-order cross partial derivative in the earnings function (F_{12}). If F_{12} is positive and F_{11} is zero, the asset effect is positive. This is the case considered by Dardanoni and Wagstaff. In my view it is not realistic because it assumes that the marginal product of the stock of health in the production of healthy time is constant. Given the more realistic case in which F_{11} is negative, the asset effect is negative if F_{12} is negative (Selden's case) and indeterminate in sign if F_{12} is positive.[48]

Dardanoni and Wagstaff (1990) introduce uncertainty into pure consumption models of the demand for health. Their study is a static one-period model. The utility function depends on the consumption of a composite commodity and health. Health is given by $H = R + I(M, R)$, where R is a random variable and $I(M, R)$ denotes the health production function. They consider two models: one in which $H = R + M$ and one in which $H = RM$. In the first model an increase in the variance of R with the mean held constant increases the quantity of medical care demanded under plausible assumptions about the utility function (superiority of the composite consumption good and

[47] A utility function $U = U(C)$ exhibits decreasing absolute risk aversion if $-U_{CC}/U_C$ falls as C rises.

[48] Chang provides more results by assuming that the earnings function depends only on "post-shock health," $f(H_2, R)$. The finding that the sign of the asset effect is ambiguous holds in his formulation.

non-increasing absolute risk aversion with regard to that good). In the second model the same effect is more difficult to sign, although it is positive if an increase in R leads to a reduction in M and an increase in the composite good and if the utility function exhibits non-increasing relative risk aversion with regard to the composite good.[49]

Liljas (1998) considers uncertainty in the context of a multiperiod mixed investment-consumption model. Uncertainty takes the form of a random variable that affects the stock of health in period t in an additive fashion. He shows that the stock of health is larger in the stochastic case than in the certainty case. Presumably, this result pertains to the expected stock of health in period t. The actual stock should be smaller than the stock with certainty given a negative shock. Social insurance that funds part of the loss in income due to illness lowers the optimal stock. Private insurance that also funds part of this loss will not necessarily lower the stock further and may actually increase it if the cost of this insurance falls as the stock rises.

To summarize, compared to a model with perfect certainty, the expected value of the stock of health is larger and the optimal quantities of gross investment and health inputs also are larger in a model with uncertainty. In a pure investment model an increase in initial assets can cause health and medical care to change, but the direction of these effects is ambiguous. Under reasonable assumptions, an increase in the variance of risk raises optimal medical care in a pure consumption model.

How valuable are these results? With the exception of the ambiguity of the asset effect, they are not very surprising. The variance of risk is extremely difficult to measure. I am not aware of empirical studies that have attempted to include this variable in a demand for health framework. The possibility that the asset effect can be nonzero in a pure investment model provides an alternative explanation of Wagstaff's (1986) finding that an increase in proxies for initial assets and lifetime earnings raise health. None of the studies has taken my suggestion to treat uncertainty in terms of a probability distribution of depreciation rates in a given period. This could be done by writing the stock of health in period t as

$$H_t = H_{t-1} - \bar{\delta}_{t-1} H_{t-1} + I_{t-1} + R_{t-1}, \tag{57}$$

where $\bar{\delta}_{t-1}$ is the mean depreciation rate and $R_{t-1} = (\bar{\delta}_{t-1} - \delta_{t-1})H_{t-1}$. I leave it to the reader to explore the implications of this formulation.

7. Health and schooling

An extensive review of the literature conducted by Grossman and Kaestner (1997) suggests that years of formal schooling completed is the most important correlate of good

[49] Garber and Phelps (1997), Meltzer (1997), and Picone et al. (1998) also introduce uncertainty into a pure consumption model of the demand for health. I do not discuss the first two studies because they focus on cost-effectiveness analysis. I do not discuss the last one because it emphasizes the behavior of individuals in their retirement years.

health. This finding emerges whether health levels are measured by mortality rates, morbidity rates, self-evaluation of health status, or physiological indicators of health, and whether the units of observation are individuals or groups. The studies also suggest that schooling is a more important correlate of health than occupation or income, the two other components of socioeconomic status. This is particularly true when one controls for reverse causality from poor health to low income. Of course, schooling is a causal determinant of occupation and income, so that the gross effect of schooling on health may reflect in part its impact on socioeconomic status. The studies reviewed, however, indicate that a significant portion of the gross schooling effect cannot be traced to the relationship between schooling and income or occupation.

In a broad sense, the observed positive correlation between health and schooling may be explained in one of three ways. The first argues that there is a causal relationship that runs from increases in schooling to increases in health. The second holds that the direction of causality runs from better health to more schooling. The third argues that no causal relationship is implied by the correlation; instead, differences in one or more "third variables," such as physical and mental ability and parental characteristics, affect both health and schooling in the same direction.

It should be noted that these three explanations are not mutually exclusive and can be used to rationalize an observed correlation between any two variables. But from a public policy perspective, it is important to distinguish among them and to obtain quantitative estimates of their relative magnitudes. Suppose that a stated goal of public policy is to improve the level of health of the population or of certain groups in the population. Given this goal and given the high correlation between health and schooling, it might appear that one method of implementation would be to increase government outlays on schooling. In fact, Auster et al. (1969) suggest that the rate of return on increases in health via higher schooling outlays far exceeds the rate of return on increases in health via higher medical care outlays. This argument assumes that the correlation between health and schooling reflects only the effect of schooling on health. If, however, the causal relationship was the reverse, or if the third-variable hypothesis was relevant, then increased outlays on schooling would not accomplish the goal of improved health.

Causality from schooling to health results when more educated persons are more efficient producers of health. This efficiency effect can take two forms. Productive efficiency pertains to a situation in which the more educated obtain a larger health output from given amounts of endogenous (choice) inputs. This is the effect that I have emphasized throughout this paper. Allocative efficiency, discussed in detail by Kenkel (2000), pertains to a situation in which schooling increases information about the true effects of the inputs on health. For example, the more educated may have more knowledge about the harmful effects of cigarette smoking or about what constitutes an appropriate diet. Allocative efficiency will improve health to the extent that it leads to the selection of a better input mix.

Causality from schooling to health also results when education changes tastes or preferences in a manner that favors health relative to certain other commodities. In some cases the taste hypothesis cannot be distinguished from allocative hypothesis, particularly when knowledge of health effects has been available for some time. But in a situation in which the new information becomes available, the allocative efficiency hypothesis predicts a more rapid response by the more educated.

Alternatively, the direction of causality may run from better health to more schooling because healthier students may be more efficient producers of additions to the stock of knowledge (or human capital) via formal schooling. Furthermore, this causal path may have long lasting effects if past health is an input into current health status. Thus, even for non-students, a positive relationship between health and schooling may reflect reverse causality in the absence of controls for past health.

The "third-variable" explanation is particularly relevant if one thinks that a large unexplained variation in health remains after controlling for schooling and other determinants. Studies summarized by Grossman and Kaestner (1997) and results in the related field of investment in human capital and the determinants of earnings [for example, Mincer (1974)], indicate that the percentage of the variation in health explained by schooling is much smaller than the percentage of the variation in earnings explained by schooling. Yet it also is intuitive that health and illness have larger random components than earnings. The third-variable explanation is relevant only if the unaccounted factors which affect health are correlated with schooling. Note that both the reverse causality explanation and the third-variable explanation indicate that the observed relationship between current health and schooling reflects an omitted variable. In the case of reverse causality, the omitted variable is identified as past or endowed health. In econometric terminology, both explanations fall under the general rubric of biases due to unobserved heterogeneity among individuals.

Kaestner and I [Grossman and Kaestner (1997)] conclude from our extensive review of the literature that schooling does in fact have a causal impact on good health. In drawing this conclusion, we are sensitive to the difficulties of establishing causality in the social sciences where natural experiments rarely can be performed. Our affirmative answer is based on the numerous studies in the US and developing countries that we have summarized. These studies employ a variety of adult, child, and infant health measures, many different estimation techniques, and controls for a host of third variables.

I leave it up to the reader to evaluate this conclusion after reading the Grossman–Kaestner paper and the studies therein. I also urge the reader to consult my study dealing with the correlation between health and schooling [Grossman (1975)] because I sketch out a framework in which there are complementary relationships between schooling and health – the principal components of the stock of human capital – at various stages in the life cycle. The empirical evidence that Kaestner and I report on causality from schooling to health as well as on causality from health to schooling underscores the potential payoffs to the formal development of a model in which the stocks of health and knowledge are determined simultaneously.

In the remainder of this section, I want to address one challenge of the conclusion that the role of schooling is causal: the time preference hypothesis first proposed by Fuchs (1982). Fuchs argues that persons who are more future oriented (who have a high degree of time preference for the future) attend school for longer periods of time and make larger investments in health. Thus, the effect of schooling on health is biased if one fails to control for time preference.

The time preference hypothesis is analogous to the hypothesis that the positive effect of schooling on earnings is biased upward by the omission of ability. In each case a well-established relationship between schooling and an outcome (earnings or health) is challenged because a hard-to-measure variable (ability or time preference) has been omitted. Much ink has been spilled on this issue in the human capital literature. Attempts to include proxies for ability in earnings functions have resulted in very modest reductions in the schooling coefficient [for example, Griliches and Mason (1972), Hause (1972)]. Proponents of the ability hypothesis have attributed the modest reductions to measurement error in these proxies [for example, Goldberger (1974)]. More recent efforts have sought instruments that are correlated with schooling but not correlated with ability [for example, Angrist and Krueger (1991)]. These efforts have produced the somewhat surprising finding that the schooling coefficient *increases* when the instrumental variables procedure is employed. A cynic might conclude that the way to destroy any empirical regularity is to attribute it to an unmeasured variable, especially if the theory with regard to the relevance of this variable is not well developed.[50]

Nevertheless, the time preference hypothesis is important because it is related to recent and potentially very rich theoretical models in which preferences are endogenous [Becker and Murphy (1988), Becker (1996), Becker and Mulligan (1997)]. Differences in time preference among individuals will not generate differences in investments in human capital unless certain other conditions are met. One condition is that the ability to finance these investments by borrowing is limited, so that they must be funded to some extent by foregoing current consumption. Even if the capital market is perfect, the returns on an investment in schooling depend on hours of work if schooling raises market productivity by a larger percentage than it raises nonmarket productivity. Individuals who are more future oriented desire relatively more leisure at older ages. Therefore, they work more at younger ages and have a higher discounted marginal benefit on a given investment than persons who are more present oriented. If health enters the utility function, persons who discount the future less heavily will have higher health levels during most stages of the life cycle. Hence, a positive relationship between schooling and health does not necessarily imply causality.

Since the conditions that generate causal effects of time preference on schooling and health are plausible, attempts to control for time preference in estimating the schooling coefficient in a health outcome equation are valuable. Fuchs (1982) measures time preference in a telephone survey by asking respondents questions in which they chose between a sum of money now and a larger sum in the future. He includes an index of time preference in a multiple regression in which health status is the dependent variable and

[50] See Grossman and Kaestner (1997) for a model in which ability should be omitted from the reduced form earnings function even though it enters the structural production function and has a causal impact on schooling.

schooling is one of the independent variables. Fuchs is not able to demonstrate that the schooling effect is due to time preference. The latter variable has a negative regression coefficient, but it is not statistically significant. When time preference and schooling are entered simultaneously, the latter dominates the former. These results must be regarded as preliminary because they are based on one small sample of adults on Long Island and on exploratory measures of time preference.

Farrell and Fuchs (1982) explore the time preference hypothesis in the context of cigarette smoking using interviews conducted in 1979 by the Stanford Heart Disease Prevention Program in four small agricultural cities in California. They examine the smoking behavior of white non-Hispanics who were not students at the time of the survey, had completed 12 to 18 years of schooling, and were at least 24 years old. The presence of retrospective information on cigarette smoking at ages 17 and 24 allows them to relate smoking at these two ages to years of formal schooling completed by 1979 for cohorts who reached age 17 before and after the widespread diffusion of information concerning the harmful effects of cigarette smoking on health.

Farrell and Fuchs find that the negative relationship between schooling and smoking, which rises in absolute value for cohorts born after 1953, does not increase between the ages of 17 and 24. Since the individuals were all in the same school grade at age 17, the additional schooling obtained between that age and age 24 cannot be the cause of differential smoking behavior at age 24, according to the authors. Based on these results, Farrell and Fuchs reject the hypothesis that schooling is a causal factor in smoking behavior in favor of the view that a third variable causes both. Since the strong negative relationship between schooling and smoking developed only after the spread of information concerning the harmful effects of smoking, they argue that the same mechanism may generate the schooling-health relationship.

A different interpretation of the Farrell and Fuchs finding emerges if one assumes that consumers are farsighted. The current consumption of cigarettes leads to more illness and less time for work in the future. The cost of this lost time is higher for persons with higher wage rates who have made larger investments in human capital. Thus, the costs of smoking in high school are greater for persons who plan to make larger investments in human capital.

Berger and Leigh (1989) have developed an extremely useful methodology for disentangling the schooling effect from the time preference effect. Their methodology amounts to treating schooling as an endogenous variable in the health equation and estimating the equation by a variant of two-stage least squares. If the instrumental variables used to predict schooling in the first stage are uncorrelated with time preference, this technique yields an unbiased estimate of the schooling coefficient. Since the framework generates a recursive model with correlated errors, exogenous variables that are unique to the health equation are not used to predict schooling.

Berger and Leigh apply their methodology to two data sets: the first National Health and Nutrition Examination Survey (NHANES I) and the National Longitudinal Survey of Young Men (NLS). In NHANES I, health is measured by blood pressure, and separate equations are obtained for persons aged 20 through 40 and over age 40 in the period

1971 through 1975. The schooling equation is identified by ancestry and by average real per capita income and average real per capita expenditures on education in the state in which an individual resided from the year of birth to age 6. These variables enter the schooling equation but are excluded from the health equation. In the NLS, health is measured by a dichotomous variable that identifies men who in 1976 reported that health limited or prevented them from working and alternatively by a dichotomous variable that identifies the presence of a functional health limitation. The men in the sample were between the ages of 24 and 34 in 1976, had left school by that year, and reported no health limitations in 1966 (the first year of the survey). The schooling equation is identified by IQ, Knowledge of Work test scores, and parents' schooling.

Results from the NLS show that the schooling coefficient rises in absolute value when predicted schooling replaces actual schooling, and when health is measured by work limitation. When health is measured by functional limitation, the two-stage least squares schooling coefficient is approximately equal to the ordinary least squares coefficient, although the latter is estimated with more precision. For persons aged 20 through 40 in NHANES I, schooling has a larger impact on blood pressure in absolute value in the two-stage regressions. For persons over age 40, however, the predicted value of schooling has a positive and insignificant regression coefficient. Except for the last finding, these results are inconsistent with the time preference hypothesis and consistent with the hypothesis that schooling causes health.

In another application of the same methodology, Leigh and Dhir (1997) focus on the relationship between schooling and health among persons ages 65 and over in the 1986 wave of the Panel Survey of Income Dynamics (PSID). Health is measured by a disability index comprised of answers to six activities of daily living and by a measure of exercise frequency. Responses to questions asked in 1972 concerning the ability to delay gratification are used to form an index of time preference. Instruments for schooling include parents' schooling, parents' income, and state of residence in childhood. The schooling variable is associated with better health and more exercise whether it is treated as exogenous or endogenous.

Sander (1995a, 1995b) has applied the methodology developed by Berger and Leigh to the relationship between schooling and cigarette smoking studied by Farrell and Fuchs (1982). His data consist of the 1986–1991 waves of the National Opinion Research Center's General Social Survey. In the first paper the outcome is the probability of quitting smoking, while in the second the outcome is the probability of smoking. Separate probit equations are obtained for men and women ages 25 and older. Instruments for schooling include father's schooling, mother's schooling, rural residence at age 16, region of residence at age 16, and number of siblings.

In general schooling has a negative effect on smoking participation and a positive effect on the probability of quitting smoking. These results are not sensitive to the use of predicted as opposed to actual schooling in the probit regressions. Moreover, the application of the Wu–Hausman endogeneity test [Wu (1973), Hausman (1978)] in the quit equation suggest that schooling is exogenous in this equation. Thus, Sander's results, like Berger and Leigh's and Leigh and Dhir's results, are inconsistent with the time preference hypothesis.

The aforementioned conclusion rests on the assumption that the instruments used to predict schooling in the first stage are uncorrelated with time preference. The validity of this assumption is most plausible in the case of measures such as real per capita income and real per capita outlays on education in the state in which an individual resided from birth to age 6 (used by Berger and Leigh in NHANES I), state of residence in childhood (used by Leigh and Dhir in the PSID) and rural residence at age 16 and region of residence at that age (used by Sander). The validity of the assumption is less plausible in the case of measures such as parents' schooling (used by Sander and by Berger and Leigh in the NLS and by Leigh and Dhir in the PSID) and parents' income (used by Leigh and Dhir in the PSID).

Given this and the inherent difficulty in Fuchs's (1982) and Leigh and Dhir's (1997) attempts to measure time preference directly, definitive evidence with regard to the time preference hypothesis still is lacking. Moreover, Sander (1995a, 1995b) presents national data showing a much larger downward trend in the probability of smoking and a much larger upward trend in the probability of quitting smoking between 1966 and 1987 as the level of education rises. Since information concerning the harmful effects of smoking was widespread by the early 1980s, these results are not consistent with an allocative efficiency argument that the more educated are better able to process new information.

Becker and Murphy's (1988) theoretical model of rational addiction predicts that persons who discount the future heavily are more likely to participate in such addictive behaviors as cigarette smoking. Becker et al. (1991) show that the higher educated people are more responsive to changes in the harmful future consequences of the consumption of addictive goods because they are more future oriented. Thus, the trends just cited are consistent with a negative relationship between schooling and the rate of time preference for the present.

Proponents of the time preference hypothesis assume that a reduction in the rate of time preference for the present causes years of formal schooling to rise. On the other hand, Becker and Mulligan (1997) argue that causality may run in the opposite direction: namely, an increase in schooling may *cause* the rate of time preference for the present to fall (may *cause* the rate of time preference for the future to rise). In most models of optimal consumption over the life cycle, consumers maximize a lifetime utility function defined as the discounted sum or present value of utility in each period or at each age. The discount factor (β) is given by $\beta = 1/(1 + g)$, where g is the rate of time preference for the present. Becker and Mulligan point out that the present value of utility is *higher* the smaller is the rate of time preference for the present. Hence, consumers have incentives to make investments that *lower* the rate of time preference for the present.

Becker and Mulligan then show that the marginal costs of investments that lower time preference fall and the marginal benefits rise as income or wealth rises. Marginal benefits also are greater when the length of life is greater. Hence, the equilibrium rate of time preference falls as the level of education rises because education raises income and life expectancy. Moreover, the more educated may be more efficient in making investments that lower the rate of time preference for the present – a form of productive efficiency not associated with health production. To quote Becker and Mulligan: "Schooling also determines ... [investments in time preference] partly through the study of history and other subjects, for schooling focuses students' attention on the future. Schooling can communicate images of the situations and difficulties of adult life, which are the future of childhood and adolescence. In addition, through repeated practice at problem solving, schooling helps children learn the art of scenario simulation. Thus, educated people should be more productive at reducing the remoteness of future pleasures" (pp. 735–736).

Becker and Mulligan's argument amounts to a third causal mechanism in addition to productive and allocative efficiency in health production via which schooling can cause health. Econometrically, the difference between their model and Fuchs's model can be specified as follows:

$$H = \alpha_1 Y + \alpha_2 E + \alpha_3 g, \tag{58}$$

$$g = \alpha_4 Y + \alpha_5 E, \tag{59}$$

$$E = \alpha_6 g, \tag{60}$$

$$Y = \alpha_7 E. \tag{61}$$

In this system H is health, Y is permanent income, E is years of formal schooling completed, g is time preference for the present, and the disturbance terms are suppressed. The first equation is a demand for health function in which the coefficient of E reflects productive or allocative efficiency or both. Fuchs assumes that α_5 is zero. Hence, the coefficient of E in the first equation is biased if g is omitted.

In one version of their model, Becker and Mulligan assume that α_6 is zero, although in a more general formulation they allow this coefficient to be nonzero. Given that α_6 is zero, and substituting the second equation into the first, one obtains

$$H = (\alpha_1 + \alpha_4\alpha_3)Y + (\alpha_2 + \alpha_5\alpha_3)E. \tag{62}$$

The coefficient of Y in the last equation reflects both the direct effect of income on health (α_1) and the indirect effect of income on health through time preference ($\alpha_4\alpha_3$). Similarly, the coefficient of E reflects both the direct efficiency effect (α_2) and the indirect effect of schooling on health through time preference ($\alpha_5\alpha_3$).

Suppose that the direct efficiency effect of schooling (α_2) is zero. In Fuchs's model, if health is regressed on income and schooling as represented by solving Equation (60) for g, the expected value of the schooling coefficient is α_3/α_6. This coefficient reflects causality from time preference to schooling. In Becker and Mulligan's model the schooling coefficient is $\alpha_3\alpha_5$. This coefficient reflects causality from schooling to time preference. The equation that expresses income as a function of schooling stresses that schooling has indirect effects on health via income. Becker and Mulligan would include health as a determinant of time preference in the second equation because health lowers mortality, raises future utility levels, and increases incentives to make investments that lower the rate of time preference.

Becker and Mulligan's model appears to contain useful insights in considering intergenerational relationships between parents and children. For example, parents can raise their children's future health, including their adulthood health, by making them more future oriented. Note that years of formal schooling completed is a time-invariant variable beyond approximately age 30, while adult health is not time invariant. Thus, parents probably have a more important direct impact on the former than the latter. By making investments that raise their offsprings' schooling, parents also induce them to make investments that lower their rate of time preference for the present and therefore raise their adult health.

Becker and Mulligan suggest a more definitive and concrete way to measure time preference and incorporate it into estimates of health demand functions than those that have been attempted to date. They point out that the natural logarithm of the ratio of consumption between consecutive time periods ($\ln L$) is approximately equal to $\sigma[\ln(1+r) - \ln(1+g)]$, where σ is the intertemporal elasticity of substitution in consumption, r is the market rate of interest, and g is the rate of time preference for the present. If σ and r do not vary among individuals, variations in $\ln L$ capture variations in time preference. With panel data, $\ln L$ can be included as a regressor in the health demand function Since Becker and Mulligan stress the endogeneity of time preference and its dependence on schooling, simultaneous equations techniques appear to be required. Identification of this model will not be easy, but success in this area has the potential to greatly inform public policy.

To illustrate the last point, suppose that most of the effect of schooling on health operates through time preference. Then school-based programs to promote health knowledge in areas characterized by low levels of income and education may have much smaller payoffs than programs that encourage the investments in time preference made by the more educated. Indeed, in an ever-changing world in which new information constantly becomes available, general interventions that encourage future-oriented behavior may have much larger rates of return in the long run than specific interventions designed, for example, to discourage cigarette smoking, alcohol abuse, or the use of illegal drugs.

There appear to be important interactions between Becker and Mulligan's theory of the endogenous determination of time preference and Becker and Murphy's (1988) theory of rational addiction. Such addictive behaviors as cigarette smoking, excessive alcohol use, and the consumption of illegal drugs have demonstrated adverse health effects. Increased consumption of these goods raises present utility but lowers future utility. According to Becker and Mulligan (1997, p. 744)), "Since a decline in future utility reduces the benefits from a lower discount on future utilities, greater consumption of harmful substances would lead to higher rates of time preference by discouraging investments in lowering these rates ..." This is the converse of Becker and Murphy's result that people who discount the future more heavily are more likely to become addicted. Thus,

"... harmful addictions induce even rational persons to discount the future more heavily, which in turn may lead them to become more addicted" [Becker and Mulligan (1997, p. 744)].

It is well known that cigarette smoking and excessive alcohol abuse begin early in life [for example, Grossman et al. (1993)]. Moreover, bandwagon or peer effects are much more important in the case of youth smoking or alcohol consumption than in the case of adult smoking or alcohol consumption. The two-way causality between addiction and time preference and the importance of peer pressure explain why parents who care about the welfare of their children have large incentives to make investments that make their children more future oriented. These forces may also account for the relatively large impact of schooling on health with health knowledge held constant reported by Kenkel (1991).

Some parents may ignore or be unaware of the benefits of investments in time preference. Given society's concern with the welfare of its children, subsidies to school-based programs that make children more future oriented may be warranted. But much more research dealing with the determinants of time preference and its relationship with schooling and health is required before these programs can be formulated and implemented in a cost-effective manner.

8. Conclusions

Most of the chapters in this Handbook focus on various aspects of the markets for medical care services and health insurance. This focus is required to understand the determinants of prices, quantities, and expenditures in these markets. The main message of my paper is that a very different theoretical paradigm is required to understand the determinants of health outcomes. I have tried to convince the reader that the human capital model of the demand for health provides the framework to conduct investigations of these outcomes. The model emphasizes the difference between health as an output and medical care as one of many inputs into the production of health and the equally important difference between health capital and other forms of human capital. It provides a theoretical framework for making predictions about the impacts of many variables on health and an empirical framework for testing these predictions.

Future theoretical efforts will be especially useful if they consider the joint determination of health and schooling and the interactions between these two variables and time preference for the present. A model in which both the stock of health and the stock of knowledge (schooling) are endogenous does not necessarily generate causality between the two. Individuals, however, typically stop investing in schooling at relatively young ages but rarely stop investing in health. I have a "hunch" that a dynamic model that takes account of these patterns will generate effects of an endogenously determined schooling variable on health in the health demand function if schooling has a causal impact on productive efficiency or time preference.

Future empirical efforts will be especially useful if they employ longitudinal databases with a variety of health outputs, health inputs, and direct and indirect (for example, the rate of growth in total consumption) measures of time preference at three or more different ages. This is the type of data required to implement the cost-of-adjustment model outlined in Section 6. It also is the type of data required to distinguish between the productive and allocative efficiency effects of schooling and to fit demand for health models in which medical care is not necessarily the primary health input. Finally, it is the type of data to fully sort out the hypothesis that schooling causes health from the competing hypothesis that time preference causes both using methods outlined in Section 7.

These research efforts will not be easy, but their potential payoffs are substantial. Medical care markets in most countries are subject to large amounts of government intervention, regulation, and subsidization. I have emphasized the basic proposition that consumers demand health rather than medical care. Thus, one way to evaluate policy initiatives aimed at medical care is to consider their impacts on health outcomes in the context of a cost-benefit analysis of programs that influence a variety of health inputs. If this undertaking is to be successful, it must draw on refined estimates of the parameters of health production functions, output demand functions for health, and input demand functions for health inputs.

References

Angrist, J.D., and A.B. Krueger (1991), “Does compulsory school attendance affect schooling and earnings?”, Quarterly Journal of Economics 106:979–1014.

Auster, R., I. Leveson and D. Sarachek (1969), “The production of health: an exploratory study”, Journal of Human Resources 4:411–436.

Becker, G.S. (1964), Human Capital (Columbia University Press for the National Bureau of Economic Research, New York).

Becker, G.S. (1965), “A theory of the allocation of time”, Economic Journal 75:493–517.

Becker, G.S. (1967), Human Capital and the Personal Distribution of Income: An Analytical Approach (University of Michigan, Ann Arbor, MI). Also available in: G.S. Becker (1993), Human Capital, 3rd edn. (University of Chicago Press) 102–158.

Becker, G.S. (1996), Accounting for Tastes (Harvard University Press, Cambridge, MA).

Becker, G.S., M. Grossman and K.M. Murphy (1991), “Rational addiction and the effect of price on consumption”, American Economic Review 81:237–241.

Becker, G.S., M. Grossman and K.M. Murphy (1994), “An empirical analysis of cigarette addiction”, American Economic Review 84:396–418.

Becker, G.S., and C.B. Mulligan (1997), “The endogenous determination of time preference”, Quarterly Journal of Economics 112:729–758.

Becker, G.S., and K.M. Murphy (1988), “A theory of rational addiction”, Journal of Political Economy 96:675–700.

Ben-Porath, Y. (1967), “The production of human capital and the life cycle of earnings”, Journal of Political Economy 75:353–367.

Bentham, J. (1931), Principles of Legislation (Harcourt, Brace and Co., New York).

Berger, M.C., and J.P. Leigh (1989), “Schooling, self-selection, and health”, Journal of Human Resources 24:433–455.

Bound, J., D.M. Jaeger and R.M. Baker (1995), “Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak”, Journal of the American Statistical Association 90:443–450.

Chang, F.-R. (1996), "Uncertainty and investment in health", Journal of Health Economics 15:369–376.

Corman, H., T.J. Joyce and M. Grossman (1987), "Birth outcome production functions in the US", Journal of Human Resources 22:339–360.

Cropper, M.L. (1977), "Health, investment in health, and occupational choice", Journal of Political Economy 85:273–1294.

Currie, J. (2000), "Child health in developed countries", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 19.

Dardanoni, V., and A. Wagstaff (1987), "Uncertainty, inequalities in health and the demand for health", Journal of Health Economics 6:283–290.

Dardanoni, V., and A. Wagstaff (1990), "Uncertainty and the demand for medical care", Journal of Health Economics 9:23–38.

Ehrlich, I., and H. Chuma (1990), "A model of the demand for longevity and the value of life extensions", Journal of Political Economy 98:761–782.

Erbsland, M., W. Ried and V. Ulrich (1995), "Health, health care, and the environment. Econometric evidence from German micro data", Health Economics 4:169–182.

Farrell, P., and V.R. Fuchs (1982), "Schooling and health: The cigarette connection", Journal of Health Economics 1:217–230.

Frisch, R. (1964), "Dynamic utility", Econometrica 32:418–424.

Fuchs, V.R. (1966), "The contribution of health services to the American economy", Milbank Memorial Fund Quarterly 44:65–102.

Fuchs, V.R. (1982), "Time preference and health: an exploratory study", in: V.R. Fuchs, ed., Economic Aspects of Health (University of Chicago Press for the National Bureau of Economic Research, Chicago) 93–120.

Garber, A.M., and C.E. Phelps (1997), "Economic foundations of cost-effectiveness analysis", Journal of Health Economics 16:1–31.

Ghez, G.R., and G.S. Becker (1975), The Allocation of Time and Goods Over the Life Cycle (Columbia University Press for the National Bureau of Economic Research, New York).

Goldberger, A.S. (1974), "Unobservable variables in econometrics", in: P. Zarembreka, ed., Frontiers in Econometrics (Academic Press, New York) 193–213.

Griliches, Z., and W.M. Mason (1972), "Education, income, and ability", Journal of Political Economy 80:S74–S103.

Grossman, M. (1972a), "On the concept of health capital and the demand for health", Journal of Political Economy 80:223–255.

Grossman, M. (1972b), The Demand for Health: A Theoretical and Empirical Investigation (Columbia University Press for the National Bureau of Economic Research, New York).

Grossman, M. (1975), "The correlation between health and schooling", in: N.E. Terleckyj, ed., Household Production and Consumption (Columbia University Press for the National Bureau of Economic Research, New York) 147–211.

Grossman, M. (1982), "The demand for health after a decade", Journal of Health Economics 1:1–3.

Grossman, M., and T.J. Joyce (1990), "Unobservables, pregnancy resolutions, and birth weight production functions in New York City", Journal of Political Economy 98:983–1007.

Grossman, M., and R. Kaestner (1997), "Effects of education on health", in: J.R. Behrman and N. Stacey, eds., The Social Benefits of Education (University of Michigan Press, Ann Arbor, MI) 69–123.

Grossman, M., J.L. Sindelar, J. Mullahy and R. Anderson (1993), "Policy watch: Alcohol and cigarette taxes", Journal of Economic Perspectives 7:211–222.

Hause, J.C. (1972), "Earnings profile: Ability and schooling", Journal of Political Economy 80:S108–S138.

Hausman, J.A. (1978), "Specification tests in econometrics", Econometrica 46:1251–1271.

Jöreskog, K.G. (1973), "A general method for estimating a linear structural equations system", in: A.S. Goldberger and O.D. Duncan, eds., Structural Equations Models in the Social Sciences (Seminar Press, New York) 85–112.

Jöreskog, K.G., and D. Sörbom (1981), LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood (International Educational Services, Chicago).

Joyce, T.J. (1994), "Self-selection, prenatal care, and birthweight among blacks, whites, and Hispanics in New York City", Journal of Human Resources 29:762–794.

Kenkel, D.S. (1991), "Health behavior, health knowledge, and schooling", Journal of Political Economy 99:287–305.

Kenkel, D.S. (2000), "Prevention", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 31.

Lancaster, K.J. (1966), "A new approach to consumer theory", Journal of Political Economy 74:32–157.

Lazear, E. (1977), "Education: Consumption or production?", Journal of Political Economy 85:569–597.

Leigh, J.P., and R. Dhir (1997), "Schooling and frailty among seniors", Economics of Education Review 16:45–57.

Liljas, B. (1998), "The demand for health with uncertainty and insurance", Journal of Health Economics 17:153–170.

MaCurdy, T.E. (1981), "An empirical model of labor supply in a life-cycle setting", Journal of Political Economy 89:059–1085.

Meltzer, D. (1997), "Accounting for future costs in medical cost-effectiveness analysis", Journal of Health Economics 16:33–64.

Michael, R.T. (1972), The Effect of Education on Efficiency in Consumption (Columbia University Press for the National Bureau of Economic Research, New York).

Michael, R.T. (1973), "Education in nonmarket production", Journal of Political Economy 81:306–327.

Michael, R.T., and G.S. Becker (1973), "On the new theory of consumer behavior", Swedish Economic Journal 75:378–396.

Mincer, J. (1974), Schooling, Experience, and Earnings (Columbia University Press for the National Bureau of Economic Research, New York).

Mushkin, S.J. (1962), "Health as an investment", Journal of Political Economy 70(Supplement):129–157.

Muurinen, J. (1982), "Demand for health: a generalised Grossman model", Journal of Health Economics 1:5–28.

Picone, G., M.U. Echeverria and R.M. Wilson (1998), "The effect of uncertainty on the demand for medical care, health capital and wealth", Journal of Health Economics 17:171–186.

Ried, M. (1996), "Willingness to pay and cost of illness for changes in health capital depreciation", Health Economics 5:447–468.

Ried, M. (1998), "Comparative dynamic analysis of the full Grossman model", Journal of Health Economics 17:383–426.

Rosenzweig, M.R., and T.P. Schultz (1983), "Estimating a household production function: heterogeneity, the demand for health inputs, and their effects on birth weight", Journal of Political Economy 91:723–746.

Rosenzweig, M.R., and T.P. Schultz (1988), "The stability of household production technology: a replication", Journal of Human Resources 23:535–549.

Rosenzweig, M.R., and T.P. Schultz (1991), "Who receives medical care? Income, implicit prices, and the distribution of medical services among pregnant women in the United States", Journal of Human Resources 26:473–508.

Sander, W. (1995a), "Schooling and quitting smoking", Review of Economics and Statistics 77:191–199.

Sander, W. (1995b), "Schooling and smoking", Economics of Education Review 14:23–33.

Selden, T.M. (1993), "Uncertainty and health care spending by the poor: the human capital model revisited", Journal of Health Economics 12:109–115.

Staiger, D., and J.A. Stock (1997), "Instrumental variables regression with weak instruments", Econometrica 65:557–586.

Stratmann, T. (1999), "What do medical services buy? Effects of doctor visits on work day loss", Eastern Economic Journal 25:1–16.

Usher, D. (1975), "Comments on the correlation between health and schooling", in: N.E. Terleckyj, ed., Household Production and Consumption (Columbia University Press for the National Bureau of Economic Research, New York) 212–220.

van Doorslaer, E.K.A. (1987), Health, Knowledge and the Demand for Medical Care (Assen, Maastricht, The Netherlands).

Wagstaff, A. (1986), "The demand for health: some new empirical evidence", Journal of Health Economics 5:195–233.

Wagstaff, A. (1993), "The demand for health: an empirical reformulation of the Grossman model", Health Economics 2:189–198.

Wu, D.-M. (1973), "Alternative tests of independence between stochastic regressors and disturbances", Econometrica 41:733–750.

Zweifel, P., and F. Breyer (1997), Health Economics (Oxford University Press, New York).

Zweifel, P., and W.G. Manning (2000), "Moral hazard and consumer incentives in health care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 8.

Chapter 8

MORAL HAZARD AND CONSUMER INCENTIVES IN HEALTH CARE*

PETER ZWEIFEL

Socioeconomic Institute, University of Zurich

WILLARD G. MANNING

Dept. of Health Studies, The University of Chicago

Contents

*The authors wish to thank Andrew Jones, Joe Newhouse, Edward Norton, and Wynand van de Ven for excellent suggestions and criticisms.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

Consumer incentives are reflected in a wide range of choices, many of which occur in both insurance- and tax-financed health care systems. However, health insurance and sick leave pay cause consumer incentives to be reflected in moral hazard effects of several types. Theoretically, ex ante moral hazard (a reduction of preventive effort in response to insurance coverage) is not unambiguously predicted, and there is very limited empirical evidence about it. The case for static ex post moral hazard (an increase in the demand for medical care of a given technology) is stronger. The empirical evidence reported comes from three sources, natural experiments, observational comparisons of individuals, and the Health Insurance Experiment (HIE). The distinguishing feature of the HIE is that participants were assigned to insurance plans, which forestalls the possibility of good risks self-selecting plans with substantial cost sharing, resulting in an overestimate of the effects of plan design on health care expenditure. While the values of estimated price elasticities vary widely among the three sources and less markedly according to the type of care (outpatient, hospital, dental, mental), the responsiveness of the demand for medical care to net price is beyond doubt. The pure price elasticity for medical care in excess of a deductible (i.e. where the marginal price is constant) was estimated by HIE at −0.2 overall. Finally, there may be a dynamic moral hazard effect (choice biased in favor of new, usually more expensive medical technology). Here, the empirical evidence is very scanty again. Another promising field for future research is the interplay between consumer incentives and rationing by the physician in managed care.

JEL classification: C12, C93, D82, G22, I11, J32, O31

1. Introduction and overview

Consumer incentives derive from the interaction of individual preferences and constraints limiting the pursuit of these preferences. Typically, economists have little to say about the formation of the preference component of incentives, and this chapter is no exception to this rule. On the constraints side, income and prices basically determine the set of feasible choices. In the context of health care, both of these quantities tend to depend on health status. Traditionally, ill health has implied a loss of income to the working individual. In present times, sick leave pay substitutes for labor income to a considerable degree in the event of illness. The higher this replacement income, the higher one would expect the propensity to initiate a sickness episode and to demand medical care to be. With regard to prices, the relative price of health care depends not only on the overall money price but also on health insurance coverage and the time costs associated with access to and use of medical services. These time costs continue to constrain consumer choice even when medical services have a zero money price, which is typical of many tax-financed and social insurance systems. Therefore, the basic tenet of this chapter is that consumer incentives matter regardless of whether health care is financed by insurance premiums, payroll contributions, or taxes.

However, in the case of health care, some may regard consumer incentives as largely irrelevant because of the physician‘s role in medical choices. Due to his informational advantage, the physician is sometimes thought not only to influence (the perception of) constraints but also preferences. The view adopted here is that physician-determined demand would reflect the extreme case of complete delegation of authority to the physician, which certainly occurs in some instances. Still, the patient can make efforts to reduce the information gap, with the magnitude of these efforts likely to depend on consumer incentives again. Delegation of authority thus becomes a matter of degree, leaving scope to consumer incentives in health care.

Against this backdrop, the plan of this contribution is as follows. In Section 2, several dimensions of consumer incentives are distinguished. In view of the importance of insurance noted above, the concept of moral hazard is introduced, to be used for structuring the main body of the paper. After discussion of the optimal amount of delegation of authority in Section 3, the contribution proceeds on the distinction of three types of moral hazard effects, viz. ex ante, static ex post, and dynamic ex post, referring to the insured's tendency to reduce preventive effort, to demand more medical services, and to opt for the newest medical technology, respectively, compared to the situation where the individual paid the full cost of his or her actions. Since these three effects are associated with differing costs and impacts on health, it would be valuable to distinguish them not only theoretically but also empirically.

However, the theory developed in Section 4 shows the response of preventive effort to health insurance coverage and sick leave pay to be ambiguous. Moreover, ample health insurance coverage serves to lower the average relative price of medical care, calculated over the universe of treatment alternatives. Thus, it is difficult to separate ex ante and ex post moral hazard empirically in the literature reviewed in this section.

Section 5 builds on the theory of decreasing multi-part tariffs to show how consumption of additional medical services during one sickness episode may lower their effective price during subsequent episodes. Results from the Health Insurance Experiment point to such anticipation effects only in the demand for dental care. Otherwise, the pure (within-block) price elasticity of the demand for medical care is around -0.2.

In Section 6, the case for dynamic moral hazard is examined. Opting for (more costly) new medical technology can in fact reduce the effective price of medical care during subsequent sickness episodes.

The contribution concludes with a few suggestions for future research (Section 7). In particular, the spreading of managed care calls for an integration of physician incentives in the determination of expenditure on medical care.

2. Dimensions of consumer incentives in health care

Knowledge of incentives is important for explaining and influencing individual behavior. Incentives derive from the interaction between objectives pursued and constraints to be observed. As to the objectives, health economics emphasizes the fact that individuals do not pursue good health only but also value other things in life (often simply called "consumption"). As to the constraints, the usual budget constraint in terms of income and prices often needs to be complemented by a time constraint. Seeking out and obtaining medical care costs time, which could be used for consumption and work and thus has an opportunity cost. Consumer incentives are reflected in decisions such as

- how much health insurance coverage to buy;
- to opt for a managed care plan rather than a fee-for-service plan;[1]
- how much preventive effort to exert;
- whether to initiate a treatment episode at all;
- to opt for school medicine or alternative medicine;
- to seek ambulatory or hospital care;
- to contact a public or a private provider;[2]
- which provider to choose.[3]

Clearly, at least some of these decisions are relevant in any health care system. Consumer incentives thus influence the size and composition of health care expenditure (HCE). In the following, the relationship between consumer incentives and individual HCE will be at the center of attention [for the determinants of aggregate HCE, see the chapter by Gerdtham and Jönsson (2000) in this Handbook]. This relationship is crucially shaped by health, sick leave, and disability insurance, private or social. This

[1] This choice is analyzed in the chapter by Glied (2000) in this Handbook.

[2] For some evidence regarding the relevance of this choice at the aggregate level, see McAvinchey and Yannopoulos (1993).

[3] With regard to hospitals, distance (which may be regarded as a proxy for time costs) consistently has been found to be the most important determinant of choice [see the survey by Porell and Adams (1995)].

statement need not be confined to those countries where HCE is predominantly financed through health insurance; even in countries characterized by a public health service, such as the National Health Service of Great Britain, insurance has an impact because a sick worker continues to receive an income in the guise of sick leave pay. Thus, insurance is important because it modifies the money price of medical care, the income of the insured, and the opportunity cost of time in the event of illness. The change in health behavior and health care consumption caused by insurance is called moral hazard. In the following, three distinctions will prove useful.

- *Ex ante vs. ex post moral hazard* [Ehrlich and Becker (1972)]: Ex ante moral hazard refers to the situation prior to the advent of illness. Here, the hypothesis is that through their preventive effort, their individuals have an influence on the probability of the occurrence of the loss. The presence of insurance coverage might undermine the individual's incentive to prevent such loss. Since this loss affects above all the health stock, over which the individual has some control over the long run, ex ante moral hazard is related to life-cycle behavior [see the chapter by Grossman (2000) in this Handbook]. Ex post moral hazard, by way of contrast, comes into play once the health loss has already occurred. At this stage, health insurance reduces the net money price of medical care, while sick leave pay reduces its opportunity cost in terms of time. Such a reduction may lead to increased use of health care or sick leave.
- *Sources of moral hazard*: While the emphasis of this chapter lies on consumer incentives and hence moral hazard effects on the insured patient, insurance also affects the behavior of agents acting on his behalf, in particular the physician. The volume and structure of services produced by hospitals and suppliers of medical innovation is likely to reflect insurance conditions, too (see below).
- *Static versus dynamic moral hazard*: Static moral hazard refers to incentives given medical technology. However, in medical care, there often is a choice between existing and new medical technology. To the extent that insurance gives access to the new technology on the same conditions as the old, it creates an incentive for the insured to ask for the latest technology, giving rise to dynamic moral hazard [Goddeeris (1984a, 1984b), Baumgardner (1991)].

Conclusion 1 *Consumer incentives in health care are reflected in a wide range of choices, many of which occur in both insurance- and tax-financed systems. To the extent that they are influenced by insurance, they are the source of moral hazard of different types.*

From a normative point of view, moral hazard can be argued to cause a negative externality to the extent that it causes the insurer (assumed unable to discriminate between moral hazard-prone and -resistant individuals) to increase premiums for everyone. Thus, moral hazard should be avoided. However, some amount of moral hazard may be deemed beneficial for two reasons. First, to the extent that physicians wield a collective monopoly, the quantity of medical care consumed falls short of the optimum. The increase in quantity caused by the moral hazard effect of health insurance can be efficiency-enhancing in this situation [Crew (1969)]. Second, moral hazard may encourage the use of a more cost-effective medical service at the expense of a less cost-effective one within an insurance scheme [Pauly and Held (1990)]. Thus, the optimal amount of moral hazard is positive rather than zero.

3. The amount of delegation of authority to the physician

There has been a debate in health economics about the importance of consumer incentives in the determination of volume and composition of HCE. At the one end of the spectrum, there is the theoretical construct of the physician-patient team, in which the physician, acting as a perfect agent, provides complete and unbiased medical information to the patient, who then decides in full sovereignty [Feldstein (1973); see Mooney and Ryan (1993) for the possible role of agency theory in modeling the physician–patient relationship]. While the physician-patient relationship is a black box, its observable output (the amount of HCE demanded in particular) would reflect consumer incentives under this construct. At the other end of the spectrum [Evans (1974)], the physician pursues his objectives, inducing demand for his services under fee-for-service payment and choking demand under capitation [see McGuire's (2000) chapter in this Handbook). In that case, observed HCE would mirror physician incentives, which would be to immunize patient demand from both money and time price, under fee-for-service payment. In the extreme, consumer incentives in health care would become a moot issue once the decision to initiate a treatment episode is made. In Section 5, an attempt will be made to integrate the physician's rationing function into the standard demand model; in empirical work, however, disentangling demand and supply side moral hazard has proved elusive.

At the heart of the debate, the degree of delegation of authority to the physician is at issue. While this delegation is usually viewed as a fact, its degree is a matter of choice, reflecting consumer incentives once again. There seem to be three factors influencing the degree of delegation, informational disadvantage, shifting of responsibility, and insurance coverage.

- *Informational disadvantage of the patient*: This is an important reason for the delegation of decision-making authority [for the general argument, see Holmström (1979)]. If the patient had the same medical knowledge as the physician, he could just have his orders executed. Clearly, the cost of gathering the relevant medical information is excessive for most patients. However, given a sufficiently high expected return, the patient may begin to bridge the information gap. This raises the issue of optimal (lack of) information on the consumer's part.

  In the context of ex ante moral hazard, an important parameter in the insured's decision with regard to preventive effort is its estimated marginal productivity, i.e. the reduction of the probability of ill health brought about. Most people know little about this parameter,[4] possibly because the expected marginal return to information about prevention falls short of its marginal cost.

[4] See Viscusi (1995) for a discussion of perception of risk and the effects of its changes on preventive effort.

- *Shifting of responsibility*: Even if well informed, the patient may wish to delegate a great deal of decision-making authority. Being ill often places quite a burden on others. Co-workers have to fill in during absence from the job, and family members are affected by the loss of labor income as well as the net cost of medical treatment. Individuals deciding about their illness and the necessary amount of medical care thus would become the source of a negative externality. By shifting the decision-making authority to the physician, they can exonerate themselves from any liability.
- *Insurance coverage*: Insurance coverage insulates the patient from the financial consequences of his choices, among them, the degree to which he delegates authority to the physician. Without insurance, the patient's willingness to engage in such delegation presumably would be reduced.[5]

Conclusion 2 *The degree of delegation of decision-making authority to the physician depends on consumer incentives. It importantly determines the degree to which observed utilization of medical care also reflects physician incentives.*

4. Incentives and ex ante moral hazard

4.1. Theoretical background

Individuals lack control over their health status to a great extent. Being sick or healthy in some future period depends on probabilities. However, this does not preclude one's capability of influencing, albeit marginally, the probability π of being ill by preventive effort V, with $\pi'(V) := \partial\pi/\partial V < 0$. While this effect is hardly noticeable at the individual level, it may cause the number of sickness episodes to differ a great deal in the aggregate. For example, in a population of 1 million, an increase of π by one percentage point translates into 10,000 additional episodes, which not only cause medical expenditure but also absence from work. Therefore, ex ante moral hazard may be of considerable relevance not only to insurers but to society in general.

For the sake of simplicity and without loss of generality, assume there is no ex post moral hazard (which will be introduced in Section 5). Accordingly, the individual considered faces an exogenous loss L in the case of illness, equivalent to the medical expenditure incurred. Labor supply W (zero in the sick state), the wage w, the amount of health insurance coverage I as well as sick pay S with their associated premiums P and R respectively have been determined previously. Therefore, the only ex ante decision variable remaining is preventive effort V, measured in time units. This formulation is an extension of the model used in Zweifel and Breyer (1997, Ch. 6.4).[6] This makes the

[5] Although this statement appears intuitively appealing, there seems to be little research into the relationship between insurance coverage and amount of delegation of authority by the patient.

[6] At this stage, the fact that the opportunity cost of time may vary between the two states is disregarded (but see Section 5.3 below).

cost of prevention wV regardless of whether the health risk materializes or not. Thus, under the expected utility hypothesis,[7] the objective would be to maximize[8]

$$\begin{aligned} EU(V) &= \pi(V) \cdot u[S - P - R - wV - L + I] \\ &\quad + \{1 - \pi(V)\} \cdot u\big[w(W - V) - P - R\big]. \end{aligned} \tag{1}$$

As a general rule, sick leave pay is proportional to labor income such that $S = (1 - s)wW$, with $(1 - s)$ denoting the replacement rate $[0 < (1 - s) \leqslant 1]$. Focusing on interior solutions $(V > 0)$ and denoting $y^s := (1 - s)wW - P - R - wV - L + I$ and $y^h := w(W - V) - P - R$, one has the first-order condition,

$$\begin{aligned} \frac{dEU}{dV} &= \pi'(V) \cdot u\big[y^s\big] - \pi(V) \cdot w \cdot u'\big[y^s\big] \\ &\quad - \pi'(V) \cdot u\big[y^h\big] - \{1 - \pi(V)\} \cdot w \cdot u'\big[y^h\big] = 0. \end{aligned} \tag{2}$$

Using the shorthand notation, $EU'(y) := \pi(V) \cdot u'[y^s] + \{1 - \pi(V)\} \cdot u'[y^h]$, Equation (2) can be written

$$\pi'(V)\big\{u[y^s] - u[y^h]\big\} = -w \cdot EU'(y). \tag{3}$$

On the left-hand side of this first-order condition is the expected marginal return from an additional unit of preventive effort. It amounts to the decreased probability of suffering the loss expressed in the brackets, which is the utility difference between the sick and the healthy state. Note that this difference depends on both health insurance benefits I and sick leave pay S. On the right-hand side is the marginal cost of prevention, given by the wage rate of the individual as a shadow price of his time and valued by the decrease of utility associated with additional preventive effort in both states of the world.

The importance of the opportunity cost of time can be illustrated by totally differentiating Equation (3) with respect to w. Since the impulse dw can be neutralized only by an adjustment dV in this model, one has

$$d\left[\frac{dEU}{dV}\right] = \frac{\partial^2 EU(V, w)}{\partial V^2} dV + \frac{\partial^2 EU(V, w)}{\partial V \partial w} dw = 0, \tag{4}$$

which can be solved to read

$$\frac{dV}{dw} = -\frac{\partial^2 EU(V, w)/\partial V \partial w}{\partial^2 EU(V, w)/\partial V^2} = -\frac{\partial EU_V/\partial w}{\partial EU_V/\partial V}. \tag{5}$$

[7] For some results that carry over to non-expected utility, see Machina (1989, 1995).

[8] In this model, utility depends exclusively on full income, i.e. wealth. For a more general formulation including medical care (derived from the demand for health) as an argument, see Section 5.1 below. Brackets symbolize the evaluation of the function at a particular value of its argument.

The sign of dV/dw depends only on the sign of the numerator of Equation (5), because the denominator must be negative by the second-order condition. Using Equation (3) and taking its derivative with regard to w, one has

$$\begin{aligned}\partial EU_V/\partial w &= \pi'(V)\big\{[(1-s)W-V]\cdot u'\big[y^s\big]-(W-V)\cdot u'\big[y^h\big]\big\}-EU'(y)\\ &\quad -w\big\{\pi\cdot[(1-s)W-V]\cdot u''\big[y^s\big]+(1-\pi)(W-V)\cdot u''\big[y^h\big]\big\}.\end{aligned} \tag{6}$$

The first term of Equation (6) reflects an income effect because an increase of the wage rate causes the income difference between the healthy and the sick state to change, with the change weighted by the shift of probability mass away from the sick state due to prevention. The second term is the substitution effect. It is unambiguously negative, indicating the increased opportunity cost of prevention that goes along with a higher wage rate. The third term reflects risk aversion, which becomes relevant because the change of the wage rate also affects the individual's expected wealth.

Now, even absent risk aversion, Equation (6) cannot be signed. With $u'[y^s] = u'[y^h] := u'$ and $u''[\cdot] = 0$, Equation (6) reduces to $\pi'(V)(-sW)\cdot u' - u'$, which may well be negative for small values of $|\pi'|$, W, and s. Thus, in particular for individuals with low marginal effectiveness of prevention, low labor supply and high replacement rate in sick leave pay (low s), $\partial EU_v/\partial w < 0$ and $\partial V/dw < 0$, indicating a moral hazard effect. With risk aversion, however, $u'[y^s] > u'[y^h]$, making the first term indeterminate; the third term turns positive as long as $(1-s)W > V$. Thus, $dV/dw > 0$ becomes a possibility. In a more general model with endogenous labor supply, the reduction in π afforded by V produces healthy time available for market work, which would then be transformed into labor income at a high rate if w is high. In such a model [see Grossman (2000)], a positive effect of w on V may obtain as well.

Interestingly enough, ambiguity also characterizes the reaction to an increase in the coverage of health insurance I. In full analogy to Equations (4) and (5), the sign of dV/dI depends on the sign of $\partial EU_V/\partial I$. Thus, one obtains from Equation (2), noting that the premium increases with benefits $[0 < P' := \partial P/\partial I < 1]$,

$$\begin{aligned}\frac{\partial EU_V}{\partial I} &= \pi'(V)\cdot\big\{\big(1-P'\big)u'\big[y^s\big]+P'\cdot u'\big[y^h\big]\big\}\\ &\quad -w\big\{\big(1-P'\big)\pi\cdot u''\big[y^s\big]-P'(1-\pi)u''\big[y^h\big]\big\}\lesseqgtr 0.\end{aligned} \tag{7}$$

The ambiguity arises because an increase in I, according to Equation (3), not only affects the utility difference on the left-hand side, but the expected value of marginal utility of income on the right-hand side as well. The somewhat counter-intuitive response $dV/dI > 0$ cannot be excluded among risk-averse high-wage individuals. Similar results can be shown to hold when the generosity of sick leave pay [indicated by $(1-s)$] increases.

Conclusion 3 *Ex ante moral hazard depends importantly on the opportunity cost of preventive effort, which in many instances is approximately proportional to the wage rate. However, the moral hazard effect may be neutralized by risk aversion, which makes the relationship between both health insurance coverage and sick leave pay ambiguous as well.*

The basic model in this section assumed that preventive effort was limited to the consumer's own time. If we extend this model to allow for the purchase of preventive health care services, then ex ante moral hazard also reflects the generosity of coverage of preventive care – the less paid out of pocket, the more will be demanded as a result of additional substitution effects. Nevertheless, this extension does not eliminate the kind of ambiguity present in Equation (7).

4.2. Empirical evidence

Empirical evidence has come from three sources, natural experiments involving an abrupt change of cost sharing, observational comparisons of individuals with differing insurance plans, and planned, randomized trials in insurance. Justification for planned trials will be provided in Section 5.2.3 below.

In order to test Conclusion 3, one would want to observe individuals with differing wages and reservation wages, health insurance plans of differing comprehensiveness, and sick leave benefits of differing generosity. The consumer's preventive effort could be measured, e.g., in terms of time spent on physical exercise or on preparing especially healthy meals. Alternatively, one might treat preventive effort as an unobserved quantity, using the likelihood of sickness π as the dependent variable. There seem to be two problems that cannot be solved easily regardless of the source of information.

(1) *Ambiguity of predictions*. As noted in the preceding section, preventive effort V may vary positively or negatively with respect to both the degree of coverage I and the wage rate w. Thus, even if the likely endogeneity of I w.r.t. wage income and hence w is corrected for in econometric work, the expected result from such a correction would be unclear.

(2) *High correlation between variables*. Theoretically, the completeness of coverage I needs to be distinguished from the generosity of insurance in the event of illness, which is given by kp, where $k := 1 - c$ is the rate of coverage, c is the rate of coinsurance, and p the (relative) price of medical treatment. However, a high value of I means that a larger subset of the universe of treatments profits from the generosity of insurance, causing I and average kp to move together. Especially when the consumer is uncertain about the physician's choice from the universe of treatments, the two variables become observationally very similar as determinants of HCE.

For these reasons, it has proved difficult to separate the effects of ex ante from ex post moral hazard. However, the response of preventive visits to consumer incentives may provide at least illustrative evidence.

Roddy et al. (1986) examined the United Mine Workers (UMW) experience by looking at the cost sharing period and a subsequent copayment period. The initial cost sharing period was characterized by an outpatient coinsurance rate of 40 per cent, a $250 inpatient deductible, and a family stop-loss of $500. It lasted until there was a UMW strike in December 1977, when the plan was suspended. Starting in March 1978, the UMW Health Plan switched to a $5 copayment for outpatient visits and a $5 copayment for prescription drugs, subject to a maximum out-of-pocket payment of $100 per family for visits and of $50 per family prescription drugs.[9] Roddy et al. compared the two periods post with the one year before the change – with the coinsurance period and the copayment period. They found that overall visit rates returned to near baseline measures by the end of the second period. However, preventive visits fell by 25 percent in the coinsurance period (relative to the pre period), and by 28 percent in the copayment period.

Lillard et al. (1986) reported estimates from the Health Insurance Experiment of the effect of cost sharing on preventive and nonpreventive visits, with preventive visits including immunizations, screening examinations and well-care. Preventive services were significantly related to family cost sharing and individual deductible plans, but the magnitude of the response was less than for general ambulatory care. For example, they found a reduction of 10 percent ($p = 0.04$) in preventive visits in going from the free plan (no out-of-pocket payment) to the 25 percent plan, compared to an 18 percent ($p < 0.01$) reduction for non-preventive visits. For the 95 percent plan the reductions were 29 percent for preventive care and 38 percent for non-preventive care visits; both results are significantly different from zero at $p < 0.01$, but not significantly different from each other. In each case, the estimate is the response to the insurance plan as a whole, not just the plan's coinsurance rate. Even with free care, all young children (age < 7) had not received recommended levels of screening, nor adults the levels of screening and physical examinations.[10] For example, with free care, only 59 percent of all children under 7 had any immunization in the three years of data studied, while 83 percent had any preventive care. The family coinsurance plans as a group had 49 and 74 percent, respectively of any immunization and any preventive care in three years. In principle, all of these children should have had at least one immunization and a well care examination.

Keeler and Rolph (1988) also reported estimates of the pure price elasticity for different types of care in the Health Insurance Experiment.[11] They found that the arc price elasticity of well care was on the order of -0.14 between the free and 25 percent plan, compared to -0.17 for total outpatient and -0.17 for inpatient care. However, in the range of the 25 to 95 percent plans, the arc elasticities for well care were larger than for total outpatient care, -0.43 versus -0.31. They provided no formal test of the differences in price response by type of service. However, the proportion of outpatient

[9] The current dollar figures would be 2.57 times higher if corrected using the US All Items Index.

[10] The recommendations were based on recommendations from the Canadian task force on preventive care, the American Cancer Society, the American College of Physicians, and the American Academy of Pediatrics.

[11] The pure price elasticity is the elasticity for the price response in the absence of deductibles or stop losses. See Section 5.1.2 below for additional details.

episodes that are for well care is not significantly related to insurance plan, suggesting a similar plan response for well care and other outpatient health care.

Cherkin et al. (1990) also looked at preventive care in their study based on data provided by Group Health Cooperative. The introduction of the \$5 copayment in July of 1985 (approximately \$7.60 in 1998 dollars) reduced physical examinations by 14 percent, which was larger than the 11 percent decrease in overall visits. There was no evidence of an effect on immunization rates for young children, cancer screening by women, or medication use by patients with cardiovascular disease. However, the lack of a statistically significant effect in some of these cases may be due to lack of precision.

Clearly, these studies reflect variations in the net price cp rather than the general benefit level I or the role of time costs. Moreover, preventive visits may serve as a substitute for patient's own preventive effort. For these reasons, the theoretical ambiguity noted in Conclusion 3 cannot be considered resolved based on these studies, which focused on out-of-pocket money prices.

Conclusion 4 *The limited available evidence indicates that the demand for preventive care services is a declining function of out-of-pocket money price. Some of the evidence suggests that preventive care is more responsive to price than is the demand for other medical services.*

5. Incentives and static ex post moral hazard

5.1. Theoretical background

5.1.1. Effects of insurance plan generosity

As before, only two states of the world are distinguished. Thus, with probability π, the insured will find himself ill, receiving net income y^s and thus subject to the constraint,

$$\pi: \qquad y^s := \big(Y_0 + S + I(pM) - P - R - pM\big) = C^s. \tag{8}$$

With probability $(1 - \pi)$, the insured is healthy with net income y^h and facing the constraint,

$$(1-\pi): \quad y^h := Y_0 + wW - P - R = C^h. \tag{9}$$

Net income y^s is composed of nonlabor income Y_0 and sick leave pay S, from which must be paid premiums for health insurance P and for disability insurance R as well as medical expenditure pM [the loss L of Equation (1)]. Here, p symbolizes the relative price of medical care, while M stands for its quantity (such as physician visits or hospital days). Health insurance benefits I depend on medical expenditure, too. The net income must be sufficient to finance consumption in the sick state, given by C^s, with the price of the consumption good normalized to one in view of the homogeneity of demand of degree zero in prices and income. With probability $1 - \pi$, labor income wW is earned, sufficient to cover premiums and consumption expenditure.

With regard to preferences, the purely financial model of Section 4.1 is now augmented to include medical care M, which can be viewed as derived from the underlying demand for health [however, see Grossman's (2000) chapter in this Handbook for the precise nature of this relationship]. Under the assumption that the individual is an expected utility maximizer, his decision problem may be formulated as

$$EU(M, C^s, C^h, I, S) \to \max . \tag{10}$$

Since in this context, the sickness episode will constitute the principal reference period for decision making, it makes sense to regard both benefits from sick leave (S) and health insurance (I) as predetermined. Conversely, the amount of medical care (M) demanded is assumed to be under the control of the individual, implying that his delegation of authority to the physician is less than complete (see Section 3). Thus the argument focuses on only three decision variables, M, C^s and C^h. In the event of illness ($\pi = 1$), the ex post decision problem boils down to

$$U(M, C^s) \to \max . \tag{11}$$

The point of prime importance is the functional relationship between insurance benefits I and medical care expenditure pM. Usually, a health insurance policy contains a deductible D and a rate of coinsurance, c. Imposing a deductible can be viewed as a second-best solution to the problem of recovering administrative cost. Absent moral hazard, it will be optimal for the insured to shift the risk entirely upon the (risk-neutral) insurer; a policy with $c = 0$ would be of the stop-loss type [Arrow (1963)]. However, as soon as there is moral hazard, the insured usually would choose to impose a self-binding constraint on himself in order to save on premiums (assuming that premiums are calculated as to recover expected cost). Thus, a typical benefit function might look like

$$I(pM) = \begin{cases} k \cdot p \cdot M - D & \text{for } pM > D \\ 0 & \text{for } pM \leqslant D \end{cases} \quad \text{with } k = 1 - c,\ 0 < c \leqslant 1. \tag{12}$$

Substitution of (12) into (8) and collecting terms in M shows that the marginal price of medical care is $(1 - k)p = cp$ for $pM > D$. The upper budget constraint in Figure 1 illustrates. Up to the amount of the deductible D, the insured faces the full relative cost of medical care, given by p. For HCE in excess of D (or M in excess of D/p, respectively), coinsurance on HCE must still be paid, giving rise to the kinked budget constraint *EFG*. The graph shows at once a problem expounded by Newhouse (1978) and also known from the literature on decreasing block pricing [as, e.g., in electricity Taylor (1975)] or progressive income taxation. While unique optima are certainly possible, there may be cases where the consumer is indifferent between a low alternative such as C_1^* and a high alternative such as C_1^{**}. On the other hand, a value in the neighborhood of the kink should obtain only under very restrictive assumptions and with a vanishingly small probability [see Keeler et al. (1977) for a formula].

Now the generosity of a typical health insurance plan can be described and its effects derived as follows, assuming that health care is not a Giffen good.

- *Low deductible D*. The lower D, the more generous the insurance plan because the price reduction from p to cp occurs at a lower value of HCE. In terms of Figure 1, a lower D would make the kink move up along *EF*, causing *FG* to shift outward. Therefore, the more elaborate treatment alternative C_1^{**} (on the shifted segment FG) would no longer be equivalent to the more modest C_1^* but dominate it. Thus, reducing the deductible should cause the amount of medical care demanded to increase, ceteris paribus. For those above the deductible, a reduction in the deductible acts like an increase in income. For this group, a change in the deductible generally has the opposite effect of a change in disposable income. A reduction in the deductible will lead to some individuals shifting from the full price segment (*HK* or *EF*) to the partial price segment (*KL* or *FG*) because their highest point of tangency is now on that segment. For more on this, see the discussion of pure price effects in Section 5.1.3 below. Except for the income effect and this group of switchers a change in the deductible has no further effect on those above the deductible.
- *Low rate of coinsurance c*. The lower c, the more generous the insurance plan because the price reduction from p to cp is more marked. In terms of Figure 1, a lower c would make *FG* rotate outward, causing C_1^{**} (not shown) again to dominate C_1^*. In the extreme, a stop-loss plan would make *FG* run horizontal, with the consequence that the optimal quantity of care is unlimited as long as non-satiation holds.[12] Reducing the coinsurance should cause the amount of medical care demanded to increase. A reduction in the coinsurance rate will also lead some consumers to switch from consuming below the deductible to above the deductible. Except for these cases, a reduction in the coinsurance rate has no effect on the behavior of those below the deductible.

Also note that the location of the budget constraint depends on premia $\{P, R\}$ and income components $\{Y_0, S\}$. In particular, the budget constraint *EFG* of Figure 1 reflects the constraint (8), which obtains in the sick state. Especially in the case of a minor illness, however, the individual may decide to refrain from initiating a treatment episode and to continue work. This would not put the individual at point E but at point E' because constraint (9) applies, with $y^h > y^s$. The distance between E' and E (with $M = 0$) is given by

$$y^h - y^s = wW - S = C^h - C^s; \tag{13}$$

thus, it is the smaller, the greater sick leave pay S. However, the closer E' to E in Figure 1, the less likely is E' to dominate any point on *EFG*. Therefore, the more generous sick leave pay, the smaller the insured's tendency to forego medical care altogether

[12] Rationing by the physician could solve this problem (see Section 5.4 below).

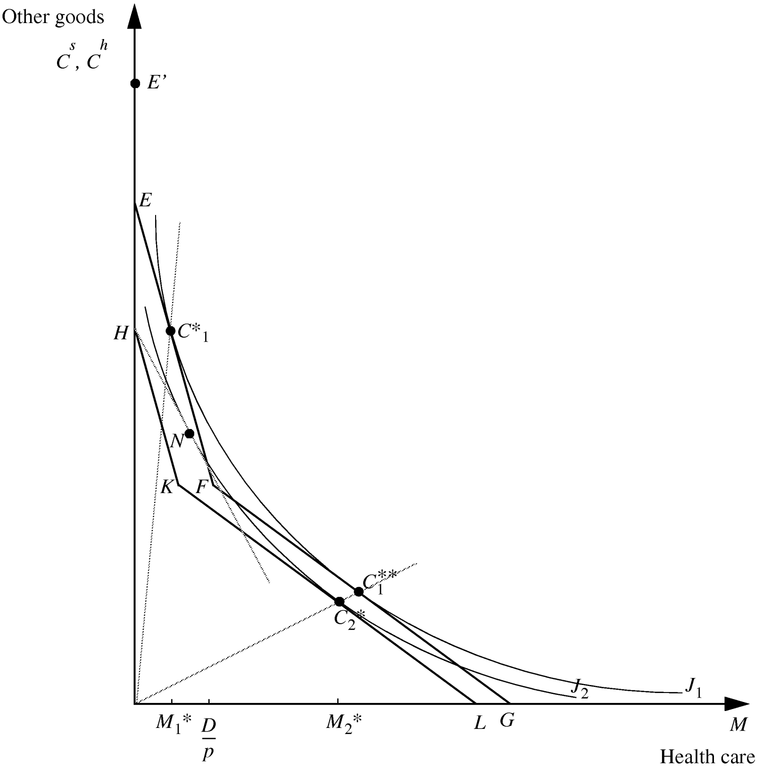


Figure 1. The effect of a deductible in a two-episode model.

and to continue work. Conversely, generous sick leave pay increases the propensity to trigger a sickness episode in the event of illness.

Conclusion 5 *The more generous insurance coverage (low deductible and rate of coinsurance in health benefits, ample sick leave pay), the larger the amount of medical care demanded as long as budget shares of premiums and out-of-pocket payments are small.*

5.1.2. The combined effect of deductible and coinsurance on demand

In this section, the interplay between deductible and rate of coinsurance in the determination of the demand for medical care is examined in greater detail. Frequently, health insurance policies do not specify a deductible per sickness episode but rather on an annual basis. Since decision-making with regard to medical care does not match the constraint imposed by the health insurance policy, measuring consumer response with regard to the coinsurance parameter (or net price) becomes a problem. As shown formally by Keeler et al. (1977), the fact that some HCE was utilized in the first period opens up the possibility of jumping beyond the kink in one of the following periods within the same accounting period (e.g., a year). They show using simulations how the net effective price of medical care decreases as a function of the number of standardized visits planned in the future.

Rather than writing out the full dynamic optimization problem, a two-period formulation is proposed here that still captures the essence of the model. Denoting by M_1 and M_2 the quantities of medical services consumed in periods 1 and 2, respectively, one has for period 2 benefits

$$I_2 = \begin{cases} 0 & \text{if } p(M_1 + M_2) \leqslant D \\ k\{p(M_1 + M_2) - D\} = kpM_2 - k \cdot (D - pM_1) & \text{if } p(M_1 + M_2) > D. \end{cases} \tag{14}$$

Thus, as long as the accumulated HCE falls short of D, there are no benefits from health insurance. On the other hand, if it exceeds D, the share k of the excess is covered. The rewriting of the second condition in terms of $(D - pM_1)$ indicates how period 1 HCE serves to lower the net price of additional medical care. Referring back to Figure 1, let the individual plan aggregate consumption only, deferred to period 2 for simplicity so that period 2 income depends on period 1 HCE only. Let the individual face exactly the same health problem a second time, reflected by the indifference curve J_2 running homothetic to J_1. On the constraint side, M_1^* units of medical care had to be paid out of pocket in the first period, resulting in a parallel inward shift of the constraint to *HKL*. The kink occurs now $(D/p)M_1^*$ units sooner than in period 1. Otherwise, the two budget constraints run parallel.

It can be easily gleaned from Figure 1 that while in period 1 the high and low cost treatment alternatives were equivalent by construction, now the high alternative C_2^* dominates because indifference curve J_2 does not ever touch the lower budget constraint, given homotheticity. Thus, while in period 1 the consumer was indifferent between high and low alternatives, he chooses the high alternative in period 2, in spite of the fact that disposable income has already been reduced by out-of-pocket payment. This means there must have been a reduction of the effective net price. In fact, this reduction is shown by the line *HN*, which indicates the relative price of medical care that allows the individual to reach indifference curve J_2, in spite of the fact that he is limited by the period 2 budget constraint.

Now let the rate of coinsurance go to zero. This would make the lines *FG* and *KL* in Figure 1 run flatter, coinciding more and more. Accordingly, the line *HN* would have to rotate upward, indicating a further reduction of the effective price of medical care in period 2.

This effect is depicted by Figure 2, taken from Keeler et al. (1977). With no coinsurance ($c = 0$), total demand aggregated over the two periods (more generally, several periods) would be at a maximum. On the other hand, if the coinsurance rate were $c = 1$, there would be no kinks at points F and K in Figure 1, causing *HN* to rotate downward

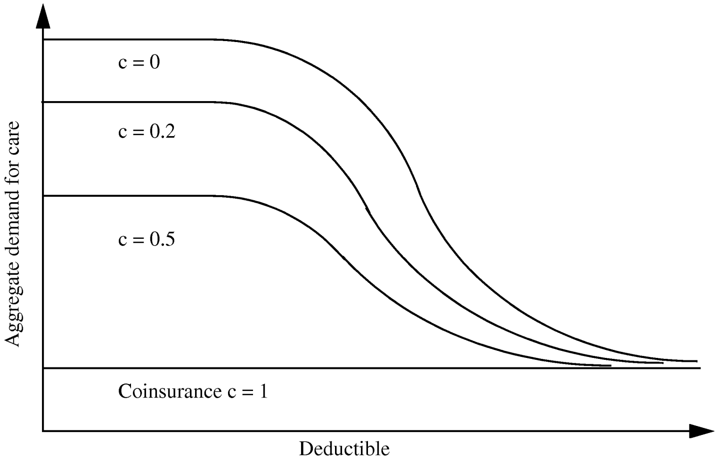


Figure 2. Aggregate demand for care as a function of deductible and rate of coinsurance. Source: Keeler et al. (1977).

and indicating a higher effective price. Thus, the reduction of effective price for medical care beyond the deductible varies inversely with the coinsurance rate.

Alternatively, let the deductible decrease. This would serve to shorten lines *EF* and *HK* in Figure 1, causing both optima C_1^{**} and C_2^* to shift outward along with the respective indifference curves. Therefore, the reduction in the price of medical care permitting the individual to still attain the shifted J_2 indifference curve would have to be greater (*HN* would have to rotate upward). Conversely, a higher deductible reduces the contribution the consumption of medical care in period 1 makes towards exhaustion of the deductible, thus causing the effective price of medical care to approach the price without insurance.

Again, this prediction corresponds with the one derived from the dynamic optimization framework of Keeler et al. (1977). Reading Figure 2 vertically, one sees that the aggregate demand for care depends very much on the net price of medical care (i.e. the coinsurance rate) as long as the deductible is low. It depends much less on the net price when the deductible is high.

Up to this point, the argument has been couched in terms of complete certainty. Thus, the utility gain from being able to attain point C_2^* in the second period (see Figure 1 again) rather than some point along *HK* accrued to the individual with certainty. Under uncertainty, it accrues to the consumer with a certain probability only, weakening his tendency to prefer the high option C_2^* over the low one. Finally, anticipated HCE may extend over several future periods rather than being concentrated in period 2.

Conclusion 6 *The fact that the natural decision period is the sickness episode, which often is shorter than the period relevant for the terms of health insurance (in particular, the deductible), causes the effective price of medical care in later periods to depend* *on medical consumption in previous periods and the probability of exceeding the deductible (or other internal limits) in the remaining part of the accounting period. The demand for medical services aggregated over the periods and its reaction to the net price needs to be conditioned on the value of the deductible.*

5.1.3. Pure price effects

The decision problem may also be couched in the following terms. Consider a consumer facing a simple health insurance policy for medical care (M) at gross price p, who pays p_1 per unit for the first M_0 units, and p_2 ($<p_1$) for each succeeding unit of the medical care good M. In the case of a health insurance policy with a deductible followed by free care, $p_1 = p$ and $p_2 = 0$. If the policy has a first dollar coinsurance rate c, followed by free care after meeting a catastrophic cap on out-of-pocket expenditures, then $p_1 = cp$ and $p_2 = 0$ [see Manning and Marquis (1996)]. However, especially with regard to pharmaceuticals, the policy may specify $p_1 = 0$ up to some value of pM or p and $p_2 = p$ beyond. The first rule reflects a Maximum Allowable Cost scheme, the second, a reference price scheme [for a description of the latter, which characterizes German social health insurance, see Zweifel and Crivelli (1996)].

The consumer has income $Y^s := S - P - R$ which is spent on medical care (M) and a composite all other good (C^s). The consumer maximizes his utility U, which has all the usual properties, subject to the budget constraint imposed by his income and the two-block insurance policy.

We will examine the consumer's choice using the indirect utility function $IU = f(p, Y^s, z, \Theta)$, where z is a vector of observed shift variables (e.g., self-reported health status) and Θ is an unobserved shift parameter (e.g., unmeasured sickliness). In the face of a two-block tariff, the individual chooses the block and consumption bundle (M, C^s) which gives him the higher indirect utility. With a constant price p_1 and income Y^s, the individual would have maximum indirect utility $IU_1^* = f(p_1, Y_1^s; z, \Theta)$ and demand $M_1^* = g(p_1, Y^s; z, \Theta)$. With constant price p_2 and net income: $Y_2^s = Y_1^s - (p_1 - p_2)M_0 < Y_1^s$, the individual would have maximum indirect utility $IU_2^* = f(p_2, Y_2^s; z, \Theta)$ and demand $M_2^* = g(p_2, Y_2^s; z, \Theta)$.

If individuals are economically rational, then they choose to consume on the block with the higher indirect utility. That is, they choose $M_2 > M_0 > M_1$ if and only if $IU_2^* > IU_1^*$. For simplicity, we can ignore the fact that there may exist some combinations of $p_1, p_2, Y_1^s(P, R),\ z$, and M_0 such that the individual is indifferent between the two blocks of a declining-block tariff. As a practical matter, such an event should occur with probability zero if the distribution of unobserved characteristics Θ has a continuous density function.[13]

[13] In contrast, for an increasing block tariff (such as mental health insurance coverage policies, which typically have a limit on covered expenditures or visits), it is possible that a non-trivial proportion of the cases may choose to consume at the kink in the budget constraint between the two blocks [see Burtless and Hausman (1978)].

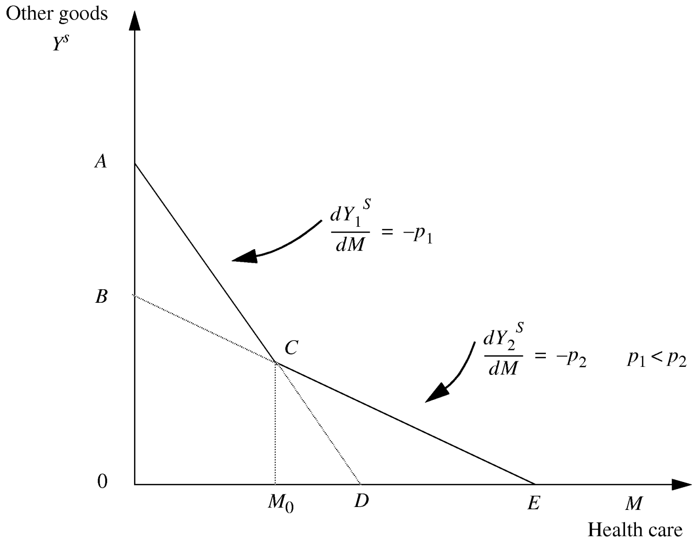


Figure 3. Budget constraint for insurance policy.

Figure 3 illustrates the choice problem assuming that the premiums $\{P, R\}$ are zero. *ACE* is the declining block budget constraint. *AD* is the constant price budget constraint for price p_1 and income Y_1^s, while *BE* is the corresponding constant price budget constraint for price p_2 and income Y_2^s. Any tangency of the indifference map along *AC* provides higher satisfaction than a tangency along *BC* by non-satiation. Similarly, any tangency along *CE* will dominate a tangency along *CD*. Hence, the choice between M_1 and M_2 always involves a choice between two alternatives that are affordable (e.g., within or on the budget constraint) given by *ACE*.

As long as the individual is not near a tangency on both segments of the budget constraint *ACE*, we can use Roy's Identity to examine the effect of a change in the out-of-pocket price on demand. For an optimum on segment *AC*, one has

$$M_1^*(p_1, Y_1^s; z, \Theta) = -\frac{\partial IU(p_1, Y_1^s; z, \Theta)/\partial p_1}{\partial IU(p_1, Y_1^s; z, \Theta)/\partial Y_s^1}. \tag{15}$$

The standard comparative statics apply as long as a fall in p_2 is not sufficient to switch an individual from a choice at price p_1 to one at p_2.

The amount of medical care chosen can also be examined by using consumer surplus to approximate the compensating variation for the arrangement in Figure 4.[14] The

[14] As long as the budget shares are small and the income elasticity is low, this approximation should be very good [see Willig (1976) and Hausman (1981)].

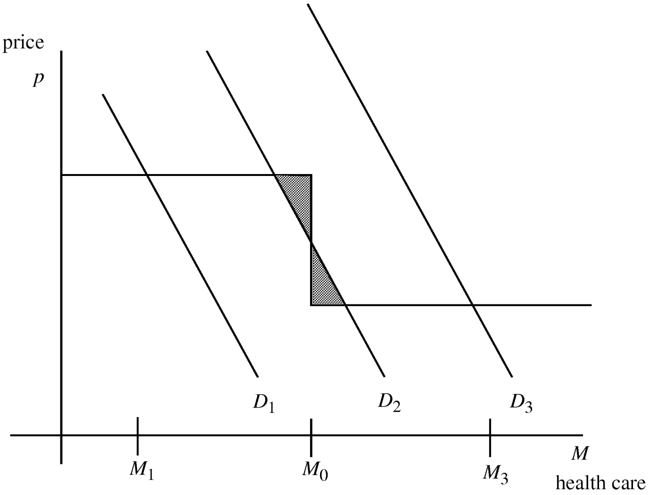


Figure 4. Consumer facing insurance policy.

demand curves $\{D_1, D_2, D_3\}$ are conventional demand curves for an individual facing constant prices p and income Y^s (but with differing values for $\{z, \Theta\}$). The declining block tariff has price p_1 up to quantity M_0 (which corresponds to the line segment AC in Figure 3), and price p_2 beyond (corresponding to segment BC). If the demand curve is similar to D_1 or to D_3, then the choice is obvious. The patient has the higher consumer surplus consuming at M_1 and at M_3, respectively. However, if the patient's demand curve D_2 cuts both sections of the price schedule then the choice is more complex. The gains from exceeding the internal limit M_0 and having some care at the lower price p_2 is the lower shaded triangle, while the cost of paying more for some infra-marginal units than they are worth is the upper shaded triangle. As long as the lower triangle is larger than the upper one, the patient will consume on the second part of the insurance plan, at an out-of-pocket price of p_2, along segment BC.

Conclusion 7 *The higher the deductible, the coinsurance rate or the gross price of care, the lower the consumption of health care.*

In the discussion so far, we have treated health care as a good or service which consumers value directly. As Grossman (1972) and others have noted, consumers value health, with health care merely a means to producing health or slowing its decline. If health status H is produced using medical care M, then we can rewrite the utility function in Equation (11) as

$$U\big(M, C^s\big) = u\big(H(M; H_0), C^s\big), \tag{16}$$

where H_0 is initial health status. As long as the health production function $H(M)$ has the usual properties, then all of the preceding results go through. Further, the solution to the first-order conditions also yields a demand function where optimal health is a declining function of the out-of-pocket price of health care.[15]

Up to this point, static ex post moral hazard effects were derived exclusively with respect to money price. Time cost, which figured prominently in ex ante moral hazard (see Section 4.1 above) will be reconsidered and the health production aspects will be further developed in Section 5.3 below.

5.2. Empirical evidence

There is substantial evidence in the literature that consumers use less health care when the level of cost sharing or copayment by the patient increases.[16] In the following section, we explore three types of evidence for medical care.[17] The first of these is from natural experiments, where there was some abrupt change in the level of copayment or coinsurance. The second is from observational comparisons of individuals with more or less cost sharing or copayment in their health insurance plan. The third is evidence from a randomized trial of insurance. Most of the discussion is focused on the effects of insurance plans on utilization or expenditures of health care; however, some also relates to the impact on health status and functioning.

5.2.1. Natural experiments

Connecticut Study. Heaney and Riedel (1970) report the results of a change in inpatient coverage. Before 1966, Blue Cross paid $15 per day to room and board charges and fully covered ancillary charges under an indemnity plan. Starting in 1966, the plan offered some group insurers the option of full coverage. Given prevailing charges for Connecticut hospitals, Phelps and Newhouse (1974) estimated that this change corresponds to a shift from a 31 percent coinsurance rate to a 0 coinsurance rate. The rate of admissions rose by 12 percent and the average length of stay by 12 percent. Data from a “control” group suggest that this before and after estimate may understate the change.

[15] This can be seen by noting that $\partial H^*/\partial p = (\partial H^*/\partial M^*) \cdot (\partial M^*/\partial p)$, where H^* and M^* are the values of health and health care that maximize utility such that $H^* = H[M^*]$. As long as $\partial H^*/\partial M^* > 0$, $\operatorname{sign}(\partial H^*/\partial p = \operatorname{sign}(\partial M^*/\partial p)$. Similarly the income response is $\partial H^*/\partial Y = (\partial H^*/\partial M^*) \cdot (\partial M^*/\partial Y)$, with $\operatorname{sign}(\partial H^*/\partial Y) = \operatorname{sign}(\partial M^*/\partial Y)$.

[16] The literature has several reviews on the effect of insurance in health care prices on the use of and expenditures for health care services. Newhouse (1978, 1980) and van de Ven (1983) review the early literature in the field. Broyles and Rosko (1988), and Rice and Morrison (1994) provide reviews of the more recent literature. A report from the US Office of Technology Assessment (1992) examines the effect of any insurance on access, utilization of health care, and health outcomes.

[17] See the chapters by Frank and McGuire (2000) for mental health and by Sintonen and Linnosmaa (2000) for dental care in this Handbook.

The Stanford University Studies. Scitovsky and Snyder (1972) examined data from a natural experiment in cost-sharing for physician services. In April 1967, there was a change in the coinsurance rate from 0 (free care) to 0.25 for faculty and staff at Stanford University who were receiving care through the Palo Alto Medical Clinic. Other insurance provisions (e.g. inpatient care) were not changed. Using data from 1966 and 1968, the authors found a 24 percent decline in physician visits, and 25 percent if age adjusted (the post sample was two years older). Although there were some shifts in the composition of care, expenditures on physician services also fell by 24 percent. Thus, the major change occurred in the quantity of care, not its composition. The results were similar by age, gender, and occupation. The arc price elasticity for this shift is -0.14. Follow-up data from 1972 indicate that the reduction in physician use was permanent, and not just a transient response to the change in coverage [Scitovsky and McCall (1977)].

There were a number of concerns with the Stanford studies. One was the simplicity of the analysis, but a subsequent re-analysis by Phelps and Newhouse (1972) indicated that the use of more sophisticated econometric methods would not alter the estimates. There was a concern that if the change in coverage was confounded with any other changes (a year of unusually high colds or flu), then the estimates could be biased depending on the event and its timing. However, there was no evidence found of unusual other changes confounded with coinsurance. The persistence of the change reduced some of these concerns. There had been some concern expressed about unmeasured shifts in out-of-plan use. But as the original investigators noted, why would patients go out-of-plan to pay 100 percent to avoid 25 percent of the bill? Finally, there was some concern about the generalizability of the results given the population studied and the fact that patients were receiving care through a large multi-speciality clinic.

Saskatchewan. The province of Saskatchewan had introduced a universal health insurance plan in the early 1960's. In 1968, the provincial health plan added a copayment of Can \$1.50 for doctor visits and a Can \$2 for home visits. Using data from 1963–1968 on nearly 40,000 individuals in 21,900 households drawn at random, Beck (1974) examined the impact of the introduction of copayment on utilization of physician services. He found a 6 to 7 percent drop in all physician services, and an 18 percent drop among the poor. The largest decreases were in general physician services. There was no significant reduction in services provided by specialists. One concern with this study has been that a province-wide change in copayment could elicit a supply side response, as well as a demand side response.[18] If there was such a response, then the demand responsiveness could be even larger.

[18] The same concern has been raised about the studies by Enterline et al. (1973) and Ricci et al. (1978) of the introduction of Medicare in Quebec in 1970. Although the number of visits did not change with the introduction of universal insurance, visits by the poor increased by 18 percent. Waiting times to get an appointment nearly doubled and waits in the office increased by ten percent. Thus, some of the rationing by price was replaced by non-price, time rationing.

United Mine Workers. In July 1977, the United Mine Workers (UMW) Health Plan introduced both inpatient and outpatient cost sharing for their members and their dependents. Before then, the plan had first dollar coverage. The cost sharing included a 40 percent coinsurance rate for outpatient care, a $250 inpatient deductible, with a family maximum of $500.[19] Scheffler (1984) used a pre-post design, with the five months before the change being the pre or no cost-sharing period, and the five months after being the post or cost-sharing period. The results indicated that there was a 28 percent reduction in outpatient visits and a 38 percent reduction in expenditures. The probability of having a hospitalization fell by over a third. However, these results may overstate the response to a permanent change in out-of-pocket expenses. To the degree that the change in cost sharing was anticipated, the short-term response may well exceed the long-term response if individuals were able to defer any discretionary use.

MediCal Studies. There have been a number of natural experiments in cost sharing or copayment for the poor enrolled in Medicaid. The most notable of these is California's introduction in January 1972 of a copayment of a $1 for each of the first two visits and $0.50 for each of the first two prescriptions filled for its Medicaid population.[20] Using aggregated quarterly data on those enrollees who were not subjected to the copayment and those who were, Roemer et al. (1975) reported that individuals with copayment had visits decline by four percent more than it did for those without copayment, while hospitalizations increased more for those subject to copayment than for those not subject to copayment. After adjustment for observed differences in the populations covered, it appears that there was a decrease of 8 percent in visits and an increase of 17 percent in hospitalizations, after some adjustment [Helms, Newhouse and Phelps (1978)]. Unfortunately, the copayment was levied only on those Medicaid enrollees with some income or property, thus confounding population covered and the price change; see Helms et al. for further details. To further complicate matters, the state had implemented a system of prior authorization for outpatient visits beyond two per month and for non-emergency inpatient care just before the beginning of the period of the price change.

In 1982, the state of California terminated medically indigent adults from the state's Medicaid program (known as Medi-Cal). A medically needy individual is one who is covered because of medical or economic need, but not eligible for one of the federal categorical aid programs [AFDC (Aid for Families with Dependent Children), aged, blind, or disabled]. Thus, this was a group that was sicker than average at the time of the change. After termination, these individuals had to rely on county facilities or charity care. In Los Angeles, the county facilities typically charged $20 to $30 for an outpatient or emergency room visit ($34 to $51 in 1998 dollars). The newly terminated did not have this fee waived. In two studies, Lurie et al. (1984, 1986) examined the

[19] These dollar amounts would be about 2.6 times larger in 1998 dollars.

[20] Medicaid is known as MediCal in California. The dollar amounts stated would be 4 times larger in 1998 dollars.

impact of this loss of health insurance on a group of individuals and compared them with their pre-termination status or a control group of Medi-Cal patients not terminated but in the same practices. Outpatient visits fell by 45 percent in the first six months after termination, and were 35 percent lower in the next six month period. General health status fell by 8 points from a baseline of 47 on a 100 point scale in the first six months and by another 2.4 points by twelve months. Diastolic blood pressure rose by 10 mm Hg. There were 7 deaths out of 186, compared to a control group which had 1 out of 109.

GHC Studies. In July 1985, Group Health Cooperative of Puget Sound (GHC)[21] began charging some of its enrollees a $5 copayment for each outpatient visit and $25 for each emergency room visit[22]. Before that time, there had been no out-of-pocket charge for visits. The population affected was employees of Washington state, and their dependents ($n = 30{,}414$). Using a control group of federal government employees and their dependents ($n = 21{,}633$), Cherkin et al. (1989) did a pre-post evaluation with a contemporaneous control group.[23] They found that there was an 8.3 percent drop in visits, mostly due to 10.9 percent drop in primary care visits; both are significant at $p < 0.001$. There were no differences in the response to copayment by age or gender, separately, but there was a larger effect for female children and younger women than for comparable aged males. The drop in visits was a larger proportion for high users (9 or more visits in the prior year) than for low users. The effect of copayment appeared to be immediate (16 percent in the first quarter) and to be sustained throughout the year after implementation (8 to 10 percent in subsequent quarters). Using income estimates based on census tract, they found no differences by income group [Cherkin et al. (1990)].

In July of 1983, GHC began to charge a $1.50 copayment for prescriptions to Washington state employees and their dependents, after having not charged them before. In July 1984, the copayment was raised to $3.00. In July 1985, a $5 copayment for office visits, and a $25 copayment for emergency room visits was introduced and coverage for certain over-the-counter drugs dropped.[24] Harris et al. (1990) conducted a pre-post comparison using a control group from large employer contracts that had no copayment throughout the study period. The introduction of copayment ($1.50) lead to a 10.7 percent decrease in the number of prescriptions, while the second raise in copayment led to an additional reduction of 10.6 percent ($p < 0.0001$). The introduction of the visit copayment and the change in the coverage of the $3 prescription copayment led to a 12 percent reduction in prescription use. The effects on drug costs were somewhat

[21] GHC is a large, staff model HMO located in Seattle, Washington.

[22] The copayment did not apply to mental health, radiology, pathology, or injection visits. Adjusted for inflation using the US Consumer Price Index, a $5 copayment in 1985 would be about $7.60 in 1998.

[23] Surveys of a subset of these two populations indicated that they were quite similar in age, gender, health status, income, and family composition. More of the state than federal employees had graduate education (19 vs. 11 percent).

[24] The 1985 dollar amounts would be about $8.20 and $41 in 1998 dollars.

lower, but still significant 6.7 percent reduction, another 5.2 percent reduction, and an 8.8 percent reduction, respectively.

Kaiser Permanente. Selby et al. (1996) examined the effect of the introduction of copayments of \$25 to \$35 for emergency room (ER) use at Kaiser, a group model HMO. Using a before and after comparison on non-elderly enrollees, they found that ER visits fell by about 15 percent relative to two control groups drawn from Kaiser enrollees. There was no statistically significant evidence of an effect on conditions that were classified as always an emergency, but declined noticeably for conditions that were less likely to be an emergency.

Taiwan. Cheng and Chiang (1997) report the results of a before and after study of the introduction of national health insurance in Taiwan in 1995. National health insurance reduced outpatient copayments to NT\$ 100–200 (roughly US\$ 4–8 at then prevailing exchange rates), and inpatient coinsurance to 10 percent. Using two-week recall by a sample of individuals interviewed twice, they found that the likelihood of utilization by the previously uninsured more than doubled for both outpatient visits and inpatient stays (0.21 vs. 0.48 for outpatient visits and 0.04 vs. 0.11 for inpatient stays, both $p < 0.05$).

5.2.2. Observational comparisons of individuals

Rosett and Huang (1973) reported results from an analysis of the 1960 Survey of Consumer Expenditures by the Bureau of Labor Statistics. Their estimates of the price elasticity varied with the level of out-of-pocket price. With an out-of-pocket price that was 20 percent of the market price, they estimated the price elasticity to be -0.35, while at 80 percent of market price, it is -1.5. One possible reason for the higher price elasticities is their construction of coinsurance as out-of-pocket spending divided by total expenditures. Large expenditures include inpatient care, which was fairly well covered, while low expenditures were largely outpatient care. Even if the true price elasticity were zero, this construction of a coinsurance variable would generate a very price responsive estimate.

Freiberg and Scutchfield (1976) reported estimates of price elasticities based on utilization rates for 13 group insurance plans offered by Blue Cross/Blue Shield of Kentucky which varied in their effective coinsurance rates. Using the average out-of-pocket amount as an indication of cost sharing, they estimated an arc elasticity of -0.23 for the inpatient admission rate and -0.07 for length of stay. Unless the plans had a constant coinsurance rate (no deductibles, stop losses or maximums on benefits), their approach is biased because price is based on observed out-of-pocket amounts. The direction of the bias would depend on how exactly the marginal price varied with quantity across plans; see Newhouse et al. (1980) for a discussion of bias in such cases.

Using data from the 1963 Center for Health Administration Studies survey, Newhouse and Phelps (1974) reported estimates of price elasticities for users of services, omitting the effect of price on the decision to use care. As such, their estimates should provide a lower bound on the overall price response. They found a range of estimates of marginal price elasticities that were less than 0.2 in absolute value; it was −0.1 for length of stay and −0.06 for physician visits. Using work group size as an instrument to deal with the endogeneity of price did not appreciably affect the conclusions due to lack of precision in the two stage least squares (2SLS) results.

Newhouse and Phelps (1976) dealt with the omission of the decision to have care by employing a two-part model of demand. For inpatient care, the own price elasticity of any inpatient stay was −0.17, while for length of stay it was −0.06. For any outpatient care, the own price elasticity was −0.11 for any outpatient care and −0.08 for number of visits. Again, 2SLS did not appreciably affect the conclusions due to lack of precision. They also found some evidence for the complementarity of inpatient and outpatient care.

Phelps and Newhouse (1974) used data on premium quotes at varying levels of coinsurance rates and deductibles to infer the underlying price elasticities of patient demand, assuming that there is no allowance in the premium for adverse selection. In the range from 20 to 25 percent, their procedure generates an estimate of the price elasticity of −0.12. In the range from 15 to 20 percent, the estimated price elasticity is −0.08. And in the range from 10 to 15 percent, it is −0.04.

A number of researchers have examined the British experience with copayment for prescription drugs to examine the effect of a series of increases in prescription copayments between 1968 and 1986. Prescription copayments were introduced in 1968. Prescription copayments varied from £0.125 in 1968 to £2.2 in 1986, which was roughly the equivalent of a coinsurance rate of 0.21 to 0.43 on costs [O'Brien (1989)]. Older adults, children, and the poor were exempted from the copayments. O'Brien found that the price elasticity for drugs subject to copayment was −0.33 ($p < 0.001$) over the whole period, but had gone from −0.23 in the first half of the period to −0.64 in the second half.[25] He also reported a positive cross price elasticity (+0.17, $p < 0.001$) between over-the-counter and prescription drugs. Hughes and McGuire (1995) also examined the effects of copayments on prescription drug use in the UK, but for the period from 1969–92, when the copayment rose to £3.75 in 1992. Using a more elaborate econometric model for that longer period, they found an own price elasticity of −0.37 ($p < 0.01$), which was becoming more elastic over time. However, it appears that the changing price elasticity over time may be more a result of functional form than any formal test of an interaction of time and price.

There is also some evidence from German private health insurance. Insurer A had conventional plans with deductibles and coinsurance, insurer B offered a fixed annual rebate for no claims, while C had an experience-rated scheme, with the bonus reaching its maximum after three years without claims.[26] Using a two-period model similar to

[25] The coefficients in the two time periods were virtually identical. The difference in elasticities probably reflects the assumption that quantity is linear in price.

[26] By honoring non-use of medical care, plans B and C are reminiscent of medical savings accounts in the United States.

that presented in Section 5.2, Zweifel (1992, chs. 5, 7, 8) predicted that plan C should be most effective in limiting moral hazard in ambulatory care. This prediction was borne out in pairwise comparisons ($p < 0.001$ for C vs. B, evidence suggestive for B vs. A) in samples of 4,700 to 9,500 individuals, with observations from 1981 and 1982. Moreover, the hypothesis that plan C might induce insureds to defer necessary care, resulting in a toothsaw pattern in HCE, was rejected. In view of the skewness of the HCE distribution, the likelihood of HCE exceeding a sequence of thresholds was estimated, which precludes a direct comparison with other studies. However, the effect of a deductible turned out to be roughly similar to that reported for the Health Insurance Experiment by Newhouse, Manning et al. (1981).

The issue of deferral of medical care was taken up explicitly by Greenwald (1987), who assessed the effect of cost sharing on the initiation of care for cancer. He compared working adults with full coverage fee-for-service versus a comparable group under an HMO with copayment. On average, those with copayment waited a statistically significant 1.25 months longer to initiate care after a suspicion of illness, of which 0.8 months was the delay between diagnosis and treatment. Given the design, however, one cannot distinguish the HMO effect from the effect of copayment.

More recently, Magid et al. (1997) reported estimates of the effect of a copayment for emergency care following a myocardial infarction. They compared enrollees of a Seattle, Washington HMO who had to face copayments of $25 to $100 with enrollees who had no copayment, using ambulance and hospital records from 1989–1994. Adjusting for known age, gender and racial differences in the two groups of enrollees, the time from onset of symptoms to arrival at the hospital was virtually the same in the two groups: median arrival times were 135 for the copayment group and 137 minutes for the no-copayment group. The 95 percent confidence interval for the difference in times was $[-19, 16]$. Additional adjustment for season, income, education, cardiac history and clinical symptoms did not alter the finding. They concluded that modest copayments did not lead to significant differences in seeking care for this particular health event.

Differences in seeking care (frequently related to the individual‘s health status) should also be reflected in differences in price elasticities. Using data from the 1980 National Medical Care Utilization and Expenditures (NMCUES), Wedig (1988) examined the connection between health status and price responsiveness of the demand for medical care. The overall price elasticity was -0.32. The price response for those in fair or poor health was almost half that of individuals with better health in terms of whether they went to the doctor at all. There were no appreciable differences in how much they visited the doctor, given any visit. This study imputed the price paid by nonusers with a regression of the average price paid by users on variables available for both groups, including indicators for type of insurance coverage. This approach assumes that conditional on the covariates, price is missing at random for the nonusers. No explicit correction was made for the fact that nonusers may have been rationally nonusers – because of (unobserved) higher prices, higher time costs of seeking care, etc. for them.

Conclusion 8 *A number of observational studies have found that the demand for health care falls with increases in out-of-pocket costs, across a variety of populations and institutional settings. The magnitude of the estimated response varies widely, however.*

5.2.3. Observational studies using aggregate data

Using aggregate state-level data from a number of sources, Feldstein (1973) estimated the effect of net price on the demand for inpatient days. Using a two stage least squares model to deal with the endogeneity of the price of insurance and of inpatient care, he found a price elasticity of -0.67.

Using state aggregate data from 1966, Fuchs and Kramer (1972) estimated a range of price elasticity for physician services. Using a net price approach, the results were in the -0.15 to -0.20 range, while based on an average price, the elasticities ranged as high as -0.36. All were significant at conventional levels.

In both of these studies, there are a number of concerns about possible biases. Relying on constructed price series may lead to a too responsive estimate of the price elasticity if there are measurement errors in the quantity estimates or a too unresponsive estimate given errors in the expenditure variables. Aggregation over individuals and services may lead to aggregation bias. And in all simultaneous equations models with instrumental variables or two stage least squares, there are concerns about the identification of the model.

5.2.4. The Health Insurance Experiment

Design. The Health Insurance Experiment was conducted to provide information on the impact of alternative cost sharing for health care on the demand for health care, on financial risk, and on the health status of a general population. In the late 1960's and early 1970's, there was no evidence on the effect of cost sharing on health status, and little on financial risk. Although there had been evidence from observational studies and natural experiments on the effect of cost sharing on health care utilization and expenditures, there were major concerns about either biases in the estimates or generalizability of the results. The particular concern was that estimates based on observational studies are often systematically biased in their estimates of the responses to insurance coverage. Sick individuals or families with sick members have an incentive to self-select better health insurance coverage if they do not have to pay an actuarially fair premium, one that fully reflects the costs that they will impose on an insurance plan. If the analyst cannot control perfectly for the sickliness of the subjects in the data set, then there will be an omitted variable bias from unmeasured or mismeasured poor health status being positively correlated with insurance generosity – the sicker the patient, the lower the deductible, coinsurance or copayment rate, and the lower the stop-loss.[27] The result could

[27] As Newhouse et al. (1989) showed, even the best available measures of health status, casemix, and severity are able to explain only half or less of systematic (non-random) differences between patients. Thus, there is

be a substantial bias in the estimate of the price elasticity or the insurance plan response that overstates the effect of price on health care demand.

The Health Insurance Experiment (HIE) was a randomized trial in alternative health insurance arrangements that was designed, in part, to use randomization to break this link between health status, income, and other factors with insurance generosity.[28] Between November 1974 and February 1977, the HIE enrolled families in four urban and two rural sites in the United States.[29] The sample was drawn from the civilian households of the sites, excluding the top three percent of the income distribution, those enrolled in the SSI/DI program, and the elderly; the elderly were excluded because the recent passage of Medicare in the United States had dealt with the issues of insurance coverage for the elderly. Families participating in the experiment were randomly assigned to one of 14 different fee-for-service insurance plans for periods of either three or five years.[30] Families were enrolled as a unit, with only eligible members participating. No choice of plan was offered; the family could either accept the experimental plan or choose not to participate.[31] The enrollment sample included 5809 individuals.

The fee-for-service insurance plans had different levels of cost sharing that varied over two dimensions: the coinsurance rate, and a stop-loss (or upper limit on out-of-pocket expenses). The coinsurance rates (percentage paid out of pocket) were 0, 25, 50, or 95 percent for all health services. Each plan had a stop-loss on out-of-pocket expenses of 5, 10, or 15 percent of family income, up to a maximum of $1000 in then-current dollars (that is, unadjusted for inflation). A stop-loss of $1000 in 1975 would correspond to about $3000 in 1998 dollars, based on the US Consumer Price Index. Beyond the stop-loss, the insurance plan reimbursed all expenses in full for the remainder of that year. One plan had different coinsurance rates for inpatient and ambulatory medical services (25 percent) than for dental and ambulatory mental health services (50 percent). Finally, on one plan the families faced a 95 percent coinsurance rate for outpatient services, subject to a $150 annual limit on out-of-pocket expenses per person ($450 per

a substantial amount of unmeasured differences in sickliness across individuals. Some of this is recognized by the individual patient and can provide a basis for risk selection; see the role of the anticipated expenditure variable in the papers by Marquis and Phelps (1987), and by Manning and Marquis (1996).

[28] See Newhouse and the Insurance Experiment Group (1993) for a fuller description of the design an the rationale for the study and more details on the study's findings.

[29] The sites were: Dayton, Ohio; Seattle, Washington; Fitchburg, Massachusetts; Franklin County, Massachusetts; Charleston, South Carolina; and Georgetown County, South Carolina.

[30] The HIE assigned families to treatments using the Finite Selection Model [Morris (1979)]. This model is designed to achieve as much balance across plans as possible while retaining randomization – that is relative to simple random sampling, it reduced the correlation of the experimental treatments with health, demographic, and economic covariates.

[31] To reduce refusals, families were given a lump-sum payment greater than the worst-case outcome in their experimental plans relative to their previous plan. The lump-sum payment was an unanticipated change in income and should negligibly affect the response to cost sharing. The family's nonexperimental coverage was maintained for the family by the HIE during the experimental period, with the benefits of the policy assigned to the HIE. If the family had no coverage, the HIE purchased a policy on its behalf. Thus, no family could become uninsurable as a result of participation in the study.

Table 1
Health Insurance Experiment means by plan

Plan	% Any medical	Visits per capita	% Any hospital	Admits per capita	Mean (free = 100)	Adjusted mean[a] (free = 100)
Free	86.8	4.55	10.3	0.128	100.0[b]	100
25	78.7	3.33	8.4	0.105	84.6	81.1
50	77.2	3.03	7.2	0.092	90.0	75.0
95	67.7	2.73	7.9	0.099	69.1	68.7
Individual deductible	72.3	3.02	9.6	0.115	81.2	80.2
p value for plans[c]	<0.0001	<0.0001	0.001	0.02	0.003	
p free vs. 95%	<0.0001	<0.0001	0.0004	0.003	<0.0001	

[a] Standardized results using 4 part model. Free mean = $1019 (in 1991 dollars).
[b] Free mean = $982 (in 1991 dollars).
[c] Test of no plan differences.
Source: Manning et al. (1987), updated in Newhouse and the Insurance Experiment Group (1993).

family); in essence, this plan had an individual deductible for outpatient care. On an insurance plan with a 25 percent coinsurance rate and a $1000 stop-loss, the family would pay 25 percent of the first $4000 in health care expenditures, and $0 beyond that for that accounting year. The experiment would pay for the remainder of the expenditures.

All plans covered the same wide variety of services. None of the HIE plans had utilization review or other forms of "managed care" that are now common in the United States, even among "fee-for-service" or managed indemnity plans; the experiment was designed before such managed care arrangements became widespread.

Demand, utilization, and expenditure results. There were two approaches to the economic analysis of the HIE utilization and expenditure data. The first examined the demand and health status response to the plan *as a whole*, while the second estimated the effect of price *per se*. In the first, there was no attempt to separate the effects of deductibles, coinsurance rates, or stop-losses. In the second, the focus was the pure price response,[32] which could then be used as the building block for assessing the impact of deductibles, coinsurance rates, or stop-losses [for an example of such a construction, see Buchanan et al. (1991)].

Plan effects. The data in Table 1 from the HIE clearly indicate that health care demand is responsive to cost-sharing arrangements across a broad range of health care [see Newhouse and the Insurance Experiment Group (1993), Manning et al. (1987)]. Plans with greater first-dollar coinsurance rates have lower probabilities of any inpatient or outpatient use, lower visit and admission rates, and lower expenditures. Subject

[32] The pure price response is what would happen if there was a change in the out-of-pocket price of medical care that was uniform across all levels of consumption.

to a stop-loss, going from no out-of-pocket expense (free care) to paying twenty-five percent out-of-pocket[33] reduces visit rates by 27 percent, admission rates by 18 percent, and expenditures by 15 percent. As the level of cost sharing rises, use and expenditure fall but at a declining rate.[34] The results are quite similar after adjustment for covariates, but they are less sensitive to the effects of a few very extreme cases.[35] The demand for outpatient health care for children is as responsive to insurance as it is for adults, but there appears to be no insurance response for children's inpatient use [Leibowitz et al. (1985a), Manning et al. (1987)]. Practically all of the observed response to cost sharing is in the quantity of services received, not in the unit costs. There was no evidence of patients switching to more expensive providers as a result of having better insurance coverage [Marquis (1985)].

The magnitude of this price response is less than that of most of the earlier US studies. The HIE response is roughly the same as that in Scitovsky and Snyder's (1972) study of Stanford University faculty and staff, which relied on a natural experiment in the cost sharing on the University's health insurance plan. There had always been some concern that the Stanford experience might not be generalizable, given the atypical population affected.

The response to income in the HIE is also less dramatic than in many of the other studies. In the HIE, the income elasticity is in the range of 0.2 to 0.4 for overall medical services, after controlling for factors that are confounded with income in this and other populations, including health status. Because of the randomization of individual families to insurance plans, income is not correlated with more generous coverage. Thus there is not the usual transmitted bias from unmeasured or imperfectly measured insurance generosity and income, when wealthier individuals tend to have better insurance coverage.

If health status is a durable good or stock, then one might expect a change in the price of health care to lead to a sudden, transitory surge in demand. In the logic of a Grossman (1972) model, an unexpected reduction in price will induce individuals to want to increase their health stock above the preceding optimal level. To bring desired and current health into alignment requires an increase in medical care demand on very generous plans greater than what current depreciation in the health stock would dictate. Once the new desired stock equals the current, demand would fall back to a lower level.[36] The data on medical and mental health demand from the HIE do not exhibit this sort of transitory shifts in the free plan relative to the less generous plans; the free

[33] Subject to a stop-loss.

[34] This declining rate is probably attributable to the effect of the stop-loss feature.

[35] The minor reversal in expenditures for the 50 percent plan is the result of a single case that contributed nearly a sixth of that plan's mean.

[36] For the less generous plans, the logic is reversed. The participant would find that his or her stock was greater than optimal given the new insurance plan. Because he cannot sell off or scrap "excess" health, the optimal course is to reduce health expenditures until the new stock equals the new desired stock and then return to a higher pattern of health expenditures.

plan provided more generous coverage than most insurance plans that the experimental subjects would have had before the experiment, while the 50 and 95 percent plans would have been less generous for most subjects. However, in the case of dental care, the HIE did find evidence of sudden, short-term surges in expenses at the outset of the experiment for participants on the free plan. There was also a statistically significant but smaller surge at the end of the experiment [Manning et al. (1985)]. Thus of all the types of health care, only dental care exhibited the kind of behavior that is implied by the health capital version of the Grossman model.

One of the implications of the standard model of patients' response to more generous insurance is that individuals will buy more care of lower marginal value as the out-of-pocket price falls. Although it is difficult to assess the marginal value directly, one might expect that well-insured patients and their doctors are more likely to purchase medically less appropriate care, have more visits, or stay more days in a hospitalization than are necessary to treat an illness. To test this conjecture, the medical records for all adult hospitalizations, other than for maternity and psychiatry, were reviewed to determine their medical appropriateness [Siu et al. (1986)].[37] Increased cost sharing reduced both appropriate and inappropriate admission rates and inpatient days. However, increased cost sharing did not reduce inappropriate care more than it reduced appropriate care. Both fell by 23 percent. Thus, it appears that cost sharing is a relatively blunt instrument for affecting the quality of care.

Emergency room utilization was also significantly related to cost-sharing [O'Grady et al. (1985)]. The overall response to insurance plans was quite similar to that for other outpatient medical care. However, ER visits for more urgent diagnoses were more responsive to cost sharing than were ER visits for less urgent diagnoses. Given the availability of office-based physicians in poorer neighborhoods, it was not surprising that the poor made heavier use of the ER than did wealthier participants. There was no evidence that the ER use by the poor was more responsive to insurance than that of wealthier participants. The lack of a significant finding could be due to either lack of precision or lack of a true effect. The HIE was not powered to provide substantial precision for comparisons of rich and poor for utilization or expenditures.

Inpatient care was less responsive to health insurance than outpatient care; see Table 1 above. It is important to remember that the response to insurance is a response to coinsurance and to stop-losses. This result could reflect either a lower price response for inpatient care, or be an artifact of the stop-loss feature of these plans. For most inpatient stays, part or all of the bill was beyond the stop-loss and free at the margin, while for outpatient care, most patients were below their stop-loss.

The demand for prescription drugs exhibited the same response to insurance coverage as did outpatient health care [Leibowitz et al. (1985b)]. However in the HIE, outpatient

[37] The available methods for evaluating the medical appropriateness of care did not cover children or psychiatry. At the time of the study, nearly all maternity cases delivered in a hospital in the United States. Because outpatient surgery was still rare, cases that could have been treated in outpatient surgery were considered medically appropriate. For the protocol used, see Gertman and Restuccia (1981).

Table 2
Arc price elasticities of medical spending (standard errors in parentheses)

Range	Acute	Chronic	Well	Total out-patient	Hospital	Total medical	Dental
0–25	−0.16	−0.20	−0.14	−0.17	−0.17	−0.17	−0.12
	(0.02)	(0.04)	(0.02)	(0.02)	(0.04)	(0.02)	(0.03)
25–95	−0.32	−0.23	−0.43	−0.31	−0.14	−0.22	−0.39
	(0.05)	(0.07)	(0.05)	(0.04)	(0.10)	(0.06)	(0.06)

Source: Newhouse and the Insurance Experiment Group (1993).

physician services and prescription drugs had the same coinsurance rate and were subject to the same stop-loss.

Pure price effects. The difficulty with the responses to the HIE plans as a whole is that they do not tell us directly about responses to other insurance plan structures. The "pure" price elasticity provides the building block for such comparisons. To estimate the effect of a change in out-of-pocket price that was independent of the level of consumption (e.g., no deductibles or stop-losses), Keeler and Rolph (1988) examined data on the timing and size of episodes of treatment as a function of the coinsurance rate and distance from the plan's stop-loss. If an individual was a substantial dollar distance away from the stop-loss, then he would be acting as if there was a relatively low probability of his exceeding the stop-loss and receiving free care for part of the year. Decisions made near the stop-loss would combine the response to price and any anticipation of possibly exceeding the stop-loss and receiving free care for part of the year. This was their way of operationalizing the dynamic optimization problem described earlier in Keeler et al. (1977) [see also Section 5.1.2 above].

The results from the analysis of the HIE episodes indicate that the pure price response for overall medical care is on the order of −0.2 [Keeler and Rolph (1988); see Table 2]. Equivalently, going from no insurance to full insurance almost doubles expenditures and the number of episodes of care. There is some variation over the range of prices and services. Total medical care demand has a pure price elasticity of −0.17 at low coinsurance rates (0–25 percent) and −0.22 at higher values (25–95 percent). The price elasticity for outpatient medical care is −0.17 at low coinsurance rates (0–25 percent) and −0.31 at higher values (25–95 percent). The principal source of the response to cost sharing is the number of episodes of illness that are treated, and not how intensively the individual episodes are treated. The size of episodes does not seem to depend on the insurance plan [Keeler and Rolph (1983)].

One of the implication of the dynamic optimization process described in Keeler et al. (1977) is that the patients respond to a price below the marginal price of care if there is some positive probability that they will exceed the stop-loss. Patients "anticipate" the lower price that might prevail if they are sick enough to carry them beyond a deductible or stop-loss. The greater the likelihood, the lower is the "effective" price below the marginal. Keeler and Rolph (1988) found no evidence of anticipation on consumption decisions, but did find an in increase in utilization rates once the stop-loss was exceeded. Thus, the consumer response to the stop-loss appears to be myopic.

Health status results. The Health Insurance Experiment also examined the effect of cost sharing on health status after the individual had been enrolled for three to five years. Because of the differences in the nature of illness and disease, the study examined children (ages <14) and adults (aged $\geqslant 14$) separately. Health status was measured by both self-report and by physical examination. For the summary measures of health status (including self-reported health status, physical health, and mental health), there were no statistically significant or appreciable differences across the fee-for-service plans for the average adult [Brook et al. (1983)]; the study had enough precision to rule out large, clinically important effects. However, there was an appreciable (but statistically insignificant) adverse effect of cost sharing for poor adults who had been in poor health at the beginning of the study; unfortunately the precision for this rare, but important, group was not sufficient to rule out large effects. Of the physiological measures, only corrected far vision and diastolic blood pressure were better in the free plan, above all among the poor at elevated risk [Keeler et al. (1987)]. If the latter result were maintained for a number of years, and if the mortality patterns followed those in the Framingham study, then those at elevated initial risk would have a lower risk of subsequent death with free care than with the cost sharing plans [Newhouse and the Insurance Experiment Group (1993); corrected from Brook et al. (1983)]. For dental care, there were no differences in overall rates of decayed, filled and missing teeth (combined), but the free plan had about one more filled and one fewer decayed teeth than did the cost sharing plans [Bailit et al. (1984)].

There was some evidence of an effect of cost sharing on symptoms. Shapiro et al. (1986) used data on symptoms collected from adults annually by survey. Symptoms were categorized as serious or minor, based on physicians' assessments of whether the symptom merited seeing a physician. Although the prevalence of minor and serious symptoms was similar at baseline for the free and cost sharing plans, there were differences later. Those with cost sharing were more likely to report a serious symptom (16.5 vs. 14.5 percent, $p < 0.05$). Minor symptoms did not change markedly over time and were not significantly different for the free and cost-sharing plans. However, adults on the cost-sharing plans were one third less likely to see a physician for a minor symptom. There were no significant differences in care seeking by insurance plan for those with serious symptoms, although individuals enrolled in cost-sharing plans were less likely to seek care (17.9 vs. 22.3 percent, $p = 0.095$).

Cost sharing reduced disability measured by days of restricted activity (RADs), but not as measured by work loss days (WLDs) [Rogers et al. (1991)]. The participants on the free plan reported 9.8 RADs per year, compared to 8.6 on the intermediate cost sharing plans (25 and 50 percent coinsurance rates), 8.2 on the 95 percent plan, and 9.0 on the individual deductible plan ($p < 0.01$). WLDs were 5.5, 4.8, 4.8, and 4.5, respectively ($p = 0.38$); however there was much less precision for WLDs than for RADs. Because the difference in RADs across insurance plans is quite similar to the difference in visit rates, one might infer that the difference in RADs is due to the time spent in seeking health care.

For the average child in the study, there was no significant or appreciable difference in health status by insurance plan [Valdez et al. (1985, 1986)]. There was enough precision in the study to rule out an effect equal to half the effect of having hay fever. There were no significant differences by subgroups of initial health status or income; but the study had much less precision for such comparisons. Children had a similar pattern to adults for oral health status.

5.2.5. Assessment of the evidence

The theoretical models presented earlier in this chapter indicate an increase in the price paid out-of-pocket by the patient should lead to a reduction in the amount of care sought if the budget share of out-of-pocket expenses is small – as it typically is. There is substantial empirical evidence supporting this proposition. The earlier literature based on natural experiments and observational studies of individuals suggested a very wide range of elasticities, from the -0.14 of the Stanford University studies to the -1.5 of Rosett and Huang. There were a number of issues raised in these studies, including questions of confounded changes, generalizability for the natural experiments, and adverse selection in the observational studies. The one study that is able to surmount these issues is the Health Insurance Experiment (HIE), which randomized non-elderly individuals in different sites to varying levels of cost sharing. Use of a random sample of individuals in sites under varying levels of stress (e.g., delays to appointment) allows for tests of differential responses to cost sharing. The HIE found a price elasticity on the order of -0.2, which did not vary appreciably by income or health status or by site. This estimate is consistent with the earlier Stanford University studies of Scitovsky and Snyder/McCall. Similar price elasticities were found in the HIE for well care, prescription drugs, emergency room visits, and other general health care. The four notable exceptions were:

(1) the demand for hospitalizations for children was insensitive to insurance plan;
(2) dental care was much more elastic in the short run than in the long run;
(3) outpatient mental health care was more responsive than outpatient medical care; and
(4) very urgent care was less responsive to cost sharing than other services.

Unexpectedly, cost sharing had no influence on the mix of medically appropriate and inappropriate inpatient care, because it reduced both by the same proportion. This indicates that cost sharing *per se* is not well suited to discouraging medically marginal care.

In contrast to a lengthy literature on the effect of cost sharing on utilization and expenditures, there is much less evidence about its effect on health status. The Medi-Cal termination studies indicate that a vulnerable population that suddenly becomes uninsured will have poorer health status. On the other hand, the HIE suggests that there are no adverse effects of cost sharing for the average individual. The difference in results could reflect differences in baseline risk, with the MediCal patients being at greater risk given their poor initial health. The differences could also reflect the presence of a stop-loss on the HIE plans that was never more than 15 percent of income. This would have shielded individuals from the financial threats of a catastrophic health care expenditure. Finally, none of the HIE plans was as difficult to use as the situation of terminees having to seek charity care or to queue at county facilities. The pattern of HIE results also suggests that individuals who are both poor and sick will do better with free care than cost sharing, which is consistent with the adverse effect on the sick poor in the MediCal termination studies.

There are a number of concerns about the Health Insurance Experiment. Although families were randomized to insurance plan, there was a higher rate of refusal and subsequent attrition on the cost-sharing plans than on the free-care plan. Nevertheless, the enrollment samples were quite similar, including pre-study utilization. Moreover, the insensitivity of the experimental results to adjustment for baseline health status and other differences suggests that the higher attrition and refusal rates are not a major concern. The HIE was such a small part of any physician's practice that his/her knowledge of a patient's insurance coverage may have been limited. This means that the HIE results may not be generalizable to a situation where one of the plans would become the standard in health insurance, which would likely evoke a behavioral response on the part of physicians. Finally, and most importantly, the estimates are now nearly 15 to 20 years old, and apply to a world with "unmanaged" third-party, fee-for-service insurance. Although the estimates may apply in the older system, it is not clear how relevant they are in a world which uses non-price rationing of demand as extensively as some of the current American systems do.

There is some limited evidence on the effect of copayments in traditional staff and group model HMOs which suggests that utilization of care continues to be responsive to price (see the two studies by Cherkin et al., and the Selby et al. study). However, the arc elasticities are as low, and in some cases lower than those from the HIE.

One other concern is with the international generalizability of findings. The response in visit and admission rates to the introduction of national health insurance in Taiwan appears to have been more responsive than the HIE experience. The UK experience with prescriptions suggests a greater response than in the United States. However, Zweifel's analysis of German health plans points to a response to deductibles comparable to HIE estimates.

5.3. Full price effects

5.3.1. Theoretical background

Up to this point, the way income is generated in the sick state was not analyzed. Yet, the fact that sick leave pay S may replace labor income wW should have important implications for the opportunity cost of time in the sick state, which should also influence the demand for medical care.

In order to understand the effect of time costs of receiving care, we will consider a variant of the one-period labor-leisure model, such as the model described by Acton (1975). Consumers have a utility function U that depends on health H, other goods C, and leisure l. Health is produced using doctor visits M, $H = f(M, \xi)$, costs p in monetary costs and takes time t per unit of health care received. Here, ξ is a shift parameter that changes the marginal productivity of health care in producing health. It symbolizes the outcome uncertainty surrounding medical interventions; however, since this complication will be neglected below, ξ is dropped in what follows. Health care is valued by consumers only for the health that it can produce. If sick leave pay is proportional to earned labor income [at a rate $(1 - s)$, see Section 4.1 above], the Lagrangian function becomes[38]

$$\begin{aligned} £^w &= U\{H(M), C, l\} \\ &\quad + \lambda\{Y_0 - P - R + (1 - s)(24 - l - tM)w - cpM - C\}. \end{aligned} \tag{17}$$

An individual who is unemployed or out of the labor force still has to observe a time constraint, which is introduced using the Lagrangian multiplier μ (the shadow price or the reservation wage, respectively). In this case, the Lagrangian function reads

$$£^n = U\{H(M), C, l\} + \lambda\{Y_0 + S - P - R - cpM - C\} + \mu\{24 - l - tM\}. \tag{18}$$

Only the first-order conditions for the variant (17) are given here:

$$\frac{\partial £^w}{\partial M} = \frac{\partial U}{\partial H}\frac{\partial H}{\partial M} - \lambda\{cp + wt(1 - s)\} = 0, \tag{19}$$

$$\frac{\partial £^w}{\partial C} = \frac{\partial U}{\partial C} - \lambda = 0, \tag{20}$$

$$\frac{\partial £^w}{\partial l} = \frac{\partial U}{\partial l} - \lambda w(1 - s) = 0, \tag{21}$$

$$\frac{\partial £^w}{\partial \lambda} = Y_0 - P - R + (1 - s)(24 - l - tM)w - cpM - C = 0. \tag{22}$$

Assuming that the second-order conditions hold [utility functions $U(\cdot)$ and production functions $f(\cdot)$ have the usual properties], then consumers act as if they face a full price

[38] The superscript s denoting the sick state is dropped here for simplicity, being replaced by w (working) and n (nonworking), respectively.

of health care of $cp + wt$, full income of $(Y_0 + S + 24w - P - R)$, and a full price of other goods and services C of 1. The resulting system of demand conditions reads,

$$\begin{aligned}
&\text{Labor:} && W^* = 24 - l - tM = g\big\{cp + wt(1-s), w, Y_0 + S + 24w - P - R\big\},\\
&\text{Health care:} && M^* = m\big\{cp + wt(1-s), w, Y_0 + S + 24w - P - R\big\},\\
&\text{Health:} && H^* = f\big(m\big\{cp + wt(1-s), w, Y_0 + S + 24w - P - R\big\}\big),\\
& && = h\big\{cp + wt(1-s), w, Y_0 + S + 24w - P - R\big\},
\end{aligned}$$

where in each demand equation the first term is the full price of medical care, the second is the full price of leisure, and the third is full income.

Thus if health care has a small budget share, we would expect health care to be a declining function of the gross price of health care p, the coinsurance rate c, and the time required for a visit t (including travel, time in the office, etc.). An increase in the health insurance premium P by itself decreases health care consumption if health is not inferior. If coupled with a reduction in the coinsurance rate c, its effect is ambiguous. However, most researchers would expect that the net effect would be positive. Similarly, a pure increase of the sick leave premium R should curtail demand for medical care at the margin. If matched with an increase of generosity of sick leave benefits [an increase of $(1-s)$], this effect may be reinforced because the full marginal price of medical care increases. This can be seen from Equation (19).[39]

Finally, as in all labor/leisure problems, a change in the opportunity cost of time – the wage rate w, or $w(1-s)$, respectively – has an ambiguous effect on all goods and services. For health and health care, there is the usual positive effect through full income, a negative substitution effect via the effect on the price of leisure, and a negative own price effect through increasing the full price of health care, $cp + wt(1-s)$.

This formulation can easily be modified to allow for income taxes, deductibles on health insurance and sick leave, stop-losses or upper limits on covered health care or time lost from work. In the case of a proportional income tax, w becomes the after-tax wage rate, and Y_0 becomes the after-tax unearned income.

Deductibles and stop-losses on any insurance arrangement follow the same logic as above for deciding which facet of the budget constraint that an individual chooses to operate on.

5.3.2. Empirical evidence

Few of the studies in the literature have attempted to empirically implement this approach. The reason is due to missing data on p, w, and t. Unless one is looking at a population that is very sick, there will always be individuals who do not use health care

[39] This price effect of sick leave pay should be distinguished from its (opposite) income effect derived in Section 5.1.1.

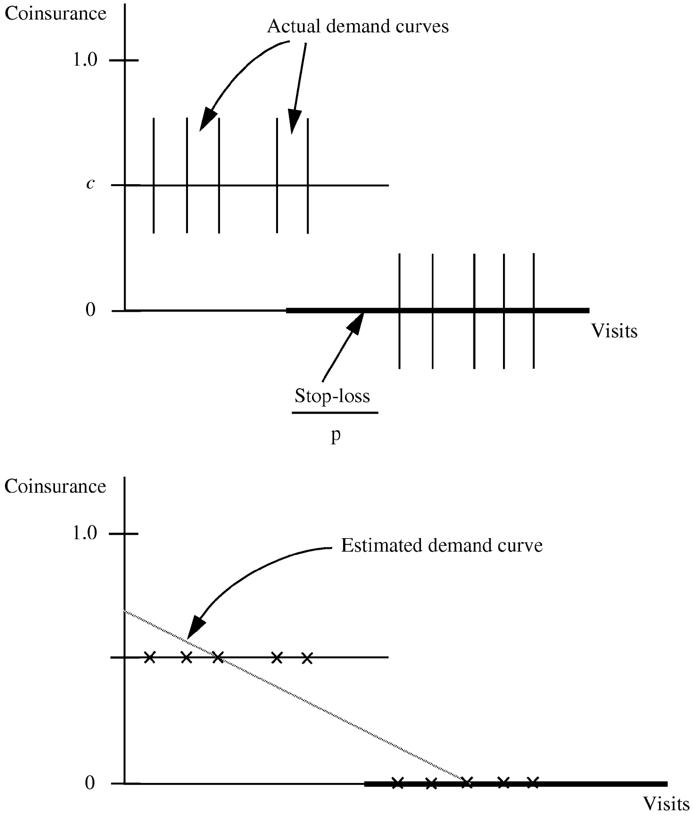


Figure 5. Declining block bias in price effects.

during the period of observation. For these individuals, we cannot observe either p or t. Not all individuals work. For these individuals, we cannot observe the opportunity cost of time w. If individuals are rational, then those with missing values are a self-selected rather than a random subpopulation. They have full prices of medical care that exceed the value of the first visit. That is, their p or t or wt or $cp + wt$ is too high relative to the value of health care, which may be very small or zero if the person is healthy. Also, they have reservation wages that are higher than their market wages. Dropping such missing cases, or using values of p, t, and w estimated from users and workers will lead to systematically biased results, unless one can employ a model like the Selection Model used in labor economics. Unfortunately in most cases, the model is not identified by exclusions, because there is no variable which affects the likelihood of any health care that does not affect the level of care (if any).

One of the classic studies of the role of time costs in health care demand is Acton's (1975) study of users of New York's "free" city outpatient departments and municipal hospitals. Building on the work by Becker (1965) and Grossman (1972), he developed a theoretical model similar to that in Section 5.3.1. Specifically he had the visit to a doctor require some of the patient's time. The result was that the opportunity cost of the patient's time was sufficient to generate a price response even when there was no out-of-pocket copayment or cost sharing. Using data from a 1965 survey he found a full price elasticity of -0.4 ($p = 0.01$) for the use of outpatient department services.[40] There was also evidence of a cross price effect on the use of private physician services.

A more recent study incorporating time costs as determinants of demand is by Leu and Doppmann (1986). Their data base consisted of 3,125 Swiss adults in 1980 whose demand for medical care was related to (among other things) the time cost of access and of treatment. The elasticity of ambulatory care visits w.r.t. the time cost of access was -0.047 ($p < 0.001$), while the elasticity w.r.t. treatment time turned out positive. The authors interpret this positive elasticity as a possible indication that treatment time may serve as an indicator of quality. Neither type of time cost proved significant in the equation for hospital days and rehabilitation spa days. Unfortunately, the qualifications formulated above apply because it was necessary to impute a wage rate to individuals not in the labor force, with no modeling of the selection mechanism resulting in their nonparticipation. Therefore, these estimates may be subject to bias.

Conclusion 9 *The way income is provided to the individual when sick importantly determines the full price of receiving medical care. If sick leave pay is proportional to earned labor income, its generosity serves to increase the opportunity cost of medical care. This effect is absent in the case of unemployed or retired individuals. However, imputing the correct opportunity costs of medical care to this group is fraught with great difficulties.*

5.3.3. A methodological issue

One of the insights in microeconomics is that we can construct the demand response to complex multi-part tariffs, such as health insurance and sick leave policies, from the demand responses to simple linear (constant marginal price) budget constraints. Individuals act *as if* they faced a constant marginal price for the block that they are on, and an appropriately adjusted net income amount.

This has led a number of applied researchers to use the observed marginal price and net income as the explanatory variables in their analyses, rather than estimating the response to the full budget constraint. This may be good economics, but it is poor and biased econometrics. The difficulty is that the observed marginal price and net income are not independent of the error term. Individuals who are sicker or who suffer from hypochondria or have a strong taste for health will face lower marginal prices, because

[40] Acton used two stage least squares to deal with the endogeneity of distance to the providers. He also reweighted the date to correct the selection bias in any such sampling due to users sampled being more frequent users than the population at large.

they are more likely to exceed the deductible or stop-loss. If individuals had perfectly inelastic demand curves, and faced an insurance policy such as in Figure 5, then the observed marginal price would be negatively correlated with unobserved differences among individuals. Least squares would produce an estimated demand curve, such as the downward sloping dashed line, for a case where the true demand curves were vertical. If the insurance plan had first-dollar coverage and a limit on the number of covered visits, such as is common for insurance coverage for psychiatric care, the estimated demand curve would be upward sloping, suggesting that this was a Giffen good.

This bias from using observed marginal price and net income will exist as long as the equation does not have a perfect set of explanatory variables or unless some sort of instrumental variables estimator is appropriately applied.

The econometrics literature does provide some alternatives for dealing with nonlinear budget constraints [see Moffitt (1986) for a review]. Hausman (1985) offers an alternative for the situation where one can make distributional assumptions. These approaches can be used to infer pure price effects without the risk of biased estimates from using observed marginal price and income. For a general review of the econometric problems in estimating demand responses for health care services, see Newhouse et al. (1980).

5.4. Effect of rationing by the physician

5.4.1. Theoretical background

Up to this point, the patient was assumed to be able to choose the amount of medical care anywhere along his budget line. However, the degree of consumer's sovereignty is often somewhat limited. First, there are the lack of information and delegation of authority discussed in Section 3. Second, in a managed care setting, the physician also acts as the insurer's agent [see the chapter by Glied (2000) in this Handbook]. His task becomes to limit moral hazard, rationing the amount of services provided below that desired by the patient under the influence of insurance. His incentive to perform this task may derive from his sharing in the cost of medical care as a provider [Ellis and McGuire (1993)].

In Figure 6, let this desired amount be M_1^*, whereas the physician imposes $\overline{M}$. Interestingly, it is conceivable that the patient does not even want to consume $\overline{M}$. Specifically, point R_1 is dominated by point C_1^r, which entails consumption below the deductible amount. Thus, given that he is rationed, the patient may decide that the savings achievable are important enough to go with much less care, at least during the first illness episode. This effect may also be interpreted in terms of Figure 4. If the physician rations a patient characterized by demand curve D_2 by imposing a quantity $\overline{M}$ slightly in excess of M_0, he reduces the area of the lower triangle. This loss of consumer surplus may cause the patient to switch back to the first block, which would render the rationing ineffective. However, this typically will not be true in the second period. In Figure 6, given that $\overline{M}$ again is available (episode-specific rationing), the associated point C_2^r dominates any other point that can be attained along the period 2 budget line *HKL*,

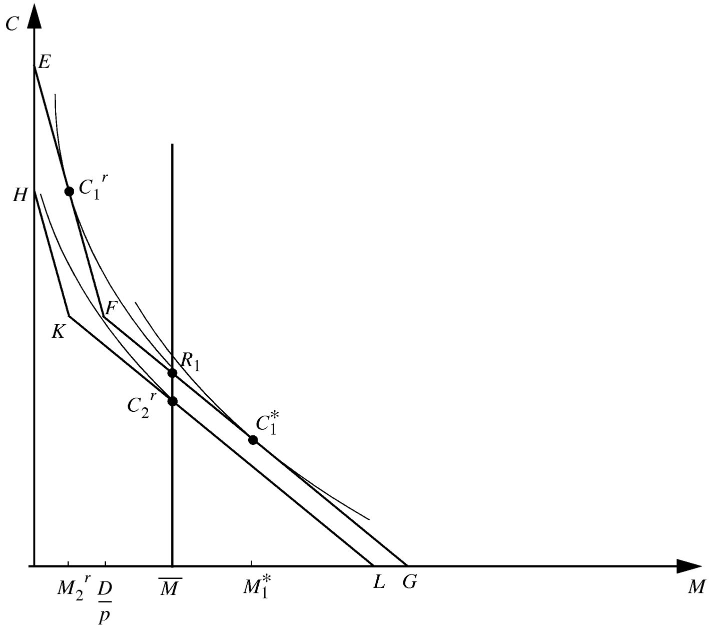


Figure 6. Effect of rationing by physician in a two-episode model.

given homotheticity. This is the consequence of the income effect caused by copayment for the preceding sickness episode.

Thus, the amount available under rationing may not even be fully used early in a sequence of sickness episodes, while becoming more and more binding as additional episodes occur. There, rationing may indeed counteract the reduction of effective price brought about by consumption of services beyond the deductible.

Conclusion 10 *Even if applied to the block with the lower marginal price, rationing by the physician may be effective by reducing demand to the block with the higher marginal price. A given episode-specific limit becomes more binding the higher the number of the sickness episode.*

5.4.2. Empirical evidence

A rather indirect test of the influence of physician rationing was constructed by Zweifel (1992, Chs. 4, 5). As noted in the preceding section, a quantity $\overline{M}$ imposed by the physician may not be demanded by the patient, although it lies beyond the deductible ($\overline{M} > D/p$). Given full consumer sovereignty, only coinsurance should be effective in limiting moral hazard beyond the kink F of Figure 6. Given rationing, however, the patient may have additional motivation to prefer a modest treatment alternative and save some money for consumption. In this way, the deductible obtains an influence beyond the kink F. Turned the other way, given rationing, a deductible should limit static ex-post moral hazard for values of HCE beyond D/p. Indeed, claims data provided by a German private insurer from 1980 to 1982 (insurer A, writing plans with deductibles or coinsurance) suggest that insureds subject to a deductible had a reduced likelihood to have ambulatory HCE in excess of a given value than those without. More to the point, this difference remains significant ($p < 0.05$) up to HCE that are triple the value of the deductible.

While these findings support Conclusion 10, they are silent about the relative effectiveness of cost sharing and rationing in the control of static ex post moral hazard. In order to sort this out, one would need to be able to distinguish rationed from non-rationed situations.

6. Dynamic ex post moral hazard

6.1. Theoretical background

At the macroeconomic level, the continuing surge of HCE in most industrial countries has been explained with reference to mainly three factors [Newhouse (1992)]:

- *Rising incomes*. To the extent that health care is a luxury good [for which there is some evidence, see, for example, Gerdtham et al. (1992) and the chapter of Gerdtham and Jönsson (2000) in this Handbook], its share in total income should increase in step with growth.
- *Demographic change*. Insurers find consistently that the average HCE increases steeply with age, at least beyond age 60. Since the share of the population 65 and older has been increasing markedly, it is tempting to conclude that aggregate HCE will follow suit [however, see e.g., Newhouse (1992), Getzen (1992), and Zweifel et al. (1999) for a critical appraisal of this conclusion].
- *Technological change in medicine*. The pace of technological change in medicine is higher than elsewhere. While lowering the unit cost of production in industry, new technology seems to drive up the cost of health care [Newhouse (1981), Zweifel (1984)].

The issue to be addressed in the present context is whether health insurance has anything to do with these influences. Health insurance and sick leave pay are unlikely to have stimulated economic growth. If they contributed to demographic change, it would presumably been through access to improved medical care. In this context, the long-run decrease of the share of uninsured in the population may seem important. However, insurance-financed health care systems outside the United States have been covering the entire population for decades, yet have been experiencing a similar cost expansion.

This leaves the suspicion that health insurance speeds up the rate of technological change in medicine by encouraging patients to opt for the latest medical technology. In the present microeconomic framework, one could refer back to Figure 1, arguing that without insurance, the individual depicted by indifference curve J_1 would never even consider the high-cost alternative symbolized by C_1^{**} but would settle for the low-cost one, symbolized by C_1^*. But then, this increase in the demand for medical services would have to be a once-and-for-all phenomenon rather than a continuing surge. Therefore, the question arises of whether insurance might somehow have a dynamic effect, affecting not only the volume of HCE but also its rate of change over time.

At first blush, the argument that insurance reduces the net price differential between the new, better and the old, standard medical service looks convincing. However, it is relative price that matters, and relative price is not affected as long as there is proportional cost sharing. Suppose, e.g., that the new procedure is three times as effective as the old, but costs 50 percent more. Thus, in terms of its benefit-cost ratio, the new procedure outperforms the old by $2:1$. With coinsurance rate c, the net ratio becomes $(2c/c) = (2/1)$, i.e. it remains unchanged.

Within the theoretical framework expounded here, there seem to be two possible reasons for an effect of insurance on technological change, even in the presence of proportional cost sharing.

- *Reduction of effective price in the presence of a deductible*: Referring to Figure 1 once more, the closer the insured gets to the deductible D in the first period, the more likely are additional sickness episodes to move him beyond the value of the deductible, into the insured area. One way to increase the value of pM_1 is to use more expensive medical technology [see Equation (14)]. However, for this explanation to hold, innovations in medical technology would have to occur mainly in the domain of small claims (below the deductible); otherwise they would not result in a lowering of effective price.
- *Money price vs. full price*: By opting for a service of higher quality and having a high unit price, the insured increases the share of money price (which is subsidized by insurance) in the full price. This may prove advantageous, especially in the case of a large consumption of medical care. In terms of Figure 3, a higher unit price p causes a sharper kink in the budget constraint because money price makes up for a larger share in the full price. Insurance coverage, by acting on the money price component (which is high to begin with since illness lowers the opportunity cost of time), therefore has a stronger impact on full price in the case of high-quality medical care than in the case of low-quality care. However, the stronger the kink, the more likely the patient is to opt for the more elaborate treatment alternative in both episodes.

To the extent that these explanations are true, suppliers of innovative products in health care will meet with enhanced demand, increasing their chances of economic success [Weisbrod (1991)]. Since many of these innovations are protected by patent, these quasi-rents are not washed away very easily. Physicians and hospitals will also be biased in their choice between two types of technology. Having the choice between process innovation (which lowers cost without affecting the characteristics of the product) and product innovation (which may appeal to the insured in spite of higher unit cost), they are likely to lean towards product innovation under conventional health insurance [Zweifel

(1995)]. Thus, insurance may go some way in explaining the difference of effects of innovations in health care as compared to the remainder of the economy.

Conclusion 11 *Health insurance may affect the pace of innovation in two ways. First, the more expensive new service may reduce the effective price of care in subsequent illness episodes, and second, it may increase the importance of insurance in lowering total price, thus reinforcing moral hazard effects beyond the deductible. Since process innovation does not have these features, insurance coverage also biases the composition of innovation in favor of product innovation.*

6.2. Empirical evidence

The one attempt at identifying the effect of dynamic moral hazard seems to be due to Newhouse (1981). He uses annual time series data to test the hypothesis that the price change (rather than the price level) for four types of health care services depends on the respective rate of coinsurance. In the case of heavily insured hospital services, he finds the predicted negative relationship ($p < 0.05$). In the case of physician fees, the evidence is suggestive, whereas the price change for dental services and pharmaceuticals does not seem to depend on coinsurance. While these findings were shown to be rather robust to specification changes, they could not be replicated with time series data covering 37 rather than 26 years [Newhouse (1988)].

Conclusion 12 *There is some tentative evidence pointing to the existence of dynamic moral hazard effects in health care in the United States.*

Future work might go beyond the reduced form approach adopted by Newhouse. On the demand side, HCE due to the use of new and existing medical technology would have to be identified and linked to the net price faced by the insured. On the supply side, the product and (cost-saving) process innovations could be distinguished and related to the net price of using them. Market equilibrium would be defined by a relative rate of product vs. process innovation and the corresponding net price of medical care, which of course depends on the rate of coinsurance.

7. Concluding remarks

This chapter deals with consumer incentives in health care. One might be tempted to argue that consumer incentives have no role to play in tax-financed health care systems. This argument is refuted here for two reasons. First, the initiation of a sickness episode frequently goes along with a loss of wage income, which depends on the generosity of sick leave payments. This income effect on the demand for medical care is present even if the money price of utilizing medical care is zero, and it deserves added emphasis in future research. Second, in health care it is full price that often matters. Tax-financed health care still burdens the consumer with time cost when he seeks care. Thus, consumer incentives prove important regardless of the way health care expenditure is financed.

With regard to insurance-based systems, it should be noted that unless physicians act as perfect agents of their patients, the scope of consumer incentives is circumscribed by the amount of authority delegated to them. This in turn depends on the informational disadvantage of the patient vis-à-vis his physician, and it is an important research question whether and how this disadvantage is influenced by consumer incentives.

Where health insurance is widely available, however, it importantly shapes the financial incentives facing consumers. The influence of insurance on behavior is commonly called moral hazard. In the health context, it is useful to distinguish between ex ante, static ex post, and dynamic ex post moral hazard effects, referring to the tendency to skimp on prevention, consume more medical care, and opt for the newest technology, respectively.

To the extent that prevention and medical technologies differ with regard to their cost and contribution to health, it would be valuable to be able to distinguish between the three types of moral hazard. Little work seems to have been devoted to this issue so far.

The bulk of the empirical evidence concerns static ex post moral hazard effects, aggregated across all types of care (preventive, curative, discretionary, or emergency). Here, the Health Insurance Experiment suggests a demand elasticity of medical care with respect to money price of around -0.2, somewhat smaller in absolute value than that derived from most observational comparisons of individuals, where consumers have the opportunity to select the insurance plan. In observational studies, there has been a concern about selection bias because those with poor health may opt for generous plans. As a result, alternatives with higher copayments may appear to "generate" lower health care expenditure not because of their superior control of moral hazard effects but because of their higher share of healthy enrollees. While this bias can in principle be avoided by specifying a sample selection mechanism, unmeasured health still is likely to influence both this mechanism and the quantity of medical care demanded.

Many insurance policies are characterized by a declining block tariff, brought about in particular by a deductible. This implies that medical care consumed during a given sickness episode increases the likelihood of profiting from a lower price during subsequent episodes.

This means that for a given deductible, the rate of coinsurance and hence the effective price of medical care depends on the quantity of care demanded – a choice variable. Unmeasured health again is likely to influence this choice, making estimation of an expected value of effective price difficult. The issue becomes even more thorny if rationing by the physician is present, which is encouraged by managed care plans, for it can be shown that the same episode-specific limit on medical services provided may not be binding during an early sickness episode but may become binding in subsequent ones.

Thus, a promising avenue for future research seems to be the modeling of the interaction of physician and consumer incentives. This would contribute to an understanding of the relative contribution the structuring of both types of incentives might make to the control of moral hazard effects in the delivery of health care.

References

Acton, J.P. (1975), "Nonmonetary factors in the demand for medical service: some empirical evidence", Journal of Political Economy 83(3):595–614.

Arrow, K.J. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53:941–973.

Bailit, H.L., et al. (1984), "Does more generous dental insurance coverage improve oral health? A study of patient cost-sharing", Journal of the American Dental Association 110:701–707.

Baumgardner, J.R. (1991), "The interaction between forms of insurance contract and types of technical change in medical care", RAND Journal of Economics 22:36–53.

Beck, R.G. (1974), "The effects of copayment on the poor", Journal of Human Resources 9:129–142.

Becker, G.S. (1965), "A theory of the allocation of time", Economic Journal 75:493–517.

Broyles, R.W., and M.D. Rosko (1988), "The demand for health insurance and health care: a review of the empirical literature", Medical Care Review 45(2):291–338.

Brook, R.H., et al. (1983), "Does free care improve adults' health? Results from a randomized controlled trial", New England Journal of Medicine 309:1426–1434. Also as (1984), "The effect of coinsurance on the health of adults: Results from the RAND health insurance experiment", RAND Publication R-3055-HHS (Santa Monica, CA).

Buchanan, J.L., et al. (1991), "Simulating health expenditures under alternative insurance plans", Management Science 37:1067–1089.

Burtless, G., and J.A. Hausman (1978), "The effect of taxation on labor supply", Journal of Political Economy 86:1103–1130.

Cheng, S.H., and T.L. Chiang (1997), "The effect of universal health insurance on health care utilization in Taiwan. Results from a natural experiment", Journal of the American Medical Association 278(2):89–93.

Cherkin, D.C., L. Grothaus and E.H. Wagner (1990), "The effect of office visit copayments on preventive care services in an HMO", Inquiry 27(1):24–38.

Cherkin, D.C., L. Grothaus and E.H. Wagner (1989), "The effect of office visit copayments on utilization in a health maintenance organization", Medical Care 27(11):1036–1045.

Crew, M.A. (1969), "Coinsurance and the welfare economics of medical care", American Economic Review 59(5):906–908.

Ehrlich, I., and G.S. Becker (1972), "Market insurance, self-insurance, and self-protection", Journal of Political Economy 80:623–648.

Enterline, P.E., V. Salter et al. (1973), "The distribution of medical services before and after 'free' medical care – the Quebec experience", New England Journal of Medicine 289:1174–1177.

Ellis, R.P., and T.G. McGuire (1993), "Supply-side and demand-side cost-sharing in health", Journal of Economic Perspectives 7:135–151.

Evans, R.G. (1974), "Supplier-induced demand: Some empirical evidence and implications", in: M. Perlman, ed. for the International Economic Association, The Economics of Health and Medical Care (Macmillan, London) 162–173.

Feldstein, M.S. (1973), "The welfare loss of excessive health insurance", Journal of Political Economy 81(1):251–280.

Frank, R.G., and T.G. McGuire (2000), "Economics and mental health", in: J.A. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 16.

Freiberg, L., and F.D. Scutchfield (1976), "Insurance and the demand for hospital care: an examination of moral hazard", Inquiry 13:54–60.

Fuchs, V.R., and M.J. Kramer (1972), "Determinants of expenditures for physicians' services in the United States, 1948–1968", Occasional Paper No. 117 (National Bureau of Economic Research, New York).

Gerdtham, U.-G., and B. Jönsson (2000), "International comparisons of health expenditure: theory, data and econometric analysis", in: J.A. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 1.

Gerdtham, U.-G., et al. (1992), "An econometric analysis of health care expenditure: a cross-section of the OECD countries", Journal of Health Economics 11(1):63–84.

Gertman, P.M., and J.D. Restuccia (1981), "The appropriateness evaluation protocol: a technique for assessing unnecessary days of hospital care", Medical Care 19:855–870.

Getzen, T.E. (1992), "Population aging and the growth of health expenditures", Journal of Gerontology: Social Sciences 47:S98–104.

Glied, S. (2000), "Managed care", in: J.A. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 13.

Goddeeris, J.H. (1984a), "Insurance and incentives for innovation in medical care", Southern Economic Journal 51:530–549.

Goddeeris, J.H. (1984b), "Medical insurance, technological change, and welfare", Economic Inquiry 22:56–67.

Greenwald, H.B. (1987), "HMO membership, copayment, and initiation of care for cancer: a study of working adults", American Journal of Public Health 77(4):461–466.

Grossman, M. (1972), "On the concept of health capital and the demand for health", Journal of Political Economy 80(2):223–255.

Grossman, M. (2000), "The human capital model", in: J.A. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 7.

Harris, B.L., A. Stergachis and L.D. Ried (1990), "The effect of drug copayments on utilization and cost of pharmaceuticals in a health maintenance organization", Medical Care 28(10):907–917.

Hausman, J.A. (1981), "Exact consumer's surplus and deadweight loss", American Economic Review 71:662–676.

Hausman, J.A. (1985), "The econometrics of nonlinear budget sets", Econometrica 53:1255–1282.

Heaney, C.T., and D.C. Riedel (1970), From Indemnity to Full Coverage: Changes in Hospital Utilization (Blue Cross Association, Chicago).

Helms, L.J., J.P. Newhouse and C.E. Phelps (1978), "Copayments and demand for medical care: the California Medicaid experience", Bell Journal of Economics 9:192–208.

Holmström, B. (1979), "Moral hazard and observability", Bell Journal of Economics 10(1):74–91.

Hughes, D., and A. McGuire (1995), "Patient charges and the utilization of NHS prescription medicines", Health Economics 4(3):213–220.

Keeler, E.B., J.P. Newhouse and C.E. Phelps (1977), "Deductibles and the demand for medical care services: the theory of a consumer facing a variable price schedule under uncertainty", Econometrica 45(3):641–656.

Keeler, E.B., and J.E. Rolph (1983), "How cost sharing reduced medical spending of participants in the health insurance experiment", Journal of the American Medical Association 249(16):2220–2227.

Keeler, E.B., et al. (1987), "Effects of cost sharing on physiological health, health practices, and worry", Health Services Research 22:279–306.

Keeler, E.B., and J.E. Rolph (1988), "The demand for episodes of treatment in the health insurance experiment", Journal of Health Economics 7(4):337–367.

Leibowitz, A., W.G. Manning et al. (1985a), "The effect of cost sharing on the use of medical services by children: interim results from a randomized controlled trial", Pediatrics 75(5):942–951.

Leibowitz, A., W.G. Manning and J.P. Newhouse (1985b), "The demand for prescription drugs as a function of cost sharing", Social Science and Medicine 21(10):1063–1069.

Leu, R.E., and R.J. Doppmann (1986), "Die Nachfrage nach Gesundheit und Gesundheitsleistungen" (Demand for health and health care services), in: G. Gäfgen, ed., Oekonomie des Gesundheitswesens (Duncker & Humblot, Berlin) 161–175, cited in: P. Zweifel and F. Breyer (1997) Chapter 4.4.

Lillard, L.A., W.G. Manning et al. (1986), "Preventive medical care: standards, usage, and efficacy", RAND Publication R-3266-HCFA (Santa Monica, CA).

Lurie, N., N.B. Ward et al. (1984), "Termination from Medi-Cal – does it affect health?", New England Journal of Medicine 311:480–484.

Lurie, N., N.B. Ward et al. (1986), "Termination of Medi-Cal benefits: a follow-up study one year later", New England Journal of Medicine 314(19):1266–1268.

Machina, M. (1989), “Comparative statics and non-expected utility preferences”, Journal of Economic Theory 33:199–231.

Machina, M. (1995), “Non-expected utility and the robustness of the classical insurance paradigm”, Geneva Papers on Risk and Insurance Theory 20:9–50.

Magid, D.J., T.D. Koepsell et al. (1997), “Absence of association between insurance copayments and delays in seeking emergency care among patients with myocardial infarction”, New England Journal of Medicine 336(24):172–1729.

Manning, W.G., H.L. Bailit et al. (1985), “The demand for dental care: evidence from a randomized trial in health insurance”, Journal of the American Dental Association 110:895–902.

Manning, W.G., and M.S. Marquis (1996), “Health insurance: The trade-off between risk pooling and moral hazard”, Journal of Health Economics 15(5):609–639.

Manning, W.G., J.P. Newhouse, N. Duan et al. (1987), “Health insurance and the demand for medical care: evidence from a randomized experiment”, American Economic Review 77(3):251–277.

Marquis, M.S. (1985), “Cost-sharing and provider choice”, Journal of Health Economics 4:137–157.

Marquis, M.S., and C.E. Phelps (1987), “Price elasticity and adverse selection in the demand for supplementary health insurance”, Economic Inquiry 25:299–314.

McAvinchey, I.D., and A. Yannopoulos (1993), “Elasticity estimates from a dynamic model of interrelated demands for private and public acute health care”, Journal of Health Economics 12(2):171–186.

McGuire, T.G. (2000), “Physician agency”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 9.

Moffitt, R.E. (1986), “The econometrics of piecewise-linear budget constraints: a survey and exposition of the maximum likelihood method”, Journal of Business and Economic Statistics 4:317–328.

Mooney, G.H., and M. Ryan (1993), “Agency in health care: getting beyond first principles”, Journal of Health Economics 12:125–135.

Morris, C.N. (1979), “A finite selection model for experimental design of the health insurance experiment”, Journal of Econometrics 11:43–61.

Newhouse, J.P., C.E. Phelps and W.B. Schwartz (1974), “Policy options and the impact of national health insurance”, New England Journal of Medicine 290:1345–1359.

Newhouse, J.P., and C.E. Phelps (1974), “Price and income elasticities for medical care services”, in: M. Perlman, ed., The Economics of Health and Medical Care (Macmillan, London).

Newhouse, J.P., and C.E. Phelps (1976), “New estimates of price and income elasticities”, in: R. Rosett, ed., The role of health insurance in the health services sector (National Bureau of Economic Research, New York).

Newhouse, J.P. (1978), The Economics of Medical Care. A Policy Perspective (Addison-Wesley, Reading, MA).

Newhouse, J.P., C.E. Phelps and M.S. Marquis (1980), “On having your cake and eating it too: econometric problems in estimating the demand for health services”, Journal of Econometrics 13(3):365–390.

Newhouse, J.P. (1981), “The erosion of the medical marketplace”, in: R. Scheffler, ed., Advances in Health Economics and Health Services Research, Vol. 2 (JAI Press, Westport, CT) 1–34.

Newhouse, J.P. (1988), “Has the erosion of the medical marketplace ended?”, Journal of Health Politics, Policy and Law 13(2):263–278, reprinted in: W. Greenberg, ed., Competition in the Health Care Sector: Ten Years Later (Duke University Press, Durham, NC).

Newhouse, J.P., et al. (1981), “Some interim results from a controlled trial in health insurance”, New England Journal of Medicine 305:1501–1507.

Newhouse, J.P., W.G. Manning et al. (1989), “Objective measures of health and prior utilization as adjusters for capitation rates”, Health Care Financing Review 10(3):41–54.

Newhouse, J.P. (1992), “Medical care costs. How much welfare loss?”, Journal of Economic Perspectives 6(3):3–21.

Newhouse, J.P., and the Insurance Experiment Group (1993), Free For All? Lessons from the Health Insurance Experiment (Harvard University Press, Cambridge).

Newhouse, J.P., E.M. Sloss, W.G. Manning and E.B. Keeler (1993), "Risk adjustment for a children's capitation rate", Health Care Financing Review 15(1):39–54.

O'Brien, B. (1989) "The effect of patient charges on the utilization of prescription medicines", Journal of Health Economics 8(1):109–132.

O'Grady, K., W.G. Manning, J.P. Newhouse et al. (1985), "The impact of cost-sharing on emergency department use", New England Journal of Medicine 313:484–490.

Pauly, M.V., and P.J. Held (1990), "Benign moral hazard and the cost-effectiveness of insurance coverage", Journal of Health Economics 9(4):447–461.

Phelps, C.E., and J.P. Newhouse (1972), "Effects of coinsurance: a multivariate analysis", Social Security Bulletin 35(6):20–29.

Phelps, C.E., and J.P. Newhouse (1974), "Coinsurance, the price of time, and the demand for medical services", Review of Economics and Statistics 56:334–342.

Porell, F.W., and E.K. Adams (1995), "Hospital choice models: a review and assessment of their utility for policy impact analysis", Medical Care Research and Review 52(2):158–195.

Ricci, E.M., P. Enterline and V. Henderson (1978), "Contacts with pharmacists before and after 'free' medical care – The Quebec Experience", Medical Care 16:256–262.

Rice, T., and K.R. Morrison (1994), "Patient cost sharing for medical services: a review of the literature and the implications for health care reform", Medical Care Review 51(3):235–287.

Roddy, P.C., J. Wallen and S.M. Meyers (1986), "Cost sharing and use of health services: the United Mine Workers' of America health plan", Medical Care 24(9):873–877.

Roemer, M.I., C.E. Hopkins et al. (1975), "Copayments for ambulatory care: penny-wise and pound-foolish", Medical Care 13(6):457–466.

Rogers, W.H., et al. (1991), "Effects of cost sharing on disability days", Health Policy 18:131–139.

Rosett, R.N., and L.F. Huang (1973), "The effect of health insurance on the demand for medical care", Journal of Political Economy 81:281–305.

Scheffler, R.M. (1984), "The United Mine Workers' health plan: an analysis of the cost-sharing program", Medical Care 22(3):247–254.

Scitovsky, A.A., and N.M. Snyder (1972), "Effect of coinsurance on use of physician services", Social Security Bulletin 35(6):3–19.

Scitovsky, A.A., and N.M. McCall (1977), "Coinsurance and the demand for physician services: four years later", Social Security Bulletin 40:19–27.

Selby, J.V., B.H. Fireman and B.E. Swain (1996), "Effect of a copayment on use of the emergency department in a health maintenance organization", New England Journal of Medicine 334(10):635–641.

Shapiro, M.F., J.E. Ware and C.D. Sherbourne (1986), "Effects of cost sharing on seeking care for serious and minor symptoms: results of a randomized controlled trial", Annals of Internal Medicine 104(2):246–251.

Sintonen, H., and I. Linnosmaa (2000), "Economics of dental services", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 24.

Siu, A.L., F. Sonnenberg et al. (1986), "Inappropriate use of hospitals in a randomized trial of health insurance plan", New England Journal of Medicine 315:1259–1266.

Taylor, L.D. (1975), "The demand for electricity: a survey", Bell Journal of Economics 6:74–110.

US Office of Technology Assessment (1992), "Does Health Insurance Make a Difference? (US Government Printing Office, Washington, DC).

Valdez, R.B., R.H. Brook, W.H. Rogers et al. (1985), "Consequences of cost-sharing for children's health", Pediatrics 75(5):952–961.

Valdez, R.B., A. Leibowitz, J.E. Ware et al. (1986), "Health insurance, medical care, and children's health", Pediatrics 77(1):124–128.

van de Ven, W.P.M.M. (1983), "Effects of cost-sharing in health care", Effective Health Care 1(1):47–58.

Viscusi, W.K. (1995), "Government action biases in risk perception, and insurance decisions", Geneva Papers on Risk and Insurance Theory 20:93–110.

Wedig, G.J. (1988), "Health status and the demand for health", Journal of Health Economics 7:151–163.

Weisbrod, B.A. (1991), "The health care quadrilemma: an essay on technological change, insurance, quality of care and cost containment", Journal of Economic Literature 29:523–532.

Willig, R. (1976), "Consumer's surplus without apology", American Economic Review 66:539–597.

Zweifel, P. (1984), "Technological change in health care: why are opinions so divided?", Managerial and Decision Economics 5(3):177–182.

Zweifel, P. (1988), "Premium rebates for no claims: the West German experience", in: H.E. Frech III, ed., Health Care in America (Pacific Research Institute, San Francisco, CA) Chapter 9.

Zweifel, P. (1992), Bonus Options in Health Insurance (Kluwer, Boston).

Zweifel, P. (1995), "Diffusion of hospital innovations in different institutional settings", International Journal of the Economics of Business 2(3):465–483.

Zweifel, P., and L. Crivelli (1996), "Price regulation of drugs: lessons from Germany", Journal of Regulatory Economics 10(3):257–274.

Zweifel, P., and F. Breyer (1997), Health Economics (Oxford University Press, New York).

Zweifel, P., S. Felder and M. Meier (1999), "Ageing of population and health care expenditure: a red herring?", Health Economics 8:485–496.

This Page Intentionally Left Blank

Chapter 9

PHYSICIAN AGENCY*

THOMAS G. McGUIRE

Boston University, Economics Department

Contents

*Financial support from the National Institute of Mental Health (K05-MH01263) is gratefully acknowledged. I am grateful to Margarita Alegria, David Cutler, Randy Ellis, Richard Frank, Jacob Glazer, Miriam Hatoum, Hsien-ming Lien, Albert Ma, Joseph Newhouse, and Mark Pauly for their comments.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This chapter reviews the theory and empirical literature on physician market power, behavior, and motives, referred to collectively as the issue of "physician agency." The chapter is organized around an increasingly complex view of the demand conditions facing a physician, beginning with the most simple conception associated with demand and supply, and building through monopolistic competition models with complete information, and finally models with asymmetric information. Institutional features such as insurance, price regulation, managed care networks and noncontractible elements of quality of care are incorporated in turn. The review reveals three mechanisms physicians may use to influence quantity of care provided to patients: quantity setting of a nonretradable service, influencing demand by setting the level of a noncontractible input ("quality"), and, in an asymmetric-information context, taking an action to influence patient preferences. The third mechanism is known as "physician-induced demand." The empirical literature on this topic is reviewed. Theories based on alternatives to profit-maximization as objectives of physicians are also reviewed, including ethics and concern for patients, and the "target-income" hypothesis. The target-income hypothesis can be rejected, although there is empirical support for non-profit maximizing behavior.

Keywords

agency, monopolistic competition, quantity-setting, physician-induced demand, target income, insurance, price regulation, networks, managed care

JEL classification: I11, L84

1. Introduction

Physician behavior is the central issue in health economics, as many writers have recognized. Arrow (1963) challenged economists to deal with asymmetric information, absence of markets for risk bearing, and the privileged social role of physicians. Making the analogy between physician payments and return to capital in the rest of the economy, Pauly (1980) observed that although profits are a small part of value added (less than the 20% of health costs represented by physician payments), return to capital directs capitalists' decisions about investment and production. Similarly, payments to physicians are laced with incentives, and these incentives, to hospitalize, to treat, to take time to diagnose carefully, and over a longer term, to select a geographic location, mode of practice or specialty of practice, direct resources in health, thus determining costs and outcomes. Fuchs (1974) aptly called the physician the "captain of the team." Managed care notwithstanding, drugs, surgery, and other health care inputs cannot be had without physician initiative and concurrence. Physicians are trained to exercise this authority.

This chapter reviews the theory and empirical research on how physicians influence the medical services used by patients. In so doing, it confronts fundamental questions about physician motives and market power, referred to together here as the issue of "physician agency." Special attention will be devoted to a new literature dealing with physician behavior in the context of incentives created by managed care. The goal of this chapter is to draw on the contributions of many writers to develop a working model of physicians that can handle the key elements of physician and patient interaction and the associated institutions.[1] Literature reviewed will be drawn from the field of health economics.[2]

In neoclassical theory, the firm sets price and quantity in order to maximize profit subject to the constraint of market demand. Every phrase in this paradigm has been questioned when it is applied to physician-firms. Do physicians maximize profit? Many have argued that physicians are motivated differently than other business people, that they are, for example, concerned for their patient's health, and make different tradeoffs when it comes to their own gain or the utility of their patient-customers. Are physicians constrained by market demand? Physicians are said to work at an advantage in relation to their customers because of their superior knowledge of the patient's medical condition and of what treatments are likely to be most helpful. According to this argument, physicians behave differently because they may exercise market power in

[1] Feldman and Sloan (1988) and Gaynor (1994) contain useful reviews of the market for physicians' services.

[2] Medical sociology also contains a very large literature on physician and patient behavior. For one overview, see Mechanic (1990). Some of the literature in medical sociology takes a different approach to modeling choice and action, rejecting the "rational choice-based approach" employed in economics and emphasizing group structure and social networks in determining actions. See Pescosolido (1992).

ways inaccessible to other sellers, controlling, not being constrained by, patients' demand. On the other hand, it is possible to pose the question in current medical markets, do physicians even set price and quantity? Ubiquitous third party payers set prices to patients through terms of coverage and to physicians through terms of reimbursement. Managed care plans seek to impose their will on physicians' decisions about treatment. In light of these complications, it is not surprising that doctors and patients are not a ready application of neoclassical economic theory.

At the same time, using currently available textbooks as a point of reference, there is no consensus about an alternative approach to physician–patient interaction. Some texts present no model at all. Among the texts committing to a model, no model or approach garners more than one vote from the electorate of authors. And no text uses a model as a general way of organizing the discussion of physician behavior.[3]

Textbook authors avoid commitment perhaps because the paradigm of the profit maximizing seller subject to a market demand constraint is not well-accepted in health, while at the same time there is no agreeable alternative. Authors of theoretical papers in journals question the neoclassical paradigm and create models to isolate the implications of some element of the physician–patient interaction (regulated prices when quality is noncontractible, for example), and while this work adds insight, the balance of the model may ignore other elements that matter for other purposes. Furthermore, fundamentals of the problem, motives, power, and imperfect information, are very complex. The social institutions overlaying doctor and patient decisions, insurance, reimbursement, and recently, managed care, themselves add to the challenge.

Empirical studies in health have a disconcerting tendency to turn up results running counter to simple neoclassical models. Writing in the inaugural issue of the *Journal of Law and Economics*, Kessel (1958) confronted the paradox of how hundreds of thousands of physicians could price discriminate, when competition among profit-maximizing firms must eliminate the practice. Kessel rejected a "charity hypothesis" as an explanation (which altered physician motives, replacing profit-maximization by

[3] Eastaugh (1992) reviews a classification of possible models based on the two dimensions of number of sellers in the market and whether the physician can induce demand, without presenting an explicit model. Feldstein (1979) uses a simple monopoly model. Phelps (1997) presents a model of monopolistic competition, but without insurance, price regulation, or incorporation of information issues. Santerre and Neun (1996) review elementary market models and relate them to physician markets, without proposing any particular conception of the working of the market. Getzen (1997) alludes to asymmetric information and agency, and imperfect competition but presents no model of the physician-firm. Folland, Goodman and Stano (1997) refer to the principal/agent problem and the concept of perfect agent (defined as what the patient would do with the information) but also have no model. The physician faces a downward demand curve and may induce demand by moving out the demand curve. There is no integration of an agency perspective. Zweifel and Breyer (1997) is the only text I am aware of that proposes a specific model with an information/agency component. Their model is very special however, since it regards outcome as contractible (that is, a basis for payment). Most approaches in the journals regard output (quantities of care), not outcome (health), as a potential basis of payment in contracts. Rice (1998) contains a critique of the welfare economics of health care, but no explicit positive model of physician price and quantity setting.

a shared ethic to provide health care even if patients lacked an ability to pay) in favor of the hypothesis that physicians set prices as a collusive oligopoly. He accused the American Medical Association (AMA) of coordinating physician pricing, threatening sanctions in the form of denial of membership in the local AMA branches and hospital privileges if patients with lower demand elasticity (the rich) were not charged a higher price than those with higher elasticity (the poor).[4] That an outside organization (even the 1950's AMA!) could coordinate the intimate economic exchange between patients and doctors seems, in hindsight, an unlikely explanation for price discrimination.

Perverse empirical results, with the signs of fundamental economic relations reversed, continued to emerge from more formal econometric research.[5] Fuchs (1978) found that an increased supply of surgeons, controlling for demand factors, *increased* market price. Rice (1983) found that a decreased price of physician services caused an *increase* in supply of services. As a possible explanation, Fuchs, Rice, and others proposed that physicians sacrifice profit to pursue a "target income." The target income hypothesis does the job in a mathematical sense, replacing the normal positive p and q relation governed by an upward-sloping supply curve by the negative relation given by the rectangular hyperbola $p \times q = T$, where T is the target.[6] The resolution comes at a cost, however, of introducing an objective for physicians that many regard as implausible.

Another set of mundane facts motivates the modeling efforts in health care. In a patient's contact with the doctor, the doctor's position is not, "here is the price of my services, how many do you want?" It is more like, "here is what you should do." Physicians set quantity, and more generally, the treatment patients should use, and may make this decision in light of the factors affecting them. Empirical studies [e.g., Gaynor and Gertler (1995)] find that when normal demand-side variables, such as demand-price, income, and clinical need are controlled for, variables affecting the supply of care, such as supply price, physician attitudes, or partnership incentives, influence what happens to the patient. How is this to be understood?

Physicians "induce demand" is one answer. The physician-induced demand (PID) hypothesis, associated with Evans (1974), is essentially that physicians engage in some

[4] Most of Kessel's fascinating article is devoted to a social history of the AMA's opposition to prepaid group practices. Kessel interprets this opposition as occurring because prepaid groups, by charging a uniform premium for membership, prevent physicians from being able to price discriminate. One can accept that the AMA's resistance to prepaid groups was economically motivated without needing to appeal to the price discrimination hypothesis. Prepaid groups reduced demand for physician services by imposing non-price rationing, thereby reducing physician income.

[5] In the first econometric test of the functioning of markets in health care, Martin Feldstein (1970) found (using a time series of national data) an upward-sloping demand and a downward-sloping supply curve, spawning another special theory of physician behavior. He proposed that physicians set prices below market clearing in order to ration demand to retain an ability to pick and choose among their patients to treat the "interesting cases." See also, Steinwald and Sloan (1974), Sloan (1976) and Dyckman (1978).

[6] In this, consider p to be the margin above cost. Speculation about a target income held by doctors can be traced back to Feldstein (1970).

persuasive activity to shift the patient's demand curve in or out according to the physician's self-interest. Patients have incomplete information about their condition, and may be vulnerable to this advertising-like activity.

In recent years, the economic theory of physician behavior has emphasized contracting and information issues. Economists have recognized for a long time that a significant market is missing in the health sector: payments or insurance based on health outcomes. Arrow (1963) observed that an efficient (first-best) health insurance policy would specify payment contingent on the individual's state of health. The moral hazard problem would be resolved if an individual who suffered a sudden health problem were paid a specified amount by the insurance company; afterwards, the individual could make his own decision to purchase health care. A state-contingent payment scheme protects the individual from the financial risk of illness *ex ante* and retains incentives for the patient to consume health care efficiently *ex post*. Nevertheless, insurance policies or physician payments contingent on health status are nonexistent because health status is too costly to verify. A second contractibility problem is also important and fundamental. Some elements of treatment are not reported. Insurance coverage and provider payment are based on reports of measures such as number of "visits" or "days" in the hospital, or accounting "costs," which only partially reveal the resources devoted to treatment. A physician or other health care provider must be relied upon to prescribe the clinical content of the services connected with a "visit" or a "day," and to invest (costly) effort into making these services productive in terms of the patient's health. Thus, the physician almost always supplies her own input into the production of health care for the patient. This input, which is often referred to as "effort," but also could be understood as "quality," is simply not contractible; the market for insurance and payment policies based on the physician's effort is also missing.[7]

The economic problem in health care transactions can be deeper than issues of contractibility. An output (change in health status) or an input (physician effort) may not even be *observable*, let alone contractible. If output or some inputs are known to the physician but unobservable to the patient (and a third-party), the problem of *asymmetric information* is introduced. Information asymmetry is used to motivate many papers in health economics, even if the information issues are not modelled explicitly.

Following the introductory material in Section 1, Section 2 begins with the simplest model of the physician market, demand and supply. Historical and current information about prices, income, supply, and specialization are presented and interpreted within a demand and supply framework. The possible collusive role of monopolistic practices, such as restrictions on entry into medical school are considered. This first economic perspective on the market remains useful for certain purposes, particularly for understanding trends in physician income and supply. As is well-known, demand and supply

[7] Ma and McGuire (1997) point out that even "visits" or "days" are not contractible, only the *reports* of visits or days are contractible. The requirement that patients and physicians report usage truthfully puts limits on the ranges of feasible insurance-payment systems.

are generated by price-taking buyers and sellers. And while the demand and supply framework has some utility, it does not come to grips with the special decisionmaking processes that characterize patients and physicians. To do so requires an explicit model of the behavior of the physician firm (and its customers). For the remainder of the chapter, the focus will be on behavior at the firm level.

Section 3 places physician-firms in a monopolistically competitive market, the versatile market structure favored by health economists. It addresses models of complete information – that is to say, models with no uncertainty on the part of either the physician or the patient about the benefits of medical care (and thus, no asymmetric information). Patients in this section have a stable set of preferences. For institutional reasons, we regard prices as being set by payers. Physician quantity-setting, however, emerges naturally within this model. The nonretradability feature of physician services implies that profit-maximizing physicians do not allow patients to choose the utility-maximizing quantity given the price the patient pays. This is the first of three ways a physician can influence quantity identified in this chapter.

Within a model of complete information, physicians can decide about quality or effort – a key input into health, which, even though observed by patients, may be impossible to verify. This is the second mechanism for physicians to influence quantity: through choice of a noncontractible input (quality) that influences patient demand. The input is observable by the patient (hence its effect on choice) but cannot be paid upon by a payer. Insurance, price regulation, physician contracting, and managed care are all introduced in Section 3.

Section 4 addresses information issues, resting, as Arrow (1963) and others have argued, at the heart of the physician–patient relationship. This section first considers the effect of shared uncertainty between patients and doctors, and then introduces the important element of asymmetry of information between patient and physician. When patients believe doctors know more than they do, doctors may be able to persuade patients to demand more or less care. For this to work, information must be asymmetric, and the physician must be taking an unobservable action to influence demand. This third mechanism for quantity determination captures the meaning of "physician-induced demand" as it is used in the literature.

As a preview to the material covered in Sections 3 and 4, Table 1 contains a summary of the three mechanisms – direct quantity setting stemming from nonretradability, altering demand by setting an observable but noncontractible input (quality), and persuasion based on asymmetric information. Some of the empirical evidence that supports the existence of these mechanisms is discussed in Sections 3 and 4, though these sections are primarily theoretical. Sections 3 and 4 study the implications of physician profit maximization within increasingly complicated demand environments.

The large body of empirical research in health economics concerned with the existence of induced demand is reviewed in Section 5. Although there has been a clear and widely accepted definition of physician induced demand (PID) for more than 25 years, the measurement and meaning of PID has been one of the most contentious issues in health economics.

Table 1
Mechanisms of physician determination of the quantity of health care

	Nonretradability allows quantity setting	Choice of noncontractible input	Persuasion
Market structure	Monopolistic competition	Monopolistic competition	Monopolistic competition
Information	Complete	Complete	Asymmetric
Physician action influencing use	NA	Not contractible	Unobservable
Main features	Supply determination within demand constraints; can explain inverse (P, Q) relationship	Demand response to "quality" or some other physician input	Physician takes action to persuade; constrained by demand response or ethics
Illustrative paper	Farley (1986)	Ma and McGuire (1997)	Dranove (1988)
Section covered	3	3	4, 5

Section 6 is concerned with physician motivation, and gives consideration to alternatives to profits as objectives guiding physician behavior. Standards of practice and ethics, concern for patient welfare, and pursuit of a target income are all considered theoretically and empirically.

In undertaking a review of physician behavior, a decision has to be made about how much respect should be paid to the conventional model of a profit maximizing firm constrained by market demand. As an expedition in "normal science," a review regards observations and empirical findings in health economics as puzzles to be solved, if possible, within the broad paradigm of neoclassical economics. On this front, Pauly (1980, p. 177) advised some time ago:

> We should not be too quick to depart from standard maximizing economic models in order to explain behavior in the medical care industry. Supposedly anomalous features of that industry sometimes vanish when more appropriate sets of data are used, while other apparent institutional differences require only redefinitions of price, ownership-entrepreneurship, and markets to make models analogous to the traditional ones applicable.

Kuhn's theory or paradigm is "an object for further articulation and specification under new or more stringent conditions" (1970, p. 23). At some point, of course, the articulations become so unwieldy that the standard paradigm should be rejected altogether. But this judgment can only be made after a systematic effort to integrate the ideas and findings within the dominant paradigm, and in the presence of an alternative theory.

2. Demand and supply for physician services

The demand and supply model of physician markets has appeal when the object of study is the industry, not the firm, and the focus is on aggregate supply, as measured by number of physicians, or on prices as measured by average wage or annual income. Acknowledging that a more complex model is necessary to explain the details of price and quantity setting at the patient level, competition and entry conditions nonetheless govern income and average compensation. Historically, the demand and supply perspective has been used to study the effects of restrictions on entry into medical school, for example. As physicians begin to sell their services to organized buyers in managed care, the forces of demand and supply can be expected to determine the terms doctors can expect in those contracts.

2.1. Prices and quantities

Physician incomes throughout most developed countries are very high, among the highest for any occupational group. In the 1970's and 1980's, physician earnings in the US grew relative to college graduates and lawyers [Gaynor (1994)]. In 1994, average physician net income, according to the AMA, was $182,400.[8] Some statistics about the distribution of physicians by specialty and their incomes are contained in Table 2. The number of physicians supplying health care is regulated by licensing laws, including provisions for students trained outside the US, and the capacity of the 126 US medical schools. About 16,000 students graduate from U.S. medical schools each year, and these are joined by 5,000 immigrant physicians. The accumulation of these flows has built the current stock of 550,000 physicians in active practice, a stock that is growing at about 1.5% per year. More than half of all physicians are part of a group practice.

In terms of physician-to-population ratios, for various years in 1990's, the US at 254 physicians per 100,000 population is higher than the UK (164), Japan (177), and Canada (221), but lower than many other high-income countries, such as France (280), Sweden (299), Germany (319), or Belgium (365).[9]

Physicians average about 55 hours of work per week. Figuring 48 weeks per year, the average physician net wage was $65 per hour in 1994. Net is about half of gross (Table 2 contains details and sources) so the average price charged by physicians in 1994 could be figured at approximately $130 for each hour of their time. This of course varied according to the specialty of the physician, and the work activity involved. The geographical distribution of physicians at the state level has been successfully studied using a basic demand and supply model [Benham et al. (1968), Fuchs (1978), Frank (1985)].

2.2. Entry conditions and monopoly profits

At one time, there were few institutional interferences in the operation of market forces in physicians' services. In the middle of the 19th century, skilled laborers earned more

[8] 1994 was the first year in which physician earnings fell in nominal terms, according to data collected by the AMA [Simon and Born (1996)].

[9] Information on physician population ratios is from various years in the 1990s, available from the World Health Organization at www.who.int/whosis.

Table 2a
Physician income 1994

	% of total	Average income
All physicians		$182,400
Specialty		
General/family practice	15%	$121,200
Pediatrics	8%	$126,200
Psychiatry	7%	$128,500
Internal medicine	28%	$174,900
Pathology	3%	$182,500
Obstetrics/gynecology	6%	$200,400
Anesthesiology	5%	$218,100
Radiology	5%	$237,400
Surgery	14%	$250,200
Type of practice		
Employee or contractor	33%	$158,400
Self-employed	67%	$210,200
Solo – 49%		
2 + MDs – 51%		

Note: Income in net of practice expenses.

Table 2b
Physician expenses 1994 (office practice costs)

Average expenses of self employed	$183,100
Distribution	
Non-physician employee wages	35%
Office rent	26%
Medical supplies	8%
Malpractice liability insurance	12%
Equipment	3%
Other expenses	16%
Total	100%

Source: Getzen (1997), from AMA publications.

than the average physician [Starr (1982, p. 84)]. Licensure was uneven among the states. Physicians came in many different flavors, with homeopaths, who believed like cured like, competing with, among other groups, allopaths, who believed in cures by opposites [Frech (1996, pp. 52–53)].

The free market in curing was eliminated by the AMA's lobbying for uniform national licensing. Between 1880 and 1890, every state licensed medicine. With the reorganization and growth of the AMA between 1900 and 1910, and a rise in membership from 8,000 to 70,000 [Starr (1982, pp. 110, 112)], membership began to be associated with hospital privileges and control of expert witnesses in malpractice cases. Reform of medical education was also a goal of the AMA. State medical boards were beginning to regulate medical schools. Flexner accelerated the process and helped transform medical training into its present form. In his 1910 report, *Medical Education in the US and Canada*, Flexner recommended an increase in the quality of medical education, uniform training (based on the Johns Hopkins model), and a decrease in the number of schools and students per school.

Since the Flexner Report, the number of medical schools and their capacity has been subject to control by the AMA. Between the period of 1910 and 1965, restrictions on the number of physicians trained decreased the supply of physicians in relation to population, from about 1.6 physicians per thousand in 1910 to less than 1.3 in 1965, a remarkable economic achievement during a period of rapid growth in income and insurance coverage driving up per capita demand for medical care.[10]

The federal government intervened with the Health Professions Educational Assistance Act of 1963, increasing the capacity of US medical schools and making immigration easier for foreign-trained physicians. Output of medical schools doubled to 15,000 between 1965 and 1975. Since 1965, the physician-to-population ratio has crept steadily upward.[11]

A demand and supply framework can be used to address whether the limited supply of physicians has protected physician incomes and generated rents. There are more than two applicants for every spot in medical school. If medical school tuition were set at cost, this in itself would be *prima facie* evidence for a restriction on entry leading to rents. However, since medical school tuitions only cover a fraction of the cost of education [Ganem et al. (1995)], excess demand for spaces at the subsidized price does not necessarily imply the number of doctors is below the quantity at which demand would equal supply at a market-clearing price for training. The number of physicians could be at the competitive level, but the cost of training is subsidized, perhaps to enable medical schools to collectively choose a pool of new physicians with desirable social characteristics ("high quality," ethnic diversity, social consciousness, geographic distribution).

Figure 1 contains a simple demand and supply diagram for medical school places. The number of places at medical schools is set by administrative policy. Tuition is subsidized. We observe then an excess demand at the subsidized price. If there were no

[10] One of the side effects of central control of medical school admissions was the spread of discriminatory admission practices. Medical schools increased discrimination against women, blacks, and Jews. Some medical schools stopped admitting women, or limited enrollment to five percent of the students [Starr (1982, p. 124)]. Black medical schools fell from seven pre-Flexner to two post-Flexner. As a result, the percentage of physicians who were women or who were black fell for fifty years after Flexner [Frech (1996, p. 54)]. Jews suffered quotas. For example, more than half of the applicants to Cornell Medical College in New York City in 1940 were Jewish, but the school limited the number of admits to 10–15 percent, making it ten times harder to get in for a Jew than for a non-Jew [Frech (1974, p. 125)].

[11] The Health Professions Educational Assistance Act of 1976 restricted the inflow of immigrant physicians. Noether (1986) has emphasized the increasing role of competition from foreign medical graduates (FMGs) in American medicine.

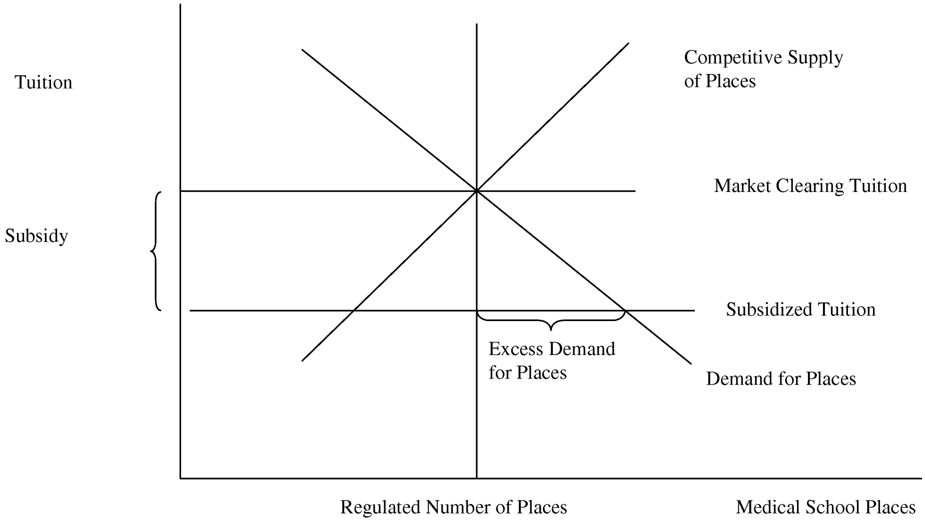


Figure 1. Demand and supply of medical school places with tuition subsidy.

subsidy, the number of places demanded would be less than are demanded at the subsidized price. It is impossible to tell if the quantity demanded at the full price is more or less than the regulated quantity. Figure 1 illustrates one possibility, that they are equal.

A large number of empirical studies have looked at the question of entry restriction, attempting to measure any "excess return" on the student's investment in medical education. In an era before federal subsidy of medical education, Friedman and Kuznets (1954) found evidence of monopoly returns that they attributed to AMA restrictions on entry. The presence of a tuition subsidy for medical schools (which does not generally exist for law or business schools) implies that, if supply were restricted to the competitive level, even the marginal physician would enjoy rents (equal to the subsidy – see Figure 1). More recent studies which evaluate the return to tuition investment contend with evaluating students' opportunity costs and correcting for physicians' long hours, without a clear finding of excess return due to entry restrictions in medicine [see, e.g., Sloan (1970), Leffler (1978), Marder et al. (1988)]. Weeks et al. (1994), for example, studied the return to education for physicians, lawyers, dentists, and MBAs, and found post-high school rates of return of 20–25% for all groups.

It seems very plausible that during the first half of this century, restrictions on the output of medical schools elevated physicians' economic position. When supply is constant and demand increases a lot, sellers benefit. As we have noted above, during the 50 years between 1900 and 1950, the number of new medical school graduates held constant.[12] Meanwhile, using figures from the *Statistical Abstract of the United States*, the population of the US doubled (from 76 million to 151 million), the average age of the population increased from 23 to 30, and real per capita Gross Domestic Product increased by a factor of six. Insurance coverage was also beginning to be prevalent by the end of this period, and doctors were becoming more competent in contending with disease. All of these factors, population, age, income, insurance, and technologically driven increments in value, increased demand.

More recent studies, using data on incomes, correcting for opportunity cost and hours worked, do not support a clear conclusion that at the current rate of production of medical schools, the output of doctors keeps physicians' incomes above a competitive level. In the last 10–20 years, outputs of medical schools have expanded considerably. In addition, market forces may be responsible for the dissipation of rents created by the original restrictions on entry. Buyers of health care – hospitals, insurers, prepaid plans, as well as individual patients – may have substituted against high-wage physicians, making more use of nurses and other "physician extenders." If physician net incomes include a rent, this will tend to attract applicants, allowing medical schools to choose students who are "high quality," exactly those who are likely to have a high opportunity cost. Some rents may have been transferred back to buyers by price restrictions. Finally, as Gaynor (1994) has argued, increasingly aggressive antitrust activity directed at physicians may have inhibited collective physician exercise of market power.

2.3. Competition among physicians

For a given number of physicians, it is in the interest of the group to minimize competition among themselves. Kessel (1958) proposed that organized medicine operated a pricing conspiracy. An alternative to price collusion is an agreement to divide the market, a generally more effective way to inhibit competition because it prevents unwanted competition on non-price as well as price dimensions. Market division maintains the inelasticity of each practitioner's demand.

The counter strategy by buyers is to increase the sellers' demand elasticity. Buyers' interest in increasing the elasticity of demand facing sellers has been appreciated in health care markets. As Dranove, Shanley and White (1993) have argued in the case of hospitals, and Scherer (2000) in the case of pharmaceuticals, when patients make choices of medical supplier, demand for suppliers is likely to be inelastic. When organized buyers (insurers, HMOs) make choices, however, demand can be more elastic, driving prices downward. The same is likely to hold true in physician markets. Organized medicine worked for many years to protect the "sacred" physician–patient relationship, ensuring that patients, not third-parties, made choices of doctors. As Reinhardt (1996, p. 9) recently put it, "Until about the mid-1980's the patient's freedom

[12] The number of other health personnel, such as nurses, did increase substantially over this period. While this had a mitigating effect on the restriction of supply of doctors, it is further evidence that there was a large increase in demand for health care services.

to choose among the doctors and hospitals *at the time of illness* was sacrosanct in the United States" [his emphasis]. This institution of patient choice was maintained by political activity by organized medicine. In a losing battle, the AMA declared "contract medicine," whereby a doctor might be hired by an employer to provide health care to its workers, unethical in 1969 [Frech (1996, p. 69)]. Kessel (1958) and Havighurst (1978) document organized medicine's longstanding opposition to prepaid group practices.[13] Scholars new to health economics may find this surprising since today, physician contracting takes so many forms, and prepaid group practices are a well-established set of institutions.

Notwithstanding attempts to prevent intraprofessional competition, there is ample evidence that competition among physicians, and between physicians and substitutable professionals, is in force. Frank (1985) estimated a demand and supply model of physician pricing and location, finding a standard relationship between price and quantity. Frank et al. (1987) studied earnings of podiatrists, foot specialists who compete primarily with orthopedic physicians. More competition from orthopedists as well as within their own specialty reduced earnings, even in a model in which any simultaneity between practitioner per population and earnings was not considered. Newhouse et al. (1982) found that the geographical distribution of physicians was consistent with competitive location theory. Benham et al. (1968) and Escarce et al. (1998) provide evidence that locational choices respond to demand conditions.

The demand and supply framework is useful for discussing aggregate trends in physician earnings, earnings by specialty, and earnings by region. [See, e.g., Simon and Born (1996).] Physicians are relying on contracts with third parties for a rapidly increasing share of their earnings [Emons and Wozniak (1997)]. In a recent survey of health plans, Rosenthal et al. (1999) reports that in 1998, 63% of HMOs in California pay most of their primary care practitioners by some form of capitation. The terms of these contracts, e.g., the level of capitation, are likely to be governed by demand and supply forces, and represent an important new area for research. As we noted in the beginning of this section, demand and supply analysis will be useful in study of the overall level of physician compensation, particularly when the buyer is a managed care plan, in a position to substitute one physician for an other, or to substitute other personnel for physicians.

Competitive analysis relies heavily on the use of a "zero-profit" condition. Entry drives profits at the margin to be equal to zero, and this analysis can be used to understand specialty and location, capitation payments, or other issues in compensation. At the level of the individual patient, however, entry conditions are not sufficient to understand the question of quantity setting. To attend to treatment determination requires us

[13] For example, the Group Health Association (GHA) in Washington, DC, an early prepaid group was deemed to violate the AMA's code of ethics. Physicians worked for a salary in an organization not controlled by doctors. Furthermore, patients' choice of doctor was restricted to a physician member of the plan. The local society expelled or disciplined physicians affiliated with GHA, and prohibited members from associating with GHA physicians. The local society was successfully prosecuted by the Justice Department for antitrust violations. See Havighurst (1978).

to move beyond demand, supply and entry conditions, and to study the implications of profit maximization.

3. Physician behavior with complete information

In this section we assume that demand conditions are given to the physician. Patients have fixed preferences for health care, and physicians can take no action to change those preferences. Furthermore, we assume in this section that information is complete. The physician recognizes patient preferences, and patients can observe and evaluate the characteristics of care supplied to them by doctors. These assumptions can be defended for at least some areas of medical care. Pauly (1988) has argued that patients can be regarded as being reasonably well-informed for about one quarter of the care they consume, such as for routine care and for care of chronic illnesses. The assumption that patients can accurately judge the value of the health care they receive is implicit in the large literature on demand which employs evidence about demand response to infer implications for optimal insurance. [See, e.g., Manning et al. (1987), Feldman and Dowd (1991); see Rice (1998) for a critique.]

We also assume in this section that physicians maximize profit. Eisenberg (1986), a physician, contends that physicians are motivated by financial self-interest, concern for their patients, and concern for the social good, devoting a chapter to each in his thoughtful book on physician decision making. His first chapter is about self-interest, and that will be our starting point too.

As we will see, in a complete information model, physicians possess two ways to influence quantity. First, with market power, the physician can set the quantity of a nonretradable service (Sections 3.1, 3.2). Second, if there is an element of noncontractible quality in medical care supply, the physician can influence quantity by choice of this input (Section 3.3).

3.1. A monopolistically competitive firm selling a service

In virtually all characterizations of physicians in economics journals and textbooks, the physician is portrayed as having some market power. Monopolistic competition is a versatile structure for representing market power, and is the expressed favorite of many writers [Frech and Ginsburg (1975), Pauly and Satterthwaite (1981), McGuire (1983), Klevorick and McGuire (1987), Dranove (1988), Dranove and Satterthwaite (1991, 1992, 1999), Getzen (1984), Zuckerman and Holahan (1991), Pauly (1979, 1991), Phelps (1997), Frech (1996), Newhouse (1978), Folland et al. (1997), Gaynor (1994)]. Monopolistic competition includes an element of monopoly (downward-sloping demand) and an element of competition (large number of competitors – each firm ignoring strategic interactions). Because of location, specialty, quality, or some other element of taste, patients do not regard physicians as perfect substitutes. Information imperfections could also generate a monopolistically competitive structure. There may in fact be good substitutes for a given physician, but if the patient doesn't know who these are, the patient may be willing to pay more for the services of the familiar doctor, an idea proposed by Pauly and Satterthwaite (1981). For now, however, we will keep informational issues in the background and simply assume that the structure is monopolistically competitive because of recognizable differences among physicians.

A point in favor of monopolistic competition is that it is an appealing alternative to models that rely on collusion among physician to explain some observed patterns of price and quantity. Consider Fuchs' (1974, p. 71) story of the "surgeon surplus:" "A comprehensive, detailed study of general surgeons in one suburban community in the New York metropolitan area revealed that the surgical workload of the typical surgeon was only about *one-third* of what experts deemed a reasonably full schedule." Fuchs further characterized the market as follows: "For most types of surgery, the quantity physicians would like to supply at the going price is far greater than the quantity demanded." How can this excess supply persist? Why does competition fail to reduce price and increase demand and workload? A conspiracy is the explanation that Kessel might have proposed. Surgeons might be colluding to keep prices high, to maximize their joint profits. But more plausibly, as Fuchs notes (p. 73), "Many surgeons believe, perhaps rightly, that demand would not increase appreciably in response to a price cut. . . ," in other words, each physician faces a downward-sloping demand. Collusion is not required to observe price above marginal cost and for there to be "excess supply."[14]

The classic evidence for a monopolistically competitive structure is a demand curve with a negative slope [Haas-Wilson (1990), Klevorick and McGuire (1987), McCarthy (1985), McLean (1980)]. If the instrument for competing for patients is quality [Gaynor and Gertler (1995)] or even the aggressiveness in "inducing demand" [Dranove (1988)], evidence that such a decision variable influences demand supports the monopolistically competitive assumption. Many other empirical features of the market are also consistent with monopolistic competition. Studies of physician practice costs conclude that physicians operate on a downward-sloping portion of their average cost curve [PPRC (1992), Escarce and Pauly (1998)]. Firm-level advertising only pays if there is imperfect competition [Feldman and Begun (1978), Haas-Wilson (1986), Rizzo and Zeckhauser (1990)]. Wong (1996) found that physicians respond to factor price changes in a manner consistent with monopolistic competition.

The basic conception of the physician–patient relationship embodied by monopolistic competition is that physicians are imperfect substitutes in the eyes of patients. A patient has a demand for the services of a particular physician, as opposed to demand for "physicians' services" in general. Although some observers have written about the "demise" of the physician–patient relationship [see Sloan et al. (1993, p. 51) for one discussion], surveys continue to show that a clear majority of patients have what they

[14] Much, but not all of surgical care was insured during the 1960's. Many patients, including the elderly, faced some price and if they were price responsive in choice of physician, would have given surgeons an incentive to reduce price to raise volume.

regard as a “regular source of care” [Moy et al. (1998)]. Even without a primary care provider, patients may rely on physicians to supervise their care during an episode of illness. The demand curve a patient has for a physician is not the same as the demand the patient has for physicians’ services, a distinction, despite the general enthusiasm for the monopolistically competitive structure that has not been given much attention in the literature.[15] Many papers motivate their model with words associated with monopolistic competition, and then analyze a single firm.

Although interaction is not strategic in monopolistic competition, actions of one physician, such as a price change, affect demand for other physicians. In what follows, we present a model of monopolistic competition in which the physician has some market power but the patient has some alternatives. We will model the patient’s alternatives as simply as possible in order to enable us to focus on the behavior of a representative physician.

Another important feature of the market will also be taken into account. Physicians sell a service; a diagnosis or treatment provided to one patient cannot be resold by that patient to some other customer. As Gaynor (1994, p. 224) observes in his review, “services are by their nature inherently heterogeneous and nonretradable.” The nonretradability of physician services has important implications for price discrimination and more generally for price and quantity setting. Farley (1986) called attention to the connection between non-retradability and price discrimination in physician markets, but the implications of nonretradability have not been fully appreciated in the context of physician markets.[16]

We can now proceed to set up a model of patients and physicians that we will build on throughout this chapter. The quantity of physician services is x. The patient benefits from services according to $B(x)$, denoted in dollars. The marginal benefit function is $b(x) = B'(x)$. $b(0) > 0$; $b'(x) < 0$. We employ a benefit function rather than a demand curve since profit maximization implies price-quantity pairs that may not be “on” the demand curve. The $B(x)$ function captures any health shocks implicitly, so that $B(0)$ may be negative. Time costs, inconvenience, and other costs and benefits of using medical care experienced by the individual are incorporated in $B(x)$.[17] By assuming that the

[15] The hospital market is also generally regarded as monopolistically competitive. See Pauly and Redisch (1973) or Dranove, Shanely and White (1993). Frank et al. (1987) estimate an earnings model for podiatrists based upon the idea that medical practitioners appropriate hospital rents. Dranove and White (1996) have another view about how doctors can secure rents from monopolistically competitive hospitals. Since each specialty group’s fees (e.g., cardiologists) fees are a small part of each hospital’s costs, and patients (or payers) decide on the hospital on the basis of total cost, each specialty will try to raise its fees as much as possible to enlarge its share of the rents.

[16] Folland, Goodman and Stano (1997, p. 377) recognize that physician services are nonretradable and support price discrimination. Dranove’s (1988) model of demand inducement implicitly recognizes this property by including the assumption that the patient’s choice is to “consent” to treatment or not. In the general literature, it is well understood that nonretradability is behind models of price discrimination [Varian (1989)].

[17] The utility function generating this benefit can be expressed as $U(y + B(x))$. Demand for x is independent of income in this formulation. See Ma and Riordan (1998) for discussion of the implications of alternative forms of utility and benefits.

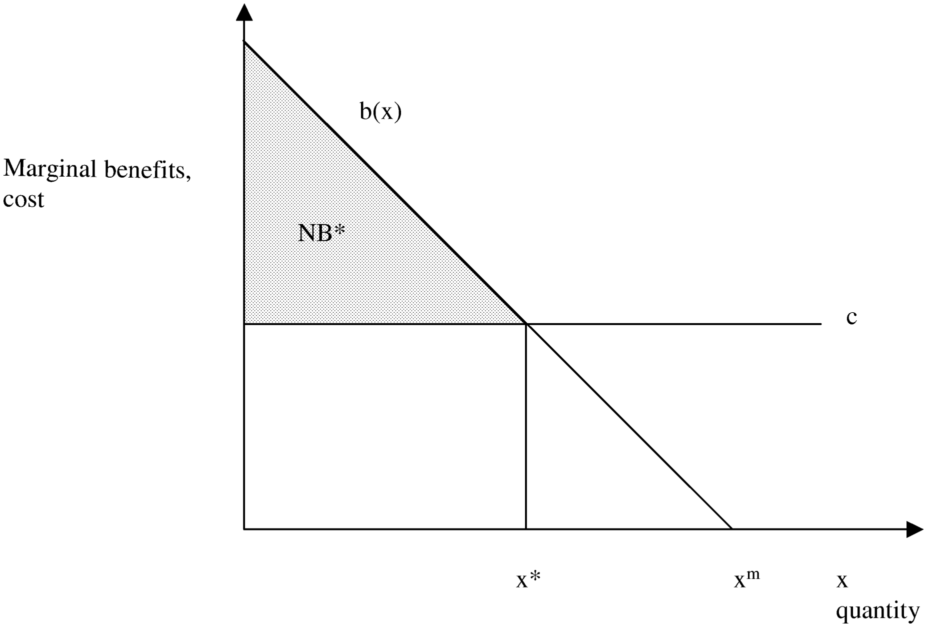


Figure 2. Benefits and costs of physician services.

benefit function depends only on the quantity of x, we abstract from the role of other goods, including income, influencing the valuation of services. Physician services are produced at constant cost per unit c.[18] If p is the price of physician services (insurance will be introduced shortly), physician profit is $\pi = px - cx$, and patient net benefit can be written $NB(x) = B(x) - px$. Define x^* as the solution to $b(x) = c$, the efficient level of x. Let $NB^* = B(x^*) - cx^*$, the maximum possible patient net benefit. Also, for purposes of reference define x^m, the level of x that maximizes $B(x)$, or, the solution to $b(x^m) = 0$. See Figure 2.[19]

In monopolistic competition, the patient has substitutes. In general, a patient could consume services of many physicians at the same time, and benefits from physicians' services would be a function of the set of services consumed. We simplify this by forcing

[18] Empirical research indicates that AC is falling. This is consistent with fixed costs and a constant marginal cost. See Escarce and Pauly (1998), and Physician Payment Review Commission (1992).

[19] It is worth calling attention to an assumption that the decision to be made by the doctor and the patient is a decision about the "quantity" of one variable, x. Often in papers on health care, this quantity is denominated in dollars. While this follows convention in economic models, the literature in medical sociology or medical decision analysis views the matter differently, with the doctor and the patient having to decide about a "treatment" which can consist of a mix of different services. Typically, there are a discrete number of treatment alternatives, described, for example in a decision tree. See, e.g., Weinstein et al. (1980).

the patient to choose a physician from whom to receive care. With this interpretation, the benefit function used here is consistent with the idea of "residual demand" in models of imperfect competition and product differentiation. We will recognize that a market gives a patient alternatives, and say that if the patient leaves this physician, he can receive net benefit NB^0 from an alternative physician. The patient then uses the current physician if and only if the net benefit he receives is no less than NB^0. By altering NB^0, this model includes perfect competition and monopoly as extremes. If $NB^0 = 0$, the patient has no alternative to this physician and the physician is a monopolist. If $NB^0 = NB^*$, the market is perfectly competitive, and the physician has no market power. In general, $0 < NB^0 < NB^*$.

The price and quantity of physician services are found by maximizing the physician's profit, subject to the constraint on patient net benefit imposed by competition with alternative physicians. We can set this up as a constrained maximization problem in Program I.

Program I:

$$L = px - cx + \lambda\big(B(x) - px - NB^0\big). \tag{3.1}$$

Maximizing L with respect to p, x, and λ, the first-order conditions (assuming an interior solution) for Program I are:

$$L_p\colon \quad x - \lambda x = 0, \tag{3.2}$$

$$L_x\colon \quad p - c + \lambda\big(b(x) - p\big) = 0, \tag{3.3}$$

$$L_\lambda\colon \quad B(x) - px - NB^0 = 0. \tag{3.4}$$

The three first-order conditions can be solved sequentially for the three variables, λ, x and p. From (3.2), $\lambda = 1$, reflecting the fact that the seller gains all the surplus above NB^0, and any relaxation of the surplus constraint goes to profits. Normally, one thinks that a two-part tariff is necessary for a seller to extract all the surplus, but here, both p and x are chosen by the physician, and two-part pricing is not required to extract surplus. Then, from (3.3), x is such that $b(x) = c$, or, as we have defined above, x is set efficiently, $x = x^*$. Finally, rewriting (3.4) as (3.4′), price is determined so as to extract all surplus above NB^0

$$p = \frac{B(x^*) - NB^0}{x^*}. \tag{3.4′}$$

Figure 3 illustrates the solution. NB^0, a given, is equal to the lightly shaded region. The combination (p, x^*) chosen by the doctor gives her profits $(p - c)x^* = NB(x^*) - NB^0$, the entire available surplus. The doctor has only to match the surplus available elsewhere to keep the patient. Quantity is always x^*, that which maximizes

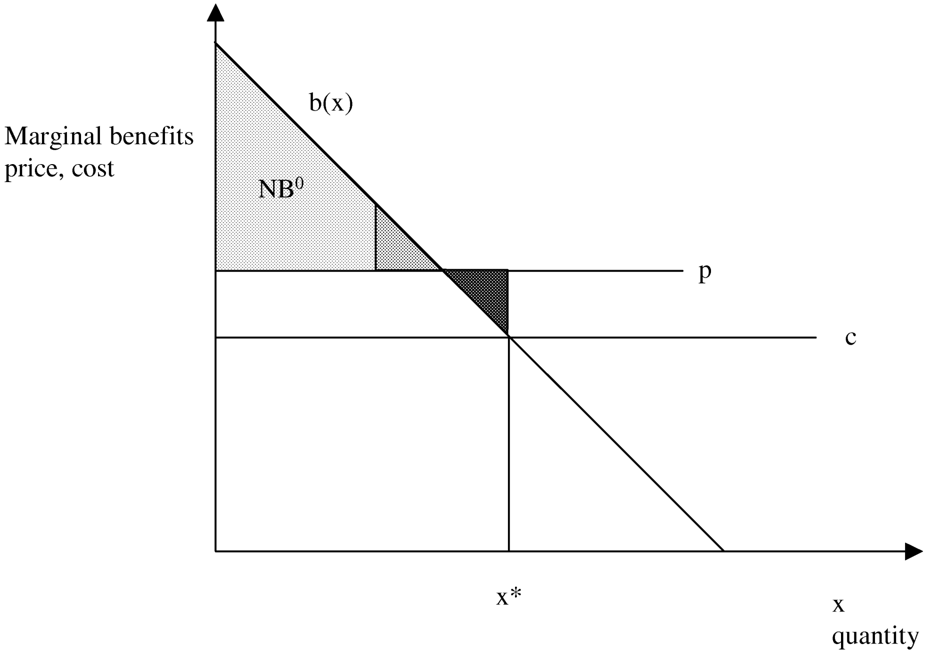


Figure 3. Setting price and quantity with net benefit constraint.

total surplus available. Note that the patient is not a price taker. At the price of p, the patient would prefer to consume fewer services than x^* but nonretradability lets the doctor set quantity. One can think of the consumer surplus gained above NB^0 by consuming up to the point where $b(x) = p$ (the moderately shaded region in Figure 3) as just being offset by the consumer surplus lost from consuming beyond this point to x^* (the dark region).[20] In effect, the physician makes an all-or-nothing offer to the patient, extracting all available consumer surplus. This is not surprising, since with market power and the nonretradability feature, the physician possesses the prerequisites for the exercise of first-degree (or perfect) price discrimination [Varian (1989)].

In Kessel's world of the 1950s, physicians could set prices (and quantities) without contending with third-party regulations. Price discrimination across patients emerges naturally from the model in Program I. Consider different patients with different benefit functions. Suppose one patient has a higher willingness to pay indicated by a higher $B(x)$. Equation (3.4′) tells us immediately that the higher willingness-to-pay patient will pay more for the same services. Nonretradability shelters the price discrim-

[20] It is evident from Figure 3 that if $NB^0 = NB^*$, the physician is forced to give the consumer x^*, at the competitive price, c.

ination. The poor paid less simply because they had a lower willingness to pay (not because they had a more elastic demand).

The model in Program I also features quantity setting by the physician. An immediate implication of profit maximization in monopolistic competition is that the physician takes advantage of the nonretradability of her services and sets both price and quantity. This quantity setting, according well with direct observation of patient–doctor interactions, emerges from the most simple model of the process of quantity determination, with (and this is worth emphasizing) perfectly rigid patient preferences not subject to manipulation by the physician. It is the first form of physician quantity setting previewed in Table 1.

Consider the effect of price regulation in this model. Suppose p is not under the control of the physician but is set by the payer. Program I could be solved again in this case dropping condition (3.2) and regarding price as fixed (one fewer unknown, one fewer equation). The net benefit constraint (3.4) still holds, and indeed, when price is fixed, it is (3.4) that can be solved for quantity, x. Note that (3.4) implies that:

$$\frac{dx}{dp} = \frac{-x}{p - b(x)} < 0. \tag{3.5}$$

In words, if price is fixed by the payer, a decrease in price will be cause an *increase* in quantity. The reason is straightforward. The physician need not give the patient any more than a fixed level of net benefit. If a payer restricts how much surplus a physician can extract by setting price, the physician can counter by extracting surplus by setting quantity higher. This yields the physician more surplus since price is fixed above cost. The patient accepts this because the price limitation increases surplus on the previously purchased units. Note that the implication of a negative derivative of quantity on price in (3.5) emerges from the very simplest model of physician quantity setting, without appeal to induced demand or target income motivation.

We now proceed to add some institutional elements to this simple model. Third-party payers insure patients against health care costs, reducing price paid by the patient at the time of service delivery to below cost. In the course of this, payers have found it necessary to constrain physicians' ability to set prices by adopting fee schedules.

3.2. A third-party payer and administered demand and supply prices

Physicians in most developed countries have lost discretion over prices. The first widespread regulation of physician charges, supported by local medical societies, was by Blue Shield plans, which systematically collected information on the prices charged by physicians in their service areas and allowed payment to a physician if the price fell within the "usual, customary, and reasonable" (UCR) fee limits. (Usual refers to what this physician regularly charges, customary to the 75th percentile of charges for similar fees in the area (last year), and reasonable to other factors, such as complicating conditions, which may justify higher fees.) The federal government's Medicare program, instituted in 1966, adopted a similarly permissive UCR system. The UCR system could limit fees paid in any one year, but gave an incentive to physicians to increase charges, eventually forcing payers to set dollar amounts to be paid by procedure for various specialists.[21]

In the US, Medicare, state Medicaid plans, Blue Cross plans, and within recent years even commercial insurance, set maximum prices they will pay physicians for each procedure [Eisenberg (1994)]. In a comprehensive study of physicians' practices in the late 1980's, Hsiao and colleagues (1988a, 1988b) devised a relative weighting scale for physician services that forms the basis of Medicare's physician fee schedule. Some other payers pay based on Medicare's schedule, while others use their own fee schedule, based on historical payment or on demand and supply conditions. Outside the US, when governments pay physicians by fees, these fees are regulated, as in Canada.

Reinhardt (1975) classified physician markets in two dimensions, price setting vs. price taking for physicians, and on the dimension of inducing demand or not. The price setting/price taking distinction is the traditional difference between a firm with market power facing a downward-sloping demand (conferring some price-setting power) and a firm with no market power facing a horizontal demand at the market price. Payers have stripped physicians of price-setting power. Does this mean that physicians have no "market power?" Equation (3.4) tells us the answer to this question is no. When payers set price, market power continues to convey an advantage to physicians, as we see in this section, because physicians retain the ability to set the quantity of their nonretradable service.

Third-party payers may also regulate the prices paid by patients. Insurance or other form of third-party payment reduces the financial price to the patient at the time services are used. The price paid by the patient is less than the price received by the physician, the difference being the amount contributed by the third-party payer. Prices patients pay can be complex, involving deductibles, copayments, and limits; here we will assume the price a patient pays is a constant share θ of the price paid to the physician. θ is the coinsurance rate. Thus, $0 < \theta < 1$. Third-party payers also set the price doctors receive, a practice that can be understood as an attempt to prevent the increase in willingness to pay of patients created by insurance from being shifted to physicians in the form of a higher price. The price paid to the physician, p, is now set by the insurer, with $p > c$, to ensure physician participation.

Physician profit depends on p, and patient net benefit depends on θp. As before, the patient has a benefit function $B(x)$ and has an alternative offering net benefit of NB^0. Physician profit maximization with a regulated price is described in Program II. Quantity x is the physician's only decision variable.

[21] Data produced as part of Hsiao's fee reform research show how out of alignment Medicare fees had become by the patchwork regulatory system in force through the 1980's. For the single highest dollar volume procedure, a cataract removal, Medicare paid more than \$1500 per hour of work in 1986. For the second highest volume procedure, an office visit, Medicare paid less than \$100 per hour. See Glazer and McGuire (1993) for more information on high-volume fees prior to reform.

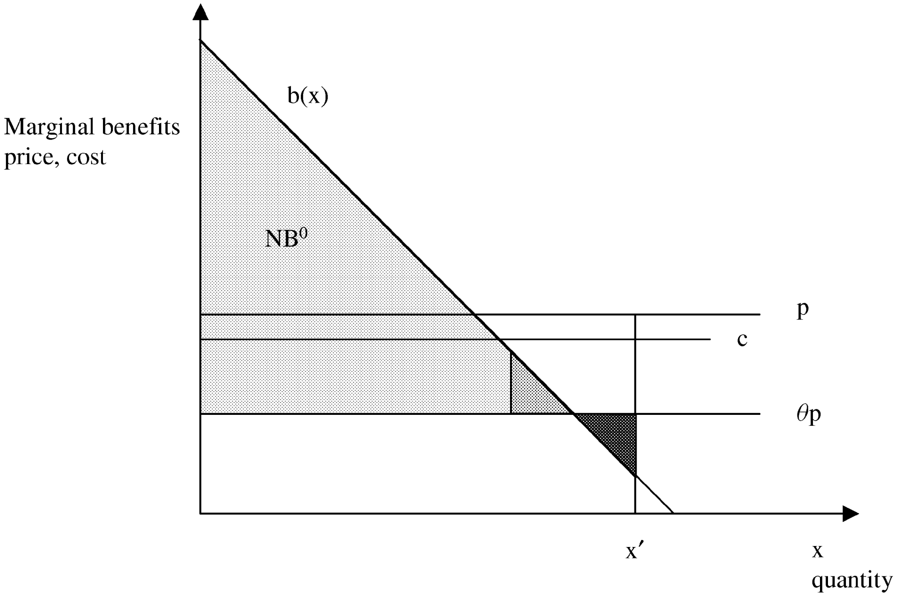


Figure 4. Setting quantity with administered prices and insurance.

Program II:

$$L = px - cx + \lambda(B(x) - \theta px - NB^0), \tag{3.6}$$

with the first-order conditions:

$$L_x: \quad p - c + \lambda(b(x) - \theta p) = 0, \tag{3.7}$$

$$L_\lambda: \quad B(x) - \theta px - NB^0 = 0. \tag{3.8}$$

So long as $p > c$, the physician profits from more x. Rewriting (3.8) as (3.8′) we see that quantity is set so as to just satisfy the net benefit constraint. We label the solution to (3.8′) as x', and illustrate it in Figure 4:

$$x' = \frac{B(x') - NB^0}{\theta p}. \tag{3.8′}$$

Equations (3.8′) and (3.7), and Figure 4 show that so long as $p > c$, $\theta p > b(x')$, since λ is positive.[22] When price is constrained, the doctor exercises market power by

[22] λ can be greater than or less than one, depending on how much p exceeds c. If profits per unit sold are very small, the value of a relaxation in the constraint to the physician will be less than a dollar. The opposite case is also possible.

setting quantity beyond the point the patient would choose given the price he faces. The physician wants to do so because she makes a profit on every unit sold, and services are nonretradable. The patient's alternative of leaving to receive NB^0 limits how far x' can be pushed. Another constraint, that $b(x) \geqslant 0$, could reasonably be added to the problem, limiting the quantity setting to quantities which convey some non-negative benefit, but this would not change the essential character of the result.

It is plausible that physicians can require patients to use more than they would like given the prices they face. A physician can put pressure on a patient to agree by conveying that if the patient does not accept the treatment, his alternative is to seek care from another practitioner. Patients in some cases may be able to avoid some overtreatment (from the patient's economic point of view) by failing to comply with prescribed treatment after some point.[23] This may be done with visits that must occur over time, for example, and after some point the patient can simply stop going to that physician. Even in the case of visits, however, the physician may be able to exert some influence. Hickson et al. (1987) found, when paid a fee with a solid margin over cost, pediatricians scheduled well-baby care in excess of that recommended by the American Academy of Pediatrics, but did not do so when they were paid a salary. Physicians recommend more treatment for insured patients, even in artificially constructed clinical scenarios [Mort et al. (1996)]. Chassin et al. (1987) found in a fee system that a sixth to a third of commonly performed procedures provided no (or negative) marginal benefit.

For many treatments, an all-or-nothing quantity setting strategy may be very effective for the physician. Suppose there are few or many tests that could be run. The patient would like "few" given the price, but wants his doctor, who insists on "many," to administer and interpret the tests. Treatment might be simple or complex, but the patient's physician might insist on complex. The many medical situations in which treatments are provided if and only if both physician and patient agree fit squarely within the model of physician quantity setting.

The welfare economics of health insurance have been based on the assumption that the patient is a price taker in medical markets. If this assumption is not correct, and as we have seen, it is contradicted by the assumption of a profit-maximizing seller of a service, the analysis of optimal insurance, and optimal health payment more generally, would require reworking. This is currently an open area for research.

A limitation of the model in Program II is the assumption that NB^0 is fixed. In a monopolistically competitive model, if an insurer reduces price, it reduces price for all physicians in the market, not just the physician described in Program II. The value of the patient's alternative, represented by NB^0, must therefore change as well with a change in the administered price.[24] The simplest complete model of imperfect competition is

[23] Patients fail to keep about 20% of scheduled visits in some studies. [See Oppenheim et al. (1979) and Smith and Yawn (1994)]. Giuffrida and Gravelle (1998) contains an economic model of patient compliance. Compliance there is regarded as desirable, though coming at a cost.

[24] See Dranove (1988) who confronted a similar issue in a monopolistically competitive model. The effect of more competition on patient quantity is ambiguous in his analysis.

two physicians who compete for patients distributed according to some dimension (e.g., distance), as has been analyzed by Glazer and McGuire (1993) and Ma and McGuire (1998). In a generalization of Program II, Ma and McGuire (1998) show the case of two physicians, competing in a Hotelling-type model, when an insurer sets price market-wide, $dx/dp < 0$ for both physicians; in other words, both physicians increase quantity in response to a regulated price fall. The reason for the positive quantity response to a fall in the administered price is the same as in the simple single physician model of Program I: lowering a regulated price channels a physician's market power to quantity.

Intuitively, the explanation is as follows: In the presence of an administered price, the physician exercises monopoly power by setting quantity. Quantity x is above the level the patient would demand at the price he pays, so increasing x represents exercise of more monopoly power. Competition with other physicians limits how much the physician wants to increase x, because an increase in x, with other physicians' behavior constant, leads some patients to leave this physician's practice. Now, reducing the price paid per unit of x reduces the penalty the physician pays for losing a patient. The loss associated with an increase in x is reduced, and the physician (and all physicians) are led to increase x in response to a regulated price fall. This result is unlikely to be fully general to models of monopolistic competition, which can have complex patterns of substitution across sellers. Furthermore, at some point, lower price must induce exit from the industry, tending to decrease quantity, at least in aggregate. It does establish, however, that a negative dx/dp is consistent with a complete information, monopolistic competition model.

A model akin to the one just discussed was used to study the effect of administered prices in Medicare. Mitchell and Cromwell (1982), and Zuckerman and Holahan (1991) assumed the demand represented demand of the group of Medicare beneficiaries, and the cost curve of the physician sloped upward. Quantity provided might be limited by demand (they assumed price-taking demand behavior), or, if the marginal cost curve cut the p line from below before the quantity consumers wanted to buy, quantity would be supply-determined.

The main application of this model of either demand or supply-limited quantity was to consider the effect of changing levels of Medicare fees and the effect of allowing physicians to "balance bill" Medicare patients. Medicare set fee allowances that determined what it paid to doctors, and determined the beneficiary's coinsurance payment. Initially, physicians did not need to limit themselves to this price. At the beginning of the Medicare program, physicians could charge any price they pleased, requiring the patient to pay the "balance bill" equal to the difference between Medicare's allowed charge and the physician's price. Medicare has steadily restricted physicians' authority to balance bill (today it is constrained to about a 10% window).

The question arises: what effect does a restriction on balance billing have on physician markets? In the Mitchell and Cromwell (1982) and Zuckerman and Holahan (1991) analysis, the physician could price discriminate, charging patients with a high willingness to pay a balance bill, but no balance bill for patients with a low willingness to pay.[25] The conclusion about balance billing from this analysis was the following: since only inframarginal patients were balance billed, balance billing functioned simply as a transfer from patients to physicians. These papers concluded that balance billing could be eliminated with no effect on quantity supplied.

This is an uncomfortable conclusion. Is nothing lost if prices paid to doctors are reduced by eliminating balance billing? The obvious concern is about the "quality" of services, in addition to quantity. If physicians have a choice about the quality of their services, we can see that the supply-constrained equilibrium is not likely to hold up. In the supply-constrained case, patients demand more (are willing to accept more) services at the (regulated) price they face, but the physician stops providing them because the marginal cost has risen to the supply price. Now, let the physician choose quality. By "quality" we mean some aspect of services that increases the value of services to consumer, is costly to the physician, and is not reimbursed by the payer. In a supply-constrained case, the physician can reduce quality, reduce marginal cost, increase profits, and then supply more services in response to the lower marginal cost. The reduction in willingness to pay stemming from the quality fall is not a problem for the doctor since the demand constraint is not binding. This process of reducing quality can continue, in fact, until demand does bind. The supply-determined case in these models will not be an equilibrium if quality is variable. When quality is variable, then, balance billing does have efficiency effects. Glazer and McGuire (1993) show that if Medicare sets fees correctly (an important proviso), all patients, those that pay as well as those that do not pay a balance bill, are better off if balance billing is permitted.[26]

In this section we have shown that physician quantity setting and an increase in quantity associated with a decrease in regulated prices both emerge from this simple model, features normally associated with special market power (inducement) and motives (target incomes) of physicians. Second, efficiency problems in the market for physicians' services take the form of too much quantity and an inefficient level of "quality." The quantity problem arises from two sources, from patient insurance and from the quantity-setting power of physicians with administered prices. The inefficient level of quality may result from quality being unreimbursed within a regulated health payment system.[27] In the next two sections, we show how managed care contains additional instruments for dealing with quantity and quality setting in physician and patient interaction.

[25] Price discrimination was between two groups. The physician set one level of balance bill applying to part of the patients, and accepted the Medicare fee, with coinsurance, for the others. Accepting the Medicare fee as full payment is referred to as "accepting assignment."

[26] In Glazer and McGuire (1993), physicians can discriminate on quality as well as quantity, between patients charged the balance bill and those who are not.

[27] It seems likely, as administered prices are pushed low to control costs, that quality will be set too low. Experience from the Medicaid program feeds this suspicion.

3.3. Noncontractible "quality," supply-side cost sharing in managed care contracts, and competition for patients

We define "quality" as a noncontractible input into the production of health for the patient. Noncontractible means that it cannot be used as a basis for payment. The care or effort that a doctor puts into a decision or treatment matters to the patient, but is difficult to incorporate into a payment system. More concretely, one could simply think of the "time" doctors spend in conducting a procedure [Glazer and McGuire (1993)]. Some physicians are paid per unit of time (e.g., psychiatrists, and anesthesiologists in Medicare). Yet actual time spent is very difficult to verify, and payments for most activities are not based on an explicit report of time. As McCall (1996, p. 51), an MD, notes, "several time-consuming activities, vital to providing good medical care, pay doctors nothing or next to nothing [including] conducting careful medical interviews, educating patients, staying up-to-date with medical advances." In recommending how patients should judge their doctors, he says, "the amount of time a doctor spends interviewing you, examining you, and explaining things reflects how genuinely concerned that doctor is for your welfare" (p. 52). "Time" is one good candidate for an observable but noncontractible input into the patient's health. Others – diligence, care, attentiveness – synonyms in this circumstance for "effort" – can be thought of as well.

Here we will use the term "quantity" to designate those physician inputs which are contractible, and "quality" to denote those which are not. From the patient's point of view, both types of inputs, quantity and quality, matter for the benefits of health care. We retain the assumption that the patient has full information about the physician, even though quality is not contractible.[28] If the level of quality ("effort") supplied by the physician is e, then the benefits to the patient of treatment by the physician can be rewritten $B(x, e)$, with derivatives $B_x > 0,\ B_{xx} < 0,\ B_e > 0,\ B_{ee} < 0$. Effort is costly to the physician so now $c(e)$, and $c_e > 0,\ c_{ee} > 0$.

In the last several years, managed care payers have become more creative in writing contracts with physicians, incorporating various incentives for physicians to be careful about quantity. These forms of contracts appeared earlier in the HMO and hospital sector, where they were called prospective payment or "supply-side cost sharing" [Ellis and McGuire (1986), Newhouse (1996)]. They also go by the name risk sharing, and take the particular form of capitation and "withholds." In a capitation contract, a physician or group of physicians is responsible for a defined set of the health care costs for a patient over a period, typically a year. Under a "withhold," a physician or group is paid according to negotiated fees, but a bonus or a withhold is paid at the end of the contract period if certain cost or other targets are met.

[28] Gaynor and Gertler (1995) suppose that a physician choses the "effort" put into work, and this will affect the quantity demanded of the physician. Medical group practices, studied by Gaynor and Gertler, reward individual work differently, according to whether payment policies are based on averaging or based on individual productivity. Gaynor and Gertler find that physicians in group practices which pool income make less effort, and this can lead to large reductions (up to 50%) in the volume of work physicians do.

Recent surveys have documented the prevalence of the changing contractual relationships between physicians and managed care payers. Emons and Wozniak (1997) use data from the 1996 AMA Socioeconomic Monitoring System to monitor contracts. Risk sharing contracts are more common with primary care physicians than with specialists, based presumably on the rationale that the primary care doctor has more control over the aggregate use of patients. In 1996, 36 percent of all physicians had at least one capitation contract, and for these physicians capitation revenues averaged 25% of the total. For some specialties, such as pediatrics, capitation revenues were 30 percent of practice totals. Withholds were about as common as capitation, with 36 percent of physicians having contracts with withholds. For these physicians, 19 percent of their revenues came from such contracts. Similar figures are reported by Remler et al. (1997) from a national survey, and from Hellinger (1996) for a slightly earlier period.

Capitation, withholds, and bonuses all fit within a framework of supply-side cost sharing. Supply-side cost sharing is present if when a service is provided (a cost incurred), some of this cost is borne by the provider because of the payment contract. Supply-side cost sharing can be low-powered as in a withhold system with a mild penalty for exceeding a cost target, or high-powered, as in a capitation system. In a fully-capitated system, the provider bears all of costs at the margin. We can write the general form of a per-patient contract with supply-side cost sharing as:

$$R + p_s x, \quad \text{with } R > 0,\ c > p_s \geqslant 0.$$

R is the portion of the payment made independent of the services provided – the capitation amount, the partial capitation amount, the bonus. p_s is the payment per unit of service. The contract features supply-side cost sharing if $c > p_s$. In a capitation contract, $p_s = 0$; in low-powered contract, p_s is closer to c. A payer can calibrate incentives to a provider by alternating the composition of a contract between payment by R and payment by p_s.[29]

To appreciate the incentive properties of these contracts we must enrich the model we have been using so far to include competition for patients. Up to now, the physician faced the potential loss of a patient if the patient were given a package of price and quantity that failed to satisfy a net benefit constraint. More generally, a physician could be considered to have a probability of keeping a patient, with the probability increasing as the physician gives the patient more net benefit. Another interpretation of such a formulation is that the physician might attract more patients of a certain type if the net

[29] One form of withhold contract would be if the physician receives a bonus B if costs do not exceed a target T. For each dollar costs are above the target, the physician losses a portion r of the shortfall. This contract can be written as $p_S x + B - r(p_S x - T)$, which fits within the $R + p_S x$ form with supply-side cost sharing. The $R + p_S x$ contract is linear, however, and cannot exactly capture nonlinear features of physician contracts, for example, if the physician "bonus" is constrained to be positive.

benefits she provides in her practice were higher.[30] With either interpretation, we can express the number of patients the physician serves as a positive function of net benefit offered: $n(NB)$, with $n' > 0$. Since p_s may be set below cost, we can no longer write demand price as a share of supply price. Instead, we will let the payer set demand price at p_d.

We now have a richer profit maximization problem for the physician, that can be summarized in Program III.

Program III:

$$\pi = n(NB)\big[R + (p_s - c(e))x\big], \quad \text{where } NB = B(x, e) - p_d x. \tag{3.9}$$

Profit is a product of the number of patients and the profit per patient. The physician's contract may have a prospective component per patient (R) and a fee (p_s) for each unit of service. The number of patients depends on the net benefit the physician supplies. Net benefit depends on both the quantity the physician supplies and the effort (quality) she puts in. Quantity is contractible, effort is not. The patient's insurance is represented by p_d, the price the patient pays for each unit of service.[31] The physician chooses x and e to maximize profit.[32]

The first-order conditions (3.10) and (3.11) describe the physician's maximization:

$$\pi_x: \quad n'(B_x - p_d)[R + (p_s - c)x] + n(p_s - c) = 0, \tag{3.10}$$

$$\pi_e: \quad n' B_e[R + (p_s - c)x] - nc_e x = 0. \tag{3.11}$$

These can be rewritten as

$$\frac{B_x - p_d}{NB/x}\left[\frac{R/x + p_s - c}{p_s - c}\right] = -\frac{1}{\varepsilon_{n,NB}} \quad \text{where } \varepsilon_{n,NB} = n' \cdot \frac{NB}{n}, \tag{3.10$'$}$$

$$\frac{R/x + p_s - c}{c} = \frac{\varepsilon_{c,e}}{\varepsilon_{n,e}} \quad \text{where } \varepsilon_{c,e} = c'\frac{e}{c},\ \varepsilon_{n,e} = n'\frac{\partial NB}{\partial e} \cdot \frac{e}{n}. \tag{3.11$'$}$$

[30] Formally, potential patients of this physician could have alternative net benefits distributed according to $NB + d$, where d ("distance") takes some distribution. A patient goes to this physician if the net benefit he receives is greater than $NB + d$.

[31] We ignore any effect of income on premiums since we assume the benefit of health care is independent of income. If we were to use this model for a formal analysis of optimal insurance and payment, risk and the insurance contract would need to be considered.

[32] The model here is closely related to Ma and McGuire (1997). In that paper, the physician chooses quality or effort and the patient chooses quantity of treatment. Price paid by the patient and paid to the physician are set by a payer. This model is also related to one studied by Feldman and Sloan (1988) where a monopoly physician sets quantity and quality to patients in the presence of price controls. Patients are price takers without insurance. The firm in Feldman and Sloan (1988) is a monopolist, and so faces no competition for patients.

We can use (3.10′) and (3.11′) to relate the policy instruments of the payer to the efficiency problems in treatment determination. There are two efficiency targets, one for x and one for e. Quantity x tends to be overused because of moral hazard and physician market power. Quality or effort e may not be set efficiently in general because it is noncontractible. The payer, equipped with a payment system that includes supply-side cost sharing, has two instruments: the overall level of payment, and the prospectiveness of the payment, corresponding roughly to R and p_s. As papers by Ma (1994) and Rogerson (1994) first made clear, prospective payment can induce a provider to undertake noncontractible effort if this effort leads to more business. Increasing the profitability per patient (say by increasing R), leads to more of the noncontractible quality.

In general, (3.10′) and (3.11′) would need to be solved simultaneously to find the physician's profit-maximizing quantity and quality decision as a function of the payment system parameters. We can see from (3.10′) and (3.11′), however, how the payment system can help solve the efficiency problems, and therefore why a managed care plan would be interested in writing a risk-sharing contract.

The first-order condition, (3.10′), describes the physician choice of quantity. $B_x - p_d$ is the marginal net benefit a patient receives from more x and NB/x is the average net benefit the patient receives. Up to this point, the marginal net benefit has been negative: the physician has pushed quantity beyond the point where $B_x = p_d$. This benefit elasticity is multiplied by a payment-system term and is equated to the negative inverse of the elasticity of the number of patients with respect to net benefit. The payment system term is the ratio of the average net revenue per unit of x, $R/x + p_s - c$, to the marginal net revenue, $p_s - c$.

Consider first the case when there is no prospective payment and $R = 0$. The payment system term becomes one and the benefit elasticity is equated to the negative inverse of a market "demand response," the change in the number of patients with respect to a change in the net benefit provided. Consider the effect of an increase in the demand response elasticity $\varepsilon_{n,NB}$. To bring the left-hand side of (3.10′) into equality with a smaller (in absolute value) negative number, x must fall to bring B_x and p_d closer together. (The marginal changes faster than the average.) In words, (3.10′) shows that the physician is restrained in pushing x too far because of the prospect of losing business. A similar idea, that physician quantity setting is restrained by market demand, was proposed by Dranove (1988) in a model of "physician-induced demand." There, patients were not sure of what they needed, but became increasingly suspicious of their physician's recommendations as the doctor's style of practice became more and more aggressive. As we show here, the same results hold in a model of complete information, recognizing that physicians have some quantity-setting ability. Market demand response may thus modify the physician's tendency to push visits beyond the point the patient would demand given his insurance. In a perfectly competitive model, this demand response is very high, and as (3.10′) indicates, the discrepancy between marginal benefit and price paid by the patient disappears altogether. In general, with monopolistic competition, and without supply-side cost sharing, there will still be some quantity setting beyond the point a price-taking patient would prefer.

Now reconsider the payment term in (3.10′), allowing $R > 0$. With supply-side cost sharing, $p_s < c$, and the payment term is negative. This has the very important effect of reversing the sign of $B_x - P_d$. Thus, with supply-side cost sharing, the physician no longer has an interest in pushing quantity beyond the point the patient may demand, and indeed, will tend to limit the quantity to less than what the patient would demand (since the sign of the marginal benefit/price difference is reversed). By making the payment system more prospective, increasing the weight on R, and decreasing it on p_s, the payer can give the doctor incentives to cut back on quantity, perhaps even hitting the first best, where marginal benefit equals cost. The promise of supply-side cost sharing is that it can compensate in this way for the insured patient's moral hazard, and lead to the efficient quantity of health care, where $B_x = c$.[33] A payer could increase the degree of supply-side cost sharing, decreasing p_s, while increasing R to maintain the overall average profitability of services. In that way, the quantity of x could be reduced towards the efficient level.

The payment system also affects the physician's choice of effort or quality, reflecting the physician's tradeoff between higher cost with higher quality, but more patients with a higher willingness to pay. Equation (3.11′) shows that in profit maximization, the physician equates the percentage markup of *average* fee over cost to the ratio of two elasticities: the cost elasticity of effort over the demand response elasticity of effort. For inducing effort, it is only the average profitability that matters. This makes sense since effort is not explicitly reimbursed. If profitability goes up, attracting new business has more value, and effort rises. An increase in effort will lead to a rise in the right-hand side of (3.11′), bringing it into equality with the left-hand side increased by the rise in profits. The basic reason is that while both $\varepsilon_{c,e}$ and $\varepsilon_{n,e}$ are positive, marginal costs are increasing in e, but marginal benefits are falling. An equation like (3.11′) has been the basis of studies of optimal provider reimbursement by Ma (1994), Rogerson (1994) and others. Since more quality leads to more customers, paying more for each customer can induce higher quality services.[34]

Many empirical studies have confirmed the effect of form of payment on doctor behavior. In a study in a position to distinguish physicians' desired supply from actual use, Hickson et al. (1987) found that pediatricians aggressively *scheduled* visits when they were paid generously by fee-for-service, in comparison to both standards of care promulgated by the American Academy of Pediatrics, and in comparison what their colleagues with comparable patients were doing when they were paid by salary: 4.9 visits per year vs. 3.8 visits per year. In both cases patients kept only some of the visits:

[33] Ellis and McGuire (1986, 1993), Newhouse (1996), Ma and McGuire (1997).

[34] The literature on quality and effort has generally regarded this to be a patient-specific variable. Papers study the problem in a single-payer, uniform patient context. If some quality dimensions are practice-wide rather than patient-specific, the analysis of the determinants of quality would need to be broadened. The classic analysis of quality determination in a firm with a single set of customers (but without a regulated price) is Spence (1975).

3.6 vs. 2.9, but the percentage kept was about 75% in both cases, suggesting that physicians could influence patient utilization. Jennison and Ellis (1987) found the same set of physicians provided more visits when paid by fees than under a capitation contract. Compared to doctors in HMOs, who do not make profits by ordering tests, Epstein et al. (1986) found that doctors in fee-for-service practice ordered 50 percent more EKGs and 40 percent more chest X-rays, the tests they perceived to be highly profitable. Rates of low-profitability tests were not elevated. At a walk-in care clinic, once salaried doctors were afforded bonuses to increase volume, lab tests went up 23 percent and X-rays, 16 percent [Hemenway et al. (1990)]. Stearns et al. (1992) found large changes in utilization in response to a shift from fee-for-service to capitated payments to a group of physicians. In a large study with extensive controls for patient characteristics, Greenfield et al. (1992) found that patients paid by fee-for-service in a large group practice were 27% more likely to be hospitalized than patients of the same group paid by capitation.[35]

The analysis in this section shows the power of payment system design for dealing with the two basic efficiency problems in physician treatment determination. As Ma and McGuire (1997) point out, however, even when the number of instruments equals the number of targets, the payment system may not be able to achieve both of these efficiency targets. For one thing, truthful reporting in payment systems constrains the choice of the payment system. In practical terms, p_s can only be decreased to 0. If it were to go below 0, both the physician (penalized for each unit consumed) and the patient (facing a positive copayment p_d) would have an interest in misreporting quantity to the payer.[36] Although the payer has two instruments, the level of payment R and the degree of prospectiveness, p_s, the second instrument is limited. The payer, because of constraints on reporting, may not be able to attain the control on moral hazard desired by use of supply-side cost sharing.

Beyond this, in reality there are more than two targets. Physicians may be risk averse, for example, and unwilling to bear the degree of supply-side incentives the payer would otherwise want to impose. Furthermore, quantity of one contractible element, x, may not be the only contractible input into treatment. Multiple noncontractible inputs may exist. Payment systems are crude in the sense that the reimbursement and insurance contract creates incentives that are uniform across a range of services. Fortunately, other powerful instruments are also available to managed care plans.

3.4. Network incentives in managed care

Managed care plans have at least two other sets of instruments in addition to supply-side cost sharing to contend with moral hazard. Utilization review allows the third party

[35] See Hellinger (1996) for a review of some of these studies.

[36] See Brundin and Ma (1998) for more extended discussion of the implications of relying on reports from patients and providers for design of payment systems.

payer to interject judgment about what services should be provided to patients. The most obvious mechanism used by managed care in this respect is denial of care. Denial countermands the decisions of patients and doctors under the sway of moral hazard. Outright denial of care, however, appears not to be very common. Remler et al. (1997) found that while managed care plans initially denied 3.4 percent of physicians' requests to hospitalize patients, $2/3$ of these denials were reversed on appeal, leading to an ultimate denial rate of only one percent.

Managed care plans also assemble a "network" of providers. Patients in a managed care plan typically must obtain covered care from a provider in a network. When an out-of-network provider is used, coverage may be significantly less, or absent. Managed care plans do not accept all doctors, and occasionally drop doctors from their networks. Emons and Wozniak (1997), using AMA data, report that 13 percent of all physicians applied for and were denied at least one managed care contract in 1996, and 6 percent were dropped involuntarily from a network during that year. Limiting patients' choice of doctors decreases a consumer's valuation of a health plan, all else equal. What does a plan gain in exchange for this restriction on patient choice?

A managed care plan seeks lower prices from physicians who are granted network privileges. If restricting the number of participating physicians is going to lead to a better price for the managed care plan, it must be that competition is imperfect; otherwise, for any quantity target, the plan would minimize price by admitting all available suppliers. In pharmaceuticals (where there can be no doubt competition is imperfect), networks are called "formularies" and by restricting choice of drugs, managed care plans and others can enhance competition within a class of drugs and bargain for a better price with manufacturers [Scherer (2000)]. Dranove et al. (1993) labeled a similar phenomenon "payer-driven competition" in the case of hospital services. When a managed care plan can direct patients to lower-priced hospitals, it increases hospitals' price elasticity of demand, eliciting lower prices.

A health plan may use a network to pursue other objectives as well. In the same way as in price competition, if the managed care plan can see and evaluate *quantity* per patient, membership in a network can be used as an incentive to control moral hazard. At the level of an individual patient, a managed care plan may have a limited ability to judge the appropriateness of utilization. But at the level of a physician's practice, on which network decisions are based, patient severity will tend to average out (if even imperfectly) and more conclusions can be drawn. A managed care plan may not know if Mrs Smith needed a Caesarean section, but it may be able to say that Dr. Jones' practice, with a 50 percent rate, is not the one the plan wants in its network. Similar remarks apply to observable elements of "quality."

Networks can also be formed by physicians in order to market themselves to managed care plans and other buyers. Provider-formed networks can interfere with managed care plans' contracts, and may pose an antitrust threat in markets where the network has market power (Haas-Wilson and Gaynor (1998), Greenberg (1998)]. "Any-Willing-Provider" laws at the state level inhibit managed care plans from establishing networks [Ohsfeldt et al. (1998)].

In terms of the model set out above, supply side cost sharing creates incentives by altering the payment system component of the physician's revenue function. A network alters the demand-response portion of the revenue function, the $n(NB)$. With a network, patients are not free to go to any doctor according to how they view the net benefits they will receive. Patient flow is at least partially controlled by the plan. In this way, the plan can feed a doctor patients if the patients are getting the care the plan wants them to get. In a regime with moral hazard, physician referral rates can go up (not down) for compliant physicians as quantity is cut back and patient net benefit falls.

Network effects are only beginning to receive explicit attention in the literature. Within a model of monopolistic competition, Ma and McGuire's (1998) managed care plan penalizes a doctor by denying some patients who would otherwise seek out the doctor if the doctor deviates from the plan's target utilization. They show that the plan may need only a small penalty to enforce the behavior it wants. They measure the network effect associated with a mental health managed care plan and find it to be associated with a large (roughly 40 percent) decrease in the quantity of visits per episode of care.

3.5. Efficient production of physicians' services

The literature on physician behavior has often examined whether the production of physician services, or the production of health services, takes place efficiently. The relation of physicians to one another (solo versus group practice) to other personnel, and to hospitals, have all been studied theoretically and empirically. Some of the motivation for this work stems from information and contracting issues.

Solo and group practice forms each have advantages and disadvantages, known in the partnership literature [Farrell and Scotchmer (1988)]. Solo practice internalizes incentives, but group practice allows more flexible allocation of "lumpy" inputs. DeFelice and Bradford (1997) test for productivity of physicians in these two modes and find them to be about equal. Gaynor and Gertler (1995) regard the question of the optimum size of the group as involving a tradeoff between risk and incentives to elicit noncontractible effort.

The economic relation between physicians and hospitals has been questioned in terms of whether it would lead to an efficient use of the two main inputs into the production of hospital care. Traditionally, physicians and hospitals have functioned as independent economic units, with the seemingly odd arrangement that a physician could admit a patient to a hospital free of charge (to the physician) and order hospital staff to incur costs as the physician saw fit. The zero price hospitals charge for "privileges" at the hospital appears not to have originated as an equilibrium, but through custom. If one regards the hospital as "capital" and the physician as "labor," the normal economic organization in a capitalist economy would be for the hospital to hire the physician, pay the physician a wage, while charging the patient a unified bill for hospital care. Pauly and Redisch (1973) suggest an opposite interpretation, that in health care, groups of physicians (the medical staff) act as a worker cooperative, hire capital, pay capital its opportunity cost only (hospitals are non-profit), and garner any surplus in the form of higher reimbursement for themselves.[37] Physicians can exercise market power by restraining the number of physicians who are members of the "cooperative." This relationship may, however, not disturb incentives for production efficiency. If physicians have a claim on any "residuals," they can be presumed to have incentives to minimize cost of treatment to patients [Pauly and Redisch (1973), Pauly (1980)].[38]

Within the conception of the labor-dominated firm, price setting and other contracting practices of third-party payers have altered the division of surplus between labor and capital. Physicians are combining with hospitals in the form of physician–hospital organizations (PHO's) or engaging in contracts, such as exclusive contracts, which, in effect, circumvent the custom of the zero price for privileges. A physician might be paid some amount, for example, in exchange for sending all her patients to a certain hospital. Under pressure from organized payers, physician–hospital relationships are changing rapidly, involving more vertical integration and closer contracting [Robinson (1997)]. The number of physician practices owned or managed by hospital-based systems increased 60% in one year between 1994 and 1995 to $11,234$ (*Modern Healthcare*, June 3, 1996.) As always, anticompetitive effects are possible in tying contracts when at least one of the parties has market power [Frech and Danger (1998)]. Formal analysis of the effect of tying contracts between hospitals and physicians has only begun to contend with an active, contract-writing, third-party payer. [See Ma (1997).]

The physician running a practice is an owner-manager, and as such must combine other inputs with her own time to produce physician services. Since physicians are scarce and highly paid, a certain amount of substitution of other personnel is required for efficient production. Reinhardt (1972) and Brown (1988) are concerned whether a physician hires the efficient number of partially substitutable personnel. In managed care plans, when physicians lose authority about hiring aides, the ratio of physician extenders to physicians goes up a great deal.[39] This could be a move to a more efficient mode of production, or a move to produce a different form of service. Econometric studies of physician cost functions have been undertaken for the purposes of evaluating efficiency [Pope and Burge (1992)] and calibrating "cost-based" payments [Escarce and Pauly (1998)]. The difficulty of measuring all inputs and outputs precludes drawing a clear conclusion about efficiency from these studies.

[37] Harris (1977), an economist and practicing physician, regarded the operation of a hospital as subject to two lines of authority: the medical staff that ran patient care and the management which made long-run resource and marketing decisions.

[38] The incentive to combine inputs efficiently brought about by this form of internalization is disturbed by differential insurance coverage for the inputs. For example, if hospital services are fully insured and physician services are not, hospital services will be overused in Pauly's (1980) framework.

[39] Jerry Cromwell (1996) considers nurse anesthetists to be "nearly perfect substitutes for anesthesiologists" but the nurses are paid at half the rate of the doctors. Substitution has been thwarted because payers, particularly Medicare, pay for inputs rather than outputs. In an integrated delivery system hiring workers according to salaries, the substitution is possible.

3.6. Summary

In a setting of complete information, recognizing that physician services are nonretradable yields the first of the three ways that a physician can influence quantity of health care used by patients. Physicians can set quantity apart from what a price taking patient would choose, subject to a constraint on patient exit. When physicians face regulated prices, market power is channeled into quantity setting, and a decrease in regulated price may lead to an increase in quantity, without a need to appeal to inducement based on asymmetric information or target income motivation. Expanding the purview to a market with many patients, the physician's quantity setting can affect the total demand at the practice level.

A complete information model can also be used to explicate the second mechanism a physician has for influencing quantity. Physician care contains observable but noncontractible elements, referred to in the literature as "effort" or "quality." These inputs may complement or substitute for the contractible inputs. By setting the level of the noncontractible input, the physician can influence demand for the contractible ones. Though noncontractible, effort or quality is not outside of the power of health plans. Partial control over the number of new patients coming to a physician provides a mechanism a health plan can use to manipulate doctors' decisions. Supply-side cost sharing can introduce a penalty for providing "too much" care. This can be balanced with a per patient payment that makes noncontractible quality worthwhile. Managed care networks allow plans to manipulate demand response and give the plan another instrument to influence doctor behavior.

The power of demand response depends, of course, on the patient's being able to observe and evaluate elements of a physician's practice. This ability is called into question in the next section.

4. Uncertainty about treatment effects and asymmetric information

Arrow (1963) titled his paper, "*Uncertainty* and the Welfare Economics of Medical Care" [italics added]. More recently, Wennberg (1985) has argued that uncertainty is the most important factor influencing physician behavior. Many writers on health care have deep misgivings about the application of simple economic models in health care, based largely on how these models ignore the uncertainty and informational asymmetries surrounding health care. In a recent paper, medical sociologist Donald Light (1997, p. 299) writes,

> "Health care is often *emergent* as diagnosis and treatment unfold. Clinical decisions are *contingent* on what is found and how the patient reacts. Cases are highly *variable*, and the course of treatment is *uncertain*." Furthermore, he stresses that, "There is great *information asymmetry*, because the clinician knows so much more than anyone else" [his emphasis].

According to Wennberg, the sources of uncertainty include the following: first, there is classification of the patient in terms of disease condition, or initial health status; second, there is uncertainty about the effects of treatment for a given condition, even in controlled conditions; and third, patient preferences may not be known (at least to the physicians). The presence of uncertainty is sometimes argued to lie behind the observed variations at the population level in the rates of treatment (Phelps, this Handbook).

In health economics, analysis has emphasized situations in which there is uncertainty, and in addition where information about effects is not shared equally, that is, situations of asymmetric information. Before discussing asymmetric information between the physician and the patient, it is useful to consider first the implications of what Pauly (1978) referred to as "irreducible uncertainty:" the absence of information about the consequences of health care treatment that is shared equally by the doctor and the patient.

4.1. Irreducible uncertainty

In Section 3, the benefits a patient receives from health care are $B(x)$. The negative second derivative of $B(x)$, $b'(x) < 0$, can be interpreted as stemming from two factors. Patients may have a declining marginal valuation of health care because (1) as more health care is consumed, the marginal impact on health status is lower, or (2) as more health status is gained, the marginal utility of health status itself falls. To capture both of these effects, we could write $B(x) = V(H(x))$ where $H(x)$ is the relationship between x and health status, and $V(H)$ is the utility of health status function. Both V'' and H'' could be less than zero. Suppose $V'' < 0$, as seems likely. We can then consider there to be *risk aversion* with respect to health status.

A simple way of introducing uncertainty is to suppose that the patient (and the doctor) are uncertain about initial health status, and this uncertainty is additive in relation to the improvement in health rendered by consumption of x. Then, the patient's benefits from treatment must be regarded as expected benefits:

$$E\big[B(x,u)\big] = E\big[V(H(x)+u)\big], \tag{4.1}$$

where u is a random variable with mean zero and variance σ_u. If the degree of absolute risk aversion is decreasing with income, it can be shown that the expected benefits from medical services x are increasing in the level of uncertainty as represented by the variance of u.[40]

Irreducible uncertainty about initial health status or the effects of treatment imposes risk on patients. A patient may want to offset this risk by consuming more health care. Judged in terms of its effect on health status, some demand may appear as not worthwhile, in the sense that the cost is high in relation to the expected increase in health status. However, from the point of view of *ex ante utility*, the utilization may be worthwhile. Patients may be rationally seeking to insure themselves against health status risk

[40] This assumption implies that the third derivative of $V(\,)$ is positive. Intuitively, as people have more income they are less risk averse. See Arrow (1971) or Pratt (1964).

by consumption of medical care. Woodward et al. (1998) show how this can work in a decision-analysis framework concerning diagnostic tests for liver cancer. Expected utility is modeled explicitly. The authors conclude that "risk aversion can increase the perceived value of diagnostic procedures and thus raise optimal diagnostic expenditures" (p. 149). (See also the discussion in Grossman's (2000) chapter of this Handbook.) We now proceed to consider cases in which there is uncertainty, and in which information is distributed asymmetrically.

4.2. *Unobservable physician actions*

Some of what the patient does not know may be known by the doctor. Economic situations involving asymmetric information are referred to as agency problems, wherein the principal (the patient) is affected by an action taken by the agent (the doctor).[41] In some cases, the asymmetry of information between doctors and patients, or indeed between doctors and payers, might be so severe that there is no way for any outside party to know what a doctor did or knew. Some patients (e.g., the very young, the very old) may not be able to report, and some aspects of treatment (e.g., pain relief) may leave no trace. Economic incentives will be little help in eliciting effort in such activities.

There will be many other cases, however, where there is asymmetric information, but mechanisms by which evidence can be collected regarding the doctor's behavior. Better health care can be expected in general to lead to better outcomes, at least probabilistically. If good outcomes can be paid upon, this can motivate doctors, a principle on which some of the burgeoning literature on "performance contracting" is based.[42] Even if outcomes cannot be paid upon, that is, are "noncontractible," outcomes may be observable to clients. It may be infeasible to pay doctors on whether they are able to cure back pain because it is too costly to validate a patient's report. Nonetheless, the patient knows if his back still hurts. If the doctor is rewarded for doing a better job, because the patient is more likely to return or to recommend this doctor to friends, the doctor is encouraged to take unobserved actions to improve quality. Note that this mechanism is similar to that studied above in the complete information case. Instead of observing effort directly, here the patient instead observes an imperfect indicator of effort, outcome.

The patient may see the outcome of the doctor's action, but because outcome is also affected by other unobservable factors (e.g., the uncertainty just discussed above), the patient does not know for sure whether the doctor's action was appropriate. In his general review of agency theory, Arrow (1986) returned to the case of doctor and patient:

[41] There is a presumption in the economics literature that more information to the patient is always good. This may not be correct. Physicians may fail to disclose some information to patients for valid paternalistic reasons. If a child might have cerebral palsy, it could be argued, why tell the parents until the situation clarifies? On the other hand, patients complain that doctors tell them too little, too late. See Sloan et al. (1993, pp. 56–67).

[42] See Lu (1999) and Shen (unpublished) for papers that include models of how providers respond to performance incentives, and measures of the consequences, intended and unintended. See Zweifel and Breyer (1997) for an analysis in which outcomes are contractible.

> The physician–patient relation is a notorious case... The very basis of the relation is the superior knowledge of the physician. Hence, the patient cannot check to see if the actions of the physician are as diligent as they could be (p. 1184).

Physician diligence, referred to in Section 3 as "effort," could take many forms. If physicians know something that could benefit the patient, e.g., initial health status, then revealing this accurately and completely to the patient would help the patient decide how much health care to seek. If physicians must find something out, diligence could consist of taking actions necessary to identify the patient's condition.[43]

Several authors have discussed the applicability of agency theory to physicians and patients [Dranove and White (1987), Mooney and Ryan (1993), Gaynor (1994)], and some papers have developed particular agency models [Blomqvist (1991), Lundback (1998), Zweifel and Breyer (1997)]. Arrow (1963, pp. 964–965) anticipated one important result in the principal-agent literature when he recognized that a potential solution to the informational asymmetry is to transfer all risk to the physician. He called this an "ideal insurance" plan that would protect the patient against the portion of health status risk that could be ameliorated by the doctor's actions. "Under ideal insurance, the patient would actually have no concern with the informational inequality between himself and the physician, since he would be paying by results anyway, and his utility position would in fact be thoroughly guaranteed."[44] In their discussion, Dranove and White (1987) note that transferring risk to the agent is infeasible because health status is in general noncontractible. In common language, patients could exploit the situation by claiming "it still hurts." Mooney and Ryan (1993) later make the same point by stressing how health markets deviate from a standard principal-agent model because of the inability to contract on outcome.[45]

Information asymmetry between buyer and seller is of course not unique to health care. The broader literature contains papers addressing the issue of "experts" consumers must rely on for advice and subsequent services. In a fashion similar to health care, experts in car repair or legal services may have an incentive to call the consumer's problem serious when it is really not, in order to increase demand for their services.[46] One theme in this literature is how competition among experts (who may or may not vary in

[43] These two aspects of diligence are referred to in the principal-agent literature as "hidden information" and "hidden action" problems. See Arrow (1986).

[44] A qualification on this is that there may be some irreducible uncertainty that cannot be ameliorated by physician actions, as discussed above.

[45] In the standard principal-agent model where outcome is observable and contractible, risk can be transferred to the agent. However, if the agent is risk averse, this would require the principal to compensate the agent for accepting the risk. Typically, then, the optimal contract is a contract in which risk is shared between both parties, the principal balancing the costs of paying a risk premium against the inefficiency created by the lower-powered incentives to the agent. See Laffont and Tirole (1993). Zweifel and Breyer (1997) apply such a model to the patient-provider case. Newhouse (1996) discusses the applicability of this type of model to hospital contracting. Paying on "outcomes" may encourage providers to select patients who are healthy and avoid the sick. See Shen (1999) for a model of this behavior and an empirical study.

[46] Gaynor's (1994) review contains an extended discussion of this literature.

skills) limits the potential exploitation of consumers [Wolinsky (1993), Emons (1997)]. Some price competition is part of these models. Market processes can be expected to work differently in health care, however, where experts cannot set prices because of the prevalence of fixed demand and supply prices.

We can add asymmetric information into the model developed so far. Assume there is an input controlled by the physician which influences health outcomes, but that is not observed by the patient. A literal "input" is one interpretation of this new element. A physician could be diligent or not in terms of putting in effort of some form not observed by the patient. Another interpretation, however, is in terms of information. A physician might reveal or not reveal what she knows, or only partly reveal information. More accurate information presumably benefits the patient, by for example, revealing to the patient ahead of time the benefits and risks of some procedure that the patient might use. To ensure that the unobserved action cannot be inferred, there must also be another input, also unobservable; otherwise, the patient could draw accurate conclusions about effort from observing outcome. Thus, write patient benefits as $B(x, e, u)$, where, as before, x represents contractible inputs. In contrast to Section 3, we will now regard e, effort, as unobservable, implying, of course, that it is also noncontractible. Furthermore, u is a random variable, also unobserved, that influences outcomes to the patient. The patient is assumed to know the functional relationship among the variables included in $B(\,)$, to observe x, and to see the outcome, the realized or observed value of health care benefits, which we can call B_o.

A simple illustration of such a model is when effort is (0, 1): the physician consults her expert colleagues or she does not. In this case, the Bayesian patient can infer a likelihood that physician took effort $e = 1$, based on his observation of B_o and x. This could be called L_1 (B_o, x). The likelihood the physician took effort $e = 0$, is simply then L_0 $(B_o, x) = 1 - L_1(B_o, x)$.[47] In a modification of Program III in Section 3.3 above, we can write the physician's profit as:

$$\pi = n(L_1)\big[R + (p_s - c)x\big], \tag{4.2}$$

where n is the number of patients the doctor sees, now a function of the doctor's likelihood of providing high effort. R is the per patient payment, p_s the payment per unit of the contractible input x, and c is cost per unit. If the patient can presume that $e = 1$ characterizes a physician in subsequent encounters, the patient's observation of B_o can be informative. A physician may regularly consult colleagues, or regularly reveal information to patients about their health status. In these cases, patients learn about what they might expect in future encounters. Demand response to this information is important, because it is the mechanism by which the physician is rewarded for high effort. The combination of prospective and per unit payment is important because a payer is interested in affecting the contractible as well as the noncontractible input.

[47] It is possible to generalize this in the direction of repeated observations in which L_0 is a function of a first and subsequent encounters, and in the direction of a continuous effort decision by the doctor.

Pauly (1980), Dranove (1988), and Rochaix (1989) have analyzed models similar to (4.2) when all payment is in the p_s component (which then must be greater than or equal to c), and there is no prospective payment. In these papers, there is a demand response influencing the physician's choice of e. Pauly's fee-paid doctors can take an action to increase the volume of contractible services used by patients, but the doctor must tradeoff the benefits of this action against the cost of discouraging patients from coming to a doctor they regard as "inaccurate." Dranove's (1988) physician makes a treatment recommendation. The patient is more likely to reject the recommendation if the physician has a "bad" reputation as being an "overprescriber." In Rochaix (1989), the threat of patient exit constrains doctors to be more conservative than they otherwise would be. These papers were oriented to the issue of "demand inducement," conceived as an activity a doctor could undertake that would not involve resource costs, but would increase quantity demanded. Said in terms of our model, the physician might observe something about the patient, and then decide whether to reveal this to the patient. Demand inducement as portrayed by Pauly and Dranove could be thought of as the doctor telling the patient they are the "type" that would benefit a lot from health care, even if they are not. This would induce the patient to be willing to consume more of the contractible input forming the basis of physician payment. This constitutes the third mechanism for quantity setting contained in Table 1. The issue of demand inducement will get more attention in the next section. We simply note here that this model covers this case, but is more general. The action could also require the physician to incur some cost. The demand-response mechanism would still reward the action and be a way to elicit the unobserved action.[48]

Generalizing the payment system to be $R + p_s x$ adds an important perspective. As Newhouse (1996) notes, if payment is pure capitation, a provider may have an incentive to use her informational advantage to discourage treatment. And as in Section 3, the physician's decision about effort will be governed by both the level of the prospective payment R, and the payment per unit of the contractible input, p_s.

As just discussed, the physician's choice of unobserved effort is put forward as a characteristic of the physician that the patient (and payer/regulator) could expect to hold true in repeated encounters. In other cases, a physician might make a choice in treatment of a particular patient that might not be a reliable indicator of a pattern of behavior. If the level of effort cannot be anticipated, demand response in terms of number of patients might not materialize. All may not, however, be lost in terms of eliciting effort. Another form of demand response is the patient's behavior in the course of treatment. Treatment is actually a sequence of actions by the patient and the doctor, as Light emphasized above. Patients may learn something over the course of treatment about their outcome, and have a chance to react to a doctor's unobserved effort. A simple example would be if the sequence were as follows: the patient seeks

[48] Chalkley and Malcolmson (1998) analyze the difficulties in writing provider contracts when the demand-response mechanism is not operative.

treatment; the doctor takes more or less care; the patient responds favorably or not to treatment. Then, if the patient's response is poor, the patient must seek more treatment. The "seeking more treatment" is a contractible action, that depends on the noncontractible outcome, that itself depends on the unobservable physician action. "Internalizing" the costs of poor outcomes is a rationale behind prospective payment and capitation, but there is a more general problem lurking here in which the payment to the provider depends on subsequent actions taken by the patient. The power of any such mechanism to redress inefficiencies in physician decisions is yet to be explored.

4.3. Unobservable physician characteristics

We can interpret e in $B(x, e, u)$ in a related way, as an unalterable characteristic of a physician. It could be "quality," but quality understood as a fixed characteristic, such as acumen in diagnosis. The economic issues involved in an unalterable characteristic are somewhat different than a behavior to be induced. If patients value this characteristic differently, efficiency requires matching patients to the right doctor. If the mix of quality can be altered in the long run by changing the composition of doctors, efficiency requires quality to be rewarded.

The physician is an "experience good," as has been noted by Gaynor (1994) among others. A patient literally has to try a doctor, and then make an inference about the doctor's quality, including any issues about a match with the patient's preferences. Because learning is imperfect and slow, the market's reward for higher quality is likely to be inadequate. Hoerger and Howard (1995) review the literature on consumer search for physicians. In their own study of search behavior, they found that only about $1/4$ of women seriously considered an alternative to their prenatal care provider. Satterthwaite (1979), Pauly and Satterthwaite (1981), and more recently Dranove and Satterthwaite (this Handbook) emphasize that asymmetric information about physician type is a basis for monopolistic competition. Pauly and Satterthwaite (1981) observe that if there are more doctors in a market, a patient, through personal contact or by information provided by friends and relatives, will tend to know less about any given doctor because the information is spread more thinly. More doctors may therefore not increase competitiveness of market, but by diluting the quality of information a patient has in some sense, and increase patient allegiance to a known doctor.[49]

A simple learning model has implications about markets and doctor quality [McGuire (1983)]. First, since patients must infer quality by good outcomes, and it is always possible that a "bad" doctor is lucky, the market's valuation of the quality of a doctor will suffer from regression towards the mean. Poor quality doctors will be valued too

[49] Wong (1996) found little support for the information-demand elasticity connection.

highly, and high quality doctors will be valued too low. Second, given uniform prices, no patient believes his doctor is less than average quality.[50]

4.4. Summary

Simple (or irreducible) uncertainty and asymmetric information both have implications for decisionmaking by doctors and patients. A model incorporating both can be built on the structure introduced in the previous section. There are many areas for potential research, along the lines of making more explicit the nature and consequences of asymmetric information. Patient observation and learning and their implications for physician behavior and market equilibrium have barely been explored in the literature so far. Viewing treatment as a sequence of patient and physician actions, for example, seems a potentially rewarding area.

In terms of physician influence over quantity, this section has added "persuasion" to the two already identified in Section 3. As we have said here, in the presence of asymmetric information, the physician can take an unobservable action (with or without incurring resource costs), that will influence patient valuation of care. This activity, denoted "demand inducement" in most of the literature on physician behavior, is the subject of many empirical papers in health economics.

5. Physician-induced demand

Evans (1974) opened his influential paper on "supplier-induced demand" with the statement (ironic in hindsight), "Everyone knows that physicians exert a strong influence over the quantity and pattern of medical care demanded in a developed economy." The meaning and measurement of supplier or physician-induced demand (PID) has in fact been one of the most contentious topics in the economics of health care. The title of Phelps' (1986) paper, "Induced Demand: Can We Ever Know Its Extent?" conveying more dismay than conviction, better captures the tone of the empirical literature.

The hypothesis of physician-induced demand, that physicians alter the patient's preferences in their own interest, threatens the economist's basic market paradigm, and undermines the normative implications that underlie economic recommendations about market policy. Positive and normative economic analysis proceed readily when consumers have stable preferences. Then, combinations of price and quantity can be interpreted in terms of the interests of the consumer and in terms of market efficiency. Throughout the 1970s, economic policy in health care focused on the demand side and

[50] A patient begins with the prior (knowledge? impression?) that the physician is average. Then, if the first outcome is good, the posterior estimate of the physician is improved, and the patient stays. If the first outcome is bad, the posterior drops. In the absence of switching costs, the patient then leaves the physician since the average doctor is always available. A survey of the (rational) patients would reveal that all thought their doctor was average or above in terms of quality.

relied on the basic demand-management paradigm. National health insurance was the main issue, and the operational question for economic policy was the degree of price subsidy that should be contained in national policy. The positive model of utilization determination was demand, and the normative theory of demand was used to derive implications for health insurance coverage. [See Newhouse and the Health Insurance Experiment Group (1993) for a report on this line of work.] Rice (1998) contains an extended discussion of the issues. But if physicians could induce demand, the policy superstructure build over the theory of consumer demand was at risk. As Reinhardt (1989, p. 339) argued in this context, "The issue of physician-induced demand obviously goes straight to the heart of probably the major controversy in contemporary health policy, namely the question whether adequate control over resource allocation to and within healthcare is best achieved through the demand side or through regulatory controls on the supply side." Dyckman (1978, ii), referring to physicians' ability to induce demand, contended that "normal market forces are weak or nonexistent."

The PID hypothesis has direct policy implications, many of which run counter to the normative implications of conventional models of demand and supply. For example, training fewer surgeons will reduce rates of unnecessary surgery in the US [Schroeder (1992)]. Prohibiting physicians from owning testing equipment will reduce excessive rates of testing [Hillman et al. (1992)]. To make physician fee policy within a fixed budget, it must be anticipated that fee reductions induce quantity increases, so fees must be reduced even more than otherwise [PPRC (1991)] to maintain a balanced budget. The Health Care Financing Administration assumed in Medicare's fee reforms that half of any payment reduction would be offset by a volume increase in Medicare [PPRC (1991)].[51] The demand-inducement assumption cost all physicians 6.5 percent of fees in 1992 (an effect compounded over the years).[52] Identifying the empirical importance of the PID effect, if any, in various policy contexts is obviously an important objective of health economics.

Fortunately, at a conceptual level at least, there is agreement about what constitutes PID. We adopt the following definition taken from the careful writing on the subject over the past two decades:

> Physician-induced demand (PID) exists when the physician influences a patient's demand for care against the physician's interpretation of the best interest of the patient.

It is important to keep two distinctions in mind when applying this definition. The first is the distinction between useful agency and inducement. Fuchs (1978, p. 36) early on defined demand inducement as above, in relation to the consumer's optimal consumption point, leaving open scope for influence in the interest of the patient distinct from

[51] The volume offset was assumed to be asymmetric: only price reductions, not price rises would be offset.

[52] See Nguyen and Derrick (1997) for discussion. The regulations are described in The Federal Register, 56, No. 227, November 25, 1991.

inducement. Thus, if a physician influenced a patient to move towards the consumer's optimal point this would not be inducement, only useful agency.[53] Eisenberg (1986, p. 57) defines inducement as "prescription of services that a well-informed consumer would not want to use." Pauly (1980) makes use of the same concept in his definition of a "perfect agent:" The physician assists the patient to demand "exactly those quantities of care of various types that the patient would have chosen if he had the same information and knowledge the physician has." (1980, p. 5). A similar idea of the "perfect agent" is contained in Culyer (1989) and Williams (1998). Even Frech's negative characterization of demand inducement is still consistent with our definition: "As more physicians crowd into a market, they give more fraudulent advice and raise the demand for health care" [Frech (1996, p. 84)]. The upshot of these definitions is that showing influence is not enough to establish "inducement." As Newhouse has said, the question of PID is not a matter of introspection, of "thinking back on one's last visit." (1978, p. 60). (After all, who has not been influenced by physician recommendations?) PID requires, in essence, a finding of "undue" influence. The empirical methods for identifying "undue influence" will be discussed below, but essentially stem from the idea that if a change in the physician's return from inducement (e.g., fees go up) stimulates a change in influence (more surgery recommended), we have evidence for PID.

The second distinction is between utilization and demand, a distinction that has become more salient with the growth in supply-side cost sharing and managed care, rationing devices that do not rely on controlling costs by decreasing quantity demanded. A physician can influence utilization without influencing demand. Here is an example. Patients treated in an HMO may receive less treatment. This could be interpreted as a PID-type mechanism – a decrease in demand caused by the physician. At the price they were paying and with a fully informed demand patients would have demanded the extra treatment but the physician influenced them otherwise and lowered their demand. Alternatively, it could be evidence of rationing – the HMO physicians simply ration the care, not allowing patients to have all they want. The HMO patients have the same demand as the non-HMO patients, it is simply unsatisfied. An empirical finding of an HMO effect or an effect of prospective payment or managed care, as is now common in the literature, is not sufficient to establish PID in the Fuchs/Pauly sense. Utilization has been affected, but it is not clear that demand – the function relating price to desired quantity – has shifted. This quantity setting is of course the first of the three ways we have identified for a physician to influence utilization, not the third way that is of immediate concern here.

The literature outside of economics is without soul searching about whether physicians influence demand. It is nearly universally considered obvious that of course they do. The concern in this literature is usually with identifying the factors, such as socioeconomic status of patients, that lead physicians to direct patients to different courses

[53] This mechanism – supplying information to the patient that changes demand – has been identified and discussed above as the third mechanism by which a physician can influence quantity used.

of therapy. A recent study by the medical sociologist John McKinlay and his colleagues [McKinlay, Potter and Feldman (1996)] described creation of a series of videotapes to study among other things the influence of insurance status, payment method to the doctor, and socioeconomic status of the patient on diagnosis and recommended treatment. The videotape method allowed the investigators to present exactly the same medical information to physicians by different patients and to physicians in two payment conditions. One of the medical conditions studied was chest pain. Insured patients were more likely to be given a cardiac diagnosis, with greater subsequent resource use than the gastrointestinal or psychogenic alternatives. Physicians practicing in an HMO setting were less likely to recommend a follow-up visit for chest pain. Diagnosis and recommended treatment were regarded as physician decisions in this study, and these were significantly influenced by nonmedical factors.[54]

Evidence such as this, while interesting for many purposes, does not make the distinctions between influence and undue influence, and between demand shifting and quantity-setting that underly the economic definition of PID. Physicians may have superior knowledge and can help the patient by conveying this information to the patients, or, perhaps, more directly by simply choosing on the patient's behalf and foregoing the effort of education and persuasion. (Some patients may not want to make choices or pay for this effort if they believe the physician is acting on their behalf.) In the McKinlay et al. study, patients with insurance are more likely to be given the higher cost cardiac diagnosis. Physicians may be acting as proper agents of the patients, giving the higher cost cardiac diagnosis more frequently to insured patients because these patients face a lower out-of-pocket price for care and may be more willing to demand more aggressive therapy. Physicians anticipate that patients have a downward sloping demand curve and recommend more aggressive treatment for the insured. The second finding illustrates the ambiguity of interpretation of findings of less quantity in an HMO. Is the lower quantity achieved through demand shifting or simple quantity rationing?

In terms of the economic view of PID, there are theoretical reasons to believe that PID, in the way we have defined it, exists to some degree. Consider a physician who is giving the "optimal" amount of information to a patient, and the patient is using his optimal quantity. An envelope theorem argument can be made that around this point, a small increase up or down in quantity has a small impact on consumer welfare, because the consumer is near his privately optimal point. The physician, by contrast, may gain or lose money (depending on the payment incentives) from inducing the patient to demand more or less. Whatever particular model we assume about physician motivation, the nature of the tradeoff presented to the physician – I can gain income by a change that has a very small effect on the welfare of my patient – implies that the physician will be doing some demand inducement. In a recent overview of health economics, Blaug (1998, p. 567) contends that, "it is only the quantitative impact … of supplier-induced

[54] Socioeconomic factors, such as race, have been shown in many studies to influence rates of treatment. See, e.g., Lee et al. (1997).

demand that is a bone of contention among American health economists." As Pauly (1980, p. 51) puts it: "Other things equal, physicians would rather tell the truth, but they would be willing to surrender some accuracy for some amount of money income." Once that tradeoff is admitted, it is hard to avoid the conclusion that the physician will be inducing some demand.

5.1. Theory of demand inducement

The theory of demand inducement has received some but not extensive treatment in health economics, surprising in light of the attention paid to the concept in the empirical and policy literature. Evans (1974) proposed that physicians maximize a utility function including income and inducement as arguments, and the disutility of inducement limited the physician's income generation. Fuchs (1978) graphically represented inducement as physicians' ability to shift a market demand curve, without addressing the mechanisms or the limits of inducement.

Any seller gains from a higher demand, and unless there is some cost to inducement, a doctor pursuing net income would induce demand to an infinite extent. It is necessary, therefore, in models of demand inducement, to introduce some limit or cost to inducement. Stano (1987a) takes one direction, making the natural analogy between inducement and advertising. He assumes that inducement has a real resource cost (like advertising) and is limited by the profit calculations of doctors in the presence of diminishing returns. More common are approaches that follow Evans, where inducement is regarded as inherently unpleasant, and limited by the psychic costs the physician bears when she gives advice to the patient slanted toward her own self interest.[55] This conception of the cost of inducement fits well with definitions of inducement we have been working with. Only influences on demand that push the patient away from the optimal consumption point impose psychic costs on the doctor.[56]

There is a distinction in the literature between models of inducement that limit inducement within a profit maximization context [Dranove (1988), Stano (1987a)] and those that incorporate a disutility of acting against the best interest of the patient [Evans (1974), Fuchs (1978), McGuire and Pauly (1991), Gruber and Owings (1996), Zweifel and Breyer (1997), Carlsen and Grytten (1998)]. Although the discussion of physician objectives other than profit maximization is primarily a subject for Section 6, we note here that the empirical literature on PID is set predominantly within models of physician utility maximization.

[55] McGuire and Pauly (1991) take this approach. Zweifel and Breyer (1997) assume physician utility depends negatively on the degree of "artificial demand creation."

[56] As we discussed in Section 4, Dranove (1988) proposed a model of inducement wherein the physician exploits her superior informational position, but is limited by the loss in credibility she suffers by being too aggressive in inducing demand. This model is useful for showing that with asymmetric information, we can expect some inducement. For many purposes we need to go beyond a model that shows there will be demand inducement in equilibrium to address whether the degree of demand inducement changes with changes in the conditions of the market for physician services.

McGuire and Pauly (1991) formalized the ideas of Evans and Fuchs in the context of a model intended to draw the implications of PID for physician response to fee changes. Inducement was limited by physician disutility. Gruber and Owings (1996) expanded on the McGuire and Pauly model by adding a parameter to capture the overall demand and supply conditions. The expanded model can be used to interpret the two main types of empirical studies on PID: physician response to changes in MD/population ratios, and physician response to fee changes.

Modifying the McGuire and Pauly model along the lines of Gruber and Owings, we can write the physician's utility maximization problem as follows:

$$\begin{aligned} &\text{Max}\, U = U(Y, I),\\ &\quad \text{where } Y = N(m_1 x_1(i_1) + m_2 x_2(i_2)),\\ &\qquad\quad I = N(i_1 + i_2). \end{aligned} \tag{5.1}$$

The physician has utility U which depends on her net income Y and the total inducement she conducts, I. $U_Y > 0$; $U_I < 0$; $U_{YY} < 0$; $U_{II} < 0$. She sees N patients who use services 1, 2. Quantity of each service is x, affected by the level of inducement i. $x' > 0$; $x'' < 0$, m is the margin for each service equal to the difference between the fee the doctor is paid and the cost of the service. Other factors influencing demand, such as patient cost sharing, are suppressed since they do not change.[57] The physician chooses i_1 and i_2 to maximize utility.[58] Deriving the first order conditions with respect to i_1 and i_2, and rewriting them, we can describe utility maximization as follows:

$$m_1 x_1' = m_2 x_2' = -U_I/U_Y. \tag{5.2}$$

The marginal (dollar) return to inducement for each service must be equated to the marginal psychic cost (in dollar terms) of inducement. The parameter N, while not explicit in (5.2) is of course one of the arguments in the function for U_Y and U_I. The pair of equations in (5.2) can now be used to interpret the effects of a change in the number of patients per doctor as a result of, say, increasing the number of physicians, or a change in fees, that is, a change in the margins.[59]

[57] The model in (5.1) could be regarded as a generalization of (4.2). Effort could be reinterpreted as "inducement," and in addition to affecting utilization, enters into utility directly.

[58] A number of elements of this model were contained in a paper by Van Doorslaer and Geurts (1987). Their paper looked at the effects of relative prices and income on physiotherapists' treatment decisions. The interpreted the effects in terms similar to McGuire and Pauly.

[59] In a regulated fee environment, Dranove's (1988) model based on physician net income maximization predicts the effect of changes in fees on inducement to be zero. Express a physician's net income as $Y = N(i)x(i)m$, and following Dranove, let the number of patients be a negative function of i. It is obvious that maximization of net income by choice of i is unaffected by m. The physician simply chooses an i to maximize the total volume of services she provides. This would be true in a multiple payer context as well, though some cross elasticities could be generated if physician services were produced at non-constant costs.

A decrease in N, as might be brought about by an increase in the supply of physicians, affects only the third term, $-U_I/U_Y$. Decreasing N decreases the quantity of inducement (in total) as well as decreasing income, thereby decreasing the marginal disutility of inducement and increasing the marginal utility of income. Together, these changes reduce the value of $-U_I/U_Y$. The amount of inducement necessary to bring the return to inducement into equality with this new value must therefore increase. Note that this effect depends on an income effect – the changing tradeoff between I and Y as income changes. Empirically then, the impact of a change in the number of physicians per capita, as studied in many papers on physician-induced demand, is essentially the product of the change in supply on income and the income effect on inducement.[60]

Suppose now the payer for service one (imagine this to be Medicare) reduces its fee, reducing m_1. Service two, paid by another payer, does not change its fee. The effect of this can be thought of in two parts, an income effect and a substitution effect. The income effect comes about because $-U_I/U_Y$ will fall because of the increase in U_Y. This will tend to increase inducement for both services 1 and 2.[61] There is also a substitution effect which comes about because a reduction in m_1 reduces the return to inducement in sector 1. This effect will tend to reduce inducement for service 1 and increase it for service 2 to restore the equality in (5.2). Thus, the effect of a fee reduction in service 1 on inducement for service 1 is ambiguous, depending on income and substitution effects, but unambiguous for service 2 – inducement should increase – because income and substitution effects work together. When service 1 is small and the margin is low compared to 2, substitution effects can be expected to dominate the income effect.[62]

The empirical literature on demand inducement can now be reviewed within this general framework, beginning with papers that study the change in inducement working through an income effect alone.

5.2. Physician-to-population ratios, income effects, and inducement

The early papers testing for PID were concerned with what Pauly (1980) labelled an "availability" effect. Does an exogenous increase in the availability or supply of physicians increase the demand for physicians' services? Demand and supply analysis implies that greater supply should increase quantity demanded through a price effect

[60] Pauly (1980, p. 51) notes that an income-maximizing physician would give the same advice irrespective of the change in the physician-population ratio. To allow for a quantity effect from inducement, "we must enlarge the set of arguments in the physician's utility function." A goal of income maximization does not admit income effects. A utility function is necessary for that.

[61] Even if the income effect was completely dominant and the physician pursued a target income, the TI model does not imply that all income will be recovered from the service experiencing the fee reduction. In general it will be distributed among all the services a physician supplies. See Section 6.

[62] McGuire and Pauly (1991) work out the comparative statics here for cases in which the income effect is all that matters (TI) and when the income effect does not matter at all, and the physician can be regarded as simply maximizing profits. This is discussed more in Section 6 below.

(money or time). Demand and supply analysis also implies through the process of geographic mobility of physicians that areas with high levels of demand ought also to have high levels of supply. Research on PID tested an additional causal effect. According to the PID hypothesis, areas in which supply is large in relation to demand, physicians' incomes are depressed, and they make seek to regain some ground by inducing demand.

Empirical tests of the effect of supply on utilization through demand inducement have taken place at the market and individual level. The first studies were at the market level, and these also suffer most from methodological limitations. From a theoretical standpoint, if the market is regarded as monopolistically competitive, the predictions of the effect of a change in a supply parameter (such as physicians per capita) is generally ambiguous, because the position of the demand curve at the physician level may affect the elasticity and the markup, with uncertain effects on the direction of price and quantity change [Reinhardt (1978), Frank (1985)]. When the physician with market power can also set "quality," the picture is even more clouded. Feldman and Sloan (1988) show that in this case, quantity can go up or down in response to supply shocks. Changes in the number of physicians per capita do not have a direct effect on the return to inducement, but only affect the utility or disutility of inducement through an income effect. Tests of PID examining physician/population ratios are in effect testing the joint hypothesis of induced demand and income effects. If the income effect itself is weak, then inducement may not change even if inducement is going on.

In terms of data, studies have been subject to a number of criticisms relating to the unobservability of key variables, and the possible correlation of these with the key independent variable, physicians per capita. Supply will tend to equal demand in a competitive equilibrium, so unless demand side variables are adequately controlled for, supply will be correlated with utilization because of market equilibrium conditions [Auster and Oaxaca (1981)]. Major determinants of demand and supply, including input prices, and even the money or time price of health care are sometimes omitted from empirical models. Most studies have used cross-sectional variation in supply, leaving open the possibility of severe bias from unobserved variables. Market definition issues (e.g., border crossing) introduce measurement error and possible bias [Dranove and Wehner (1994)]. Using a physician-to-population ratio as a measure of an exogenous income shock in a cross section is dubious [Gruber and Owings (1996)].

From the first, it should be noted, researchers were aware of many of these limitations and took what steps they could to contend with the difficulties. Fuchs (1978) in the first major paper of this type, studied the effect of the supply of surgeons in 22 metropolitan areas in 1963 and 1970. He chose surgery because time price is presumably less of an issue for these serious procedures, and quantity can be fairly readily measured. He used a TSLS regression to replace the actual number of surgeons by a fitted valued using metropolitan/nonmetropolitan area, hotel receipts, and percent white in the population in the first stage regression. A ten percent increase in the number of surgeons increased the rate of surgery by three percent, according to his estimates.

Cromwell and Mitchell (1986) address the same question as Fuchs, using the same methodology with more years, more and smaller geographic areas and better controls, and find results consistent with the earlier analysis, though with a reduced estimated inducement effect. Following up on an idea of Green's (1978) that some physician markets might be in shortage or surplus, and that we ought to be able to see an availability effect more readily in a shortage area, Cromwell and Mitchell found a much larger effect of supply on use in areas of high surgeon workload (elasticity of 0.28) than in areas with low workloads (elasticity of 0.09). Carlson and Grytten (1998) point out that a bigger availability effect in "shortage" areas may be more consistent with a rationing effort than with PID.[63] Using a similar methodology, Birch (1988) and Grytten et al. (1990) find that the number of dentists per capita has a strong effect on the volume of dental visits.[64] Cross-sectional variation in physician-to-population ratios used to estimate a demand inducement effect have left skeptics unpersuaded on theoretical [Feldman and Sloan (1988)] and empirical grounds [Phelps (1986)].

Two recent papers have enlivened the literature on PID and market-level effects of physicians per capita. Dranove and Wehner (1994) tested for "induced demand" in a case where it surely does not exist: the effect of the obstetrician-to-population ratio on the volume of births. They reasoned that if they found evidence for induced demand using the techniques common in the earlier studies of surgery, the results of those studies would be suspect. Mimicking the Fuchs and the Cromwell and Mitchell methodology, Dranove and Wehner indeed did find evidence that obstetricians appear to induce births (in an economic sense), substantiating the omitted variable and border-crossing criticisms of the earlier studies.

Gruber and Owings (1996) also looked at births, studying the income shock to obstetrician/gynecologists resulting from the 13.5% fall in fertility among US women occurring over the period 1970–1982. They reasoned that an income effect should lead obstetrician/gynecologists to induce demand for the more lucrative Caesarian section procedure over vaginal deliveries. The timing and magnitude of the decline differed across states, and Gruber and Owings found a strong correlation between within-state declines in fertility and within-state increases in Caesarean section rates, though the absolute magnitude was small. A 10 percent fertility drop corresponded to an increase of 0.6 percent in the probability of a C-section. Gruber and Owings (1996, p. 113) calculate that obstetrician/gynecologists replace about 10% of the fertility-caused income drop by an increase in C-sections. (These results do not preclude other income-recovery effects. Obstetrician/gynecologists could have changed the level of demand inducement for the gynecologist side of their practice as well. Births account for about half the income of obstetrician/gynecologist specialists, though physicians tend to specialize in either obstetrics or gynecology.)

[63] In the case of the correlation between quantity of physicians and use, a number of authors [Frech (1996)] have speculated that this can be explained by rationing. If prices are controlled, e.g., in Medicaid or in other plans, use is not demand-determined. An increase in supply increases use but by decreasing rationing, not through a PID effort. Comanor (1980) found a strong availability effect (Pauly's term) in Ontario between 1972 and 1975. Adding specialists to an area increases use of specialty care almost one-for-one. A result like this would be consistent with demand rationing as well as PID.

[64] For discussion of European empirical studies, see Zweifel and Breyer (1997).

Gruber and Owings make a good case that within-state fertility declines can be regarded as an exogenous shock to demand and income, and therefore represent a good test of demand inducement occurring through an income effect. Extensive specification checks attempting to rule out alternative explanations for the correlation between within-state fertility declines and C-section rates lend confidence to the demand-inducement interpretation.[65]

Some research uses individual level data for which market level measures of supply could be taken as exogenous.[66] Pauly (1980) used data from the 1970 Health Interview Survey from the National Center for Health Statistics, with information on over 100,000 persons. He predicted that inducement effects should be largest for the least well-informed consumers, those with low income in big cities (big cities because information on each physician is less available). He split his sample and found an availability effect for ambulatory care (not hospital care) for the group he predicted, but this was small. Pauly was explicit in recognizing that he only measured a change in inducement with a change in physician supply. "If they [physicians] do manipulate information in an effective way, they do so to approximately the same extent regardless of how many of them there are in an area, and regardless of how busy they are" (p. 16).

In a pair of papers, Rossiter and Wilensky (1983, 1984) distinguished between physician-initiated and patient-initiated visits, and between visits for more or less discretionary procedures. When health insurance and other patient level controls were taken account of, the effect of physicians per capita on use was small, and statistically significant only for more discretionary procedures. [See Stano (1987a) for critique and discussion of these studies.] Scott and Shiell (1997) later studied GPs in Australia and found spotty and small effects of GP density on physician-initiated follow up visits.

The PID hypothesis can be looked at for its implications for supply per physician as well as demand per patient. Viewed from the supply side, a strong PID effect should allow a physician to insulate herself from competition by inducing more to make up for the fewer patients to go around. The Newhouse et al. (1982) results discussed in Section 2 on location are inconsistent with full insulation. McCarthy (1985) studied the number of visits supplied by primary care practitioners in 1975 using data from the AMA about price, waiting time, and other information about the physicians' practice, as well as market-level data on physicians per capita. In most specifications, full insulation (consistent with powerful PID) could not be ruled out, but his estimates of this effect were not very precise.[67] McCarthy (1985) also estimated that demand for individual primary care practitioners to be highly price responsive, around −3. If demand is very elastic, the physician's price-cost margin is also likely to be low. In this setting, inducing demand has little payoff to the physician in comparison to where marginal induced visits bring in high profits [Stano (1987b)].

[65] For a corroborating study using fees but in the same clinical area, see Keeler and Brodie (1993).

[66] The argument is that individual level unobservables are less correlated with overall market demand.

[67] He noted that density of primary care practitioners was highly correlated with other important explanatory variables.

The basic premise behind PID is that physicians may exploit the information gap between themselves and patients. If so, as Pauly noted in his study of high- and low-educated consumers, more PID (e.g., more surgery) should be observed where the information gap is greater. Bunker and Brown (1974) reasoned that the smallest information gap should be between physicians and patients who were themselves physicians or their families. Rates of surgery, including "discretionary" procedures like appendectomies and hysterectomies were actually higher among Stanford University Medical School faculty and their spouses than among a group of other professionals and their spouses, controlling for age and sex. It may have been, as Hay and Leahy (1982) speculate, that the physician families may have faced lower prices because of professional courtesies or better insurance, or had easier access to better services, perhaps explaining their higher rates of use. To address this, Hay and Leahy used survey data, more extensive controls, but also found the physician-families used more services. (When professional courtesy was reported these few observations were dropped but may not have been fully reported since there were only four occurrences in 7800 respondents.)

With the exception of the recent study by Gruber and Owings, the evidence for PID with the "availability effect" is equivocal. Recall, however, what those studies test: that there is an income effect on PID. Absence of an income effect on PID does not, of course, imply that PID is not taking place, or may not respond to other exogenous market changes.[68] By the late 1980's, physician fees were being set by payers, and changes in fees were being examined for their effect on PID.

5.3. Fees and inducement

After the success of Medicare's price control program for hospitals embodied in the DRG system in the early 1980's, the largest payer in the US turned its attention in the later part of the decade to developing a price-setting policy for physicians [Pauly (2000)]. Medicare had imposed various price controls in its program over the years, but had never undertaken full-scale "rationalization" of prices. The existing pattern of prices were regarded as irrational and distortionary, contributing to the over use of invasive procedures [Hsiao et al. (1988a)]. Physicians, by and large, went along with fee reforms because Medicare promised to conduct the reforms in a revenue-neutral fashion (like DRG reform), to rationalize, not reduce, prices, at least on average.

The threat of PID in response to fee changes made Medicare actuaries skittish, so much so that they interpreted "revenue neutrality" to require a cut in average fees. The problem was that when the anticipated (by the actuaries) demand inducement was figured into the revenue-neutral calculations, fees had to be reduced by an extra 6.5 percent to keep Medicare's books in balance. The actuaries figured that each 1% reduction in a fee would lead to 0.5% increase in volume. The actuaries' predictions were not borne

[68] As Phelps (1997, p. 214) points out, "none of these studies could show that inducement *was not* occurring, but only that an alternative explanation existed."

out. The dreaded "volume offset" failed to materialize at least at the aggregate level. During the period 1991–1996, when the new fees were phased in, the price-adjusted volume increase for surgery (where fees were reduced) was lower than for primary care or evaluation and management services (where fees were raised) [Medicare Payment Advisory Commission (1998)].

In terms of research, studying PID in the context of changes in regulated fees has some advantages over the earlier literature concerned with an availability effect. As we noted above in Section 5.1, a change in a regulated fee changes the physician's fee/cost margin and directly changes the incentives to induce, without relying on the transmission of an impact through a potential income effect. Even profit-maximizing physicians may respond to margin changes by changing inducement. Income effects may matter, of course, but may not be as empirically important as substitution effects. In the case of an own-fee change, income and substitution effects of a fee change work in opposite directions, a fee reduction tending to increase inducement for all services because of an income effect, but making inducement less remunerative tending to decrease inducement for the service directly affected. Income and substitution effects work in the same direction for cross-fee effects, suggesting such cross-effects as a promising area for empirical research.

Fee change studies also have advantage over availability effect studies from an empirical perspective. Fee changes follow from regulatory policy that can more be readily regarded as exogenous to the physician's practice or the consumer's demand.

Hadley and Lee (1978) report that the utilization growth in Medicare during the price freeze of 1972–74 was so great that the rate of growth of total costs exceeded the growth after prices were unfrozen in 1975.[69] The alternative explanation, common to a number of such studies, is that a price freeze reduces the price to the consumer facing a given circumstance so that it is a demand response. In Medicare, balance billing was falling over this period as well, possibly accounting for some rise in demand.

In another market-level study of provincial billings in Ontario, Canada over the period 1975–1987, Hurley et al. (1990) studied fees and rates for 28 procedures. No discernible pattern in the negative and positive responses to own-fee changes was found. Hurley and Labelle (1995) later found no evidence for a utilization response to fee changes in Canada. Escarce (1993b) also found a mix of positive and negative responses when responses were studied procedure by procedure. This may reflect the ambiguity of theoretical predictions about the direction of change for an own-price effect. The Canadian data have the advantage of being generated in a single payer system with no balance billing. Billing data could be aggregated to the practice level, and income and substitution effects could perhaps be separately identified, by for example, variation in the baseline composition of certain procedures at the physician level. Rochaix (1993) found that GPs in Quebec subject to quarterly income thresholds at which fees were lowered, responded by changing the mix of services they supplied to patients.

[69] See also Holahan and Scanlon (1978), Feldman and Sloan (1988).

In a well-known study, Rice (1983) examined rates of procedures per encounter with physicians in Colorado following administered price changes. If we view the encounter as an episode of care, the experiment is meant to investigate how fee changes affected the physician's "practice style." In 1977, Medicare in Colorado began to set fees according to state-wide averages, which had the effect of reducing fees for the previously higher paid physicians in the Denver–Boulder area, and increasing fees for physicians located elsewhere. Although Rice did not couch his model in terms of income effects and own- and cross-price effects, he did regress measures of the quantity of surgical services, medical services, and ancillary services, on changes in prices paid for each of these services at the physician level. In some regressions he included only own-price effects and in others included some cross effects. For surgical services, the own-price effect was negative and smaller in absolute value than the also negative cross-price effect from medical services. This pattern is consistent with income and substitution effects working in the same direction for the cross effect, and in opposite directions (but the income effect winning out) for the own-price effect. For surgical and ancillary services, the pattern of own and cross effects (where estimated) did not appear to be consistent with an income and substitution effect model.[70]

Nguyen and Derrick (1997) studied "overpriced procedures" for which Medicare reduced fees in 1990. They improved on the earlier literature by aggregating effects to the medical practice level. Using 1989 quantities as weight, they constructed a price index for each physicians practice and examined the impact of this price index on an index of Medicare volume. They did not disaggregate the income and substitution effects of the price change but interpreted their results in these terms. Overall there were no significant volume responses (income effects just balanced by substitution effects) but for the 20% of physicians who experienced the largest price reductions, there was a significant negative net income effect. For these physicians, a one percent reduction in price led to an increase in volume of about 0.4%.

Surgeons were most adversely affected by Medicare fee reform. Thoracic surgeons were projected to lose 26% of their income [assuming constant volume; see Levy et al. (1990)], making them one of the hardest hit groups, and also, therefore, one of the groups best to study to look for income effects. Yip (1998) applied the McGuire and Pauly model to thoracic surgeons in New York and Washington state.[71] She measured the total income impact of a set of Medicare fee reductions for "overpriced procedures" and included this in a series of procedure-level regressions for Medicare and private insurance patients along with measures of procedure price and other covariates. She found strong evidence that Medicare fee cuts led to increased volumes by thoracic surgeons

[70] In a companion paper, Rice and McCall (1983) study the effect of the Medicare rate changes on physician pricing behavior and the willingness of the physicians to accept the Medicare payment as full payment. There, the responses were consistent with conventional economic behavior, rate increases led to price increases and increased willingness to accept Medicare fees as full payment (accept "assignment").

[71] These two states include physician identifiers in these discharge abstract files. Also, thoracic surgeons practice is primarily hospital-based.

to both Medicare and private payers, and that their effect worked through the income effect.[72] Taking responses across the board into account, Yip estimates that thoracic surgeons recouped 70% of the income lost by price inductions by a volume increase.[73]

5.4. Other evidence bearing on PID

In addition to changes in numbers of physicians per capita and changes in fees, other research on physician behavior also bears on physicians' persuasive powers. A difficulty in interpreting these studies is in determining which of the three quantity-setting mechanisms is operative.

5.4.1. Defensive medicine

"Defensive medicine" is when a doctor conducts procedures in order to protect herself against litigation [Danzon (2000)]. A economically pure instance of defensive medicine would be when a procedure provides no benefit to the patient or involves risk to the patient, but the physician recommends the procedure anyway for selfish reasons. Obviously, a fully informed and price-taking patient paying any part of the cost or submitting to the risk would not agree to such a procedure. If such procedures are observed, some physician quantity setting power must be in place. Note that this is not necessarily demand inducement, but could be the simply the first quantity setting mechanism identified in Section 3.

Most malpractice goes undetected, unpunished, and uncompensated. In one well-known study [Localio et al. (1991)], 98 percent of patients injured by negligence (malpractice) did not sue. At the same time, physicians perceive the risk of malpractice to be a significant one. Lawthers et al. (1992) report that doctors overestimate the risk of being sued by a factor of three, and the risk of being sued contingent on committing malpractice by a factor of 30. In qualitative responses, doctors regularly report that fear of malpractice motivates their actions. Blendon et al. (1993) found that 32 percent of US physicians admit that they "often" do more than is clinically appropriate; another 29 percent admit to "sometimes" doing too much. Four of five doctors from the Lawthers et al. study agree that they order extra procedures to protect themselves against malpractice. Some older empirical studies of lab tests asking doctors to respond quantitatively to how much of lab activity is due to malpractice get estimates of from nothing to negligible amounts [Garg et al. (1978), Lusted (1977), Werman et al. (1980), Hirsh and Dickey (1983)]. More recently, Kessler and McClellan (1996) studied the effect of malpractice

[72] In a similar study, Tai-Seale et al. (1998) also used hospital level data in Medicare and private insurance to study volume responses to Medicare fee-reductions to "overvalued procedures." The directions of effects were generally consistent with the income and substitution effect framework, although there were many data limitations, and many estimated effects were insignificantly different from zero.

[73] Yip (1998) studied both the increase in volume for a given procedure and the switch between less and more generously reimbursed procedures.

liability reform on the treatment of heart disease among Medicare beneficiaries. They concluded that reforms, hypothesized to reduce fear of liability among doctors, caused a 5–9 percent reduction in medical expenditures.

5.4.2. Self referrals

When physicians have financial ownership in testing and therapy facilities to which they can refer patients, they refer more often [Hillman et al. (1990, 1992)], and patients may be treated more intensively [Mitchell and Sass (1995)]. Self ownership in a regulated price environment is like a fee increase, the interpretation in terms of PID is thus similar to Section 5.3. Extremely high rates of use of prescriptions in connection with physician office visits in Japan is partly attributable to physician dispensing authority and how regulated office visit fees there [Scherer (2000)]. The temptations physicians face in generating income at patients' expense when physicians own ancillary facilities is recognized by restrictions in federal statutes [see Getzen (1997, pp. 146–148)]. American physicians are prohibited from owning pharmacies, and are restricted in their rights with respect to other types of referral destinations.

5.5. Summary comments on PID

Returning to the Fuchs/Pauly definition of PID, two things must be established for evidence on physician control or influence to support the PID hypothesis. First, the exercise of control must be in the interest of the physician, not the patient. This criterion, even without a gold standard of what patients ought to get, seems to have been met by the studies reviewed here. Adding up the evidence, on obstetricians doing more C-sections, surgeons doing more bypass operations, physicians referring more frequently to their own labs, and other studies, makes a convincing case that doctors can influence quantity and sometimes do so for their own purposes.

The second criterion from the Fuchs/Pauly definition is that the physician exercise quantity control by influencing the patient's demand, not by quantity setting through rationing, or via the quantity setting power available to the monopolistic competitor. Consider first the availability-effect literature. The profit-maximizing, quantity-setting monopolistic competitor is not influenced by income effects. If adding more doctors to a market is associated with more aggressive quantity setting as in Gruber and Owings, the net income or profit-maximizing model is contradicted, and the PID model is supported. It is possible to come up with other income-effect mechanisms to explain the availability effect finding, but all of these must also sacrifice the objective of income maximization.[74] It also is possible to construct other explanations for the association

[74] There are many possibilities. The physician might, for example, choose quantity trading off her own interest with the interest of the patient (Section 6.1). The income effect would then have the effect of alternating this tradeoff more in favor of her own interests. The general idea is that the physician has some other objective that becomes less important in relation to her own interests when her income falls.

between availability and quantity per patient, involving time price/access to patients, or even of effects via the market power of physicians being enhanced by more physicians in a region [Pauly and Satterthwaite (1981)]. Without more direct evidence for these possibilities, the income effect-inducement hypothesis gains support from this segment of the availability-effect literature. Much of the availability-effect literature is, however, vulnerable to the statistical artifact argument [Dranove and Wehner (1994)].

In the fee-effect literature, if fees fall for a small part of a physician's practice, the physician's response will not be revealing regarding PID. If quantity falls, this might be PID or standard supply response. If quantity goes up, it is impossible to interpret this as an income-effect driven increase in inducement – the income effect would play out for all services, and for the segment of the practice where fee fell, would have to work against the substitution effect [McGuire and Pauly (1991)].[75] There are a few papers [e.g., Yip (1998)] for which Medicare fee changes had a large enough income impact on physicians for there to be credible income effects at work. What about demand-shifting versus quantity-setting? For a fee fall, the quantity-setting model (or even a simpler price-taking consumer model) generates a negative fee-quantity prediction if the fee reduction reduces the price to the patient and loosens the demand constraint). This could have been operative in some of the earlier work on Medicare fees [Rice (1983)], but is unlikely in the later period studied by Yip (1998) when there was more supplemental insurance by the elderly, and for procedures where the surgeons' fee was a small part of the financial and personal cost.

In sum, there is a large volume of research on PID that is supportive of PID but not highly discriminatory between the PID hypothesis and theories with fixed patient preferences. Some recent papers have refined the tests for PID using income effects and have provide more direct support for the PID idea. It is worth recalling at this juncture, as Phelps (1997) and others have pointed out, that the studies discussed are about *changes* in the intensity of PID following income changes. PID could well be empirically important but simply not respond to income effects. A no-effect finding in the literature therefore does not contradict the existence of PID. It is hard to dispute Fuchs (1996, p. 3), in his presidential address to the AEA reviewing the state of health economics, where he says that, "Despite many attempts to discredit it [citing Dranove and Wehner (1994)], the hypothesis that fee-for-service physicians can and do induce demand for their services is alive and well [citing Gruber and Owings (1996)]."

The "fee-for-service" physicians in Fuchs' quote are now increasingly being paid by managed care plans, and their response to this environment presents new opportunities to test PID. Empirical cases are appearing in which insurance benefits to patients are improved, and use falls due to other managed care changes implemented at the same time [Ma and McGuire (1998)], strong evidence for the existence of some physician control over quantity. Determination of the discharge status of the patient ("satisfied,"

[75] This leaves open the issue of how such findings are to be interpreted. One possibility is suggested in Section 6.3.3 below.

"rationed"), and how this changes with managed care would be one fruitful direction for distinguishing between PID and rationing hypotheses.

It seems appropriate to return to the question of what difference it makes if physician quantity control is exercised through manipulation of demand/preferences or a more direct quantity control without changing demand. For some purposes it does not matter. From the normative standpoint of evaluating whether a change in fees exacerbates overutilization, the comparison is with respect to some standard of efficiency – a cost-effectiveness standard from clinical research or the mystical perfect-agent demand curve. If a physician increases quantity beyond the desired point, it is irrelevant for the planner whether the mechanism is preference shift or direct quantity setting.

From the standpoint of running a market in health plans, however, the distinction between PID or direct quantity setting does matter. Much of contemporary health policy is based on the assumption that a market of competing health plans can be the basis of an efficient health care system. For this strategy to succeed, consumers must be able to evaluate and choose plans in their self-interest. If plan A rations me too tightly, I leave it and go to plan B. This process will work if I can evaluate when I am rationed (or given too many things I know I don't really need). It works, in other words, if physician quantity setting is seen for what it is by consumers. If, however, physicians use PID in competing health plans, and patients are persuaded to like what they get, the basic mechanism for encouraging efficiency in rationing among competing managed care plans breaks down. Given the direction of health policy in the US and many developed countries, it is more important than ever to know if, when physicians do less in managed care, they persuade patients that this is good for them, or if patients are knowingly choosing a plan that rations them to efficiently contend with the prisoners' dilemma of moral hazard in health insurance.[76]

In his presidential address, Fuchs (1996, p. 8) also reported the results from a sample of American health economists, economic theorists, and physicians. 68 percent of health economists, 77 percent of theorists, and 67 percent of practicing physicians agreed with the following statement:

> Physicians have the power to influence their patients' utilization of services (i.e. shift the demand curve), and their propensity to induce utilization varies inversely with the level of demand.

The majority of respondents agree that two things are true, that physicians induce demand and they do so in a way that does not reflect simple profit or income maximization. In the next section, we turn to the question of the objectives being pursued in physician decisionmaking.

[76] Of course, alternative directions for policy, such as relying on a "single-payer" system without competition, also have their own set of distortions and inefficiencies.

6. Other physician objectives

Many years before his survey of economists and physicians, Fuchs warned in his influential book (1974, p. 60), *Who Shall Live?*: “A common mistake is to think that the behavior of physicians can be understood only in terms of their desire to maximize income.” A “charity hypothesis” [Kessel (1958)], concern for medical ethics [Arrow (1963)], desire for interesting cases [Feldstein (1970)], and target income [Newhouse and Sloan (1972), Evans (1974)] had all been proposed by the time Fuchs was writing. One way or another, these alternative hypotheses incorporated a concern for patient health or economic welfare into physician objectives.

Neoclassical economic analysis derives predictions about behavior of suppliers by assuming that suppliers make decisions in order to maximize the profit of the firm. Reservations about the profit maximization assumption are of course not confined to physicians and the field of health economics. “Alternative theories of the firm” relying on utility maximization, bargaining models, or behavioral models, with parallels in the health literature, have been around for many years. While this literature has had only a limited impact on mainstream economics, physicians seem a particularly good candidate for alternative objective functions. Physician-firms are most often “owner-operated” permitting the ready indulgence of utility-related objectives. Physicians enjoy very high incomes, and if a falling marginal utility of income renders income less attractive in relation to other objectives, physicians are likely to be susceptible to an income effect. Physician decisions affect their customers in a profound way – the “cost” to the client of physicians’ opportunistic behavior may be much higher than it is for other suppliers with comparable informational advantages. Therefore, physicians, facing a tradeoff between what is good for my customer and what is good for me, may be induced to sacrifice more income because of the steep tradeoff. Finally, as Arrow (1963) pointed out, the social norm shared in large degree by the buyers and the sellers of health care is that physicians enjoy special independence in decisionmaking in exchange for an understanding that they act in the patient’s best interest.

6.1. Medical ethics as a constraint on choices

Medical ethics involve a wide range of issues from confidentiality, duty to inform patients, end-of-life treatments, as well as our concern about the role of ethics in routine decisions about health care use [Anderson and Glesnes-Anderson (1987)]. Medical ethics command physicians to *primum no nocere* (first, do no harm), and more actively, to “act in the patient’s best interest” [Hiller (1987)]. Arrow (1963) thought that the efficient and fair allocation of health care depended on physicians behaving ethically, not exploiting patients’ vulnerability. Mechanic (1998) argues that the reciprocal relationship of ethical behavior and trust contribute to effective medical care. The imposition of an active third-party payer into clinical decisionmaking through managed care has been alleged to threaten physician’s loyalty to patients [Rodwin (1993), Emanuel and

Goldman (1998), Mechanic (1998)], raising new questions in the debate about the role of ethics in medical care.[77]

One interpretation of medical ethics is that in a situation in which there are many choices of how to treat a patient, ethics dictate that the physician chooses the "medically correct" way to proceed. If ethics operated in this way, health care choices would not respond to economic incentives on either the demand and the supply side, as they evidently do. Ethics could fit within a model in which incentives matter by serving to cull some but not all alternatives from what a physician could choose. It might be unethical, for example, for a physician to conduct a Caesarean section in some situations, unethical not to in others, and in some third set of conditions, "ethics" might not come into play in the decision yes or no. This view of the role of ethics seems to fit Hillman's (1990) meaning: "Whereas most physicians will act in the patients' best interest when the medical decision is clear-cut, the effect of financial incentives may be more important in areas where the correct decision is not clear."

Formal treatments of ethics in medical decisionmaking are few. Ma and McGuire (1997) study a model in which ethics are represented by a lower bound on the health benefits a physician is willing provide to a patient. Physicians control one input (non-contractible effort) and the patient controls the other (contractible visits). If the physician anticipates that a patient will not choose extensive treatment, perhaps because of the prices the patient is paying, the physician feels compelled to make up for this by putting in more of the input she controls. A payer can take advantage of a physician's ethical constraint by setting up a payment system that puts the physician in the position of being forced to take more effort to make sure the patient attains an acceptable outcome.

6.2. Utility and the patient's best interest

An "ethic" has the flavor of a dictate or a constraint – once the constraint is binding, other objectives of the physician become irrelevant. Perhaps for this reason, most papers in health economics do not use a constraint to represent ethics, but instead represent physician concern for patients with a utility function including as an argument something valued by the patient (quantity, quality) or the patient's utility itself. In this construction, the physician's ethically driven concern for patients is subject to being traded off against self-interest.[78] Models with physician induced-demand reviewed above in Section 5 often use physician concern for patients, Pauly's "internal conscience," as a brake on inducement. Eisenberg (1986, p. 57) expresses the role of patient welfare in

[77] Mechanic and Schlesinger (1996) speculate that financial incentives managed care plans give physicians incentives to reduce treatment interfere with the "trust" between physicians and patients that contributes to effective medical care. A physician with incentive to do less than necessary may, however, be as trustworthy as one who has incentives to do too much.

[78] For an early model of this type, see Woodward and Warren-Bolton (1984). There, "ethics" took the form of a cost felt by the physician as the actual treatment diverged from the "medically appropriate" treatment.

a way consistent with the argument-in-the-utility-function interpretation, “A substantial part of the physician’s satisfaction with practice is fulfilled by serving successfully as a patient’s advocate.” He reviews evidence to show that prices patient’s pay affect the physician’s behavior, implying a concern for patients’ overall well-being (not just health) on the part of the physician.

Other papers endow physicians with a utility function of the form $U(\pi, B)$, where π is the physician’s net income, and B is the benefits or utility the patient receives [e.g., Chalkley and Malcolmson (1998), Ellis and McGuire (1986), Ma and Riordan (1998), Rosenthal (1998)].[79] An attractive feature of such a formulation is that it can be used to derive a “supply curve” of services to a patient as a function of the degree of supply-side cost sharing. Substituting a payment system with a parameter for supply-side cost sharing readily generates comparative statics [Rosenthal (1998), see also Jennison and Ellis (1987)]. With assumptions about the form of $U(\ )$, supply-side payment systems can be solved for the form which maximizes social surplus (normally, simply $\pi + B$).[80]

Another interpretation of $U(\pi, B)$ is that it represents a Roth–Nash solution to a cooperative game between a patient and a physician disputing what quantity the patient should consume. The patient demands x_d on the basis of his insurance coverage, the physician desires to supply x_s on the basis of her payment. Maximizing $U(\ )$ with respect to x represents the axiomatic solution to the bargaining [Ellis and McGuire (1990)]. This interpretation can be regarded as a generalization of Program II in Section 3. The maximand expressed in Equation (3.5) maximized profit subject to a constraint on consumer utility. In terms of a bargaining model, the consumers NB^0 is the “point of minimal expectations,” and all the bargaining power rests with the physician [Roth (1979)]. More generally, bargaining power is shared between the two parties and the quantity settled upon will fall between the quantities desired by the two parties.[81]

6.3. *Target incomes*

A radical alternative to the assumption of profit maximization in the theory of the firm is the so-called “behavioral theory,” pioneered by Simon (1958). Behavioral theory claims firms operate by “rules of thumb” rather than by maximization, with targets for rates of return or markups over cost. In health economics, a prominent behavioral theory proposes that physicians make decisions to maintain a “target income.”

[79] This formulation of utility has parallels in the hospital literature. See Newhouse (1970), Feldstein (1971), Frank and Salkever (1991).

[80] In Ellis and McGuire (1986), the utility function is additive, although the weights on profit and patient benefit need not be equal.

[81] In the simple case of quadratic utilities of both the patient and the physician and equal bargaining power, the Roth–Nash solution turns out to be the simple average of the two quantities [Ellis and McGuire (1990)].

6.3.1. Background

The target income (TI) hypothesis explains why higher physician-to-population ratios (presumably a measure of supply in relation to demand) can be associated with a *higher* price of physician services, not a lower one, as simple price theory would suggest. Suppose physicians only set a price high enough so as to attain some target. They could make more by charging a higher price, but choose not to, perhaps because of concern for patients' welfare. As more physicians appear in a market and patients are spread more thinly among the available suppliers, physicians must raise prices to maintain the target income. TI behavior, in the 1970s, reflected *restraint* in pricing. Physicians were "humanitarian" in Farley's (1986) term, not fully exploiting their price-setting power unless they were forced to by competitive pressures. The TI hypothesis was taken very seriously by health economists and policy makers. The federal government sponsored a conference and published a volume on the supply and pricing issue, *The Target income Hypothesis and Related Issues in Health Manpower Policy* [Department of Health, Education and Welfare (1980)].

TI behavior, when used to explain physician response to competition from more physicians was not connected to demand inducement (as later it came to be). TI behavior was not about quantity setting, it was about pricing, an explanation for unexploited monopoly power.[82] Indeed, if physicians could induce demand, they could have done more for the fewer patients competition left them, and not raised prices.

In the 1980's, when direct fee-setting replaced increased supply as the mechanism used by regulators to limit physician prices, TI was used to explain another empirical anomaly, the negative correlation between fees physicians were paid and the quantity supplied [Rice (1983)]. If physicians wanted a target income and income was $P \times Q$, if you reduced P, then they would increase Q. During the 1980s, writers proposing TI explanations linked it to PID. Physicians could set quantity because they could induce demand. Interestingly, TI behavior was no longer a form of benevolence. Mr Hyde took over from Dr Jeckyl. TI frustrated policies designed to contain health care costs. Physicians were using their power to influence patient utilization for their own (the physicians') interests in order to counter the well-intentioned regulation of high fees.

6.3.2. Target income or income effects?

In the context of fees and demand inducement, simple TI theory is generally presented as a behavioral, i.e., a nonmaximizing, theory. In the terminology used in this chapter, physician net income is: $(p_s - c)x$. Call $p_s - c$ the margin, m, on services. If the physician chooses x by inducing demand so as to hit a target, T, then

$$x = T/m. \tag{6.1}$$

[82] An exception in Feldstein's (1970) discussion of the reasons why he found unconventional relations among price and quantity when attempting to estimate demand and supply curves for physicians.

It is obvious that $dx/dm < 0$; indeed the implication of (6.1) is that the elasticity of x with respect to the margin is -1.

There are two severe problems with this theory [McGuire and Pauly (1991)]. First, the idea of a "target" is problematic. It is difficult to explain why physicians would pursue a target income in the first place, difficult to explain how targets are set, and difficult to explain the evident differences in targets across individuals and over time. No one has established that the distribution of incomes cross-sectionally and over time among physicians are any different from what would be expected with conventional maximizing behavior.

Second, the formulation of target income theory in (6.1) is simply conceptually inadequate for explaining physician behavior in typical US markets in which physicians supply services to many payers. Suppose a physician has a target T, but supplies services to payer 1 and payer 2, and we designate the margins and quantities for each by subscripts. Then, the combinations of x_1 and x_2 that satisfy the target are:

$$T = m_1x_1 + m_2x_2. \tag{6.2}$$

This equation does not yield a unique solution for x_1 and x_2. There are an infinite number of combinations of quantities that satisfy a target for any pair of margins. The TI theory based on (6.1) does not generalize: it is incapable of generating quantity predictions for more than one payer. There is no unique solution, and no testable comparative statics from (6.2) regarding the effect of fees on quantities in a context like the US. The behavioral TI theory does not tell the physician how to choose among the many combinations of x_1 and x_2 (and more generally the large set of services she supplies) to hit a target.

McGuire and Pauly (1991) propose a utility maximizing framework in which a physician can set quantities for multiple payers through demand inducement. They show that target income behavior and profit (or income) maximization lie at opposite ends of a spectrum of income effects. When income effects are all that matter around a certain point, physicians act so as to pursue a target. When income effects are absent, physicians maximize income.[83] Furthermore, this framework can be used to generate comparative statics. When one price falls, income generation will be pursued differently with respect to all services supplied, along the lines of the income and substitution effect discussed above in Section 5.3.

It is unnecessary to view TI as an "alternative" to profit maximization. The idea of a "reference" income goes back to Feldstein (1970) who discussed the impact of a

[83] Income effects on behavior are generated by changes in the marginal utility of income. Targeting behavior emerges when these changes are very drastic, that is, when the derivative of the marginal utility of income with respect to income is minus infinity. This accords with the sense of a TI. When income is less than the target, its marginal utility is very high, when income is above the target its marginal utility is very low. Thus, around the target, the marginal utility must fall steeply. And as is well-known, when there are no income effects, firm behavior is net income maximizing.

proposed price ceiling imposed by regulation on physician fees in terms of income and substitution effects. Evidence from the literature can be assessed from the point of view of what it says about the magnitude of the income effects, not with regard to a yes/no issue of TI behavior.

Taking this perspective, the recent demand inducement literature [Gruber and Owings (1996), Yip (1998), Tai-Seale et al. (1998)] provides evidence for and explicit discussion of an income effect. The early fee and availability effect literature was debated in terms of a target, not an income effect [see Wedig et al. (1989), Rice and Labelle (1989), Feldman and Sloan (1988)]. Income effects on physicians could be estimated in the same way as in labor economics, with information on non-employment income, such as assets or a spouses' income. Sloan (1975) and Hurdle and Pope (1989) studied physician supply decisions and the effect of non-practice income, both studies finding evidence for small income effects. Rochaix (1993) found no effect of outside income on supply.

Rizzo and Blumenthal (1996) analyze a survey of young physicians in which questions were asked that could shed light on physician motivations. The young doctors were asked what income they considered to be "adequate," considering the stage they were at in their career. Rizzo and Blumenthal treated this reported income as a target, and found that when physicians were away from this target, they pushed prices higher, tending to move in the direction of the "target." This paper recalls the earlier spirit of the TI literature in which physicians exercise restraint in pricing (not pushing as far as they might) as their income reaches the "adequate" range.[84]

6.3.3. Revenue targeting from a participation constraint

"Targeting" behavior can stem from another source that does not require an assumption that physicians pursue a "target" income. The targeting discussed now can also explain targeting within a single payer. Suppose, following the presentation in Section 3.3 above that a physician's cost depends on some activities which are reimbursed in a payment system and some which are not directly reimbursed. In the notation introduced above, cost is $C(x, e)$ where x is paid upon and e is not. The revenue function can be expressed as $R(x)$. The physician must make a decision about whether to accept a certain patient, or patients from among a class, perhaps defined by a payer. To accept a patient, a *participation constraint*, as it is referred to in the contracting literature, must be satisfied. We can normalize the required profit to be 0, and express the participation constraint as:

$$R(x) - C(x, e) \geqslant 0. \tag{6.3}$$

[84] Rizzo and Zeckhauser (1997) reexamine the same data and find that these young doctors increase their hourly earnings more the farther they are away from the target on the downside, but above the target, the distance from the target does not matter. Although the findings are cast in these papers in terms of TI behavior, evidently this is not literally correct since doctors are not at the target, but making tradeoffs to get closer, behavior that can be understood within the more conventional approach of income effects.

Note that this constraint (6.3) has elements of a "target." A physician must gain a certain profit per patient to justify taking on the patient. This follows simply from recognizing that a physician has a certain opportunity cost of time. The (unreimbursed) effort that goes into caring for this patient could be spent elsewhere.

Suppose there is a class of patients (e.g., Medicaid birth-related clients) seen by a physician for whom the fees are reduced by regulation, violating the physician's participation constraint at the old values of e and x. The physician has several choices. She can drop the patients and refuse to participate in Medicaid. The physician can "upcode," labeling the procedures she does in a more elaborate fashion, perhaps risking censure or penalty. She can cut back on the time she spends per visit (reduce e). If some prices are unregulated but associated with the use of this patient, these prices can be raised to satisfy a participation constraint. Finally, she can generate more billing by increasing the quantity of reimbursed services supplied (increase x, if $R' > C_x$). Gabel and Rice (1985) refer this set of physician responses as the "price of paying less." Medicaid and Medicare experience less physician participation as fees are reduced. Physicians upcode in response to fee regulation [Berry et al. (1980), Yip (1998)]. Danzon et al. (1984) found evidence directly consistent with the operation of participation constraint: when physician fees for an office visit were limited, physicians compensated by raising fees on the associated lab tests to retain the overall net revenue associated with a visit.

Quantity response at the episode or payer level can be understood as one of the set of things a physician can do to satisfy a participation constraint. Rice's (1983) empirical work on Medicare, for example, is essentially an episode-level analysis. The fee-effect on the participation constraint expressed here is another explanation for the observed P-Q relation.

The participation constraint route to targeting avoids an implausible assumption about motivation. It also avoids the logical gap in TI theory in terms of multiple payers. A participation constraint applies to each payer. Therefore, a "target" behavior observed for one small payer can be explained. Note that unexploited income generation must be available to a physician for any of this set of responses to emerge (except for dropping the patient). Fraud (upcoding), price, or quantity-setting, must be limited by other forces, such as disutility, as in the Evans (1974) or McGuire and Pauly (1991) framework.

7. Conclusion

In the neoclassical theory of the firm, the firm sets price and quantity in order to maximize profit subject to the constraint of market demand. Every phrase in the paradigm has been questioned in the course of this chapter. Do physicians maximize profit? There is abundant evidence that in some circumstances physicians are prepared to trade off income against welfare of the patient. Furthermore, this tradeoff is affected by income effects, in a manner consistent with conventional views of labor supply.

There is not enough evidence, however, to justify keeping the "target income" theory alive. The theory is logically incomplete in a multiple payer environment; there is no evidence to support this extreme version of income effects; there is a theoretically superior way to generate target-like behavior even within a single payer by invoking a participation constraint.

Following from the view that physicians' tradeoff of other (patient-centered) objectives with income depends on their level of income, the weight on income in physicians' utility may be changing. As managed care plans succeed in limiting the prices charged and quantities supplied, physician income objectives may become paramount. The profit-maximization assumption may be becoming more applicable to physician behavior.

Are physicians constrained by market demand? The answer to this is "yes," even while noting that there are several mechanisms physicians have to influence quantity provided. The understanding of "market demand" must first of all extend beyond the conventional demand curve. In general, even in the most simple models of physicians with some market power, the demand curve, relying as it does on price-taking patients, does not describe prices and quantities in this market. While the "demand curve" has limited use, market demand is still an essential concept. If physicians set quantity only by virtue of the nonretradability of their services, patient benefits (market demand) constrains this activity. If physicians move demand by undertaking nonreimbursed activities perceived as "quality" by patients, demand considerations, in particular how the physician attracts patients, remain relevant. If physicians "induce" demand through a persuasive activity when patients have less information than the physician, market demand response can still limit what even the most self-interested physician can get away with.

A large body of credible research establishes that physicians set quantities, and they do so partly in response to self-interest. An important question for research is to decompose the source of the quantity setting. The welfare economics of quantity setting due to nonretradability within a fixed demand, observable quality or effort, or unobserved persuasive activity are very different. Interpreting the consequences of quantity-setting for purposes of policy depends on assessing the relative strength of the three potential sources.

Do physicians even set price and quantity? Prices for "visits" are easily observable and contractible, and within the grip of third-party payers. Physicians are subject to market forces like other workers, so the prices chosen by health plans are probably best regarded as being determined by demand and supply. Quantities are another matter. The difficulty of third parties' contracting on outcomes (even if the patient observes a signal related to outcomes) means that physicians are certain to remain with discretion about quantities, even when measured in simple terms like "visits" of various types. Economic models, abstracting the complexity of medical decisions into a single dimension of "quantity," give the impression that treatment decisions are more easily monitored and controlled than they really are. Acknowledging the many elements composing treatment – the many inputs, the sequence of events, the observable and the behind-the-scenes activities – leads inevitably to the conclusion that the simple monitoring and incentives devices used by payers leave a great deal of authority about treatment with the physician.

References

Anderson, G., and V. Glesnes-Anderson (1987), Health Care Ethics (Aspen Publication, Rockville, MD).

Arrow, K.J. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53:941–973.

Arrow, K.J. (1971), Essays in the Theory of Risk Bearing (Markham, Chicago).

Arrow, K.J. (1986), "Agency and the market", in: K.J. Arrow and M.D. Intriligator, eds., Handbook of Mathematical Economics, Vol. 3 (Elsevier Science Publishers, North Holland) 1183–1195.

Auster, R.D., and R.L. Oaxaca (1981) "Identification of supplier induced demand in the health care sector", Journal of Human Resources 16:327–342.

Benham, L., A. Maurizi and M. Reder (1968), "Location and migration medics: physicians and dentists", Review of Economics and Statistics 50:332–347.

Berry, C., P.J. Held, B. Kehrer, L. Markheim and U. Reinhardt (1980), "Canadian physicians' supply response to universal health insurance: The first years in Quebec", in: J. Gabel, J. Taylor, N. Greenspan and M. Blaxall, eds., Physicians and Financial Incentives (US GPO, Washington, DC) 57–59.

Birch, S. (1988), "The identification of supplier-inducement in a fixed price system of health care provision: The case of dentistry in the United Kingdom", Journal of Health Economics 7:129–150.

Blaug, M. (1998), "Where are we now in British health economics?" Health Economics 7:563–579.

Blendon, R.J., et al. (1993), "Health reform lessons learned from physicians in three nations", Health Affairs (Fall):194–203.

Blomqvist, A. (1991), "The doctor as double agent: information asymmetry, health insurance, and medical care", Journal of Health Economics 10:411–432.

Brown, D.M. (1988), "Do physicians underutilize aides?" Journal of Human Resources 23:342–355.

Brundin, I., and C.A. Ma (1998), "Moral hazard, insurance, and some collusion", Boston University Industry Studies Program Working Paper #89.

Bunker, J.P., and B.W. Brown Jr. (1974), "The physician–patient as an informed consumer of surgical services", New England Journal of Medicine 290:1051–1055.

Burstein, P.L., and J. Cromwell (1985), "Relative incomes and rates of return for US physicians", Journal of Health Economics 4:63–78.

Carlsen, F., and J. Grytten (1998), "More physicians: improved availability or induced demand?", Health Economics 7:495–508.

Chassin, M., et al. (1987), "Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures", Journal of the American Medical Association 258(18):2533–2537.

Chalkley, M., and J.M. Malcolmson (1998), "Contracting for health services when patient demand does not reflect quality", Journal of Health Economics 17(1):1–20.

Comanor, W.S. (1980), National Health Insurance in Ontario: The Effects of a Policy of Cost Control (American Enterprise Institute, Washington, DC).

Cromwell, J. (1996), "Health professions substitution: a case study of anesthesia", in: M. Osterwies, C.J. McLaughlin, H.R. Manasse Jr. and C.L. Hopper, eds., The US Health Workforce: Power Politics, and Policy (Association of Academic Health Centers, Washington, DC).

Cromwell, J., and J.B. Mitchell (1986), "Physician-induced demand for surgery", Journal of Health Economics 293–313.

Culyer, A.J. (1989), "The normative economics of health care finance and provision", Oxford Review of Economic Policy 5:34–58.

Danzon, P.M. (2000), "Liability for medical malpractice", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 26.

Danzon, P.M., W.G. Manning and M.S. Marquis (1984), "Factors influencing laboratory tests and prices", Health Care Financing Review 5:23–32.

DeFelice, L.C., and W.D. Bradford (1997), "Relative inefficiencies in production between solo and group practice physicians", Health Economics 6:455–466.

Department of Health, Education and Welfare (1980), The Target-Income Hypothesis and Related Issues in Health Manpower Policy (Bureau of Health Manpower, DHEW (HRA), Washington, DC) 80–127.

Dranove, D. (1988), "Demand inducement and the physician/patient relationship", Economic Inquiry 26:251–298.

Dranove, D., and M.A. Satterthwaite (1991), "The implications for resource-based relative value scales for physicians' fees, income and specialty choices", in: H.E. French III, ed., Regulating Doctors' Fees: Competition, Benefits and Controls under Medicare (AEI Press, Washington, DC) 52–70.

Dranove, D., and M.A. Satterthwaite (1992), "Monopolistic competition when price and quality are imperfectly observable", The Rand Journal of Economics 23(4):518–534.

Dranove, D., and M.A. Satterthwaite (1999), "The industrial organization of health care markets", Chapter 20, this Handbook.

Dranove, D., M. Shanley and W.D. White (1993), "Price and concentration in hospital markets: the switch from patient to payer driven competition", Journal of Law and Economics 36:179–204.

Dranove, D., and P. Wehner (1994), "Physician-induced demand for childbirths", Journal of Health Economics 13:61–73.

Dranove, D., and W.D. White (1987), "Agency and the organization of health care delivery", Inquiry 24:405–415.

Dranove, D., and W.D. White (1996), "Specialization, option demand, and the pricing of medical specialists", Journal of Economics and Management Strategy 5:277–306.

Dyckman, Z. (1978), "Physicians: a study of physicians' fees", Staff Report, Council of Wage and Price Stability (US Government Printing Office, Washington, DC).

Eastaugh, S. (1992), Health Economics: Efficiency, Quality and Equity (Auburn House, Westport, CT).

Eisenberg, J.M. (1986), Doctors' Decisions and the Cost of Medical Care (Health Administration Press, Ann Arbor, MI).

Eisenberg, J.M. (1994), "Economics: physicians income and set-fee structures", The Journal of American Medical Association 271:1663–1666.

Ellis, R.P., and T.G. McGuire (1986), "Provider behavior under prospective reimbursement", Journal of Health Economics 5:129–151.

Ellis, R.P., and T.G. McGuire (1990), "Optimal payment systems for health services", Journal of Health Economics 9:375–396.

Ellis, R.P., and T.G. McGuire (1993), "Supply-side and demand-side cost sharing in health care", Journal of Economic Perspectives 7:135–151.

Emanuel, E.J., and L. Goldman (1998), "Protecting patient welfare in managerial care: six safeguards", Journal of Health Politics, Policy and Law 23:635–659.

Emons, W. (1997), "Credence goods and fraudulent experts", Rand Journal of Economics 28:107–119.

Emons, D.W., and G.D. Wozniak (1997), "Physicians' contractual arrangements with managed care organizations", in: Socioeconomics of Medical Practice (American Medical Association, Chicago, IL).

Epstein, A.M., et al. (1986), "The use of ambulatory testing in prepaid and fee-for-service group practices: relation to perceived profitability", New England Journal of Medicine 314:1089–1093.

Escarce, J. (1993a), "Effects of lower surgical fees on the use of physician services under medicare", Journal of the American Medical Association 269(19):2513–2518.

Escarce, J. (1993b), "Medicare patients' use of overpriced procedures before and after the Omnibus Budget Reconciliation Act of 1987", American Journal of Public Health 83(3):349–355.

Escarce, J. (1993c), "Effects of the relative fee structure on the use of surgical operations", Health Services Research 28:479–502.

Escarce, J., and M.V. Pauly (1998), "Physician opportunity costs in physician practice cost functions", Journal of Health Economics 17:129–151.

Escarce, J., D. Polsky, G. Wozniak, M. Pauly and P. Kletke (1998), "Health maintenance organization penetration and the practice location choices of new physicians", Medical Care 36:1555–1566.

Evans, R. (1974), "Supplier-induced demand: some empirical evidence and implications", in: M. Perlman, ed., The Economics of Health and Medical Care (Macmillan, London) 162–173.

Farley, P.J. (1986), "Theories of the price and quantity of physician services", Journal of Health Economics 315–333.

Farrell, J., and S. Scotchmer (1988), "Partnerships", Quarterly Journal of Economics 103:279–298.

Feldman, R., and J. Begun (1978), "The effects of advertising restrictions: lessons from optometry", Journal of Human Resources 13(Suppl.):247–262.

Feldman, R., and B. Dowd (1991), "A new estimate of the welfare loss of the health insurance", American Economic Review 81:297–301.

Feldman, R., and F. Sloan (1988), "Competition among physicians, revisited", Journal of Health Politics, Policy and Law 13:239–261.

Feldstein, M. (1970), "The rising price of physicians' services", Review of Economics and Statistics 52(2):121–133.

Feldstein, M. (1971), "Hospital cost inflation: a study of non-profit price dynamics", American Economic Review 61:853–872.

Feldstein, P. (1979), Health Care Economics (John Wiley and Sons, New York).

Folland, S., A. Goodman and M. Stano (1997), The Economics of Health and Health Care (Prentice Hall, Upper, Saddle River, NJ).

Frank, R.G. (1985), "Pricing and location of physician services in mental health", Inquiry 38:115–133.

Frank, R.G., and D. Salkever (1991), "The supply of charity services by non-profit hospitals: motives and market structures", Rand Journal of Economics 22:430–445.

Frank, R.G., J.P. Weiner, D.M. Steinwachs and D.S. Salkever (1987), "Economic rents derived from hospital privileges in the market for podiatric services", Journal of Health Economics 6:319–337.

Frech III, H.E. (1974), "Occupational licensure and health care productivity", in: J. Rafferty, ed., Health Manpower and Productivity (Lexington Books, Lexington, MA).

Frech III, H.E. (1996), Competition and Monopoly in Medical Care (AEI Press, Washington, DC).

Frech III, H.E., and K.L. Danger (1998), "Exclusive contracts between hospitals and physicians: the antitrust issues", Health Economics 7:175–178.

Frech III, H.E., and P.B. Ginsburg (1975), "Imposed health insurance in monopolistic markets: a theoretical analysis", Economic Inquiry 13:55–70.

Friedman, M., and S. Kuznets (1954), Income from Independent Professional Practice (National Bureau of Economic Research, Basic Books, New York) 45.

Fuchs, V.R. (1974), Who Shall Live? (Basic Books, New York).

Fuchs, V.R. (1978), "The supply of surgeons and the demand for operations", The Journal of Human Resources 35–56.

Fuchs, V.R. (1996), "Economics, values, and health care reform", American Economic Review 86:1–24.

Gabel, J.R., and T.H. Rice (1985), "Reducing expenditure for physicians services: the price of paying less", Journal of Health Politics, Policy and Law 9:595–609.

Ganem, J., J. Krakower and R. Beran (1995), "Review of US medical school finances, 1993–1994", Journal of the American Medical Association 274(9):723–730.

Garg, M.L., W.A. Gliebe and M.B. Elkhatib (1978), "The extent of defensive medicine: some empirical evidence", Leg. Aspects Med. Practice 6:25–29.

Gaynor, M. (1994), "Issues in the industrial organization of the market for physician services", The Journal of Economics & Management Strategy 211–255.

Gaynor, M., and P. Gertler (1995), "Moral hazard and risk spreading in partnerships", The RAND Journal of Economics 26:591–614.

Getzen, T.E. (1984), "A 'brand name firm' theory of medical group practice", Journal of Industrial Economics 33:199–215.

Getzen, T.E. (1997), Health Economics: Fundamentals and Flows of Funds (John Wiley, New York).

Giuffrida, A., and H. Gravelle (1998), "Paying patients to comply: an economic analysis", Health Economics 7:569–579.

Glazer, J., and T.G. McGuire (1993), "Should physicians be permitted to 'balance bill' patients?", Journal of Health Economics 12:239–258.

Green, J. (1978), "Physician-induced demand for medical care", The Journal of Human Resources 13:21–33.

Greenberg, W. (1998), "Marshfield clinic, physician networks, and the exercise of monopoly power", Health Services Research 33(Part II):1461–1476.

Greenfield, S., E.C. Nelson, M. Zubkoff, W. Manning, W. Rogers, R. Kravitz, A. Keller, A. Tarlov and J. Ware (1992), "Variations in resource utilization among medical specialties and systems of care: results of the medical outcome study", Journal of the American Medical Association 267:1624–1630.

Grossman, M. (2000), "The human capital model", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 7.

Gruber, J., and M. Owings (1996), "Physician financial incentives and Cesarean section delivery", RAND Journal of Economics 27:99–123.

Grytten, J., D. Holst and P. Laake (1990), "Supplier inducement: its effect on dental services in Norway", Journal of Health Economics 9:483–491.

Haas-Wilson, D. (1986), "The effect of commercial practice restrictions: the case of optometry", Journal of Law and Economics 29:165–185.

Haas-Wilson, D. (1990), "Consumer information and providers' reputations: an empirical test in the market for psychotherapy", Journal of Health Economics 9:321–333.

Haas-Wilson, D., and M. Gaynor (1998), "Physician networks and their implications for competition in health care markets", Health Economics 7:179–182.

Hadley, J., and R. Lee (1978), "Toward a physician payment policy: evidence from the economic stabilization program", Policy Sciences 10:105–120.

Harris, J. (1977), "The internal organization of hospitals: some economic implications", Bell Journal of Economics 8:467–482.

Havighurst, C.C. (1978), "Professional restraints on innovation in health care financing", Duke Law Journal 2:303–387.

Hay, J., and M.J. Leahy (1982), "Physician-induced demand", Journal of Health Economics 2:231–244.

Hellinger, F.J. (1996), "The impact of financial incentives on physician behavior in managed care plans: a review of the evidence", Medical Care Research and Review 53:294–314.

Hemenway, D., et al. (1990), "Physician responses to financial incentives: evidence from a for-profit ambulatory care center", The New England Journal of Medicine 322:1059–1063.

Hickson, G.B., et al. (1987), "Physician reimbursement by salary or fee-for-service: effect on a physician's practice behavior in a randomized prospective study", Pediatrics 80:744–750.

Hiller, M.D. (1987), "Ethical decision making and the health administrator", in: Anderson and Glesnes-Anderson, eds., Health Care Ethics (Aspen Publication, Rockville, MD).

Hillman, A. (1990), "Health maintenance organizations, financial incentives and physician judgements (editorial)", Annals of Internal Medicine 112:891–893.

Hillman, A., et al. (1990), "Frequency and costs of diagnostic imaging in office practice – a comparison of self-referring and radiologist-referring physicians", New England Journal of Medicine 323:1604–1608.

Hillman, A., et al. (1992), "Contractual arrangements between HMOs and primary care physicians: three-tiered HMOs and risk pools", Medical Care 30:136–148.

Hillman, A., et al. (1992), "Physicians' utilization and charges for outpatient diagnostic imaging in a Medicare population", Journal of the American Medical Association 268:2050–2054.

Hirsh, H.L., and T.S. Dickey (1983), "Defensive medicine as a basis for malpractice liability", Trans. Stud. Coll. Physicians Phila. 5:98–107.

Hoerger, T.J., and L.Z. Howard (1995), "Search behavior and choice of physician in the market for prenatal care", Medical Care 33:332–349.

Holahan, J., and W. Scanlon (1978), Price Controls, Physician Fees, and Physician Incomes from Medicare Medicaid (The Urban Institute, Washington, DC).

Hsiao, W.C., et al. (1988a), "Estimating physicians' work for a resource-based relative-value scale", New England Journal of Medicine 319(13):835–841.

Hsiao, W.C., et al. (1988b), "Results and policy implications of the resource-based relative-value scale", New England Journal of Medicine 319(13):881–888.

Hurdle, S., and G. Pope (1989), “Physician productivity: trends and determinants”, Inquiry 26:100–115.

Hurley, J., and R. Labelle (1995), “Relative fees and the utilization of physicians’ services in Canada”, Health Economics 4:419–438.

Hurley, J., R. Labelle and T. Rice (1990), “The relationship between physician fees and the utilization of medical services in Ontario”, Advances in Health Economics and Health Services Research 11:49–78.

Jennison, K., and R.P. Ellis (1987), “Comparison of psychiatric service utilization in a single group practice”, in: McGuire and Scheffler, eds., The Economics of Mental Health Services: Advances in Health Economics and Health Services Research, Vol. 8 (JAI Press, Greenwich, CT).

Keeler, E., and M. Brodie (1993), “Economic incentives and the choice between vaginal delivery and Cesarean section”, Milbank Quarterly 71:365–404.

Kessel, R. (1958), “Price discrimination in medicine”, Journal of Law and Economics 1:20–53.

Kessler, D., and M. McClellan (1996), “Do doctors practice defensive medicine?” Quarterly Journal of Economics 111:353–390.

Klevorick, A.K., and T.G. McGuire (1987), “Monopolistic competition and consumer information: pricing in the market for psychologists’ services”, Advances in Health Economics and Health Services Research 8:235–253.

Kuhn, T. (1970), The Structure of Scientific Revolutions, 2nd edn. (Univ. of Chicago Press, Chicago).

Laffont, J.J., and J. Tirole (1993), A Theory of Incentives in Procurement and Regulation (MIT Press, Cambridge, MA).

Lawthers, A., et al. (1992), “Physicians’ perceptions of the risk of being sued”, Journal of Health Politics, Policy and Law 17(3):463–482.

Lee, A.J., S. Gehlbach, D. Hosmer, M. Reti and C. Baker (1997), “Medical treatment differences for blacks and whites”, Medical Care 35:1173–1189.

Leffler, K.B. (1978), “Physician licensure: competition and monopoly in American licensure”, Journal of Law and Economics 21:165–188.

Levy, J.M., et al. (1990), “Impact of the medicare fees schedule on payment to physicians”, Journal of the American Medical Association 264:717–722.

Light, D.W. (1997), “From Managed competition to managed cooperation: theory and lessons from the British experience”, The Milbank Quarterly 75:297–341.

Localio, A.R., et al. (1991), “Relation between malpractice claims and adverse events due to negligence: results of the Harvard medical practice study III”, New England Journal of Medicine 7:370–376.

Lu, M. (1999), “Separating the ‘true effect’ from ‘gaming’ in incentive-based contracts in health care”, Journal of Economics and Management Strategy 8:383–432.

Lundback, M. (1998), “Quality and efficiency in health care production”, CEFOS Report 10 (Center for Public Sector Research, Goteborg University, Sweden).

Lusted, L. (1977), “A study of the efficacy of diagnostic radiologic procedures: final report to the National Center for Health Services Research” (National Center for Health Services Research, Rockville, MD).

Ma, C.A. (1994), “Health care payment systems: cost and quality incentives”, Journal of Economics & Management Strategy 3(1):93–112.

Ma, C.A. (1997), “Option contracts and vertical foreclosure”, Journal of Economics and Management Strategy 6:725–753.

Ma, C.A., and T.G. McGuire (1997), “Optimal health insurance and provider payment”, American Economic Review 87(4):685–704.

Ma, C.A., and T.G. McGuire (1998), “Network effects in managed health care”, unpublished (Boston University).

Ma, C.A., and M. Riordan (1998), “Health insurance, moral hazard, and managed care”, unpublished (Boston University).

Manning, W.G., et al. (1987), “Health insurance and the demand for medical care: evidence from a randomized experiment”, American Economic Review 77:251–277.

Marder, W.D., et al. (1988), “Physician supply and utilization by specialty: trends and projections” (American Medical Association, Chicago).

McCall, T.B. (1996), Examining Your Doctor (Citadel Press, Seacaucus, NJ).

McCarthy, T.R. (1985), "The competitive nature of the primary-care physician services market", Journal of Health Economics 4:93–117.

McGuire, T.G. (1983), "Patient's trust and the quality of physicians", Economic Inquiry 21:203–222.

McGuire, T.G., and M.V. Pauly (1991), "Physician response to fee changes with multiple payers", Journal of Health Economics 385–410.

McKinlay, J.B., D.A. Potter and H.A. Feldman (1996), "Non-medical influences on medical decision-making", Social Science and Medicine 42:769–776.

McLean, R.A. (1980), "The structure of the market for physicians' services", Health Services Research 15:271–280.

Mechanic, D. (1990), "The role of sociology in health affairs", Health Affairs 9:85–87.

Mechanic, D. (1998), "The functions and limitations of trust in the provision of medical care", Journal of Health Politics, Policy and Law 23(4):661–686.

Mechanic, D., and M. Schlesinger (1996), "The impact of managed care on patient's trust in medical care and their physicians", Journal of the American Medical Association 275:1693–1697.

Medicare Payment Advisory Commission (1998), Health Care Spending and the Medicare Program: a Data Book (The Commissian, Washington, DC).

Mitchell, J.B., and J. Cromwell (1982), "Physician behavior under the Medicare assignment option", Journal of Health Economics 2:245–264.

Mitchell, J.M., and T.R. Sass (1995), "Physician ownership of ancillary services: indirect demand inducement or quality assurance?" Journal of Health Economics 14:263–289.

Mooney, G., and M. Ryan (1993), "Agency in health care: getting beyond first principles", Journal of Health Economics 12:125–135.

Mort, E.A., et al. (1996), "Physician response to patient insurance status in ambulatory care clinical decisions-making", Medical Care 34:783–797.

Moy, E., B. Bartman, C. Clancy and L. Cornelius (1998), "Changes in usual sources of medical care between 1987 and 1992", Journal of Health Care for the Poor and Underserved 9:126–138.

Newhouse, J.P. (1970), "Toward a theory of nonprofit institutions: an economic model of a hospital", American Economic Review 60:64–74.

Newhouse, J.P. (1978), The Economics of Medical Care: A Policy Perspective (Addison Wesley, Reading, MA).

Newhouse, J.P. (1996), "Reimbursing health plans and health providers: efficiency in production versus selection", Journal of Economic Literature 34:1236–1263.

Newhouse, J.P., and the Health Insurance Experiment Group (1993), Free For All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge).

Newhouse, J.P., et al. (1982), "Does the geographical distribution of physicians reflect market failure?", Bell Journal of Economics 13:493–505.

Newhouse, J.P., and F. Sloan (1972), "Physician pricing: monopolistic or competitive: reply", Southern Economic Journal 38:577–580.

Nguyen, N.X., and F.W. Derrick (1997), "Physician behavioral response to a Medicare price reduction", Health Services Research 32:283–298.

Noether, M. (1986), "The effect of government policy changes on the supply of physicians: expansion of a competitive fringe", Journal of Law and Economics 29:231–262.

Noether, M. (1986), "The growing supply of physicians: has the market become more competitive?", Journal of Labor Economics 4:503–537.

Ohsfeldt, R., M. Morrisey, L. Nelson and V. Johnson (1998), "The spread of state any-willing-provider laws", Health Services Research 33(Part II):1537–1555.

Oppenheim, G.L., J.J. Berman and E.C. English (1979), "Failed appointments: a review", The Journal of Family Practice 8:789–796.

Pauly, M.V. (1968), "The economics of moral hazard", American Economic Review 49:531–537.

Pauly, M.V. (1978), "Is medical care different?" in: Greenberg, ed., Competition in the Health Care. Sector: Past, Present and Future (Aspen Systems Corporation).

Pauly, M.V. (1978), "Medical staff characteristics and hospital costs", Journal of Human Resources 13:77–114.

Pauly, M.V. (1979), "The ethics and economics of kickbacks and fee-splitting", Bell Journal of Economics 10:344–352.

Pauly, M.V. (1980), Doctors and Their Workshops: Economic Models of Physician Behavior (University of Chicago Press, Chicago).

Pauly, M.V. (1988), "Is medical care different? Old questions, new answers", Journal of Health Politics, Policy and Law 13:227–237.

Pauly, M.V. (1991), "Fee schedules and utilization", in: H.E. Frech III, ed., Regulating Doctors' Fees: Competition, Controls, and Benefits under Medicare, 288–305.

Pauly, M.V. (2000), "Insurance reimbursement", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 10.

Pauly, M.V., and M. Redisch (1973), "The not-for-profit hospital as a physicians' cooperative", American Economic Review 63:87–99.

Pauly, M.V., and M.A. Satterthwaite (1981), "The pricing of primary care physicians' services: a test of the role of consumer information", Bell Journal of Economics 12:488–506.

Pescosolido, B.A. (1992), "Beyond rational choice: the social dynamics of how people seek help", American Journal of Sociology 97(4):1096–1138.

Phelps, C.E. (1986), "Induced demand: can we ever know its extent?" Journal of Health Economics 5:355–365.

Phelps, C.E. (1997), Health Economics, 2nd edn. (Harper Collins, New York).

Physician Payment Review Commission (1991), Annual Report to Congress (The Commission, Washington, DC).

Physician Payment Review Commission (1992), "Practice expenses under the Medicare fee schedule: a resource-based approach", Technical Report 92-1, PPRC (Washington, DC).

Pope, G.C., and R.T. Burge (1992), "Inefficiencies in physician practices", Advances in Health Economics and Health Services 13:129–164.

Pratt, J. (1964), "Risk aversion in the small and in the large", Econometrica 32:122–136.

Reinhardt, U. (1972), "A production function for physician services", Review of Economics and Statistics 54:55–65.

Reinhardt, U. (1975), Physician Productivity and the Demand for Health Manpower (Ballinger Publishing, Cambridge, MA).

Reinhardt, U. (1978), "Comment on monopolistic elements in the market for physician services", in: Greenberg, ed., Competition in the Health Care Sector (Aspen Publications) 121–148.

Reinhardt, U.E. (1989), "Economists in health care: saviors or elephants in a porcelain shop?", American Economic Review 79:337–342.

Reinhardt, U. (1996), "The economic and moral case for letting the market determine the health workforce", in: M. Osterweis, C.J. McLaughlin, H.R. Manasse Jr. and C.L. Hopper, eds., The U.S. Health Workforce: Power Politics, and Policy (Association of Academic Health Centers, Washington, DC).

Remler, D.K., et al. (1997), "What do managed care plans do to affect care? Result from a survey of physicians", Inquiry 34:196–204.

Rice, T. (1983), "The impact of changing Medicare reimbursement rates on physician-induced demand", Medical Care 21:803–815.

Rice, T. (1998), The Economics of Health Reconsidered (Health Administration Press, Chicago, IL).

Rice, T., and R.J. LaBelle (1989), "Do physicians induce demand for medical services?" Journal of Health Politics, Policy and Law 14:239–261.

Rice, T., and N. McCall (1983), "Factors influencing physician assignment decisions under Medicare", Inquiry 20:45–56.

Rizzo, J.A., and D. Blumenthal (1996), "Is the target income hypothesis an economic heresy?", Medical Care Research and Review 243–293.

Rizzo, J., and R. Zeckhauser (1990), "Advertising and entry: the case of physician services", Journal of Political Economy 98:476–500.

Rizzo, J., and R. Zeckhauser (1997), "Income targets and physician behavior", unpublished.

Robinson, J.C. (1997), "Physician-hospital integration and the economic theory of the firm", Medical Care Research and Review 54:3–24.

Rochaix, L. (1989), "Information asymmetry and search in the market for physician services", Journal of Health Economics 8:53–84.

Rochaix, L. (1993), "Financial incentives for physicians: the Quebec experience", Health Economics 2:163–176.

Rodwin, M.A. (1993), Medicine, Money, and Morals: Physicians' Conflicts of Interests (Oxford University Press, New York).

Rogerson, W. (1994), "Choice of treatment intensity by a nonprofit hospital under prospective pricing", Journal of Economics and Management Strategy 351:7–51.

Rosenthal, M. (1998), "Treatment intensity under prospective payment for outpatient mental health care", unpublished (Harvard School of Public Health).

Rosenthal, M., R. Frank, A. Epstein and J. Buchanan (1999), "Doctors, dollars and delegation", unpublished (School of Public Health, Harvard University).

Rossiter, L.F., and G.R. Wilensky (1983), "A clarification of theories and evidence on supplier-induced demand for physicians' services", 611–627.

Rossiter, L.F., and G.R. Wilensky (1984), "Identification of physician-induced demand", Journal of Human Resources 19:231–244.

Roth, A. (1979), Axiomatic Models of Bargaining (Springer, Berlin).

Santerre, R., and S. Neun (1996), Health Economics: Theory, Insights and Industry Studies (Irwin, Chicago).

Satterthwaite, M. (1979), "Consumer information, equilibrium price and the number of sellers", Bell Journal of Economics 10:483–502.

Scherer, F.M. (2000), "The pharmaceutical industry", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 25.

Schroeder, S. (1992), "Physician supply and the US medical market place", Health Affairs 11:235–243.

Scott, A., and A. Shiell (1997), "Do fee descriptors influence treatment choices in general practice: a multi-level discrete choice model", Journal of Health Economics 16:323–342.

Scott, A., and A. Shiell (1997), "Analyzing the effect of competition on general practitioners' behavior using a multi level modeling framework", Health Economics 6:577–588.

Shen, Y. (1999), "Selection incentives in a performance-based contracting system", Boston University, unpublished.

Simon, H. (1958), "Theories of decision-making in economics and behavioral science", American Economic Review 49:253–283.

Simon, C., and P. Born (1996), "Physician earnings in a changing managed care environment", Health Affairs 15(3):124–133.

Sloan, F.A. (1970), "Lifetime earnings and physician's choice of specialty", Industrial and Labor Relations Review 24:47–56.

Sloan, F.A. (1975), "Physician supply behavior in the short run", Industrial and Labor Relations Review 29:549–569.

Sloan, F.A. (1976), "Physician fee inflation: evidence from the late 1960's", in: R. Rosett, ed., The Role of Health Insurance in the Health Services Sector (National Bureau of Economics Research) 321–354.

Sloan, F.A., et al. (1993), Suing for Medical Malpractice (University of Chicago Press, Chicago).

Smith, C.M., and B.P. Yawn (1994), "Factors associated with appointment keeping in a family practice residency clinic", The Journal of Family Practice 38:25–29.

Spence, M. (1975), "Monopoly, quality and regulation", Bell Journal of Economics 6:417–429.

Stano, M. (1987a), "A further analysis of the physician inducement controversy", Journal of Health Economics 6:229–238.

Stano, M. (1987b), "A clarification of theories and evidence on supplier-induced demand for physicians' services", Journal of Human Resources 22:611–620.

Starr, P. (1982), The Social Transformation of American Medicine (Basic Books, NY).

Stearns, S., B. Wolfe and D. Kindig (1992), "Physician responses to fee-for-service and capitation payment", Inquiry 29:416–425.

Steinwald, B., and F. Sloan (1974), "Determinants of physicians' fees", Journal of Business 47:493–511.

Tai-Seale, M., et al. (1998), "Volume responses to Medicare payment inductions with multiple payers: a test of the McGuire–Pauly model", Health Economics 7:199–219.

Van Doorslaer, E., and J. Geurts (1987), "Supplier-induced demand for physiotherapy in the Netherlands", Social Science and Medicine 24:919–925.

Varian, H. (1989), "Price discrimination", in: R. Schmalensee and R. Willig, eds., Handbook of Industrial Organization (North-Holland, Amsterdam) 598–654.

Wedig, G., J. Mitchell and J. Cromwell (1989), "Can optimal physician behavior be obtained using price controls?", Journal of Health Politics, Policy and Law 14:601–620.

Weeks, W.B., et al. (1994), "A comparison of the educational costs and incomes of physicians and other professionals", New England Journal of Medicine 330:1280–1286.

Weinstein, M., H. Fineberg, H. Elsten, H. Frazier, D. Neuhauser, R. Neutra and B. McNeil (1980), Clinical Decision Analysis (W.B. Saunders, Philadelphia).

Wennberg, J.E. (1985), "On patient need, equity, supplier-induced demand, and the need to assess the outcome of common medical practices", Medical Care 23:512–520.

Werman, B.G., S.V. Sostrin, Z. Pavlova and G.D. Lundberg (1980), "Why do physicians order laboratory tests? A study of laboratory test request and use patterns", Journal of American Medical Association 243:2080–2082.

Williams, A. (1998), "Medicine, economics, ethics and the NHS: a clash of cultures?", Health Economics 7:565–568.

Wolinsky, A. (1993), "Competition in a market for informed expert services", Rand Journal of Economics 24:380–398.

Wong, H.S. (1996), "Market structure and the role of consumer information in the physician services industry: an empirical test", Journal of Health Economics 15:139–160.

Woodward, R.S., M.A. Schnitzler and L.K. Kuols (1998), "Reduced uncertainty as a diagnostic benefit", Health Economics 7:149–160.

Woodward, R.S., and F. Warren-Bolton (1984), "Considering the effect of financial incentives and professional ethics on 'appropriate' medical care", Journal of Health Economics 3:223–237.

Yip, W. (1998), "Physician responses to medical fee reductions: changes in the volume and intensity of supply of Coronary, Artery Bypass Graft (CABG) surgeries in the medicare and private sectors", Journal of Health Economics 17:675–700.

Zuckerman, S., and J. Holahan (1991), "Medicare balance billing: its role in physician payment", in: H.E. Frech III, ed., Regulating Doctors' Fees: Competition, Controls, and Benefits under Medicare 143–169.

Zweifel, P., and F. Breyer (1997), Health Economics (Oxford University Press, New York).

Chapter 10

INSURANCE REIMBURSEMENT

MARK V. PAULY

Wharton School, University of Pensylvania

Contents

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This paper discusses theoretical and empirical findings concerning insurance reimbursement of patients or providers by insurers operating in private markets or in mixed public and private systems. Most insurances other than health insurance do not "reimburse"; instead they pay cash to insureds conditional on the occurrence of a prespecified event. In contrast, health insurance ties the payment to medical expenditures or costs incurred in some fashion, often making payments directly to medical providers. These differences are caused by a much higher degree of moral hazard and the dominant effect of insurer demand on provider prices. Health insurances also often prohibit "balance billing," provider charges in excess of some prespecified amount. Such prohibitions are related to patient inability to shop or bargain, and to insurer market power.

Empirical evidence suggests that some versions of physician and hospital reimbursement have increased the level of medical spending relative to the level that would be experienced under prospective payment. In particular, cost-based reimbursement raises total spending. Optimal reimbursement, with balance billing prohibited, may also be chosen to control moral hazard; payment will generally involve a mix of fee-for-service and predetermined (salary or capitation) payment, and may well involve positive patient cost sharing. Monopsony behavior by dominant insurers is possible, and may improve consumer welfare but not total welfare.

JEL classification: I11, G22

1. Introduction

In this essay I address two broad questions. The first question is that of the determinants of the form and level of reimbursement insurers do offer or should offer for covered medical services. I will deal with this question primarily in the context of medical services markets with multiple insurers, private and sometimes public, rather than the context of a single government insurer procuring services.[1] The second question is the more general one of the effects of various types or levels of reimbursement, whether optimal or optimally chosen or not, on total spending, unit prices, quantity, quality, and outcomes of care.

"Reimbursement" affects total spending in two ways, which may be described as "supply side" or "demand side". At one extreme, if a single insurance shields the insured entirely from out of pocket payment ("free care") and every provider experiences excess demand (no consumer choice of provider), the general principles of supply procurement apply, suitably modified for the quality variation and lifetime health aspects of medical care services. In this case, the degree of risk aversion of consumers (and even whether or not they are risk averse) is largely irrelevant. In this essay I will focus more on the other extreme, in which consumers may pay something out of pocket and/or have the option of selecting among a set of suppliers who will profit from additional business. In this case, the supply incentives must be specified as well, but it will be the combinations of the two sets of incentives that will be of primary interest.

Health services markets in the United States, with multiple private insurers and the public Medicare and Medicaid programs for subsets of the population, clearly exemplify the kind of setting I will be discussing; the bulk of the research in this area has also been concerned with this setting. However, such multiple-insurer arrangements do exist in other countries, even when there is some form of national health insurance, so the applicability is more general.

2. Reimbursement in the theory of insurance

In the general theory of insurance, there is virtually no role for something called "reimbursement". The classic insurance purchaser buys a contract for a price called a premium. That contract promises to pay a certain amount of money if a specified event occurs. For instance, if I buy $1 million of term life insurance, the contract specifies that certain individuals will be paid $1 million in the state of the world "I die". If I die, the result is only a benefit payment, not reimbursement of any previous outlay, although there is I hope a utility loss felt by my heirs. In this most simplified case, both key aspects of the insurance contract are easy to observe or determine afterwards: what state of the world has occurred, and what payment is to be made conditional on the occurrence of that state.

[1] For more on the later issue, see the paper by Chalkley and Malcomson (this volume).

In property-casualty insurance, the situation is somewhat more complex. The contract usually ties the amount to be paid to the amount of reduction in value of the pre-specified insured asset; fire insurance on my home will reimburse such a loss in wealth up to a maximum amount (after a deductible). If the damage is less than total, the payment is tied in some fashion to an estimate of the amount of damage. Usually the benefit equals an estimate of the cost or expenditure for repairing the damage, as long as that amount is less than or equal to the differences in value between the damaged and undamaged asset. It may be reasonable then to say that I am reimbursed for my loss, but there is usually no requirement that I actually use the insurance proceeds to pay to repair the fire damage. I could just deposit the check in my bank account, and no suppliers of any specific commodity are explicitly reimbursed. Compared to the life insurance case, the number of states covered by the contract is much larger (every possible amount of loss up to the maximum covered), and determination of which state has occurred and how much is to be paid is correspondingly somewhat more complex, but the fundamental idea is the same: insurance pays cash spendable on anything when a particular event occurs.

The only imposed limits on reimbursement in these contexts occurs at the time of insurance purchase; too large a requested payment (enormous life insurance benefits, or fire damage benefits in excess of the value of the house) raises insurer concerns about possible moral hazard (here often correctly interpreted as criminal behavior), but this is usually a relatively minor item. Beyond this, the amount of benefit payment is a choice variable for the insurance purchasers; insurers can and do sell policies with a wide variety of different levels of payment, conditional on a wide variety of events. Other than the (net) income effect of receipt of the benefit, there is usually little economic impact. It is possible, in the case of a widespread natural disaster, such as a hurricane or earthquake, that the price of building materials might be affected by the presence of insurance, and so some insurers might arrange for services rather than cash payment but, again, this phenomenon is limited in scope and relatively rare.

3. Medical services are different

Reimbursement for medical services differs from these classic insurance indemnity models for two reasons I will explore in this essay. One is the greater likelihood of moral hazard, the stimulus insurance coverage can offer to medical care spending. The other is the practical possibility that a private insurer might need to take account of and be able to affect the marginal price of services. Since moral hazard can matter even when services prices are taken by insurer and insured as given (whether or not they are at the competitive level), I will treat the case of moral hazard only first. Since I will argue that one strong normative and positive rationale for an insurer to affect or manipulate the marginal supply price is also because of moral hazard, I treat such supply-side effects second.

4. Indemnity insurance and the theory of health insurance benefits

The classic version of health insurance reimbursement closest to the life and property – casualty cases just discussed arises when health insurance takes the form of so-called indemnity insurance. The insured demands and obtains a medical service, for which he receives and pays a bill. Then (and only then), he turns to his insurer in order to be reimbursed for this bill. The key point in this classic case is that the reimbursement is triggered by, and to some extent depends on, the provider's bill. However, the insurer is initially assumed to take the existence and size of this bill into account only to the extent that its size affects benefit payments and therefore breakeven premiums. In this simple (but historically relevant) case, health insurance functioned like a credit card, paying bills but not questioning them.

The existence and importance of moral hazard relevant to reimbursement is related in a crucial way to the question of how benefit payments might feasibly be determined and why the indemnity form was supplanted. There are two distinguishable dimensions of moral hazard. One is shared by health insurance with many other kinds of insurance [Pauly (1968), Ehrlich and Becker (1972)]. The presence and magnitude of insurance coverage may cause insureds to alter the level of protective or preventive measures which affect the probability of a loss-producing event. It is the other aspects of moral hazard which, by an order of magnitude, is virtually unique to health insurance: because of the difficulty of measuring or specifying whether a loss has occurred, the size of the loss is affected by health insurance as well. For almost all other insurances, it is relatively easy to set up procedures to assess or estimate what the size of the loss is, and to agree on and specify those procedures in the insurance contract. For example, no one would imagine that the amount or type of reimbursement for tornado insurance would affect the number of tornadoes or the damage they do, the damage is usually localized enough that the prices of repair materials and services are also unaffected, and an insurance adjuster can easily determine the amount of damage. The only impact of different levels of benefit payment is to alter the amount of loss retained by the insured, in the sense that, as long as benefit payment is less than the amount of the loss, the insured is not fully "reimbursed" for the loss. Health insurance is different. How sick the insured has become is very difficult to verify objectively, especially before care (or repair) has begun. In one sense, the difference in difficulty of identifying for health insurance which state of the world has occurred, compared to other insurances, is a matter of degree, but it is a very large difference in degree indeed.

Partial reimbursement of total health care expenditures is actually the rule rather than the exception in market-based health insurance. Early in the history of private health insurance payment took the form of a pure indemnity – so much money for the loss of a foot, for a birth, and so forth. Such payment need not equal all the medical care expenditure that people choose to make, and it rarely covered the total loss in overall wellbeing associated with reduction in health status. However, pure indemnity payment is difficult in connection with reductions in health status because of the difficulty of verifying, in a contractually feasible way, what the adverse health event was [Ma and

McGuire (1997)]. So the basis for payment gradually turned to easier-to-verify events presumed to be positively correlated with reductions in health status: medical services spending, and time absent from work. In both cases some evidence of a threat to health was required, but what might be called the severity of the loss was generally measured by the amount of spending or lost wages. In either case, moral hazard is generated, since even persons with mild illness might attach positive marginal values to additional medical services spending or paid time off from work (up to a point), and therefore might increase the loss if insurance covers it.

In health insurance, the existence of this type of moral hazard is well documented [Newhouse and the Insurance Experiment Group (1993)]; people do use different quantities and types of medical care when they become ill and are insured, compared to when they are not insured. Control of such moral hazard can be based on demand side influences, based on only partial reimbursement of the full medical bill. Reimbursement affects demand because partial reimbursement of the bill will ordinarily leave the patent/consumer at risk for out of pocket payment, and it is this out of pocket price which determines quantity and quality of services demanded [Pauly (1968)].

5. Optimal reimbursement in price-taking markets with and without moral hazard

Consider first the simplest case in which utility depends only on income y, and in which there are random reductions in income. Absent insurance administrative cost, the optimal insurance policy makes payments conditioned on the state of the world such that the "expected marginal utility" of income is the same in all states of the world. If the utility function is $U = U_i(y_i)$, maximization of expected utility with actuarially fair insurance implies that the following must be satisfied for all states i and j

$$U_i'(y_i) = U_j'(y_j) \tag{5.1}$$

where $U_i'(y_i)$ is the marginal utility of income in state i and y_i is money income in that state.

In the case of diminishing marginal utility of income and no state dependence (i.e., $U_i'(y_i) = U_j'(y_i)$) this condition would imply that optimal insurance in the absence of administrative costs will equalize ex post income or wealth in all states of the world. The closest analogy to this classic case for health insurance would be as follows: suppose after an illness occurs there is one and only one treatment possible and that all consumers would choose that treatment whether insured or not. That is, there is no moral hazard. Moreover, after the treatment health is returned to its initial level. Although the person's utility may depend on health, in fact the ex post level of health does not vary. If the cost of treating the reduction in health is represented by $c(\Delta H)$, then we can write the utility function as $U = U(y - c(\Delta H))$, and the first order conditions are the same as before.

However, in the general case, the assumptions in this example need not hold, for at least two reasons:

1. As already noted, it may be difficult to establish or define the amount of the health loss in money terms.
2. The health state may affect the marginal utility of other consumption ("money income").

In the case of conventional insurance, the dollar payment by the insurance in the event of a loss should equal the value of the lost asset, if the asset is completely destroyed, or the lesser of the cost of repair or the value of the asset, if the damage is partial. The analogy for health is obviously imperfect, since few consumers would decide to accept death because the cost of a treatment was greater than the value of staying alive. A more useful point is that, since survival is generally necessary to enjoy income, the value of survival will be nearly as large as the total value of one's (lifetime) consumption [Linnerooth (1979)]. In short, the optimal reimbursement for care even of relatively high cost and low health impact could be quite large, if survival is at stake.

Could optimal benefit payments be greater than the cost of care? The general answer clearly is affirmative: if care is ineffective but money payments can increase utility (by substituting for lost health), optimal insurance may involve an indemnity payment in excess of the cost of care.

A key issue here is how (or whether) health state affects the marginal utility of money. Consider an illness that reduces health but for which no effective treatment exists. If we write the utility function in the general form

$$U = U(y, H), \tag{5.2}$$

where H is the level of health, or "health stock". Suppose that U_y' is directly related to the level of H. In such a case, optimal benefits will be lower than if the marginal utility of income was independent of the level of health.

For example, voluntarily purchased insurance rarely pays for pain and suffering, even though such payments often form an important part of damages paid in medical negligence suits. One explanation for the absence of insurance reimbursement is that the marginal utility of money is lower in states of reduced health. If sick people cannot enjoy what money will buy, and if there is no way for insurance to pay benefits in improved health, there may be little or no utility gain from buying insurance that will pay cash on the occasion of an illness. Add to this the exacerbation of moral hazard and the difficulty of writing verifiable contracts, and it is understandable why this type of insurance is rare.

Figure 1 outlines other possible cases. The rows characterize the effectiveness of care for the person's illness. Care might be entirely ineffective, might be effective but of low effectiveness, or might be highly effective. In the first row, with no cure possible, health is, in effect, the "irreplaceable" commodity discussed by Cook and Graham (1977); they show that the optimal benefit payment will be positive but less than the person's value or willingness-to-pay for lost health.

In the second row, care is sufficiently ineffective that the person will not optimally choose to return his health to its initial level. Here optimal payments depend on the

Health productivity of medical services	Utility function	
	General $(U = U(y, H))$	Additively separable $(U = U(y) + U(H))$
Zero (irreplaceable)	Benefits less than value of lost health	Benefits equal value of lost health
Low	Indeterminant	Benefits greater than $c(\Delta\overline{H})$
High	Benefits equal $c(\Delta\overline{H})$	Benefits equal $c(\Delta\overline{H})$

Figure 1. Optimal reimbursement as a function of utility function and health productivity of medical services.

form of the utility function. In the additively separable case (middle column), payments should equal the cost of care plus the remaining value of lost health as long as the cost of care is less than the value of the reduction in health. In the first column, with a general utility function, benefits may exceed, equal, or fall short of the cost of care depending on how changes in health affect the marginal utility of income. Finally, the last row describes the case where it is efficient to consume enough care to return health to its initial level; the person can be made "as good as new". Here the benefit payment will equal the cost of care regardless of the utility function.

If moral hazard occurs, benefit payments may differ from these first best levels, and so may the levels of use of medical care. It will usually not be efficient to set benefits equal to the cost of care; "free care" will lead to rates of use in excess of the first and second best options.

The first-best insurance "reimbursement," as noted, is a lump sum payment that does not depend on the cost or amount of care received and which can be either larger or smaller than the cost of care received. How does the amount of care in the second best world with moral hazard differ from the first best amount of care? The general answer to this question turns out to be surprisingly complex. We know that the level of use under zero cost sharing (and no rationing) will always be more than the first best optimum. However, because the first best might involve a lump sum payment in excess of the cost of care, it is possible that the income effect of such a payment could be so great that the resulting volume of care, even with full cost sharing at the margin, would exceed the amount associated with the second best optimal level of coverage [Ma and Riordan (1997)]. Whether this will happen depends on whether the first best payment exceeds the cost of care by a substantial amount. (As already noted, this will not happen if the marginal utility of income is reduced in states of poor health, or if health care is sufficiently productive that "good as new" levels of use are approximately optimal.)

If optimal benefits do exceed the cost of care, this result is more than a theoretical curiosity, because it implies that supply-side effects to reduce moral hazard could lead to an expansion of coverage sufficiently generous that income effects would increase optimal use. That is, discovery of supply-side incentives to get closer to the first best outcome could well lead to increased medical care spending, compared to the second best optimum with moral hazard. (To my knowledge, however, no one ever claimed that moral hazard at the second best optimal level of coverage would lead to greater use than would have occurred under ideal insurance; virtually all discussions of the welfare cost associated with moral hazard assume levels of coverage that are either assumed to be high or hypothesized to be increased by distortions like the tax subsidy.) Finally, there are almost no voluntarily purchased private insurances, even associated with the most aggressive managed care plan, that pay cash benefits in excess of the cost of care.

If we add the possibility that insurance carries a positive administrative loading, the conclusion from insurance theory that optimal coverage is full coverage above a deductible holds as well for health insurance as for other insurances. A deductible, at least initially, reduces well-being of a risk averse person only slightly, but does save on administrative costs, the more so if it avoids the expense of processing and paying many small claims.

If moral hazard is present, it will usually be optimal to have some positive cost sharing, and the most frequently analyzed version of cost sharing is in the form of coinsurance, in which the insured pays a given percentage (e.g., commonly 20%) of some approved charge level. A crucial point is that, compared with full coverage, this kind of reimbursement gives providers an incentive to be technically efficient since greater efficiency embodied in lower market prices will attract price-sensitive patients. Indeed, in the absence of limits on the total fee or change, historically the most important out-of-pocket payment was not the 20 percent coinsurance but the rather buyer responsibility for charges in excess of the optimal maximum reimbursement level. The precise form that cost sharing should take depends in turn on the precise form of the demand function as well as on the utility function.

6. Service benefit insurance

While health insurance has sometimes taken exactly the form just described, the most common form of market insurance did not pay simple per unit indemnity benefits with deductibles. Instead, in the United States and in other countries with private health insurance, benefits often were stated in physical rather than monetary terms, such as payment for a semiprivate hospital room, for a doctor office visit, and so forth.

Such "service benefit" reimbursement is explained in three ways. First, in a world of imperfect capital markets, consumers may find it difficult to pay in advance of reimbursement. Second, unit prices may vary over the time period of the insurance contract in unpredictable ways or across markets the insured might use; service benefit reimbursement protects against that risk. Finally, the insurer may be able to negotiate a price lower than the price charged to individual consumers who would pay out of pocket, with the reward to the provider being a larger volume of customers attracted by the service benefit feature. In many (though by no means all) ways, provider reimbursement in managed care plans is an extension and modification of these service benefit arrangements.

Even ignoring the possibility of discounts, and regardless of whether the consumer pays the provider first or the insurer pays first to the provider, one notes that the key aspect of reimbursement here is whether or not the provider is required to accept the level of insurance reimbursement, or the notional price on which that reimbursement is based, as payment in full. That is, a key aspect of patient/consumer reimbursement is whether providers are allowed to "balance bill", to collect from the consumer some amount in excess of the insurance reimbursement and the insurer-determined copayment. It is, after all, the out-of-pocket payment by the consumer which determines (in a neoclassical model) the level of consumer demand, either in total or from one provider compared to another.

The level of payment for which a consumer might be balance billed therefore depends on the procedures adopted by the insurer to determine the size of the insurance benefit. In US health insurance, public and private, three factors potentially limited reimbursement to a level below the total amount billed. For reimbursement to be made, the payment had to be usual, customary, and reasonable (UCR). While the terminology here is somewhat variable, these terms usually meant that the price had to be no higher than charged to other payers (usual), at or below some percentile of the distribution of prices in the market area (customary), and only moderately increased over the previous years' charges (reasonable).

Research did emphasize the "inflationary" character of these arrangements, arising largely from the phenomenon that an increase in prices charged to other payors could provide the rationale for an increase in payments from any given payor in subsequent periods [Frech and Ginsburg (1975)]. Of course, because of patient cost sharing (or due respect for the welfare of mankind) there would eventually be a limit to unit price increases.

For example, if the initial price was $\$P$, the initial level of reimbursement would be $0.8P$. With insurance reimbursement, price would rise, triggering a further rise in UCR reimbursement. If aggregate supply is fixed, the process of "inflation" will eventually stop, but only after reimbursement has risen to five times its initial value (since $0.2(5P) = P$).

These provisions were, however, an attempt to provide protection against risk in a way that avoided a problem (a special kind of moral hazard) associated with choosing among providers. In medical services markets, well before insurance came to dominate, price charged for apparently similar services varied widely. Consumers are often unaware of all prices being charged in the market, there are variations in the degree of local monopoly, there can be perceived subtle variations in quality, and medical services quality is difficult to determine in advance. An insurance which reimbursed on the basis of whatever charge happened to be rendered would offer the fully insured consumer little incentive to search for lower priced providers. Insurers therefore sought to provide incentives for consumers to search and providers to restrain prices charged to insured customers by setting per unit upper limits to the charge that would form the basis of the reimbursement benefit. There was, however, a tradeoff: some unlucky consumers would experience the risk of being required to pay high balance bills, and an upper limit would leave them exposed to that risk.

In addition to these influences based on the theory of optimal insurance, providers themselves had obvious reasons to be concerned about the conditions that determined reimbursement. In virtually all countries provider trade associations asked or were asked to be consulted on determination of reimbursement levels.

In the United States, this link was strong and explicit: provider associations founded and owned the dominant health insurers. Early on, the physician-owned insurer actually set a payment schedule with limits on balance billing for low income insureds, but most insureds received benefits based on a schedule of maximum payments. After Medicare introduced a UCR-type basis of payment for all services, such fee schedules disappeared.

7. Balance billing

If the physician is permitted to balance bill, what are the effects? There are two effects in a price taking market. Most obviously, the balance bill amount serves as additional cost sharing for those insureds who use physicians who balance bill. Secondly, the level of prices overall, and therefore the amount balance billed, may be affected by the average or typical level of reimbursement in a market. There is empirical evidence in support of the first proposition. When the US Medicare system changed its procedures to reward physicians who "accepted assignment" and eschewed balance billing, quantities of services demanded jumped upward, exactly as theory predicted.

The first phenomenon actually triggers the second. As the level of reimbursement is raised, the amount balance billed will shrink, and the quantity demanded will rise. If the market supply curve is upward sloping, the gross price will rise, increasing total spending, the balance billed amount and any coinsurance. This is one of the reasons why insurance without prohibition of balance billing is thought to be "inflationary".

The balance billing issue in physician payment has much in common with the reference pricing system used in Germany and elsewhere to pay for outpatient drugs. In physician services insurance with balance billing permitted and in reference pricing systems, a fixed insurance payment is set, with sellers free to change higher amounts than the reimbursement if buyers are willing to pay that price. However, the reaction of markets has been quite different; sellers reduced prices to the reimbursement level for drugs; they voluntarily chose not to balance bill. In contrast, physicians appeared to be eager to raise prices above the reference level for physician services.

What determines whether or not an insurer prohibits balance billing? Physicians will accept such limits, even in competitive markets, as a way of signaling their lower prices. However, as a recent debate in the US Medicare system has shown, limits on balance billing do constrain insureds' ability to contract with higher priced but possibly higher quality physicians. More generally, however, bans or limits on balance billing are greatly assisted by insurer (buyer) market power.

If balance billing is prohibited or limited, what should determine the level of reimbursement? The simplest answer is: reimbursement should be set at the level of the competitive price at the first-best quantity. This is the proper rule to follow when the service in question is of homogeneous quality and is supplied according to an upward sloping for horizontal supply curve. However, results are different if quality is variable, if higher quality costs more, and if consumers differ in their demand for quality. Then the optimal level of reimbursement is the solution of a kind of two-stage problem. In stage one, the level of quality supplied at each reimbursement level is determined. The consumer then chooses in stage two the point on this quality supply curve that corresponds to *optimum optimorum*. (The analysis of balance billing when the provider market is not competitive will be considered in the next section.)

8. Substitutes and complements

Studies of the impact of varying levels of reimbursement on the use of related service find results consistent with both economic theory and common sense: from the consumer's perspective, some services are substitutes (inpatient and outpatient care [Davis and Russell (1972)] while others are complements (doctor visits and prescription drugs [Hillman et al. (1999)]).

Supply side adjustments more consistently suggest substitution. It appears, for example, that the showdown in hospital inpatient spending growth after the Medicare DRG system was introduced led to a substitution of physicians' services [Wedig et al. (1989)], and study of limits on pharmaceuticals reimbursement (in a way that directly penalized physicians) led to greater use of hospital resources [Von der Schulenburg and Schoeffski (1993)]. More general results from American and European settings have yet to be obtained.

9. Alternatives to reimbursing market-level fee for service

We now turn to the question of the determination of payment rates by insurers who do not base reimbursement on market-level fees. We first consider the fundamental question: if reimbursement to providers is not set based on these market prices, how is it to be determined?

We look at which kinds of payers choose which methods, and then consider the effects of both the methods chosen to set reimbursement and the level of the reimbursement itself.

There are two broad classes of methods used to set provider reimbursement rates: cost-based methods and administered price methods. Administered pricing here means only that the insurer eventually sets a reimbursement level independent of a specific provider's cost or posted price. Often administered prices are uniform across classes of providers, but in the general case they need not be. A third method, which I will discuss later, is the use of bidding or bargaining methods.

We assume that all insurers forbid balance billing, and that each insured has only one insurance policy. There are four behaviors that reimbursement policy might affect. First, as already noted, it might affect the quantity and quality of care a consumer obtains. Second, if consumers differ by severity or risk in ways unknown to the payer, a kind of adverse selection can occur. Third, providers may or may not choose to minimize costs of whatever care they provide. And finally, variations in reimbursement policy, by affecting patient cost sharing, may affect risk reduction. The economic problem here, as usual, is that these four behaviors trade off. Achieving the optimum on any one dimension usually involves sacrificing one or more of the others. With at least four dimensions to the tradeoff, it is obvious that generalizeable propositions are going to be few and complex.

Complexity is compound because the appropriate behavioral model of the service's provider probably varies and surely is not definitely known. Here I give a brief catalog of the kinds of models that have been considered.

One model is that of a utility maximizing nonprofit firm. Nonprofit hospitals are assumed to maximize the utility they obtain from output quantity, output quality, and a host of other possible objectives, all subject to a breakeven constraint. Physicians, on the other hand, are most parsimoniously modeled as interested in a real revenue or utility which is affected by money income and hours of leisure. Sometimes other motives, such as patient health or wellbeing, or the accuracy of advice, are added. The third model is the profit-maximizing firm, whose net income is assumed to be both maximized and accurately measured.

One simplifying factor is that cost-based reimbursement has never been used for physician providers, presumably because there is no hope of measuring the true cost of the physician's own input. An issue in both hospital and physician context has been the definition of the unit of output or the basis of which payment is made: should it be finely disaggregated individual services (fee for service) at one extreme, all services received by an insured person per time period (capitation), at the other, something in between, or some mix of all of the above [Ma and McGuire (1997)]? In addition to these options, there is the added possibility of paying on the basis of observed inputs or their costs, rather than on the basis of output.

The longest history and the greatest amount of research applies to cost-based reimbursement, so I begin with an informal discussion of it.

There were two major problems with cost-based reimbursement, primarily used for hospital care. First, there were debates on exactly what the cost was. Measurement of cost in an unequivocal way, especially when overhead costs had to be allocated, could be contentious. Danzon (1982) has even argued that the rules for cost measurement constitute a kind of revenue function to the provider.

The second issue is whether the level of cost at a given level of output – even with perfect measurement – was indeed fixed. One reason it might not be is because of the possibility of producers deviating from cost-minimizing behavior [Chalkley and Malcomson (2000)]. Doing so makes no sense if firm managers have only profits as an objective and they are paid exactly costs; this point is discussed in more detail below. But perhaps the presence of cost-based reimbursement assists the emergence of firms with other objectives, ranging from charity care to on-the-job leisure, which require higher costs to be pursued. The other provider objective frequently discussed is quality. It too may be higher, even excessively so, under full cost-based reimbursement.

Let us first assume that no insurer has so large a market share that it expects to be able to affect provider price or income. All providers are currently paid fee for service, but an insurer contemplates paying providers for its insureds on a capitation basis. Since the marginal monetary revenue from capitation is zero, and the marginal real provider net income is negative (accounting for the implicit and explicit cost of additional output), it is easy to see that the unconstrained utility maximizing output for the profit-maximizing firm is zero, or else the bare minimum required by liability laws. Since this very low output has so far not been observed in regimes of capitation payment, there must be some other influences which change the desired output. The two influences discussed in the literature are market competition and provider concern for patient health and well being [Pauly (1970)].

The competition explanation is usually rather informal; it captures the idea that a provider who takes a capitation payment for a set of patients and then refuses to provide any services will develop a reputation for under service that will cause patients to select other providers. One limiting case here is the following: Suppose groups or networks of providers decide to furnish positive amounts of services even though they are capitated; they will therefore require a capitation payment large enough to cover the cost of the volume of services they do provide. Suppose consumers are perfectly informed about the quantities being furnished by each competing set of providers, and finally suppose the premium charged for coverage linked to any set of providers exactly reflects the capitation rate being paid to that set. Then it is easy to see that the competitive outcome will be optimal. A network which decides to provide the first best quantity (the quantity that would be demanded in the absence of moral hazard), and charges a capitation rate linked to it, will be preferred by all consumers to any other. If consumers have different demands, they will sort themselves among providers choosing various quantity-premium combinations. The competitive equilibrium will be ideal.

One problem with this attractive solution is that it is not unique. Suppose a payment system with positive marginal revenue also is offered. With perfectly informed consumers choosing among sets of providers, a set which, in return for a combination of fixed and per-service payments that just covered costs, selects the optimal volume would be able to offer the same premium and likewise attract customers. Even under the initial fee-for-service only regime, a group which chose to limit services to avoid moral hazard would be preferred. If competition is sufficiently strong, the choice across plans rewards the optimal volume, regardless of how it is achieved. If competition prevents underservice with capitation, it should be equally effective in preventing overservice under fee for service. The key is whether the quantities of services supplied can be linked to a provider-specific premium.

As we will note below, the main issue is in identifying a mechanism that will induce providers to provide the first best quantity, above and beyond their need to attract patients to their capitated list. An incentive mechanism which reinforces the desire to provide the optimal quantity might be more likely to represent a stable outcome [Pauly (1980)].

The other motivation for providing positive amounts of services for capitated patients who have presented themselves for treatment is a postulated concern physicians have for patient health. The usual way to model this is to assume that patient health enters the physician's utility function as well as the patient's utility function. Concern that over-treatment might harm health may even help to inhibit it under high levels of fee for service. However, the key implication is that as marginal revenue increases, physicians are willing to supply larger quantities. Under capitation, physicians usually do fail to provide all services that might add to health, but they still provide some services. Increase the marginal revenue, and they will provide more. Then we can derive a very simple but powerful proposition. If the quantity under full capitation (with payment yielding the physician as much real net income (including disutility from unnecessarily sick patients) as the full fee for service alternative) is less than the first best optimum, and the quantity under full fee for service is greater than the first best optimum, there must be some marginal price and associated partial capitation payment in between that will yield the first best outcome. Moreover, since the provider incentives fully determine the per patient quantity, no patient cost sharing is necessary.

This is a simplified version of the models constructed by Ellis and McGuire (1986, 1990, 1993); further elaborations allow for the quantity to be determined by bargaining between patient and provider, or for physicians to influence the output of hospitals as well as their own services. The less concern providers have for patient well being (e.g., perhaps providers are owners of clinical labs or for-profit hospitals), or the less well providers' concerns match patient demands (e.g., because providers do not accurately know patient-perceived severity or patient values), these models become less applicable. Also, if the insurer does not know or is unable to select the optimal rate of use, such powerful reimbursement tools cannot be used.

However, the Ellis–McGuire model deals with reimbursement of care for patients who have already presented for treatment. The level of reimbursement (and therefore any effect on user prices) may affect the initiation of care. Since this decision is crucial in determining total expense, the relative neglect of this point may be important.

10. Monopsony and provider market power

Now suppose, in contrast to the earlier assumption, that an insurer has a large enough buyer market share to affect provider price or net income. For this to happen, it must also be the case that the market supply curve is upward sloping. Two issues are relevant to reimbursement in this case: the possible use by a welfare-maximizing insurer of a fee schedule to control moral hazard, and possible monopsony behavior by an insurer interested in net income and uninterested in overall economic welfare [Pauly (forthcoming)].

The argument about optimal price setting in the presence of market power is really an even simpler version of the "concerned doctor" model above. Suppose an insurer has a monopsony in the purchase of some service, and suppose the insurer is able to determine the optimal volume of that service. Then there exists a point on the supply curve for that service at which the price calls forth just the optimal quantity [Pauly (1995)]. The difference is that the upward sloping supply curve does not require that the provider be concerned about the patient, only that marginal (implicit) cost be increasing. To close the model, we need to assume a very slight reputation effect or concern about patients to make sure that the optimal quantity of services get allocated to the right patients.

In this case, the insurer could set socially optimal reimbursement, but will it? The answer obviously depends on insurer objectives and market structure. On the structure side, if the insurer is a monopsonist, in a position to be able to buy almost all of the medical services in the market, it is almost surely also a monopolist, able to sell all of the insurance in the market. If it is a profit maximizing firm, it will maximize monopsony profits by choosing not the socially optimal price and volume but rather by considering as its input cost the price marginal to the supply price on the standard services supply curve. This firm is likely to pay a low price for services and charge a high price for insurance, limiting both the amount of services per insured and the total number of insureds [Pauly (forthcoming)].

What if the insurer is a private nonprofit firm operated in the interest of consumers of insurance? It will set a premium that will just cover the cost of the benefits it pays out. But it will still behave like a monopsonist, in this case one controlled by a buyers' cartel, and therefore will still restrict supply in the interest of holding down unit prices and total premiums.

What if the insurer with monopsony power is administered by the government – for example, as in the case of US Medicare or most national health insurances? There will surely be temptation to act like a monopsonist – to go further than just forcing provider price down to the competitive level. There is little more that we can say in theory. We can note that such public systems do tend to pay significantly lower prices and wages than in the private market in the United States, including paying significantly lower wages to nurses and technicians and other specialized workers who have traditionally not had much market power themselves [Pauly (1993)].

Finally, what if the provider also has market power and the insurer has a sizable but not dominant market share? In such a case the insurer may set a reimbursement level below the provider's posted change, prohibit balance billing, and yet expect some services to be supplied to its insureds.

The model for this case was first developed from the US Medicaid program by Sloan, Cromwell, and Mitchell (1978) and by Lee and Hadley (1979). Under the assumption that the demand for care by those covered by insurance is unlimited at the reimbursed price, the model is effectively one of discriminating monopoly, with the dominant insurer's demand curve treated as perfectly elastic at the reimbursed price. The main conclusion from this model is that a reduction in reimbursements will cause declines in both the volume of services supplied to the dominant insurer's clients and in the price the provider charges to others in the market (as the provider seeks to attract customers to replace the less profitable clients of the dominant insurer). If, in contrast, demand is limited (for example, by time cost), there is no effect of marginal changes in reimbursement on either quantity or price; the only effect is lower provider profit.

11. Reimbursement and productive efficiency

As already noted, reimbursement, especially for hospital services, has sometimes been based on accounting cost. In addition, especially for physician services, the number of separate specialized measures of output is large, reaching into the thousands. For both reasons, reimbursement has been thought to contribute to productive inefficiency.

Firm-specific cost-based reimbursement is essentially payment on the basis of inputs, rather than on the basis of output. The major analytic problem is that, if reimbursement were truly equal to cost (including normal return on equity), the profit-maximizing firm would not have determinate input or output decisions, since its profit would in theory be the same regardless of what it did. Since some insurers have found it difficult or impolitic to specify what services patients should receive, policymakers have advocated and tried to design systems with neutral financial incentives, so that output decisions can be made by health professionals on the basis of patient health alone. Strange as such motivation may seem to economists, it was part of the rationale for cost based reimbursement in the private Blue Cross hospital insurance in the 1930s.

If payment literally just equaled true cost, there would be no reason to deviate from cost minimization, but no incentive to seek it either. However, very slight deviations from reimbursement of the full true cost can cause substantial deviations from cost minimization; the neutrality property is very knife-edged. On the one hand, if true costs can be less than reimbursed costs, the literal solution for a profit-maximizing firm is to maximize costs. On the other hand, cost based reimbursement always fails to reimburse managerial "effort" to minimize costs, because there is no way to measure this input; it therefore leads to suboptimal levels of effort [Chalkley and Malcomson (2000)].

In both cases, all that is needed to offer an incentive to minimize cost is replacement of firm-specific cost based payment by a fee or price schedule. Any schedule will do the job; the only requirement is that reimbursements be independent of inputs and depend only on outputs. This proposition is not well understood. Even before changes in the 1990's, because of administratively imposed limits on the rate of growth, the old Medicare physician payment scheme had become a de facto fee schedule for almost all physicians. It may have been unfair, but it was not inefficient in the sense of disincentives for (short run) cost minimization. Likewise, hospitals traditionally did maintain list prices for their services along with cost based reimbursement, and a fee or price schedule applied to those services would have offered an incentive to minimize the cost of producing each service.

This brings us to the second issue: how finely should services be defined? The objective of Medicare's hospital payment reform was not simply to offer incentives to produce hospital days, laboratory tests, and the like at minimum cost. It was to try to influence the volume of these inputs per hospital discharge. The reason why the level or mix of such services was not ideal had to do primarily with moral hazard (and the very low level of patient cost sharing for inpatient care). The per admission payment system based on diagnosis related groups adopted by US Medicare in 1983 is therefore as close to the capitation payment analyzed above as to the economic literature on price incentives and government procurement, which is sometimes cited [Newhouse (1996)]. For example, in that literature it is important to keep a certain set of suppliers in business, while inducing them to be efficient. Medicare, in contrast, did not use its general reimbursement policy for such an objective.

To take just one example, length of stay fell dramatically after the new system was introduced. Was that an improvement in productive efficiency, or a reduction in moral hazard? I think it was the latter. The lost extra days of stay almost surely provided some perceived benefit to patients (even if the benefit was less than cost) and there is little evidence that excessive levels of inputs were used to produce them. That is, the days may have been "medically unnecessary", but they were not produced inefficiently. That there was some positive benefit is the burden of much of the backlash against managed care's shortening of stays. This was not "high level" pricing but rather the creation of supply side reimbursement incentives to limit the amount of services per admission. Even the presence of (sometimes) empty beds can be viewed as provision of a safety margin against (rare) demand surges, not as pure waste. As the discussion of capitation indicates, the optimal reimbursement generally will pay the hospital some positive marginal revenue for more services. However, the best way to do this is not to make the payment depend on the cost, but rather to make it depend on the volume of services. The real issue then is not a tradeoff between selection and productive efficiency (given quantity and quality of output), but rather is a tradeoff between control of moral hazard (by means of a fixed, "lumped" price) and selection caused by legitimate variation in the ideal volume of services for heterogeneous patients.

12. Heterogeneity in non-competitive markets

Suppose the insurer does set an administered price for a service as indicated above which would be optimal if all patients consuming that service generated the same cost, holding quality constant. However, suppose patients differ in severity in ways known to patient and provider but not to the insurer, and those differences would affect the optimal quality-constant cost. A simple assumption is that the cost of providing some defined "services" to sicker patients is higher than for less sick patients. The insurer may know the average cost, or even the distribution of costs, but it cannot know which patients are sicker. Then if a fixed prospective price is paid, the sicker patients either will not be treated or will not be treated with the same quantity. Under competition among providers, the less sick patients will more receive beneficial services (or higher quality) worth less than their cost as providers "bid away" rents competing for them. The alternative method of payment would be to pay providers the cost of whatever they provide. Then the sicker patients will not be discriminated against, but we are back to the warped incentive world of cost based reimbursement. Newhouse (1996) has recently argued that this type of tradeoff is both common and important in health services reimbursement; the second best solution is a mixed payment method in which part of the payment is fixed and part depends on cost.

There is a subtle but potentially important issue here. If the provider really does not know who is sicker (because the transactions costs of finding out are too great), or if the provider does not have the power to refuse treatment (as hospitals do not), suppliers will accept sicker patients at the same rate as less sick persons, because they do not know who is who, or cannot discriminate based on what they do know. But if the provider never learns about this higher cost until after the fact, there will be no need to offer additional revenues at the margin for incentive reasons. That is, if the provider does not know who is sicker, the provider cannot reject the sicker patient. Alternatively, if the provider is not allowed to reject the sicker patient, that patient will be treated. However, it may be plausible that the provider, having accepted for treatment a patient who is sicker, then discovers that fact in the course of treatment, and seeks to avoid further treatment or to cut treatment back to a level that is less costly than the reimbursement. In this case, the patient will not be "shunned", but the patient might be dumped (or dumped on). (One unexplored issue is where the dumped patients are to go; they almost always must be referred to another hospital. Usually it is assumed that there is a public hospital which will take them, raising the hypothesis that public hospitals encourage dumping, or, more generally, that they encourage inefficiency that would not be present if there were no hospital to accept dumped patients.)

The key point is that this information is developed over time. There is no need to make a marginal payment for sicker patients in the first period, but once the insurer learns something about the patient's risk (even if it is only a noisy, not a perfect signal) and starts to suspect that this may be a sicker patient, then a positive marginal payment may be a helpful incentive. Even this conclusion depends on there being another transaction cost, since the extra-sick patient (or the high risk consumer) could move from seller to seller, staying one step ahead of the accumulation of information.

Of course, if patients know they are sicker than average, a provider may be able to get them to self select. By offering an inferior quality of a service sicker patients need, a provider may discourage them from seeking care. In contrast, adjusting the reimbursement for higher costs or higher severity will make such a strategy unnecessary. The main point here is that, for self selection to happen, patients must know their severity and there must be a feasible managerial strategy for providers that will discourage them. Ellis (1998) provides an example of such a model, but one should not suppose that these conditions always hold.

Where does this leave the argument for mixed payment systems? Some positive marginal revenue triggered by actions which proxy greater severity would be efficient (if administratively complex), since the first deviations from a fixed price offer little incentive for inefficiency but great rewards for treating sicker patients. Of course, if the main consequence is only lower quality for high risks (and higher quality for low risks), not refusal to treat, the real question is the value of the lost or gained quality. All of these considerations cease to apply if balance billing is allowed: then sicker people could pay extra, though they would bear out-of-pocket risk for doing so. Indeed, patient willingness to pay an "extra care" surcharge would be a good signal of high risk, and could itself trigger adjustments in reimbursement.

Reimbursements should therefore be (at least partially) independent of firm-specific costs. But should they be based on some measures of average (or efficient) costs, or should they be based on the value of a service to the insurer and its customers?

The US Medicare system opted for relative weights based on a labor theory of value in setting physician reimbursement, and on reimbursement based loosely on benchmark costs for hospital payment, but managed care firms have been much more aggressive in talking about rewarding providers for good outcomes. The simple model discussed earlier shows that, in a very real sense, this is or ought to be a false dichotomy.

With increasing costs (or variable quality), there is no such thing as setting price to equal "the" cost [Held and Pauly (1983)]. Instead, depending on where the price is set, quantity and quality will be adjusted until marginal cost becomes equal to it [Gertler and Waldman (1992)]. On the one hand, the challenge to an insurer is to determine the whole shape of the cost curve, but, on the other hand, the reimbursement level can control quantity and quality.

There is a way to avoid the problem of undertreatment or dumping more severe cases under a fixed per unit payment mechanism: just set a high payment rate. Since providers will accept all patients whose cost is less than the payment rate, there is some payment high enough that all (or enough) of the high risks will be treated. Of course, providers will earn profits, and the taxes on premiums needed to raise the funds will cause distortions, but if the costs of treating patients at each level of severity are given, those profits themselves will only represent a transfer, not a reduction in efficiency. However, public insurers interested more in their budgets than in economic welfare will not find this argument convincing. It remains true that there is a tradeoff among incentives, selection, and the overall level of payment.

13. Empirical results on reimbursement

The conversion of the US Medicare system from cost-plus to fixed price per admission provides the best evidence we have of the effect of reimbursement on hospital care. When the new system was introduced in the mid-80s, length of stay, profit margins, and the rate of growth of costs all eventually fell. (So did admissions, to everyone's surprise.) Such a reaction would be predicted either for a profit-maximizing firm, or a non-profit utility-maximizing firm subject to a budget constraint. Under either model, what happened was what one would expect to happen [Hodgkin and McGuire (1994)].

There even was a fall in length of stay for non-profit hospitals with below average costs, although it appears to have been smaller than that for high cost hospitals; this phenomenon is not consistent with some models of the non-profit hospital [Phelps (1997)], but it can easily be made consistent by assuming that new Medicare profits can be used to cross-subsidize other objectives [Newhouse and Byrne (1988)].

The evidence is also fairly strong, however, that there were no significant adverse affects on health outcomes, a result inconsistent with the heterogeneity models (since the high risk patients should have suffered), but a result that is not definitive because of the imprecision of measurements of health. McClellan (1997) has also argued recently that the actual Medicare payment system had many administrative exceptions, and therefore does not really offer the fixed price incentives just discussed.

For physician reimbursement, as discussed elsewhere, the results are more ambiguous. Very low relative fees, such as those historically paid by the Medicaid program, do cause doctors to decline to furnish services [Sloan, Mitchell and Cromwell (1978)]. Payment by capitation or salary leads to less supply than under fee-for service. Higher marginal revenues earned by the doctor for recommending non-physician services (such as imaging in a physician-owned imaging facility) lead to larger volumes of such services [Hillman, Pauly and Kerstein (1989)]. However, small variations in fees around the high levels at which fees have historically been set do not have predictable relationships, with higher fees sometimes being associated with larger volumes, sometimes no difference, and sometimes lower volumes [Rice (1998)].

Similar effects from a transition from cost-based to charge-based reimbursement have been found for nursing home care [Cohen and Dubay (1990), Thorpe, Gertler and Goldman (1991)]. For example, conversion of nursing home payment to a method of payment based on resource groups slowed cost growth, but did so to the greatest extent for those producers who were most constrained by the new payment system [Thorpe, Gertler and Goldman (1991)].

14. Bidding models

When the insurer does not know the cost, it can in principle cause providers to reveal their costs. Under certain circumstances, it can even cause providers to reveal the costs associated with whatever level of selection insurers experience.

One difficulty in a bidding approach for health care is dealing with the variation in severity discussed earlier. To bid for a hospital admission or a capitated life, the bidder must estimate the average severity of those who will be attracted. Another difficulty is that, if there is a single winner, access to services may be limited.

For both of these reasons, the case of bidding to set reimbursement has thus far largely been limited to standardized, portable products and services such as disposables and lab tests on already-collected specimens. Medicare has been trying to mount a so called "bidding demonstration" to set per-capita payments to managed care plans [Dowd, Feldman and Christianson (1996)], but this mechanism would not award a single contract to the low bidder. Instead, the low bid would set the level of reimbursement for the higher bidders, who then would be forced to charge higher premiums to obtain their originally proposed price. Issues of quality adjustment and selection have barely been investigated.

There are really two issues here. One, given most prominence in the discussion, is whether the lowest bid will approximate the efficient cost of production. The second, largely ignored, question is what happens to firms who submit higher bids and therefore are forced to charge consumers' higher premiums than other firms are allowed to charge.

If the firm level price elasticity of demand is high, there is a strong incentive to each firm to bid near its cost of production. It cannot lose money by doing so, and is guaranteed to have an attractive low price in the market. However, the story may change substantially if quality is variable. Charging a higher than minimum premium may actually attract business if it allows the profitable provision of highly valued quality. The analysis here is complex, since optimal bidding strategy depends on the quality levels other providers choose.

15. Conclusion

Arms-length indemnity insurance sets reimbursement levels according to some fairly simple principles, but is limited in its ability to control moral hazard through reimbursement policy alone. If it is desired to limit patient cost sharing, an attractive alternative is to control moral hazard through supply-side reimbursement policy. However, because of asymmetric information and variable quality, that task becomes complex.

In principle a sufficiently complex policy, involving fixed and variable payments, and adjustments for noisy measures of quality and quantity and for often – speculative and peculiar provider objective functions, can achieve a second best optimum. The practical problem is how to tell empirically whether one has specified the ideal fine-tuned system. Research and experience can help.

References

Arrow, K.J. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53:941–973.

Chalkley, M., and J.M. Malcolmson (2000), "Government purchasing of health services", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 15.

Cohen, J.W., and L.C. Dubay (1990), "The effects of medicaid reimbursement method and ownership on nursing home costs, case mix, and staffing", Inquiry 27(2):183–200.

Cook, P.J., and D.A. Graham (1977), "The demand for insurance and protection: the case of irreplaceable commodities", The Quarterly Journal of Economics 91(1):143–156.

Danzon, P.M. (1982), "Hospital 'profits': the effects of reimbursement policies", Journal of Health Economics 4(4):309–331.

Davis, K., and L.B. Russell (1972), "The substitution of hospital outpatient care for inpatient care", Review of Economics and Statistics 54:104–107.

Dowd, B., R. Feldman and J. Christianson (1996), Competitive Pricing for Medicare (AEI Press, Washington, DC).

Ehrlich, I., and G.S. Becker (1972), "Market insurance, self-insurance, and self-protection", Journal of Political Economy 80(4):623–648.

Ellis, R.P. (1998), "Creaming, skimping, and dumping: provider competition on the intensive and extensive margins", Journal of Health Economics 17(5):537–555.

Ellis, R.P., and T.G. McGuire (1986), "Provider behavior under prospective reimbursement", Journal of Health Economics 5:129–152.

Ellis, R.P., and T.G. McGuire (1990), "Optimal payment systems for health services", Journal of Health Economics 9:375–396.

Ellis, R.P., and T.G. McGuire (1993), "Supply-side and demand-side cost sharing in health care", Journal of Economic Perspectives 9:135–151.

Frech, H.E., and P. Ginsburg (1975), "Imposed health insurance in monopolistic markets: a theoretical analysis", Economic Inquiry 55–70.

Gertler, P., and D. Waldman (1992), "Quality-adjusted cost functions and policy evaluation in the nursing home industry", Journal of Political Economy 100(6):1232–1256.

Held, P.J., and M.V. Pauly (1983), "Competition and efficiency in the end stage renal disease program", Journal of Health Economics 2(2):95–118.

Hillman, A.L., M.V. Pauly and J. Kerstein (1989), "How do financial incentives affect physician's decisions, resource use and financial performance in health maintenance organizations?", New England Journal of Medicine 321:86–92.

Hillman, A., et al. (1999), "Effects of physician and patient financial incentives on prescription drug spending in managed care plans", Health Affairs 18:189–200.

Hillman, B., et al. (1990), "Frequency and costs of diagnostic imaging in office practice – a comparison of self-referring and radiologist-referring physicians", New England Journal of Medicine 323(23):1604–1608.

Hodgkin, D., and T.G. McGuire (1994), "Payment levels and hospital response to prospective payment", Journal of Human Resources 13(1):1–29.

Lee, R., and J. Hadley (1979), "Physicians' fees and public Medicare programs", Working paper No. 998-17 (The Urban Institute).

Linnerooth, J. (1979), "The value of human life: a review of the models", Economic Inquiry 17(1):52–74.

Ma, C.A. (1994), "Health care payment systems: cost and quality incentives", Journal of Economics and Management Strategy 3(1):93–112.

Ma, C.A., and T.G. McGuire (1997), "Optimal health insurance and provider payment", American Economic Review 87(4):685–704.

Ma, C.A., and M.H. Riordan (1997), "Health insurance, moral hazard, and managed care", Industry Studies Working Paper (Department of Economics, Boston University).

McClellan, M. (1997), "Hospital reimbursement incentives: an empirical analysis", Journal of Economics and Management Strategy 6(1):91–128.

Newhouse, J.P. (1996), "Reimbursing health plans and health providers: efficiency in production versus selection", Journal of Economic Literature 34:1236–1263.

Newhouse, J.P., and D.J. Byrne (1988), "Did Medicare's prospective payment system cause lengths of stay to fall?", Journal of Health Economics 7:413–416.

Newhouse, J.P., and the Insurance Experiment Group (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge and London).

Pauly, M.V. (1968), "The economics of moral hazard", American Economic Review 58:533–539.

Pauly, M.V. (1970), "Efficiency, incentives, and reimbursement for health care", Inquiry 7:114–131.

Pauly, M.V. (1980), Doctors and Their Workshops (University of Chicago Press).

Pauly, M.V. (1993), "U.S. health care costs: the untold true story", Health Affairs 12:152–159.

Pauly, M.V. (1995), "Paying physicians as agents: fee-for-service, capitation, or hybrids?", in: T. Abbott, ed., Health Care Policy and Regulation (Kluwer Academic Publishers) 163–174.

Pauly, M.V. (forthcoming), "Monopsony, managed care, and medical markets", Health Services Research.

Phelps, C.E. (1997), Health Economics, 2nd edn. (Addison-Wesley, Reading, MA).

Rice, T. (1998), The Economics of Health Reconsidered (Health Administration Press, Chicago).

Shleifer, A. (1985), "A theory of yardstick competition", Rand Journal of Economics 16(3):319–327.

Sloan, F.A., J. Mitchell and J. Cromwell (1978), "Physician participation in state Medicaid programs", Journal of Human Resources 13(Suppl.):211–245.

Thorpe, K.E., P.J. Gertler and P. Goldman (1991), "The resource utilization group system: its effect on nursing home case mix and costs", 28(4):357–365.

Von der Schulenburg, J.M., and O. Schoeffski (1993), "Implications of the structural reform of Healthcare Act on the Referral and Hospital Admission Practice of Primary Care Physicians", University of Hannover, Institute of Insurance.

Wedig G., et al. (1989), "The Medicare physician fee freeze: what really happened?", Health Affairs 8(1):21–33.

PART 3

INSURANCE MARKETS, MANAGED CARE, AND CONTRACTING

This Page Intentionally Left Blank

Chapter 11

THE ANATOMY OF HEALTH INSURANCE*

DAVID M. CUTLER and RICHARD J. ZECKHAUSER

Harvard University and National Bureau of Economic Research

Contents

*We are grateful to Dan Altman for research assistance, to Jon Gruber, Tom McGuire, Joe Newhouse, and Alexandra Sidorenko for helpful comments, and to the National Institutes on Aging for research support.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This article describes the anatomy of health insurance. It begins by considering the optimal design of health insurance policies. Such policies must make tradeoffs appropriately between risk sharing on the one hand and agency problems such as moral hazard (the incentive of people to seek more care when they are insured) and supplier-induced demand (the incentive of physicians to provide more care when they are well reimbursed) on the other. Optimal coinsurance arrangements make patients pay for care up to the point where the marginal gains from less risk sharing are just offset by the marginal benefits from reduced provision of low valued care. Empirical evidence shows that both moral hazard and demand-inducement are quantitatively important. Coinsurance based on expenditure is a crude control mechanism. Moreover, it places no direct incentives on physicians, who are responsible for most expenditure decisions. To place such incentives on physicians is the goal of supply-side cost containment measures, such as utilization review and capitation. This goal motivates the surge in managed care in the United States, which unites the functions of insurance and provision, and allows for active management of the care that is delivered.

The analysis then turns to the operation of health insurance markets. Economists generally favor choice in health insurance for the same reasons they favor choice in other markets: choice allows people to opt for the plan that is best for them and encourages plans to provide services efficiently. But choice in health insurance is a mixed blessing because of adverse selection – the tendency of the sick to choose more generous insurance than the healthy. When sick and healthy enroll in different plans, plans disproportionately composed of poor risks have to charge more than they would if they insured an average mix of people. The resulting high premiums create two adverse effects: they discourage those who are healthier but would prefer generous care from enrolling in those plans (because the premiums are so high), and they encourage plans to adopt measures that deter the sick from enrolling (to reduce their overall costs). The welfare losses from adverse selection are large in practice. Added to them are further losses from premiums that vary with observable health status. Because insurance is contracted for annually, people are denied a valuable form of intertemporal insurance – the right to buy health coverage at average rates in the future should they get sick today. As the ability to predict future health status increases, the lack of intertemporal insurance will become more problematic.

The article concludes by relating health insurance to the central goal of medical care expenditures – better health. Studies to date are not clear on which approaches to health insurance promote health in the most cost-efficient manner. Resolving this question is the central policy concern in health economics.

Keywords

adverse selection, agency problems, HMOs, indemnity insurance, intertemporal insurance, managed care, moral hazard, pooling equilibrium, separating equilibrium, supplier-induced demand

JEL classification: I10

Insurance plays a central role in the health care arena. More than 80 percent of health care expenditures in the United States are paid for by insurance, either public or private, with an even greater percentage supported in most other developed nations. Insurance thus provides the money that motivates and supports the health care system.

This paper describes the anatomy of health insurance. At the micro level, it details why individuals seek insurance, and the challenges in structuring insurance policies. At the macro level, it explains the role of health insurance in the medical care sector. The medical care triad (Figure 1) depicts that sector in a fundamental fashion. Insurers mediate between individuals[1] and their providers. Often times, the flow of funds is more roundabout: governments or employers nominally pay insurers, but these costs are then passed on to individuals, via increased taxes or lower wages.

The insurer intermediary must design a policy to pay for (and possibly provide) care. This is a treacherous task. Designing a health insurance policy is not nearly so challenging technologically as, say, designing a personal computer system, but it must still overcome some distinct and substantial economic obstacles. The most important of these obstacles are *agency problems*. Insurers cannot get relevant parties to do what efficiency requires. Thus, people with generous insurance spend more on medical care than people with less generous insurance (moral hazard), and providers paid on a fee-for-service (piece-rate) basis may provide more care due to supplier-induced demand than they would if they were not paid per task. In a situation where agency relationships are imperfect, insurance is necessarily second-best. Insurers must trade off the benefits from more generous insurance – primarily the reduction in risk it affords – against the costs of more generous insurance – moral hazard or supplier-induced demand. Throughout this chapter, we highlight central lessons about health insurance, which are then collected in Table 10. This clash between risk sharing and incentives is Lesson 1 about health insurance.

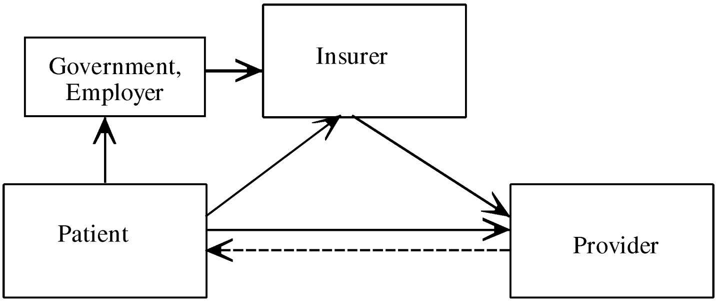


Figure 1. The medical care triad. Solid lines represent money flows; the dashed line represents service flows.

[1] Throughout the paper, to facilitate exposition, we mostly refer to patients or insureds as individuals, although most health insurance is purchased on behalf of families.

Agency problems in health care can be alleviated in two ways. The demand-side approach discourages excessive utilization by making people pay something when they consume medical care. Demand-side rationing is epitomized in the traditional indemnity insurance plan, which prevailed in the United States for a half century. The supply-side approach discourages utilization by monitoring providers carefully, penalizing them if they are profligate, and giving them financial incentives to provide only essential care. Increasingly, supply-side limitations are fostered by integrating insurance and provision. Some HMOs, for example, are both insurers and providers of care. Integration of the insurance and provision functions is unique to medical care, and results from the fundamental difficulties with solely demand-side rationing. The integration of health insurance and provision of medical services is Lesson 2 about health insurance. Sections 3 through 5 of the chapter lay out the issues involved in demand- and supply-side rationing.

We then move from these micro relationships to the broader arena of the market for health insurance. People have preferences for different types of health insurance, and those preferences should be accommodated to the extent possible. In addition, competition in health insurance can encourage production efficiency, driving down overall costs. But competition in health insurance produces results unlike competition in other markets, for a fundamental reason: the costs of providing insurance, as opposed to say computers or food, depend on the characteristics of the buyer. People with a poor medical history will benefit more from and cost more to insure than those with a healthy past. Thus, the sick will sort themselves into more generous plans than will the healthy. This process, called *adverse selection*, can substantially limit the benefits of health plan choice. Individuals will have incentives to choose less generous policies over more generous ones (to pool with the healthy instead of the sick) and insurers will have incentives to reduce the generosity of their benefits (to attract the healthy instead of the sick). Lesson 3 describes the consequences of competition when buyer identity affects costs. Section 6 discusses adverse selection and approaches to deal with it.

The natural tendency of insurers to charge the sick greater premiums than the healthy presents a further challenge to health insurance: lack of coverage against the long-term risk of becoming sick and having higher expected costs in the future. Using the thought experiment of individuals making choices behind the veil of ignorance, they would choose to insure their risk of becoming sicker than average – a multi-year risk – just as individuals in any year wish to insure their medical costs that year. Markets for multi-year insurance do not exist, however, for understandable reasons, and in practice individuals are left without this insurance. The kernel of the problem is that information on risk levels becomes available before insurance contracts are drawn. Lesson 4 is that early information dries up insurance markets. Long-term insurance is taken up in Section 7 of the chapter.

However effectively health insurance controls costs or spreads risks (the focus of most of this chapter), its key goal is to promote health. In Section 8 we examine the relationship between health insurance and health. Variations in insurance generosity have relatively little impact on health outcomes among those with insurance. This finding is consistent with the idea that insurance generally restricts care offering relatively low value. But the time frame over which these issues has been examined is not large. We know less about the long-run effect of different health insurance arrangements on health than we should. We mark the centrality of health as opposed merely to financial transfers and the lack of clear evidence on the relative benefits of different systems as Lesson 5 about health insurance.

At the outset, it is important to take account of the distinctive role health insurance plays in society. Economists traditionally measure value by willingness to pay, and the value of health insurance, or its byproduct medical care, is calibrated in dollar terms – the same as apples or television sets. In much of the world, however, particularly outside the United States, medical care and medical insurance are treated differently. Medical care is often viewed as a right, for which market-based allocation is not appropriate. For some, the right is absolute; markets should play no role in the allocation of medical services. More moderate positions assign government a special responsibility for medical care, which leads to a government insurance system or set of subsidies. Rights-oriented sentiments show up even in the United States. The United States subsidizes medical insurance directly for poor people and old people, and indirectly for the working-age population (through the exclusion of health insurance from individual taxable income). While some such subsidies may be justified on externality grounds (when people get medical care, they are less likely to spread infectious diseases to others), merit-good arguments, or fiscal externality arguments (when people are healthier, they earn more, pay more in taxes, and receive less in public benefits), we suspect that a right to medical care is the more basic motive.

But the rationale for subsidizing health insurance, as opposed to medical care, is less clear. The government could promote consumption of medical care through direct delivery of services or by subsidizing inputs, without intervening in the medical insurance market. We thus focus primarily on the economic analysis of health insurance, leaving aside normative views about access to basic medical services [Hurley (2000), Wagstaff and van Doorslaer (2000), and Williams and Cookson (2000)]. We come back to the access issue in the last section.

In this essay, we follow common parlance by [primarily] using the terms health care and health insurance, although the terms medical care and medical care insurance might be better descriptors. Health status cannot be insured. The costs of medical care can be, and are, albeit often bearing the label health insurance.

We begin in the first section by discussing the provision of health insurance around the world and in the second with a review of the principles of insurance. We then examine the micro and macro issues in health insurance.

1. Health insurance structures in developed nations

Health insurance is common to all developed countries, but the mechanism for obtaining insurance differs from country to country. In most countries, health insurance is universal; everyone is entitled to coverage and is required to purchase it.[2] In some nations, such as Canada, the financing is through taxation; people pay an income or payroll tax, and the proceeds are used by the government to purchase or provide health insurance. In other nations, the financing is through private insurance; individuals or their employers contribute to health insurance companies, which then provide insurance for the population. While the payment for any individual may differ in these two systems (a tax-financed system generally imposes relatively more on the rich), the implications for the provision of health insurance are generally slight. Governments in both systems are intimately involved in determining what services are covered, the cost sharing that patients face, and the restrictions imposed on providers.

The specifics of health insurance structures differ significantly across developed nations. Countries such as the UK and Italy finance health insurance through general taxation and (at least historically) provide services publicly.[3] Countries such as Canada and Germany finance insurance publicly but contract for services through private providers.

1.1. Health insurance in the United States

Describing the detailed structures for health insurance in different nations would take an entire volume. We focus our attention primarily on the United States. The United States is distinctive among OECD countries because health insurance is not universal.[4] Table 1 shows the sources of health insurance in the United States. About one-quarter of the United States population is insured through the public sector. The primary public programs are Medicare, which mostly insures the elderly, along with the disabled and people with kidney failure; and Medicaid, which insures younger women and children, the elderly (for services not covered by Medicare such as nursing home care), and the blind and disabled. Other public programs, primarily for veterans and dependents of active-duty military personnel, insure another 1 percent of the population.

Another 60 percent of the population has private health insurance. Most of this insurance is provided by employers; less than 10 percent of the population purchases insurance privately. The predominance of employer-provided insurance results from the favorable tax treatment of that method of payment. Compensation to employees in the form of wages and salaries is taxed through federal and state income taxes, and through the federal Social Security tax. Compensation paid as health insurance, in contrast, goes untaxed. Since marginal tax rates range from 15 to 40 percent for most employees,[5] the

[2] In some countries, such as Germany, temporary workers do not receive health insurance, but they comprise a small part of the population. All citizens are entitled to insurance.

[3] Countries such as the UK have moved to more of a decentralized provision system in recent years. Hospitals have been set up as private trusts, for example, and physicians are no longer salaried.

[4] Since 1996, health insurance coverage has been required in Switzerland, but before then it was subsidized so heavily that essentially everyone purchased it.

[5] Income tax rates can range as high as 40 percent, but the income level at which these rates are reached are past the cap on earnings subject to the payroll tax.

Table 1
Sources of health insurance coverage for the United States population

Source	Groups insured	Share of total population (%)	Share of total payments (%)
Public			
Medicare	Elderly; disabled; end-stage renal disease	13	22
Medicaid	Elderly; blind and disabled; poor women and children	10	15
Other*	Military personnel and their dependents	1	8
Private			
Employer sponsored	Workers and dependents	56	53
Nongroup	Families	6	
Uninsured		16	2

* Other public spending includes non-insurance costs such as public hospitals, the Veterans Administration, etc.

Source: Authors' calculations based on data from Department of Health and Human Services, National Health Accounts (medical spending), and from Employee Benefit Research Institute (insurance coverage).

subsidy to employer-provided insurance, as opposed to individually-purchased insurance, is substantial. The subsidy to employer-provided health insurance generally does not extend, however, to out-of-pocket payments made by employees. As a result, there are incentives to have generous insurance, paid for by employers, with few individual copayments. We return to the effects of this subsidy structure below.

The remaining 16 percent of the United States population is uninsured. The implications of being uninsured are a subject of vigorous debate [Weissman and Epstein (1994)]. Some of the uninsured (perhaps 4 percent) are eligible for public insurance (particularly Medicaid) but have chosen not to take up that insurance. Presumably, if these people become sick they will enroll in Medicaid.[6] Others will receive "uncompensated" care if they become sick – they will get emergency care if they need it, but they will not pay for it. The costs of uncompensated care then get shifted to people with insurance, for whom payments made exceed the cost of services provided. In this sense, the United States has a form of universal insurance coverage for catastrophic care, although the patchwork nature of that coverage is undoubtedly suboptimal. It also limits primary and preventive care for those without health insurance.

The last column of Table 1 shows the share of total payments that each group makes. As in any insurance policy, people may use more or less of the service than they pay

[6] Since it is difficult to deny treatment, providers have a strong interest to enroll eligible people in Medicaid, so that they can receive some payment for them.

for. This is particularly true for the uninsured, whose out-of-pocket payments are much lower than the cost of services they receive. The table reports the share of total payments made by each group; the share of services that is used by each group will be somewhat different. Because people insured through the public sector are older and sicker than people insured privately, and because some of the costs of the uninsured are passed on to the public sector, the public sector accounts for much more of medical spending than its demographic share of insurance coverage. Close to half of medical spending in the United States is paid for publicly. While this amount is extremely high relative to most goods and services in society, it is low by international standards for medical care. In OECD nations, governments generally pay for 75 to 90 percent of medical care.

Whether run publicly or privately, health insurance encounters fundamental problems that any insurer must face. Adverse selection, though diminished for government since some of its programs are so heavily subsidized that the vast majority choose to participate, still exists, and moral hazard affects governments no less than private insurers. Thus, when we discuss the optimal design of health insurance policies, we do not distinguish between public or private insurers. We return to public versus private insurance issues in the conclusion.

2. The principles of insurance

In this section and the next three, we discuss the optimal design of health insurance policies. Our perspective is that of an insurer – public or private – wanting to optimally insure its enrollees against the costs of treating adverse health outcomes.

The value of health insurance is rooted in the unpredictability of medical spending. While individuals know something about their need for medical services, the exact amount they will spend on medical care is to a significant degree uncertain. Medical spending is extremely variable. Table 2 shows the distribution of medical spending in the United States in 1987 [Berk and Monheit (1992)]. The top 1 percent of medical care users consume an average of nearly $50,000 each in a year (in 1987 dollars), and

Table 2
Distribution of medical spending, 1987

Share of distribution	Cumulative share of spending (%)
Top 1 percent	30
Top 5 percent	58
Top 10 percent	72
Top 50 percent	98
Total population	100

Source: Berk and Monheit (1992).

account for 30 percent of medical spending. The top 10 percent of users account for nearly three-quarters of total medical spending. The shorter the time period, of course, the greater is the percentage disparity in medical spending among individuals. But even looking over several years, the skewness of medical spending is substantial [Roos et al. (1989), Eichner, McClellan, and Wise (1998)]. In such a situation, insurance can significantly spread risks.

Risk-averse individuals will want to guard against the potential of requiring a substantial amount of medical care. One way to do this is to wait, borrow money for treatment should they get sick, and then repay the money when well. But borrowing when debilitated is difficult, since the individual may not live long enough or be healthy enough to repay the loan. The borrowing process, moreover, may also take more time than the sick individual has available. A reasonable alternative might be for individuals to save money when they are healthy to pay for medical care should they get sick. But some sicknesses are significantly more expensive than others. The substantial expenses of very severe illness make saving prior to illness impractical as a protective measure. All of us would have to significantly curtail consumption to save up for expenses that would be borne by only a few. The natural solution is to insure against the possibility of medical illness by pooling risks with others in the population. Annual consumption would be reduced only by the premium, the average cost of care.

Risks to health have always been with us, but health insurance is a relatively new phenomenon, only becoming economically significant in the postwar era. Fire and life insurance were well developed by the end of the 19th century, and marine insurance was already being written in the 12th century. There was little role for health insurance in earlier eras, however, since expensive medical treatments could accomplish little for health. Insurers also feared they could not control individual use of medical services if the services were insured. Once effective hospital care – an extremely expensive commodity – became possible, significant health insurance became desirable and inevitable.

2.1. Insurance with fixed spending

The simplest insurance situation is one where sickness entails a fixed cost and insurance is priced at its actuarial cost. Imagine a situation where initially identical individuals are either healthy or sick in a period of one year. There is one disease. People are healthy with probability $1 - p$, in which case they require no medical care. People get sick with probability p. Let $d = 0$ or $d = 1$ indicate whether absent medical care the person is healthy or sick. Treatment of a person who is sick requires medical spending of m. The after-expenditure health of a sick person is $h = H[d, m]$. To simplify exposition, we assume that medical spending restores a person to perfect health, so that $H[1, m] = H[0, 0]$.

Before proceeding, we alert the reader to our use of mathematics. We use mathematics to derive statements precisely. We also endeavor to explain all of our results intuitively. Thus, readers who wish to skip the mathematical portions of the chapter can still follow the central arguments.

Individuals receive utility, u, which depends on their consumption, x, and their after-treatment health, h. Thus we have $u = U(x, h)$. Assume, for simplicity, that people have exogenous income endowments, y; and that they can neither borrow or lend. Thus, an individual's consumption is what is left over after paying medical expenditures, or if insured, his insurance premium, π. Thus, for uninsured people, $x = y$ when healthy and $x = y - m$ when sick. For insured people, $x = y - \pi$ whether healthy or sick. We use the subscripts I and N to indicate whether the individual is insured or not insured.

Let $U(x) \equiv U(x, H[0,0])$; i.e., it is the reduced form utility function for consumption given perfect health. In the absence of insurance, an individual's expected utility is given by:

$$\begin{aligned} V_{\mathrm{N}} &= (1-p)U\big(y, H[0,0]\big) + pU\big(y-m, H[1,m]\big), \\ &= (1-p)U(y) + pU(y-m), \end{aligned} \tag{1}$$

where the second equality follows from the assumption that medical care restores the person to perfect health.[7] We assume that U has the standard property that utility is increasing in consumption albeit at a declining rate: $U' > 0$ and $U'' < 0$. We further assume that medical expenditures are worthwhile even if the individual is not insured.

Suppose the individual purchases insurance against the risk of being sick. For an insurance company to break even, the fair insurance premium would have to be $\pi = pm$. The insurance company collects the premium each year and pays out m when the individual is sick. If an individual chooses this policy, his utility would always be:

$$V_{\mathrm{I}} = U(y - \pi). \tag{2}$$

Using a Taylor series expansion of Equation (1),[8] we can approximate that equation as:

$$V_{\mathrm{N}} \approx U(y-\pi) + U'\big(U''/2U'\big)\pi(m-\pi). \tag{3}$$

Therefore,

$$\text{Value of Insurance} = (V_{\mathrm{I}} - V_{\mathrm{N}})/U' \approx (1/2)\big(-U''/U'\big)\pi(m-\pi). \tag{4}$$

[7] Assuming that medical expenditure is worthwhile, this analysis actually requires a less stringent condition. The same equation would apply if restored health imposed a fixed utility cost, k, relative to initial perfect health, so that $U(c, H[0,0]) = U(c, H[1,m]) + k$ for all c.

[8] The Taylor series is taken about the level of income net of insurance premiums. From Equation (1), $V_{\mathrm{N}} \approx (1-p)[U(y-\pi) + U'\pi + (1/2)U''\pi^2] + p[U(y-\pi) - U'(m-\pi) + (1/2)U''(m-\pi)^2]$. Collecting terms, this simplifies to $V_{\mathrm{N}} \approx U(y-\pi) + U'\{(1-p)\pi - p(m-\pi)\} + (1/2)U''\{(1-p)\pi^2 + p(m-\pi)^2\}$. The term $(1-p)\pi - p(m-\pi)$ is zero. The term $(1-p)\pi^2 + p(m-\pi)^2$ can be expanded as $(1-p)\pi^2 + pm^2 - 2pm\pi + p\pi^2$. Since $pm = \pi$, this simplifies to $pm^2 - \pi^2 = \pi(m-\pi)$.

The left hand side of Equation (4) is the difference in utility from being uninsured relative to being insured, scaled by marginal utility to give a dollar value for removing risk. The right hand side is the benefit of risk removal. Here, $(-U''/U')$ is the *coefficient of absolute risk aversion*; it is the degree to which uncertainty about marginal utility makes a person worse off. Because $U'' < 0$ and $U' > 0$, this term is positive. The term $\pi(m - \pi)$ represents the extent to which after-medical expenditure income varies because the person does not have insurance. It too is positive. The product of terms on the right hand side of Equation (4), therefore, is necessarily positive, implying that fair insurance is preferred to being uninsured. The dollar value of risk spreading increases with risk aversion and with the variability of medical spending.

The intuition supporting this result is that risk averse individuals would like to smooth the marginal utility of income – to transfer income from states of the world where their marginal utility is low to states of the world when their marginal utility is high. In the absence of insurance, a person's marginal utility of income when healthy is $U'(y)$ and when sick is $U'(y-m)$. Since marginal utility falls as income increases, marginal utility is lower when healthy than when sick. Transferring income from healthy states to sick states until marginal utility is equalized maximizes total utility, assuming fair insurance. Health insurance carries out this transfer, charging premiums up front and reimbursing expenditures later.[9]

There is a diagrammatic way to make the same point; it is shown in Figure 2. We think of the two states of the world – being sick and being healthy – as if they were two goods. Individuals would like more consumption in each state. In the absence of any probability of being sick, people would be able to consume y in each state. Because of required medical spending, however, people can only consume $y - m$ when sick. This is shown as point E in the figure.

[9] The situation is more complex when medical spending fails to restore the person to perfect health, and the marginal utility of income is affected by health status. Suppose that when sick a person still needs medical spending of m, but that his after-expenditure health remains below what it would be had he never got sick; i.e., that $H[1,m] < H[0,0]$. Expected utility for people without insurance is given by $V_{\mathrm{N}} = (1-p)U(y, H[0,0]) + pU(y-m, H[1,m])$, and the marginal utilities of income are $U_x\ (y, H[0,0])$ when healthy and $U_x(y-m, H[1,m])$ when sick, where the subscripts indicate partial derivatives. Because the marginal utility of income may be affected by health and health varies across sickness states, it is not clear how much insurance the person will want. If people attach little value to money when sick – for example, if there are few pleasurable activities they can engage in – they may not want any health insurance at all. Alternatively, if the value of money when sick is particularly high, say because aides are needed to carry out the activities of daily life, people may want more than full insurance against medical expenditures.

This example highlights the difference between *medical care insurance* and what, if we used a strict interpretation, would be labeled *health insurance*. Health insurance transfers money across people – generally from the healthy to the sick. The money can be used to purchase medical services the individual otherwise could not afford, or to allow the individual to purchase more of other goods and services after medical care has been paid for. But health insurance cannot guarantee that an individual's health will be unaffected by outside factors. Insuring one's health is technologically infeasible.

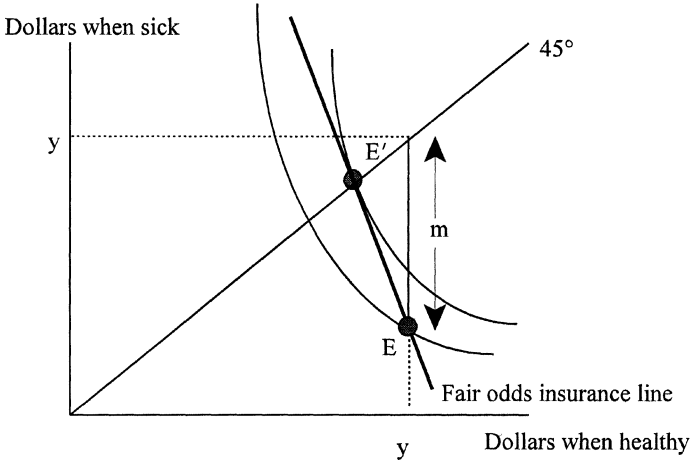


Figure 2. The welfare gains from health insurance.

The fair odds insurance line is the individual's implicit budget constraint. It is drawn for the case where $p = 0.2$. The slope of the line is $-1/p$, or -5.[10] The indifference curve for consumption is also steeply sloped, recognizing that the sick state is unlikely to arise. Thus, people are not willing to give up much consumption when healthy to get consumption when sick. A person can trade consumption when sick for consumption when healthy, at a rate given by the insurance premium. People will choose to purchase some insurance. If insurance is priced actuarially fairly, individuals will choose to be fully insured – they will have the same consumption when sick as when healthy. This optimum is shown as point E′ in the figure. People are better off at E′ than they are at E; they have moved to a higher indifference curve.

In our simplified world, the optimal insurance policy is an *indemnity* policy – it pays a fixed amount of money for a particular condition when the individual is sick. The amount paid equals the cost of the appropriate treatment for the person's disease; if there is more than one disease, the payments vary. Since each disease requires a fixed amount of care – there is no more nor less that a person can consume – there are no wasted resources in the policy; the indemnity insurance plan is efficient. Beyond its efficiency properties, the indemnity policy is the simplest health insurance policy. In effect, it operates as a contingent claims market; people get paid a specified amount depending on which contingency occurs [Zeckhauser (1970)].

Health insurance started off as a quasi-indemnity policy – in most cases paying a fixed amount per day in the hospital. The first Blue Cross policies, for example, were

[10] A fair insurance policy that charges \$1 each year and pays an amount k when sick is defined by: $pk + 1 = 0$. Thus, $k = -1/p$. Some authors assume the insurance payment is made only when the person is healthy, in which case the fair odds policy is defined by: $pk' + (1 - p) = 0$, or $k' = -(1 - p)/p$.

developed just before and during the Great Depression. These policies, run by hospitals, guaranteed a certain number of hospital days per year (for example, 21 days) for an annual premium (for example, \$5 to \$10 in the early 1930s). After World War II, life insurance companies entered the health insurance market, driven by the profits of Blue Cross policies and the expanding demand for health insurance resulting from its favorable tax treatment. These nascent health insurers offered indemnity policies as well, limiting their potential losses by fixing the maximum amount they would pay per hospital day.

3. Moral hazard and principal-agent problems

Health insurance must address several problems beyond risk spreading. We now turn to some of these challenges.

3.1. Moral hazard

Moral hazard refers to the likely malfeasance of an individual making purchases that are partly or fully paid for by others [Arrow (1965), Pauly (1968, 1974), Zeckhauser (1970), Spence and Zeckhauser (1971), Kotowitz (1987)].[11] He will overspend; i.e., he will use more services than he would were he paying for the medical care himself. Since insurance is an arrangement where others pay for the lion's share of one's losses, it creates a moral hazard to use additional medical resources. The designation moral hazard, a disquieting term, frequently connotes some moral failure of individuals, but this is not meant to be so. Indeed, Kenneth Arrow (1985) employs the less judgmental and more informative term "hidden action" for moral hazard.

Moral hazard is a concern because it conflicts with risk-spreading goals. Insurance is valuable because it allows people to transfer income from when they need it less to when they need it more. But this transfer is not perfect because people increase their consumption of medical care when it is subsidized. This creates an inherent second-best problem in designing insurance policies: insurers must trade off the benefits from spreading more risk against the cost of increased moral hazard. We formalize this Lesson 1 about health insurance:

> *Lesson* 1: *Risk spreading versus incentives*. Health insurance involves a fundamental tradeoff between risk spreading and appropriate incentives. Increasing the generosity of insurance spreads risk more broadly but also leads to increased losses because individuals choose more care (moral hazard) and providers supply more care (principal-agent problems).

[11] The theory of moral hazard, if not the words, goes back at least to Adam Smith: "The directors of such companies, however, being the managers rather of other peoples' money than of their own, it cannot well be expected, that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own... Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company" [Smith (1776, p. 700)].

Moral hazard, or hidden action, emerges in one form in the risks that individuals choose to take. People may take worse care of themselves when they have insurance than if they do not. If their actions were readily observed, the insurance company would merely not pay off were they reckless or negligent. But individual actions are difficult to observe; they are hidden. The extent of moral hazard in terms of actions that affect health may not be large for health insurance in most instances, since the uncompensated loss of health itself is so consequential.[12] Thus, it would be surprising if people smoked because they knew health insurance would cover the costs of lung cancer.

Hidden action also arises because individuals may get treatments they would not pay for themselves. Though the action itself (seeking medical care) is not hidden, the motivation behind it is.[13]

Optimal insurance plans would pay for treatment only if the individual would have chosen the same treatment had he borne the full bill. The thought experiment here is whether the person would pay for the medical expenditure in expectation, before he knew his condition. For example, suppose that a person has income of $25,000, and faces a 1 percent probability he will have a serious illness. If he could commit in advance, he would agree to receive $50,000 of medical care when sick in exchange for a $500 premium. If fully insured, however, the individual will choose to consume $60,000 of care. The moral hazard in this example is $10,000 – the additional spending beyond the optimal amount of care he would contract for in advance of being sick.

In the terminology of demand theory, moral hazard is the *substitution effect* of people spending more on medical care when its price is low, not the *income effect* of people spending more on medical care because of insurance, by efficiently transferring resources from the healthy state to the sick state, makes them richer when sicker [De Meza (1983)]. In the example considered, say the individual would have spent half his income, $12,500, on medical care in the absence of insurance. Insurance thus raises medical spending by $47,500, but only a fraction of this increase is due to moral hazard.

If some fixed m were the known optimal medical expenditure for any sick person, insurance plans would experience no moral hazard. They could simply pay m in medical expenditures to or for those who are sick. Moral hazard arises because medical needs are not fully monitorable, and different people with the same condition have different optimal expenditures, at least as best the insurance company can determine. Suppose that the optimal medical expenditure for treating a particular condition is m_i, which varies across people, indexed by i. The insurance company requires the individual to pay a coinsurance amount $c(m)$ for medical care received. The rest of the care, $m - c(m)$, is paid by the insurer. In effect, the insurer takes the individual's medical expenditure

[12] This does not mean that people will not smoke or faithfully take their medications. But there is no moral hazard if their actions would be the same if they had no health insurance, i.e., if these health-harming behaviors are inelastic with respect to cost sharing.

[13] Moral hazard also results from patients making less effort to search for low-cost providers. For example, when patients pay but one-fifth of the cost of their drugs, they will have weak incentives to switch to generic brands or stray beyond the local pharmacy.

to be a signal of his true medical needs; the coinsurance payment creates the necessary costs to have signaling operate.

Two polar extremes for the form of $c(m)$ are commonly found. The first is the indemnity policy discussed above: the insurer pays a fixed amount, call it m^*, and the individual pays $c(m) = m - m^*$. The second is full insurance: the insurer pays the full costs of medical care, regardless of its cost, and the individual pays nothing (i.e., $c(m) = 0$). The full insurance policy removes all risk from the insured, but engenders greater moral hazard.

To understand the optimal insurance policy, consider a case where an indemnity policy is not optimal. Suppose that rather than being healthy or sick, the individual has a range of potential illness severities, s, with s distributed with density function $f(s)$. Health is given as before by $h = H[s, m]$. The patient's s will determine the optimal treatment. The insurer cannot observe s, however. Thus, making a fixed indemnity payment to anyone sick is not optimal. The *ex ante* utility function for the insured consumer is:

$$V_{\mathrm{I}} = \int U\big(y - \pi - c(m(s)), H[s, m(s)]\big) f(s)\,\mathrm{d}s, \tag{5}$$

where $m(s)$ tells how much medical care an individual with condition s chooses to receive.

We consider first the optimal policy – the amount of medical services the person would like to contract for if he could write a perfect state-contingent contract and thereby eliminate moral hazard. When s is observable, the coinsurance rate depends only on s, hence can be written as $c(s)$. The individual will choose $m^*(s)$ maximum feasible utility:

$$\mathrm{Max}_{m(s)} \int U\big(y - \Pi - c(s), H[s, m]\big) f(s)\,\mathrm{d}s, \tag{6}$$

where $\Pi = \int (m(s) - c(s)) f(s)\,\mathrm{d}s$. The solution to this problem sets

$$H_m U_H = E[U_x], \tag{7}$$

where the subscripts denote partial derivatives and $x = y - \Pi - c(s)$. The left-hand side represents the gain in utility from spending another dollar on medical care; it is the product of the effect of medical care on health and the effect of health on utility. The right hand side is a weighted average expectation of the marginal utility of consumption in different illness states, namely:

$$E[U_x] = \int U_x\big(y - \Pi - c(s), H[s, m]\big) f(s)\,\mathrm{d}s. \tag{8}$$

Equation (7) says that with the optimal first-best policy, the expected marginal utility gained from an additional dollar of medical care in each state of the world equals the utility cost of a dollar.[14]

In the case where the marginal utility of income does not depend on the health state,[15] imposing a coinsurance payment in any health state, i.e. a variable $c(s)$, increases the variability of income and thus reduces expected utility. The optimal policy for this commonly studied case in thus no coinsurance, and a payment $m^*(s)$ that fully reimburses optimal spending in each state.[16]

Now consider a situation where severity of illness is not monitorable, hence the optimal policy just discussed cannot be implemented. At the time the consumer is seeking medical care, he alone knows his severity. We assume the consumer treats the insurance premium as fixed – nothing he does will raise or lower his insurance premium that year. Further, we assume for now that individuals are not penalized in future years for additional medical spending this year, because expected future changes in costs are spread equally over everyone in the group. The cost to the consumer of another dollar of medical expenditure will be $c'(m)$.[17] The sick consumer will therefore choose medical care utilization to maximize utility when sick. Thus, he will choose $m^{\#}(s)$ as the m which maximizes utility given knowledge of s:

$$\mathrm{Max}_{m(s)} U\big(y - \Pi - c(m), H[s,m]\big) \quad \text{for each } s. \tag{9}$$

The solution to this problem will depend on the specific s the individual has realized, and is given by the first order condition:

$$H_m U_h = c'(m) U_{zx} \quad \text{for each } s. \tag{10}$$

The left-hand side once again represents the gain in utility from spending another dollar on medical care. The right-hand side is the utility cost to the individual from spending

[14] This assumes that these functions are well behaved, hence that local optima are global optima. Some medical expenditures may offer increasing returns over a relevant range. For example, it may cost $200,000 to do a heart transplant, with $100,000 accomplishing much less than half as much. Efficiency then requires the insurance program spend at least to the minimum average cost of benefits point, or not at all.

[15] This case would arise if utility is additively separable between income and health.

[16] If utility does depend on the health state, for example, if a disabled person needs more non-medical services, then optimal coincurance will actually pay the individual when disabled.

[17] The structure of the insurance plan may present the insured with a range of decreasing marginal cost. Say a plan has a deductible of $600 with a copayment of 20% beyond that point, a common structure. The insured can receive $600 of benefits for $600, but $1200 of benefits for $720. Say the individual solves, and finds a $400 expenditure is locally optimal. He must also look globally to the optimal expenditure beyond $600, which may be superior. Recognizing that using up a deductible gets one to a range of lower costs, gives the insured an interesting dynamic optimization program where there are two benefits from spending below the deductible: (1) the health care itself, and (2) the increased potential for getting to the low-cost range [Keeler, Newhouse, and Phelps (1977)].

that dollar; it is the product of the out-of-pocket cost of medical care and the utility loss from losing that dollar for consumption.

Comparing Equations (7) and (10) shows the loss due to moral hazard. When $c'(m) < 1$, as it will be when marginal spending is in any way insured, people will overconsume medical care when sick and thus pay more for health insurance than is optimal.[18]

3.1.1. Evidence on the price elasticity of medical care demand

How does an individual's demand for medical care respond to his required out-of-pocket expenses? Economists used to differ on this question. Table 3 details estimates of the elasticity of demand for medical care.[19] A substantial literature in the 1970s estimated the elasticity of demand for medical care using cross-sectional data, or cross-sectional time series data. Pre-eminent among these papers are Feldstein (1971), Phelps and Newhouse (1972b), Rosett and Huang (1973), and Newhouse and Phelps (1976). Feldstein (1971) was the first statistically robust estimate of price elasticities using time-series micro data, in this case on hospitals. Feldstein identified the effect of coinsurance rates on demand using state-variation in insurance coverage and generosity, estimating a demand elasticity of about -0.5. The subsequent papers use patient-level data and more sophisticated study designs. The elasticities that emerged from these papers ranged from as low as -0.14 [Phelps and Newhouse (1972b)] to as high as -1.5 [Rosett and Huang (1973)]. The implication of this range of elasticity estimates was that moral hazard was likely a significant force.

This estimation literature suffered from two major difficulties, however. First, the generosity of health insurance at either the state or the individual level might be endogenous. Generous insurance might boost utilization of medical services, as posited; or alternately, areas where people desire or need more medical care may also be areas where people demand more health insurance. One cannot separate these two effects statistically without an instrument for the rate of insurance coverage in an area, but such instruments were not easy to find. Second, the studies typically failed to distinguish average and marginal coinsurance rates. Usually for data reasons, most of these studies related medical spending to the *average* coinsurance rate in an area. But theory predicts that medical spending should relate to the *marginal* coinsurance rate. Because insurance policies are non-linear, average and marginal prices may differ substantially.[20] As a result of these problems, as late as the 1970s many critics still believed that medical care was determined by "needs" and no other economic factors, i.e., that demand was totally

[18] This can be derived by taking expectations of both sides of Equation (10) and comparing to Equation (7). There is also a risk-bearing loss when severity, is not monitorable, as reflected by the term U_z in (10) as opposed to $E(u_x)$ in (7).

[19] Zweifel and Manning (2000) discuss the elasticity of demand for medical care in more detail.

[20] Of course, if individuals are appropriately forward looking, it is the *expected* marginal coinsurance rate at the end of the year that should affect behavior, rather than the ostensible marginal coinsurance rate at the time services are used.

Table 3
Estimates of the elasticity of demand for medical care

Paper	Data	Restrictions	Estimation method	Total price elasticity	Visits price elasticity	Quality price elasticity
Feldstein, P.J. (1964)	1953, 1958 Health Information Foundation and NORC surveys	general care	cross-section estimates of physician visits	−0.19 (physician visits)		
Feldstein, M.S. (1970)	BLS survey; NCHS 1963–1964 survey; physician interviews	aggregated physician service data	time-series regression	1.67 (physician services)		
Rosenthal (1970)	1962 sample of New England hospitals	68 of 218 general, short-term hospitals	univariate estimates for short-term care categories	0.19 to −0.70		
Feldstein, M.S. (1971)	AHA survey of hospitals, 1958–1967, NCHS 1963–1964 survey	all hospitals, aggregated by state	time-series regression	−0.49 for total bed days	−0.63 for visits to hospital	
Davis and Russell (1972)	1970 guide issue of "Hospitals"	aggregated hospital outpatient care; 48 states' not-for-profit hospitals	cross-sectional estimates	−0.32		
Fuchs and Kramer (1972)	1966 Internal Revenue Service tabulations	physician services, aggregated into 33 states	TSLS: IVs are number of medical schools, ratio of premiums to benefits, and union members per 100 population	−0.10 to −0.36		
Phelps and Newhouse (1972a, 1972b)	Palo Alto Group Health Plan, 1966–1968	physician and outpatient ancillary services	natural experiment: introduction of coinsurance	−0.14* OLS, −0.118 Tobit (physician visits)		

continued on next page

Table 3, *continued*

Paper	Data	Restrictions	Estimation method	Total price elasticity	Visits price elasticity	Quality price elasticity
Scitovsky and Snyder (1972)	Palo Alto Group Health Plan, 1966–1968	physician and outpatient ancillary services	natural experiment: introduction of coinsurance	−0.060* (ancillary)	−0.14* (physician visits)	
Phelps (1973)	verified data from 1963 CHAS (University of Chicago) survey	hospitalization and physicians' services	cross-sectional Tobit estimates	not significantly different from zero		
Rosett and Huang (1973)	1960 Survey of Consumer Expenditure	hospitalization and physicians' services	cross-sectional Tobit estimates	−0.35 to −1.5		
Beck (1974)	random sample of poor population of Saskatchewan	physicians' services	natural experiment; introduction of co-payments	−0.065*		
Newhouse and Phelps (1974)	1963 CHAS survey	employeds' hospital stays within coverage	cross-sectional OLS (TSLS estimates insignificant)	−0.10 (length of stay)	−0.06 (physician visits)	
Phelps and Newhouse (1974)	insurance plans in US, Canada, and UK	general care, dental care, and prescriptions	arc elasticities across coinsurance ranges	−0.10		
Newhouse and Phelps (1976)	1963 CHAS survey (larger sample than in previous work)	employeds and non-employeds	cross-sectional OLS (TSLS estimates insignificant)	−0.24 (hospital), −0.42 (physician)		
Scitovsky and McCall (1977)	Palo Alto Group Health Plan, 1968–1972	physician, outpatient ancillary services	natural experiment: coinsurance increases	−2.56* (ancillary)	−0.29* (physician visits)	
Colle and Grossman (1978)	1971 NORC/CHAS health survey	pediatric care	cross-sectional estimates	−0.11	−0.039	

continued on next page

Table 3, *continued*

Paper	Data	Restrictions	Estimation method	Total price elasticity	Visits price elasticity	Quality price elasticity
Goldman and Grossman (1978)	1965–1966 Mindlin–Densen longitudinal study	pediatric care	hedonic model		−0.060 (compensated −0.032)	−0.088 (compensated −0.085)
McAvinchey and Yannopoulos (1993)	waiting lists from UK's National Health Service	acute hospital care	dynamic intertemporal model	−1.2		
Newhouse et al. (1993)	RAND Health Insurance Experiment	general care	randomized experiment	−0.17 to −0.31 (hospital), −0.17 to −0.22 (outpatient)		
Bhattacharya et al. (1996)	1990 Japanese Ministry of Health and Welfare survey	outpatient visits	Cox proportional hazards model	−0.22		
Cherkin et al. (1989)	Group Health Cooperative of Puget Sound	non-Medicare HMO patients	natural experiment: introduction of copayments	−0.035* (all visits), −0.15* to −0.075* (preventive)		
Eichner (1998)	1990–1992 insurance claims from employees and dependents of a Fortune 500 firm	employees aged 25 to 55	one-and two-stage Tobit regressions of out-of-pocket costs	−0.32		
SUMMARY				−0.20	−0.05 to −0.15	

* Elasticities computed according to appendix of Phelps and Newhouse (1972b).

inelastic, although others believed that the demand elasticity was substantial – perhaps -0.5 or more.

To address these problems, the United States government funded a social insurance experiment, designed to estimate the demand elasticity for medical care. The Rand Health Insurance Experiment [Newhouse et al. (1993), Zweifel and Manning (2000)] randomized nearly 6,000 people in 6 areas to different insurance plans over a 3- to 5-year period in the early 1970s. The insurance plans varied in contractual levels of cost sharing. Elasticity estimates were formed by comparing utilization in the different plans. The Rand Experiment found an overall medical care price elasticity of about -0.2. This elasticity is statistically significantly different from zero, but noticeably smaller than the prior literature suggested. Sound methodology, supported by generous funding, carried the day. The demand elasticities in the Rand Experiment have become the standard in the literature, and essentially all economists accept that traditional health insurance leads to moderate moral hazard in demand. The Rand estimates are also commonly used by actuaries in the design of actual insurance policies.

3.1.2. Coinsurance in practice

The indemnity policy, which characterized health insurance at its inception, became outdated over time. With increased medical technology, the range of optimal spending within a given condition became great. Indemnity policies left individuals bearing too much risk. As a result, insurance structures moved from indemnity payments to a service benefit policy – a policy that covers all medical expenses, with some cost sharing. Service benefit policies grew steadily in importance in the post-war period, reaching their height in the early 1980s.

Service benefit policies use three cost-sharing features, sometimes in concert: the *deductible*, the *coinsurance rate*, and the *stop loss* amount. Figure 3 and Table 4 show how these cost sharing features operate. The deductible is the amount that an individual must pay before the insurance company pays anything. The deductible is usually set annually; the typical deductible in 1991 was about $200 for an individual and $500 for a family. Consumers pay the full price for care consumed under the deductible. The coinsurance rate is the percentage of the total bill above the deductible that a patient pays. Nearly all indemnity plans had a coinsurance rate of 20 percent. The coinsurance is paid until the patient reaches the stop loss – the maximum out-of-pocket payment by the person in a year. A typical stop loss in an indemnity policy was about $1,000 to $1,500 in a year.

In addition to these features, many policies impose further cost sharing through caps on various types of expenditures. For example, policies may permit 8 mental health visits per year, or have a $1 million lifetime limit on overall medical expenditures. Such provisions discourage use, and may deter high cost users from selecting the insurance plan, and providers from turning expensive cases into subsidized meal tickets. Table 4 details the frequencies with which various policy features were found in insurance policies in 1991.

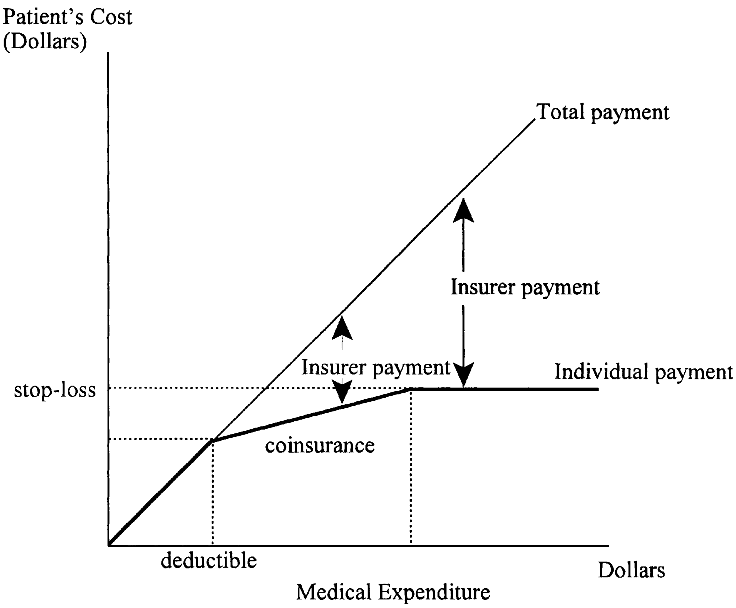


Figure 3. Cost sharing under indemnity insurance.

Table 4
Risk-sharing features of indemnity insurance policies, 1991

Characteristic	Average/percent
Deductible	
Individual	$205
Family	$475
*Coinsurance rate**	
<20 percent	13%
20 percent	78%
>20 percent	4%
Stop loss	
⩽$500	21%
$501–$1,000	30%
$1,001–$2,000	32%
>$2,000	17%
Maximum lifetime benefit – individual	
⩽$250,000	9%
$250,001–$999,999	6%
⩾$1,000,000	85%

Source: HIAA Employer Survey, 1991.
* Remaining responses are "rate varies" and "other".

Somewhat misleadingly, the service benefit policy is frequently called an "indemnity insurance plan" by economists, with the system that developed to provide this policy termed the "indemnity insurance system". In fact, true indemnity health insurance policies (a fixed payment per disease) had existed but were largely replaced by the service benefit policy. For consistency with other literature, we follow this nomenclature despite its inaccuracies. This nomenclature is particularly unfortunate since recently insurance has been moving back towards the indemnity model, frequently with the risk of above-average spending being placed on the provider rather than the patient. We discuss regimes of provider responsibility in Section 4.

3.1.3. Optimal insurance given moral hazard

Knowledge of the utility function and the parameter values that determine medical spending elasticities can be combined to design the optimal insurance policy – the actuarially fair policy that maximizes expected utility subject to the constraint that individuals will act in a self-interested fashion, i.e., that moral hazard will operate. Such a policy is inherently second-best; in calibrating its level of generosity, it balances the utility benefits of greater risk-sharing across people against the moral hazard costs incurred. The insurer's challenge is to define the function of risk sharing by insureds, the $c(m)$ function, that maximizes expected utility.

To analyze the optimal policy, we assume patients differ in the severity of their illness.[21] The insurer will seek to find the $c^*(m^\#)$ function that produces the maximum possible expected utility with:

$$E[U^*] = \mathrm{Max}_{c(m\#)} \int U\big(Y - \pi - c^*\big(m^\#\big), H\big[s, m^\#\big]\big) f(s)\,\mathrm{d}s, \tag{11}$$

where $m^\#$ is defined as the solution to Equation (9). Because insurers cannot determine an individual's health state, the insurance policy cannot differentiate payments on the basis of illness severity.

An additional constraint operates on the insurance company: premiums must cover expected costs. Thus,

$$\pi = \int \big[m^\#(s) - c\big(m^\#(s)\big)\big] f(s)\,\mathrm{d}s. \tag{12}$$

The optimal insurance policy can be formally written as a problem in dynamic optimization [Blomqvist (1997)].[22] Alas, this is a complicated problem, whose algebra is

[21] Moral hazard arises, let us remember, because individuals differ in unmonitorable ways. Thus it could be on income, on health status, or on some aspect of preferences.

[22] The problem is formally analogous to the optimal tax problem in public finance when ability is unobservable [Mirrlees (1971), Diamond (1998)].

Table 5
Estimates of the optimal insurance policy

Author	Optimal policy
Feldstein and Friedman (1977)	58 percent coinsurance rate
Buchanan, Keeler et al. (1991)	\$200 deductible; 25 percent coinsurance rate

Newhouse et al. (1993)*	\$200 to \$300 deductible;
	25 percent coinsurance rate;
	\$1,000 stop loss (assumed)

Manning and Marquis (1996)	25 percent coinsurance rate; >\$25,000 stop loss
Blomqvist (1997)**	Cost sharing declines from 27% at roughly \$1,000 of spending to 5% above roughly \$30,000

* Amounts are in 1983 dollars.
** Amounts are based on the Rand Health Insurance Experiment data.

not particularly revealing. The analytic solution balances two factors. The first is the reduced overconsumption from making people pay more out of pocket for medical care. If the coinsurance rate is increased in some range, people in that range pay more for medical care, as do people at all higher levels of spending (because their coinsurance rates have been increased). This increase boosts the efficiency of provision. Countering this, however, is a loss in risk spreading benefits. As people are made to pay more out of pocket, they are exposed to more risk, and this reduces their welfare. The optimal coinsurance rate balances these two incentives.

A small literature has simulated optimal insurance policies using this framework. Table 5 shows the results of these simulations. Table 5 reveals a wide range of disparities in optimal insurance policies. Some of the studies find that simulated insurance policies are substantially less generous than actual indemnity policies of the past 20 years [Feldstein and Friedman (1977), Blomqvist (1997)], while other studies find that they are about the same [Buchanan et al. (1991), Newhouse et al. (1993), Manning and Marquis (1996)].[23] The difference between these various estimates has not been fully reconciled, although one suspects that differing degrees of risk aversion and moral hazard are important. One suspects that real world policies will be more generous than optimal policies because of the tax distortions favoring more generous insurance: payments to insurance which are then made to providers are not taxed as income to employees, while wage and salary payments, which might be used to pay for medical care out-of-pocket, are. Indeed, other research shows that the benefits that employer health insurance policies offer are sensitive to employee tax rates [Pauly (1986)].

[23] The implication of the Blomqvist estimates for health insurance cost sharing depend on whether income losses are compensated or not.

A second important difference between real world and optimal policies is that the former almost invariably have a constant coinsurance rate, i.e., linear structures, whereas the latter do not. The optimal policy can be substantially superior. Blomqvist (1997), for example, finds that coinsurance rates should range from over 25 percent at low levels of spending to 5 percent at high levels of spending. There is likely a tradeoff between optimality and simplicity. Optimal policies can be very complicated, while real world situations are characterized by relatively simplistic structures.

If services or diseases differ in the degree of moral hazard they entail, the optimal insurance policy will differ by service or disease as well. Suppose, for example, there is a fixed number of diseases that a person can have and that moral hazard varies by disease. The insurance company can observe the disease of the person (e.g., cancer or appendicitis) but not the severity of illness within the disease. Then, the optimal insurance policy will have different cost sharing by disease [Zeckhauser (1970)]. Coinsurance formulas could just as easily depend on service (e.g., outpatient psychiatry) or locale of medical care (e.g., hospital care).[24] In practice, elasticity estimates do differ across services. The Rand Health Insurance Experience found higher demand elasticity for outpatient care than for inpatient care, and within outpatient care a greater demand elasticity for mental health care. Most health insurance policies, including Medicare, draw distinctions between services in their coinsurance schedules. Thus, Medicare has a separate hospital deductible, and private insurance plans frequently cover a fixed number of psychiatric visits.

Moral hazard is a significant concern in insurance policies but it is not one that necessarily argues for government intervention. Government insurance policies, after all, may engender just as much moral hazard as private insurance policies. There is a rationale for government to be involved in goods subject to moral hazard only if the government is better able to monitor or punish moral hazard than the private sector. This is not obviously the case in medical care.

3.2. Patients, doctors, and insurers as principals and agents

Thus far, we have implicitly assumed that patients choose the amount of medical care they want, knowing their illness, the range of possible treatments, and the prices of the treatments to them. But few patients are so well informed. In most cases of serious expenditure, it is the doctors who make the resource-spending decision, with patients and insurers bearing the costs; patients usually do not know the charge until the bill comes. Patients, physicians, and insurers are in a *principal–agent* relationship: the patient (principal) expects the doctor (agent) to act in his best interest when he is sick. Similarly, the

[24] This is analogous to the Ramsey rule of optimal taxation. The Ramsey rule states that optimal taxes on a set of commodities should be inversely related to the elasticity of demand for each commodity – in minimizing inefficiency, inelastic factors should be taxed more. The statement here is the equivalent but for subsidies instead of taxes.

insurer would like the doctor to act in its interests. Of course, patients also bear the insurance costs for seeking care, so that *ex ante* the patient's incentives and the insurer's incentives line up. But once the patient becomes sick and requires care, the parties' incentives diverge.

This three-player agency problem creates substantial problems for health insurance. To the extent that medical treatments are decided upon jointly by physicians and patients, the *supply side* of the health insurance policy (the rules about paying physicians) will matter along with the *demand-side* of the insurance policy (the rules about cost sharing for patients).

With the traditional service-benefit insurance policy, doctors and patients frequently have relatively congruent interests, which may differ from those of the insurer. Patients who face but a fraction of the costs they incur will desire excessive treatments. Service-benefit insurers usually pay more to physicians who provide more medical services. The result is that patients and physicians want essentially all care that improves health, respectively ignoring and welcoming resource expenditures. The view that physicians should do only what is best for the patient is codified in the Hippocratic Oath – providers should promote the best medical outcomes for their patients. Hippocrates said nothing about providing care the patient or society would have deemed *ex ante* to be wasteful.

Plato anticipated the application of agency theory to the health care arena by a goodly margin. He wrote that, "No physician, insofar as he is a physician, considers his own good in what he prescribes, but the good of his patient; for the true physician is also a ruler having the human body as a subject, and is not a mere moneymaker" (*The Republic*, Book 1, 342-D).[25] With the passage of 2,000+ years, fidelity to principals has slipped a bit, and new participants – insurers, government, employers, and provider organizations – have strode into the arena. But the principles are very much the same.

A more sinister view of the principal–agent problem contends that physicians manipulate patients into receiving more services than they would want, so that physicians can increase their income. This has been termed *supplier-induced demand* in the literature. An enormous amount of work in health economics has been devoted to the question of whether and to what extent suppliers induce demand. The empirical evidence on this issue is discussed by McGuire (2000). Lesson 1 notes the tradeoff between risk spreading and appropriate incentives applies on both the demand- and supply-sides of the market.

Increasingly, the arrows of responsibility among the players – who is agent, who principal – now point in all directions. For example, doctors now have responsibilities to other providers and insurers, not just to patients. Such added doctor responsibilities, primarily to hold down expenditures, ultimately enhance patient welfare, at least on an expected value basis, if not when the patient is sick. Insurers, acting for their customers as a whole, want to limit spending to only that care that is necessary; i.e., the care

[25] One might instead heed the warning of George Bernard Shaw nearly a century ago: "That any sane nation, having observed that you could provide for the supply of bread by giving bakers a pecuniary interest in baking for you, should go on to give a surgeon a pecuniary interest in cutting off your leg, is enough to make one despair of political humanity" [Shaw (1911)].

patients would select given a lump-sum transfer that depends on their condition and making them pay all costs at the margin. With patients, physicians and insurers pulling in different directions, a conflict over what care will be provided frequently results.

3.3. Transactions costs

Processing claims costs money; the more claims processed the more it costs. National estimates of medical expenditure suggest that 15 percent of insurance premiums are devoted to administrative expense.[26] Someone must read the bill, approve the spending, and pay the claim. Insurance companies seek to control these costs, and policies are designed accordingly.[27]

A major part of claims processing costs – monitoring, transferring money, and the like – are invariant to the size of the claim. Size-invariant costs are a greater percentage burden for small bills than for larger bills. This suggests limiting health insurance to larger claims and having individuals pay directly smaller expenses [Arrow (1963)]. This insight gives further justification to the widespread use of deductibles and coinsurance for small bills, and for the fact that historically insurance developed first for inpatient doctor and hospital charges, where bills are the largest.

4. Relationships between insurers and providers

The medical care system is a network, with patients, monies and information flowing from one party to another. The information flow to insurers, however, is not so rich that they can guarantee that only cost-effective care will be provided. Their monitoring difficulties provide the motivation for cost-sharing in insurance policies. But cost sharing has limited value: Patients do not make the most costly decisions, the Hippocratic Oath does not extend to conserving society's resources, and risk spreading considerations severely limit what charges can be imposed.

Return now to Figure 1, the Medical Care Triad. Working solely on the left side of the triangle, the demand side, these arguments suggest that passive insurers are unlikely to be able to limit utilization appropriately. Recognizing this, insurers also work the right side of the triangle – the supply side. Increasingly, insurers attempt to provide incentives for providers to limit spending. The incentives may be imposed at arm's length, as Medicare does with its DRG system: treat a simple heart attack, and a hospital gets paid a flat

[26] This includes the expenses of paying bills as well as marketing. Divisions between these sources of administrative expense are not very precise.

[27] Of course, individuals must also bear some costs in paying bills on their own, so it is not self evident which method of payment, individually or through insurance, is cheaper. But most people implicitly assume that insurers have additional transactions costs for paying bills beyond what individuals face. Thus, there is likely to be a net transactions cost to purchasing insurance. There are also transactions costs associated with selling the policy, but they do not vary with the magnitude of claims.

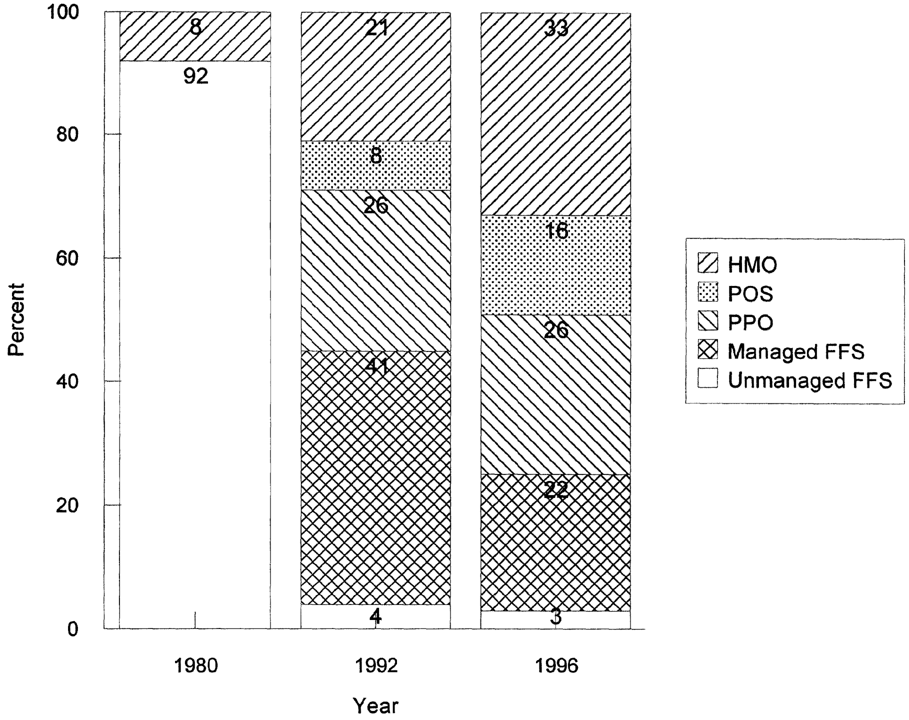


Figure 4. Changes in health plan enrollment. The sample is people with private (employer or individual) insurance. Source: Data are from Lewin-VHI.

amount, roughly $5,000. Or the insurer may form a contracting alliance with providers, as it does say with network HMOs. At the extreme, insurers and providers merge into a single entity. Uniting disparate organizations in this way enhances monitoring possibilities and better aligns incentives, but it also creates the potential for diseconomies of scope, e.g., requiring another layer of management when care is delivered.

The sweeping nature of insurer–provider interactions is indicated by Figure 4 [see also Glied (2000)]. In 1980, over 90 percent of the privately insured – i.e., employer- or self-paid – population in the United States was covered by "unmanaged" indemnity insurance. By 1996, that share had shrunk to a mere 3 percent.

Table 6 provides a taxonomy of different insurance-provider arrangements. The most limited arrangement is a "managed" indemnity insurance policy. It bundles a traditional indemnity policy with limited utilization review, for example requiring that non-emergency hospital admissions be precertified. At the most intrusive, insurers can seek to monitor care on a retail basis through tissue review committees, or on a statistical, wholesale basis by monitoring a physician or hospital's overall utilization. Such reviews can be used to refuse or reduce payment. Such intrusiveness by insurers may be unhelpful and, coming after-the-fact, may be ineffectual. It certainly is not welcome

Table 6
Key characteristics of insurance policies

Dimension	Indemnity insurance	Managed care		
		PPO	IPA/network HMO	Group/staff HMO
Qualified providers	Almost all	Almost all (network)	Network	Network
Choice of providers	Patient	Patient	Gatekeeper (in network)	Gatekeeper (in network)
Payment of providers	Fee-for-service	Discounted FFS	Capitation	Salary
Cost sharing	Moderate	Low in network; High out of network	Low in network; High out of network	Low in network; High/all out of network

Roles of insurer	Pay bills	Pay bills;
		Form network
			Pay bills;
			Form network;
			Monitor utilization
				Provide care

Limits on utilization	Demand-side	Supply-side (price)	Supply-side (price, quantity)	Supply-side (price, quantity)

to physicians. As Figure 4 shows, managed indemnity insurance, though non-existent in 1980, claimed a 41 percent share by 1992, but has fallen to 22 percent today.

Preferred Provider Organizations (PPOs), a second type of managed care, form a network of providers, including physicians, hospitals, pharmaceutical purveyors, and others, and control costs by securing discounts from them. The *quid pro quo* for the discounted fee is that insureds are steered to in-network providers. Out-of-network providers may get reduced coverage or no coverage at all. More typically, the patient's coinsurance or copayment rates are merely set lower for in-network providers. In 1991, for example, the typical PPO had an in-network coinsurance rate of 10 percent and an out-of-network coinsurance rate of 20 percent. PPOs usually impose preauthorization requirements as well, though they are rarely especially strict. As Figure 4 shows, PPO enrollment, zero in 1980, now makes up about one-quarter of the privately insured population.

Full integration creates the strongest link between insurance and provision. In the United States, these merged entities are called health maintenance organizations (HMOs). They sell their services directly to employers or individuals on an annual fee basis, and then they deliver care. There are three major types of HMOs. Within a group/staff HMO – the most common form, with Kaiser being the best known example – physicians are paid a salary and work exclusively for the HMO. The HMO may have hospitals on contract, or may run its own.

HMOs employ a range of mechanisms to limit utilization. They reflect the traditional economic instruments of regulation, incentives, and selection of types. HMOs frequently regulate physicians' practices, for example limiting the referrals they can make or the tests they can order. But the efficiency benefits of HMOs arise much more from aligning the incentives of provider and insurer, rather than through direct regulation. Some group/staff HMO physicians are salaried; as a result, they have a weaker incentive to provide marginal care than their fee-for-service counterparts. Moreover, HMOs monitor the services that physicians provide. They may reward parsimonious resource use directly with compensation, though more likely with perks or subsequent promotion. Extravagant users are kicked out of the network. Finally, since physicians differ substantially in their treatment philosophies, HMOs can select physicians whose natural inclination is toward conservative treatment.

Given the ability of HMOs to limit utilization on the supply-side, price-related demand-side limitations can be less severe. Cost-sharing to enrollees is generally quite low – typically about $5 to $10 per provider visit, although other forms of demand-side limitation survive (for example, patients may have to get approval from their internist before seeing a specialist).

Independent Practice Associations (IPAs), or Network Model HMOs, represent a more recent innovation in managed care.[28] These plans neither employ their own physicians nor run their own hospitals. Instead, they contract with providers in the community. By limiting the size of the network, the plans secure lower costs from willing providers. In addition, these plans employ stringent review procedures. For example, patients may need approval to receive particular tests. Finally, IPAs often provide financial incentives to limit the care that they provide. For example, some plans pay physicians on a "capitated" basis. The physician receives a fixed payment per patient per year. Out of this capitated stipend, the physician must pay for all necessary medical services, possibly including hospital services and prescription drugs. The physician's incentives for cost control become even more significant when all expenditures come out of his own pocket.

In many HMOs, patients can go outside of the network and still receive some reimbursement. This is termed a Point of Service (POS) option. But reimbursement out-of-network is not as generous as reimbursement within. Use of non-network services, for example, frequently requires a deductible followed by a 10 to 40 percent coinsurance payment.

As Figure 4 shows, HMO enrollment of all forms (including POS enrollment) has increased from 8 percent of the population in 1980 (then predominantly group/staff model enrollment) to nearly half of the privately insured population today.

This vertical integration in managed care, with insurers and providers linked or united, is virtually unheard of in insurance of other types. Auto insurance, for example, is an indemnity policy. People choose what coverage they will have, what deductibles will be in force, etc. When there is a crash, the insured and the adjuster get together, perhaps at the repair shop, to negotiate the cost of the repair. The insured or the repair shop, entities having no particular relationship to the insurer, are paid that amount, less

[28] Some IPAs are older, but their form gained popularity only recently.

any deductibles, which are the responsibility of the insured. After major crashes, cost-ineffective repairs are avoided by declaring the car a total loss, giving the wreck to the insurance company and reimbursing the owner.

But such a contingent claims system could not work with health care. The claims are more frequent and uncertainties much greater, making costs much harder and more expensive to estimate. “Scrapping” a human body is rarely an inexpensive or palatable proposition. The burgeoning links between insurers and providers in health care, we believe, are a response to the *a priori* difficulty of writing contingent claims contracts in the medical sector.

Vertical integration is also important because it can elicit price discounts. Managed care partly represents a price club. In exchange for an up-front fee, the patient gets to purchase goods at a significant discount. The discounts are secured through bulk purchase bargaining, or by directly hiring the sellers. In exchange for lower prices, patients precommit to receive care from a limited set of providers, or to pay harshly for the privilege of going elsewhere.

Finally, vertical integration is important because it fundamentally transforms the principal-agent conflicts in the medical system. Physicians no longer look out for the interests of just their patients, or perhaps their patients’ interests and their own. Now, physicians must watch out for the insurer as well. And patients must be more attuned to the incentives their physician is under. We note the integration of insurance and provision as the second lesson of health insurance:

> *Lesson* 2: *Integration of insurance and provision*. With medical care, unlike other insurance markets, insurers are often directly involved in the provision of the good in addition to insuring its cost. The integration of insurance and provision, intended to align incentives, has increased over time. Managed care, where the functions are united, is an extreme version. Under it, doctors have dual loyalties, to the insurer as well as to the patient.

4.1. Equilibrium treatment decisions in managed care

One can understand the impacts of managed care using a framework similar in spirit to what we described for patient cost sharing, only the physician’s choices are targeted. A typical physician payment, for example, is

$$\text{Payment} = R + r \cdot \text{Cost}. \tag{13}$$

Here, R is the prospective amount and r is the retrospective amount. A fully capitated system sets $r = 0$ and $R > 0$, while a fully retrospective system sets $R = 0$ and $r \geqslant 1$. Thus, the capitated system focuses solely on incentives; the retrospective system removes all risk from the doctor.

Changing to a capitated system might affect treatment decisions in several ways. One effect is to raise the physician’s “shadow price” for providing treatment – physicians might require a greater expected health benefit before providing care under managed

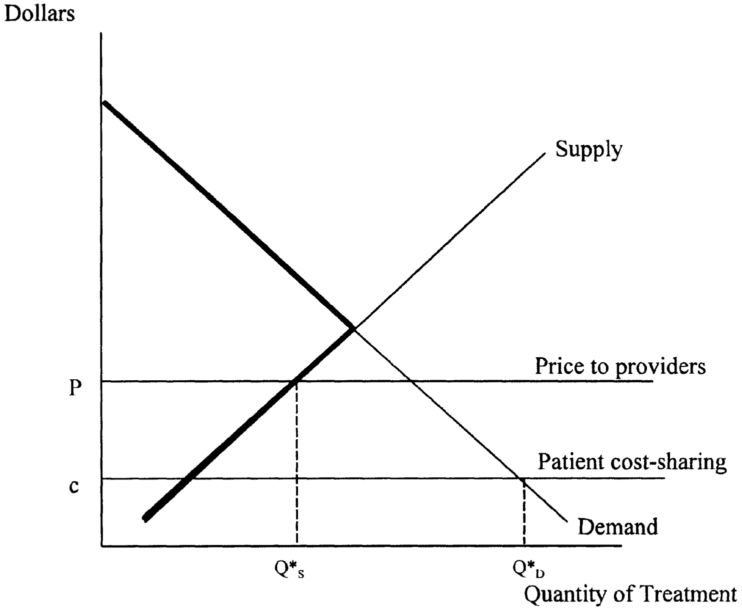


Figure 5. Conflict in quantities desired between providers and patients.

care than under traditional indemnity insurance [Frank, Glazer, and McGuire (1998), Keeler, Carter, and Newhouse (1998)]. This effect is particularly strong when the physician is capitated, and thus bears the marginal cost for providing additional care.

In addition, managed care might harmonize treatment decisions across patients. Protocols in managed care, for example, encourage or require physicians to treat patients with the same condition similarly.

In both of these circumstances, the physician's views about optimal treatment may differ from the patient's. Doctors may want to limit care while patients may want more. This divergence is particularly likely if patients pay little at the margin for medical care, as they do in many managed care plans (at least for in network services). The conflict of incentives between physicians and patients in managed care contrasts with the situation in traditional indemnity insurance, where the incentives of patients and physicians are generally aligned (although both differ from the incentives of insurers).

Figure 5 shows a potential conflict; at the prices each faces, the patient demands much more care (Q_D^*) than the physician wants to provide (Q_S^*).[29] Which level of care will ultimately be provided? Knowing how treatment decisions will be made in such an environment is difficult, as economic analysis of rationed goods in general does not

[29] We have drawn the physician's supply curve as upward sloping. This needn't be. It could be vertical or backward bending. Our point would carry through, nevertheless.

reach uniform conclusions. The situation is particularly severe in the medical care market because patients do not pay substantial amounts at the margin for medical care; thus, willingness to pay is not an accurate way of gauging individual value of services. There are several possible outcomes. One possibility is that the short-side principle, applies, which predicts that the equilibrium quantity will be the lesser of demand and supply. This is shown as the thickened line in Figure 5 and corresponds to a situation where treatment decisions in managed care are made predominantly by physicians. The short-side principle underlies much of the work on managed care [see, e.g., Baumgardner (1991), Ramsey and Pauly (1997), Pauly and Ramsey (1998)].

But the short-side principle assumes patient wishes play no role when demand exceeds supply. Treatment decisions may come out of a "bargaining" process that balances the wishes of physicians and patients [Ellis and McGuire (1986)]. One can interpret this bargaining either as an explicit process between the parties, or as the physician balancing his own self-interest (or the insurer's profits) with the best interests of the patient.

The actual level of service delivered is likely to vary with the particular medical situation. Patients with chronic conditions may know a great deal about their treatment options; the outcome may thus be close to the patient's demand. In emergency situations, the opposite may be true. The effectiveness of managed care in limiting medical spending may thus differ across settings.

4.2. Evidence on supply-side payment and medical treatment

A substantial literature examines the role of supply-side payment systems in influencing medical treatments. A change from *retrospective*, or cost-based reimbursement, to *prospective* reimbursement is typically analyzed.

Table 7 presents studies on this topic. It documents the impact of prospective payment on four aspects of hospital care: the number of admissions and transfers; average length of stay or other inputs; hospital profits; and quality of care. Prospective payment might increase or decrease hospital admissions. On the one hand, sick people might be less likely to be admitted to a hospital under prospective payment, since reimbursement for these individuals falls short of expected cost. On the other hand, hospitals might be more eager to admit healthy patients, for whom reimbursement exceeds costs. As Table 7 shows, admissions generally declined with the implementation of prospective payment.

While one might worry about whether care for the sick is excessively rationed in such a system, the literature on whether patients are being "dumped" under prospective payment (e.g., sent to public hospitals) is not particularly clear. A loose consensus is that there is some dumping of patients under PPS, but the magnitude is not particularly great [Morrisey, Sloan, and Valvona (1988), Newhouse (1989), Newhouse (1996)].

Studies examining the effect of prospective payment on average lengths of hospital stay and other inputs find nearly uniformly that average hospital stays fell with the reimbursement change. This is what theory predicts: hospitals no longer paid for each additional service will cut back on marginal care, which is expensive relative to health benefits. The effect of prospective payment on hospitals stays is uniformly strong and

Table 7
Prospective payment and medical treatments

Paper	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Frank and Lave (1986)	1981 NIMH discharge data and AHA surveys	OLS regression		LOS for psychiatric patents fell 0.3 days more in states with PPS in Medicaid		
Guterman and Dobson (1986)	HCFA in-house statistics	comparison of means and other descriptive statistics		LOS dropped 13% from 1981–84 vs. 4% in previous four years combined		
Sheingold and Buchberger (1986)	1981 PPS cost reports	simulation of provision of free care by hospitals under PPS rules				each 1% decrease in financial margin leads to 0.3–0.5% less free care provision
Carroll and Erwin (1987)	1982–85 patient records from non-random sample of 10 Georgia long-term care facilities	comparison of means				patients dying within 30 days of entering long-term care facility dropped from 14.7% to 8.3% under PPS
Feder, Hadley, and Zuckerman (1987)	1982 and 1984 AHA surveys	comparison of means			total margins for hospitals under PPS rose 2.9%, compared to no change for hospitals under TEFRA	

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Fitzgerald et al. (1987)	patients with hip fractures admitted to a municipal hospital from 1981–85	comparison of means		LOS fell from 16.6 to 10.3 days		percent in nursing home at six months after discharge rose from 13 percent to 39 percent
DesHarnais, Chesney, and Fleming (1988)	1980–85 Professional Activities Study of CPHA hospitals	comparison of means	discharges dropped 3% from 1980–85	LOS dropped 20% from 1980–85		no significant adverse effect on quality of care
Fitzgerald, Moore, and Dittus (1988)	elderly patients admitted for new hip fracture at large, mid-western community hospital,	comparison of means		LOS dropped from 21.9 to 12.6;		percent discharged to nursing home rose from 38 to 60; percent in nursing home at one year rose from 9 to 33 percent
Lave, Frank, Taube, Goldman, and Rupp (1988)	1984 Medicare PATBILL file, 1984 NIMH psychiatric discharges, HCFA, AHA, CHPS	comparison of means		LOS for psychiatric patients at PPS hospitals fell 23% under PPS; charges fell 20% under PPS		
Morrisey, Sloan and Valvona (1988)	1980, 1983–85 sample of 501 CFHA hospitals	multinomial logit and OLS for post-hospital care selection	probability of transfer increases significantly after PPS	LOS decreased in almost all major DRGs after advent of PPS		

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Newhouse and Byrne (1988)	1981, 1984–85 20% sample of Medicare claims from non-waiver states	comparison of means		LOS rose at long-term hospitals (not on PPS) relative to short-term hospitals (on PPS)		
Sloan, Morrisey, and Valvona (1988)	1980, 1983–85 sample of 501 CFHA hospitals	comparison of means		ICU/CCU days rose less in PPS states than non-PPS states from 1980–83		
Frank and Lave (1989)	1981–84 National Hospital Discharge Survey	hazard model		LOS fell 17% with per case payment for psychiatric patients		
Gaumer, Poggio, Coelen, Sennett, and Schmitz (1989)	1974–83 AHA surveys and standardized mortality rates	comparison of means				mortality rates 1% to 2% higher than predicted for urgent care patients in PPS states
Gerety et al. (1989)	Chart review of patients with hip fracture before and after prospective payment	comparison of means		LOS fell by 1.4 days		poorer discharge ambulation; no effect on nursing home residence at 1 year

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Hadley, Zuckerman, and Feder (1989)	1983–85 AHA surveys	comparison of means		LOS fell by 10.3% under PPS		
Newhouse (1989)	1983–84 5% random sample of PPS hospital bills	comparison of means; OLS regression	1/4 of cases in unprofitable DRGs move to city/ county hospitals			
Palmer et al. (1989)	patients with hip fractures admitted to a private, suburban, teaching hospital from 1981–87		LOS fell from 17.0 to 12.9 days			No effect on nursing home residence or ambulation at 6 months
Russell and Manning (1989)	annual reports of trustees of federal Hospital Insurance Trust Fund	comparison of cost projections before and after PPS			Medicare costs for 1990 reduced by $18 billion compared to projections	
Sager, Easterling, Kindig, and Anderson (1989)	1981–85 age-specific national mortality data	comparison of means				deaths in nursing homes rose by 2.6% in PPS states; no change in non-PPS states
Sheingold (1989)	1983–87 Medicare Cost Reports	comparison of means	Discharges dropped 6% in 1983 and 1984		PPS margins fell from 14.7% to 7.9% between 1983 and 1985	

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
DesHarnais, Wroblewski, and Schumacher (1990)	1980–87 Professional Activities Study of CPHA hospitals	comparison of means	admissions of psychiatric patients fell under PPS			
Folland and Kleiman (1990)	1980–85 stock market returns	seemingly unrelated regressions of excess returns			no significant excess returns to hospital management firms after PPS	
Guterman, Altman, and Young (1990)	1983–86 AHA and Healthcare Financial Management Association surveys	comparison of means			teaching hospitals have highest but fastest falling PPS margins	
Kahn et al. (1990) [series of articles using the same data in a single-volume series]	1981–82 and 1985–86 Medicare records from 297 for patients with six conditions	comparison of means	patients were 1% to 1.6% sicker at admission			no significant effects on quality: – patients admitted from home, not discharged home fell 4%; – likelihood of instability at discharge rose 22%; – patients receiving poor care fell 13%; – in-hospital mortality fell 3%; – 30-day mortality rose 1.6%

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Menke (1990)	1983–86 Medicare Parts A and B claims data	OLS regression		LOS for stroke patients fell by 2.4 days		
Ray, Griffin, and Baugh (1990)	Medicare enrollees with hip fracture in Michigan			LOS fell by 4.4 days		mortality at 1 year was unchanged
Cutler (1991)	1984, 1988 Massachusetts inpatient data	OLS regression		LOS and inpatient procedures fell under PPS		
Willke, Custer, Moser and Musacchio (1991)	1983–87 AMA Socioeconomic Monitoring System telephone surveys	comparison of means		LOS dropped by 0.6 days from 1983–87; doctors' practice hours per week rose 2.0 hours per week		
Fisher (1992)	1985–90 Medicare Cost Reports, AHA employee data	comparison of means			hospitals with Medicare profits dropped from 84.5% to 40.7% from 1985–90	
Eze and Wolfe (1993)	1982–86 Dept. of Veterans Affairs Patient Treatment Files, Medicare discharge data	ANOVA	discharges to VA hospitals rose, by 135.6% for serious cases			

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Hodgkin and McGuire (1994)	1983–90 ProPAC Medicare extracts	comparison of means	admissions fell 11% from 1983-90	LOS for Medicare fell and then rose from 1984–89 to levels consistent with other payers	hospital margins projected to drop from 14.5% to −10.2% under PPS	
Cutler (1995)	1981–88 New England Medicare admissions, 1981–89 Social Security death records	hazard models for readmission and mortality				compression of mortality into immediate post-admission period
Staiger and Gaumer (1995)	1984–87 25% random sample of AHA hospital file, Medicare MEDPAR discharge data	beta-logistic model of mortality				reduced payments compress mortality into period just after discharge
Ellis and McGuire (1996)	1988–92 New Hampshire Medicaid Services psychiatric discharges	simultaneous equations treatment of panel data		LOS for psychiatric patients fell 25% under PPS		
SUMMARY			admissions fall; moderate dumping from PPS to non-PPS hospitals	LOS falls significantly; other inputs fall as well	initially higher Medicare profit margins reduced over time	effects on quality ambiguous for average patient, adverse for marginal patient; lower in-hospital mortality

impressive; many studies find reductions of 20 to 25 percent over a period of 5 years or less. These studies provide among the clearest evidence that supply-side reimbursement changes do affect medical treatments.

Despite the reduction in average lengths of hospital stay, a number of studies find that profit margins fell under prospective payment. This reduction in profits came largely from a reduction in revenues. As the reduction in length of stay indicates, costs fell with the introduction of prospective payment.

In addition to examining the effect of prospective payment on quality, the literature has also examined how managed care as a whole affects medical spending. Studies of this question are summarized elsewhere, including in this Handbook [Miller and Luft (1997), Congressional Budget Office (1992), Glied (2000)]; we discuss it only cursorily here.

Virtually all studies find that managed care insurance reduces medical spending in comparison to traditional indemnity insurance. The consensus estimate would be that patients under managed care spend about 10 percent less than patients in indemnity plans, adjusted for differences in the underlying health of the two groups. The effect is somewhat greater for inpatient hospital spending, but is offset by some additional outpatient utilization in managed care insurance. Overall, therefore, incentives on the physician side clearly have an effect on overall utilization.

5. Optimal mix of demand- and supply-side controls

Given the availability of both demand- and supply-side controls, which should be employed? A first pass suggests that supply-side limitations are preferable, since providers are relatively less risk averse than are patients. In practice, however, plans with both types of limitations are sold, and indeed most plans available have a mix of demand- and supply-side cost containment features (for example, capitation with high cost sharing on out-of-network use, or indemnity insurance with utilization review).[30]

Both demand- and supply-side controls may be desirable in the presence of the other. First, patients and providers may control different features of the medical interaction. For example, the Rand Health Insurance Experiment found that patient cost sharing had a substantially greater impact on the probability that a patient uses services than on the level of services provided conditional on use [Newhouse et al. (1993)]. One can interpret this as saying that cost sharing affects insureds, but not their physicians. The evidence cited above shows that managed care can limit the level of services provided, however. An insurer or provider facing this situation might then want to combine demand- and supply-side cost sharing, the former to limit the initiation of visits and the latter to control the intensity of treatment provided within visits [Ma and McGuire (1997)].

[30] The coexisting prevalence of both types of plans may be transitional, since managed care is still relatively new. But managed care plans have increasingly been incorporating more consumer choice and cost sharing (for example, in out-of-network use). This suggests the combination is not just transitional.

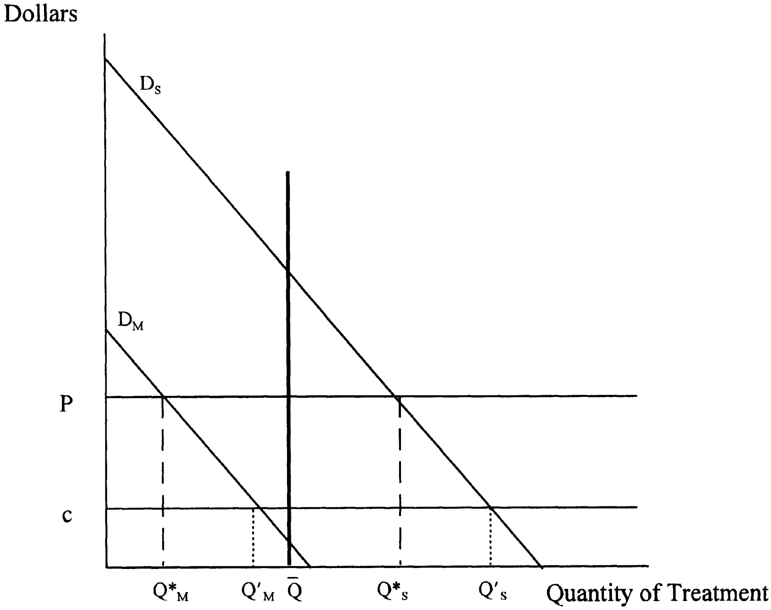


Figure 6. Demand and supply side expenditure controls.

Combining demand- and supply-side controls can also promote flexibility in types of treatment. Consider the situation in Figure 6 [Baumgardner (1991), Ramsey and Pauly (1997), Pauly and Ramsey (1998)]. There are two types of patients: those who are moderately ill (denoted M), and those who are more seriously ill (denoted S).[31] Moderately ill patients demand less medical care at any price than severely ill patients. We assume the insurer cannot distinguish the two groups, however; thus, cost sharing or quantity restrictions must apply equally to the two.

Given a price of medical care P, the optimal amounts of medical care to receive are Q^*_M and Q^*_S respectively for the moderately and severely ill. With a coinsurance policy that requires the patient to pay c for each unit of care, the equilibrium will be medical care levels of Q'_M and Q'_S. Because of moral hazard, medical care demand will be too high. Insurers might alternately adopt a fixed quantity constraint, for example $\overline{Q}$ for each patient.[32] At $\overline{Q}$, the right amount of medical care is provided in total, but not for each patient; the moderately ill patient will receive too much medical care, while the severely ill patient will not receive enough. Thus, neither demand- nor supply-side cost containment by itself yields an optimal allocation.

[31] Note that this may apply conditional on a diagnosis. For example, the conditions could be severe and moderate heart attacks.

[32] We assume that managed care features this type of restriction.

Combining demand- and supply-side cost containment can improve the situation, however. For example, starting from $\overline{Q}$, raising coinsurance will discourage utilization by the moderately ill person before the severely ill person (because the marginal value of care is much lower for the former). If the coinsurance rate necessary to deter low value utilization is small, the risk spreading loss from such coinsurance will be small, and the net welfare consequences of deterrence will be positive. The ability to limit demand by the moderately ill person, in turn, allows an increase in $\overline{Q}$, since this constraint applies only to the severely ill person. Indeed, if demand for the moderately ill person is fully constrained by the cost sharing, $\overline{Q}$ could be increased to the optimal level of care for the severely ill person. More generally, coinsurance and constraints can be combined to enable rationing in more than one dimension when there is heterogeneity of optimal treatment. A combination of the two systems may be superior to using either system alone.

A third rationale for combining demand- and supply-side controls is to limit selection behavior by providers. Providers paid on a capitated basis will have incentives to attract healthy patients and "dump" sick ones, since the provider's payment is the same with the two patients but the costs are much greater in treating the sick patient. Incorporating patient cost sharing into the insurance policy can relax the supply-side constraints and thus limit the incentives to dump patients [Ellis (1998)]. We return to this type of adverse selection in the next section.

Theoretical results to date generally suggest a combination of demand- and supply-side controls may offer significant advantages. Moreover, with so many differing incentives in the medical care system, optimal reimbursement schemes undoubtedly differ across specialties (for example, in response to moral hazard propensities) and groups of providers (for example, if the ability to bear risk differs with group size), which increases the potential for working both sides of the market. The way demand- and supply-side systems interact with each other, however, is not well understood; neither is the tradeoff between a fine-tuned system and a system that is simple and comprehensible. Real world structures suggest simplicity has its virtues. It is noteworthy, for example, that virtually all coinsurance operates at a flat rate between the deductible and any stop loss amount.[33]

6. Markets for health insurance: plan choice and adverse selection

To this point, we have talked of the design of a single health insurance plan. Most private insurance in the United States is offered on a menu basis, however, with different insureds selecting different plans. Health insurance choice is a natural way to meet differing individual preferences. Some people will prefer managed care insurance, which

[33] Simplicity and transparency may be a handicap. Conceivably insureds and doctors, not understanding what they will be respectively charged or paid for something, may behave more reasonably. For example, a low but complex coinsurance rate might be the best way to discourage utilization. It imposes less financial risk than, say, a higher flat rate, but might be just as effective in controlling use.

limits utilization but costs less, while others will opt for a more open-ended indemnity-style policy. Within indemnity insurance policies, some will be willing to bear more financial risk than others. Having these preferences reflected in market outcomes is beneficial.

In addition, health insurance choice is important to promote efficiency. Customers shopping for the lowest prices drive costs to their lowest level. Moreover, product characteristics will be shifted and new products introduced to meet consumer demands. These benefits of competition for health insurance are analogous to the benefits competition yields in other markets.

But health insurance is fundamentally different from other markets in ways that create harmful side effects from competition. The key problem is that with health insurance, unlike other services or commodities, the identity of the buyer can dramatically affect costs. Insuring a 60 year old costs 3 times as much as insuring a 30 year old, and among 30 year olds, some will have far higher costs than others. Whom one pools with in health insurance dramatically affects what one has to pay.

Generally, the sick are drawn to more generous plans than the healthy. Those expecting to use more services will, all else equal, want more generous policies than those expecting to use fewer services. If plans could charge individuals their expected cost for enrolling in each plan, the market would efficiently sort people. Such charges are generally not imposed, however, since it is widely believed that it is not fair to make people pay a lot more just because they are sick. Knowing the individual-specific prices may also not be technically feasible.

When plans can only charge average prices, generous plans will disproportionately attract sicker people, and more moderate plans will disproportionately attract healthier ones. This phenomenon is termed *adverse selection* [Akerlof (1970), Arrow (1985)]. As a result of adverse selection, generous plans will have to charge premiums above moderate plans not only because they offer more benefits but also because they attract a worse mix of enrollees. These premium differentials, if passed on to insureds, will tilt unfairly against generous plans.[34]

Adverse selection into more generous plans leads to two fundamental difficulties. First, people will choose to be in less generous plans, so that they can avoid paying for the higher costs of very sick people. Second, plans will have incentives to distort their offerings to attract the healthy and repel the sick. Since no plan would like to enroll the sickest people, *all* plans will find it profitable to distort their benefits. Indeed, even innovations that improve quality of health care may be unattractive to plans even if they come without additional cost, if they attract the wrong people. The distortion in plan provisions resulting from adverse selection is variously termed plan manipulation, skimping [Ellis (1998)], or stinting [Newhouse (1996)].

[34] This would happen, for example, if employers make a fixed dollar contribution to the premiums of each plan offered to their employees. The converse is also true; if employers heavily subsidize the difference between plan costs, employees will choose the generous plan too often.

Table 8
Benefits and costs for HIGH and LOW risk individuals

	Generous plan		Moderate plan		Basic plan	
	Benefits	Costs	Benefits	Costs	Benefits	Costs
HIGH risk	$33	$16	$20	$4	$14.00	$2.80
LOW risk	$6	$4	$5	$1	$3.50	$0.70

The consequences of these undesired side effects of competition are felt in market equilibrium. The equilibrium with adverse selection may be inefficient; it may not even exist. We express this as the third lesson of health insurance:

> *Lesson* 3: *Competition when consumer identity matters*. When consumer identity affects costs, competition is a mixed blessing. Allowing individuals to choose among competing health insurance plans can allocate people to appropriate plans and provide incentives for efficient provision. But it can also bring with it adverse selection – the tendency of the sick to differentially choose the most generous plans. Adverse selection induces people to enroll in less generous plans, so they can be in a healthier pool, and gives plans incentives to distort their offerings to be less generous with care for the sick.

Many models of adverse selection have been developed. We start with simple models and then present more advanced models.

6.1. Equilibrium with adverse selection – the basics

To understand the patterns in adverse selection, we start with the simplest possible situation [Rothschild and Stiglitz (1976), Wilson (1980)]. Assume there are two individuals, one HIGH risk and one LOW risk, and two plans, a generous plan and a moderate plan. Table 8 gives the hypothetical benefits and costs for the generous and moderate plans. We suppose that the generous and moderate plans are what HIGH and LOW respectively would design for themselves, assuming that each had to pay his own costs.[35] Note that HIGH costs more in either plan and both people use more services in the generous plan than in the moderate plan.

Equilibrium. Efficiency requires people to be in the generous plan if the additional benefits of that plan to them are greater than the additional costs they incur. In this case HIGH should be in the generous plan, and LOW should be in the moderate plan, since the additional value to HIGH of the generous plan ($13) relative to the moderate plan is greater than its additional cost ($12), while the converse is true for LOW (a benefit

[35] This assumption of respective optimality facilitates exposition, but is not required.

of $1 compared to an additional cost of $3). The efficient outcome thus separates the insureds.

Were separation to happen, the premiums would be $16 for the generous plan (the cost of HIGH) and $1 for the moderate plan (the cost of LOW). At these prices, however, HIGH would select the moderate plan; the $15 savings are greater than the $13 loss. Of course, once HIGH joins the moderate plan costs escalate, but they are still only $2.50 (the average of 4 and 1). HIGH's cost savings by enrolling in the moderate plan ($13.50) are still greater than his loss in benefits ($13). LOW will also prefer the moderate plan.

The market equilibrium will thus have both individuals in the moderate plan, a pooling equilibrium. This is not efficient, however. The reason this inefficiency arises is that individuals do not pay their own costs in each plan, but rather the average cost of the plan. Hence, HIGH mimics LOW so that he can share his costs with LOW.

There are a variety of ways to struggle back towards efficiency. Two logical candidates, assigning people to plans or charging people on the basis of expected cost, are undesirable because they respectively override free choice or sacrifice risk spreading.[36] Two additional possibilities would be to cross-subsidize the generous plan by the moderate plan [Cave (1985)], or to distort the plan offerings [Rothschild and Stiglitz (1976)].

Cross-subsidy. Suppose the moderate plan is taxed an additional $1.25 per capita, which is used to offset the premium of the generous plan. In the separating equilibrium, the premiums in the two plans will be $14.75 and $2.25, and HIGH will now prefer the generous to the moderate plan. Both insureds are better off with the subsidy than without. HIGH clearly prefers a subsidy to no subsidy. LOW also prefers the subsidy, because he pays only a $1.25 subsidy, compared to an additional $1.50 premium if he pooled with HIGH in the moderate plan.

Plan manipulation. A second mechanism to induce a separating equilibrium is to replace the moderate plan with something stingier. When faced with a stingier plan, HIGH might choose the generous plan over pooling with LOW. Making the moderate plan stingier is distasteful to LOW, but the cost to HIGH is substantially greater. This disparity in costs is what allows "hurting" the plan to produce separation.

Consider a plan called basic, also detailed in Table 8, which gives both HIGH and LOW 70 percent of the benefits and costs they would receive from the moderate plan. Thus, LOW would receive benefits of $3.50 at a cost of $0.70 were he in the basic plan and HIGH would receive benefits of $14, incurring a cost of $2.80. If basic and generous were the two plans offered, LOW would select the basic plan. If HIGH selects the basic plan as well, his premium, i.e., average cost, would be $1.75. He'd prefer the generous plan, which offers an additional $19 in benefits, but would cost only $14.25 more. LOW

[36] Partial measures are possible. For example, many employers "carve out" mental health benefits from all plans and provide those services using one insurer. Adverse selection is one rationale for this [Frank, McGuire, and Newhouse (1995)].

prefers the basic plan to pooling with HIGH in the moderate plan. Plan manipulation sacrifices efficiency, since LOW generates more net benefits in the moderate plan.

In practice, plan manipulation can take many forms. Aerobics programs, for example, will attract the vigorous healthy while spinal cord injury or high-tech cancer care facilities pull in the costly sick. There are generally more opportunities to trim a high cost-attracting service than to add aerobics equivalents.[37] Thus, we expect plans to be ungenerous with services for conditions that will predictably have high costs.

Market competition will lead to the manipulated equilibrium. Assume that the moderate and generous plans were the only offerings. All participants would pool in moderate. Introducing the basic plan would then attract LOW, HIGH would go off to generous, and the moderate plan would be abandoned.[38]

In practice, plan manipulation and cross-subsidy of premiums can be combined to promote separation. The market equilibrium will have two plans. One will be the optimal plan for HIGH, given whatever subsidy he is receiving. The other plan, which will enroll LOW, will be the plan as close as possible to moderate whose combination of subsidy and manipulation just makes HIGH prefer his optimal plan.

We show this graphically in Figure 7(a), assuming there is a continuous choice of plans.[39] We array the plans in Figure 7(a) from least to most generous – in this case variability among plans is due to differences in the percent of expenses covered. The figure shows the expected utility of LOW (the upper two lines) and HIGH (the lower two lines) at each possible level of generosity, and for both the pooling and separating equilibria. LOW does better than HIGH, since he has a lesser chance of incurring the cost of sickness. HIGH is better off pooling than separating for it allows him to shed costs; the opposite is true for LOW. For both LOW and HIGH, their optimal separating equilibrium offers less than full insurance. This might be because of moral hazard or administrative costs; without these factors each in isolation would want the most generous policy. We show HIGH as wanting full insurance in the pooled equilibrium; in our example, the benefits from the subsidy in that plan are greater than the moral hazard or administrative cost loss. In the least generous plan (no insurance), both HIGH and LOW are indifferent between pooling and separating equilibria. In the most generous plan (full insurance) the two pay the same price and get the same utility in the pooling equilibrium.

Consider the situation if HIGH and LOW are initially at A, the full insurance pooling equilibrium. An insurer that offered a plan with generosity G_1 would attract LOW,

[37] However, the Harvard University Group Health Plan – an option for Harvard faculty – offers a $50 wellness payment, which can be used say for sneakers, as an attractor.

[38] The efficiency costs of separation produced through plan manipulation may be small. That is because the moderate plan was designed for LOW. Assuming smoothness, the costs of moving away from the optimal plan are initially trivial. But the costs to HIGH, who is already far from his optimum, may be great. This disparity allows cheap distortion to produce target efficiency [Nichols and Zeckahuser (1982)].

[39] The classic diagrammatic presentation of plan manipulation (dating from Rothschild and Stiglitz) uses indifference curves. We present this in the Appendix.

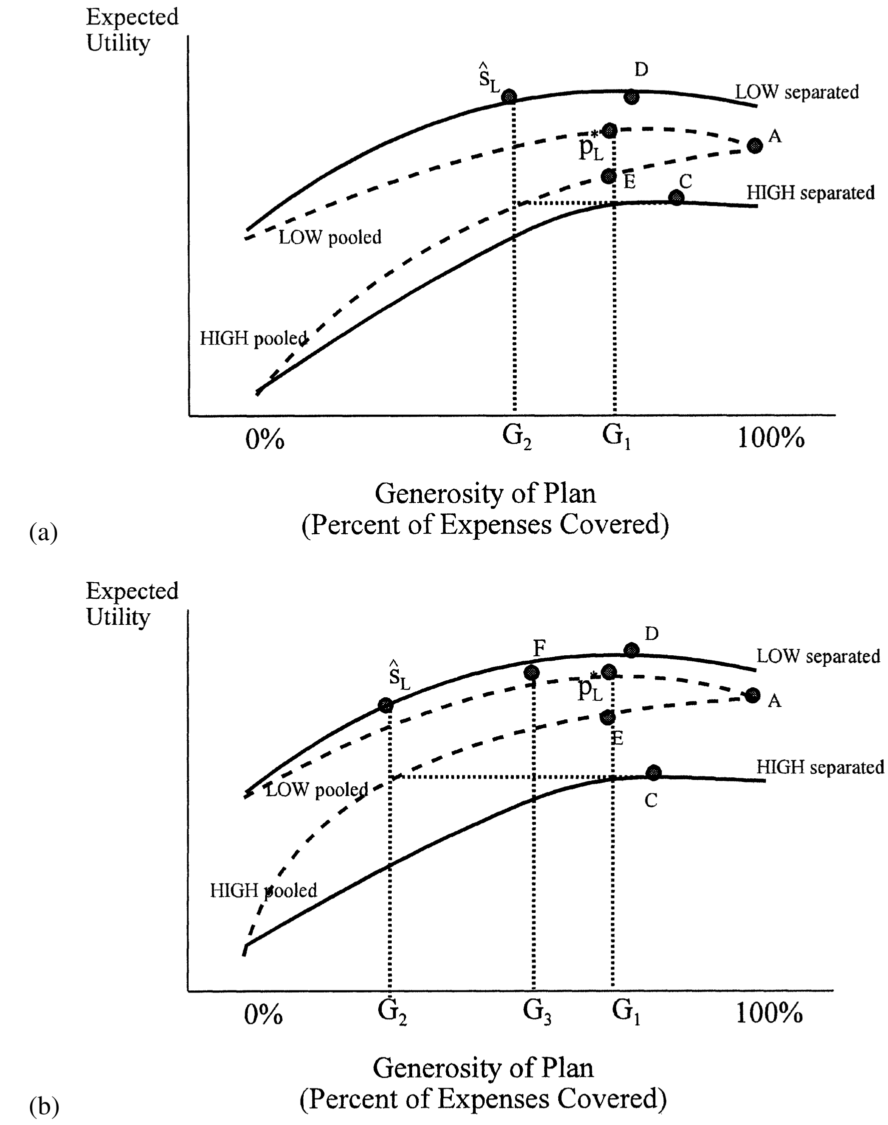


Figure 7. Reduction in insurance to separate HIGH and LOW. (a) Stable separating equilibrium. (b) Unstable separating and pooling equilibria.

since LOW prefers G_1 to A. HIGH would then move to G_1, because E is preferred to C, the separating equilibrium if only HIGH is in the generous plan. As the pooled policy becomes less generous, its attractiveness to HIGH falls. Policy G_2 makes HIGH just indifferent between pooling with LOW and the separating equilibrium at C. The stable equilibrium will thus have two policies: LOW will be at point $\hat{s}_L$ with policy G_2 and HIGH will be at point C.

With slight changes in the curves, however, the situation at G_2 may not be stable either. Consider the situation in Figure 7(b), drawn for the case where the risk difference between HIGH and LOW is less than in Figure 7(a). Here LOW's preferred pooling equilibrium is superior to his best sustainable separating plan, $\hat{s}_L$. Thus, the separating equilibrium at G_2 will be broken by the pooling equilibrium at G_1. But the converse is also true; the pooling equilibrium at G_1 is broken by a plan say at G_3, with a price just low enough to attract LOW at F, whereas HIGH would prefer to stick with the premium and coverage at E. Once LOW went to F, however, the premium at E would have to rise, and HIGH would chase LOW to G_3. Thus, there is no stable equilibrium in Figure 7(b).

The model underlying Figure 7 assumes a frictionless world, where individuals shuttle costlessly between plans and there are no costs involved in establishing new plans. If such costs play a role, they may enable otherwise breakable equilibria to survive. For example, if establishing a plan entails high fixed costs, but individuals' transit costs remain low, p_L^* becomes stable, since breaking p_L^* with G_3 is costly but yields only temporary profits. Interestingly, greater transit costs for individuals may promote instability, since a temporary period for attracting individuals to an unstable equilibrium may last longer, and therefore be more attractive despite the fixed costs of establishing a plan. Even in this simple model, the ultimate outcome of markets with adverse selection is uncertain.

6.2. *Equilibria with multiple individuals in a risk group*

The simple model of adverse selection had a single HIGH and LOW risk. The lumpiness of movement implied by this specification is an important limitation of the model. With multiple individuals of a given risk type, there can also be a third class of equilibria, a "hybrid" equilibrium, to join the pooling and separating equilibria. We now show this equilibrium.

Imagine that there are now many HIGHs and LOWs, with similar tastes for insurance within each group.[40] Our example uses the parameter values from Table 8, with the \$33 benefit for HIGH under the generous plan changed to \$34. Suppose we start in the separating equilibrium, with HIGHs in the generous plan and LOWs in the basic plan. The expected utility in this equilibrium is shown by the points A and B in Figure 8. Recall that the LOWs all prefer the moderate plan to the basic plan. Imagine that they all enroll in that plan. Now suppose that instead of all the HIGHs choosing the moderate plan, only a share of them choose it. Figure 8 traces expected utility for HIGHs and

[40] A more general formulation would allow individuals within a cost class to differ on such factors as risk aversion, or in tastes for plans. Then the division of HIGHs between the moderate and generous plans would reflect the individuals' preferences.

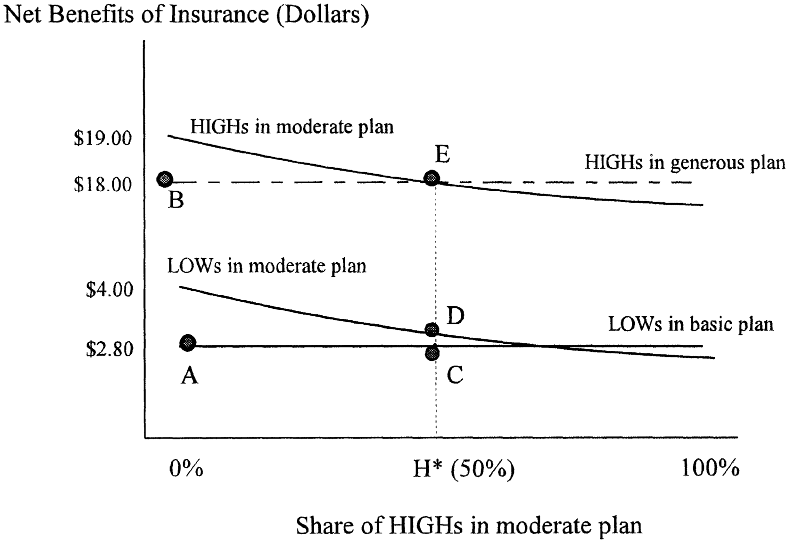


Figure 8. Hybrid equilibria with adverse selection. Note: Dashed lines assume all LOWs choose the moderate plan. The figure uses the values in Table 9, assuming the benefits to the HIGH risks in the generous plan is $34 instead of $33.

LOWs as a greater share of the HIGHs choose the moderate plan. Once H^* of the HIGHs have enrolled in the moderate plan – the number is 50 percent for our parameters – HIGHs will be indifferent between the two plans. No additional movement of HIGHs will occur.

The LOWs in the moderate plan are worse off pooling with some of the HIGHs than they would be if they had the moderate plan to themselves. But that does not indicate whether the LOWs prefer to separate themselves in basic. Indeed, in Figure 8, expected utility for the LOWs given a share H^* of HIGHs in the moderate plan (point D) is greater than expected utility in the basic plan (point C). The equilibrium with all of the LOWs[41] and a share H^* of the HIGHs in the moderate plan – what we term the "hybrid equilibrium" – is stable.

The hybrid equilibrium need not be stable, however. If the HIGHs are sufficiently costly, the LOWs will prefer the separating equilibrium to the hybrid equilibrium (point C will be above point D) and thus the two groups would separate completely.

[41] The LOWs will never end up split between the basic and moderate plan. Say the basic and moderate plans were equally attractive with a fraction of the LOWs in the moderate plan. As more LOWs moved to the moderate plan it would become more attractive. Hence, the equilibrium would tip all the LOWs into the moderate plan.

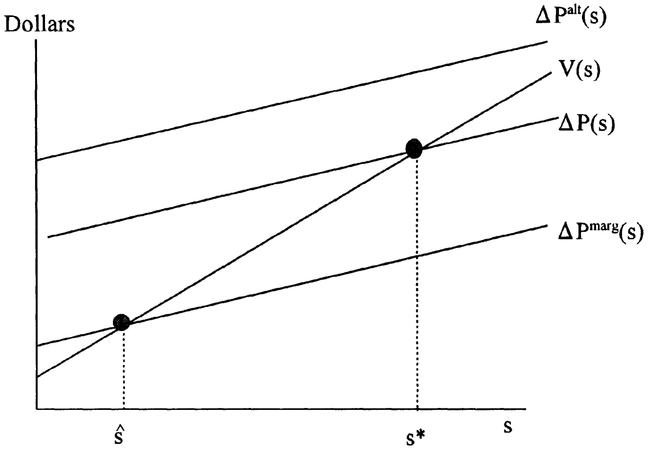


Figure 9. Enrollment consequences of adverse selection.

6.3. Continuous risk groups

Our two-risk-types model suggests that at least some high risks will enroll in their most preferred plan while low risks may be distorted into less generous plans. In situations with more than two risk groups, however, this situation may be reversed; the low risks may be in their preferred plans but the high risks may not. We show this using a model developed by Feldman and Dowd (1991), Cutler and Reber (1998), and Cutler and Zeckhauser (1998). The model assumes there are two pre-established plan types.

Suppose there is a continuous distribution of risks in the population, denoted by s. For simplicity, we normalize s to be the person's expected spending in the generous policy. There are two plans, one generous and one moderate. The value of more generous insurance to an individual is $V(s)$, where $V' > 0$ (the sick value generous policies more than the healthy). Figure 9 shows $V(s)$. At any additional cost for choosing the more generous policy, people will strictly divide into plans. If s^* is the sickness level of the person indifferent between the two policies, people with $s > s^*$ will choose the generous policy, whereas people with $s < s^*$ will choose the moderate policy. Average sickness in the generous policy is $s_G = E[s \mid s > s^*]$, and average sickness in the moderate policy, is $s_M = E[s \mid s < s^*]$.

Plan premiums, in turn, depend on who enrolls. We assume people in the moderate policy cost a fraction α ($\alpha < 1$) of what they would cost in the generous policy.[42] In a competitive insurance market, premiums will equal costs: $P_G = s_G$, and $P_M = \alpha s_M$. The premium difference between the two plans is therefore:

$$\Delta P(s) = P_G - P_M = (1-\alpha)s_M + [s_G - s_M]. \tag{14}$$

[42] The literature reviewed above suggests that $\alpha \approx 0.9$ for an HMO relative to an indemnity policy.

The first term in the final expression is the cost savings the moderate plan offers to its average enrollee. The second term is the difference in the average sickness level in the two plans; it is the consequence of adverse selection.

As marginal people move from the generous to the moderate plan, the average sickness in each of the plans will rise. Depending on the distribution of s, the price difference between plans may widen or narrow. Because medical spending in practice is significantly right-skewed (Table 2), it is natural to conjecture that the premium in the generous plan will rise by more than the premium in the moderate plan. Figure 9 reflects this expectation as an upward sloping $\Delta P(s)$ curve.

The guideline for efficiency is that the price differential must be appropriate for the individual at the margin in choosing between plans. All other people would be appropriate sorted, with sicker people choosing the generous plan and healthier people choosing the moderate plan.[43] The price for the marginal individual is given by:

$$\Delta P^{\text{marg}}(\hat{s}) = (1-\alpha)\hat{s}, \tag{15}$$

where $\hat{s}$ is the person for whom Equation (15) holds. We show this schedule in Figure 9 as lying below the $\Delta P(s)$ line. $\hat{s}$ optimally delineates people in the moderate and generous plans.

Comparing Equations (14) and (15) shows that only by coincidence will the equilibrium be efficient. Suppose that the efficient allocation prevailed. From Equation (14), the price difference between the two policies will differ from this amount for two reasons. The first term in Equation (14) is generally below the efficient differential; it represents the savings from the moderate plan for the *average* person in the moderate plan, not the *marginal* person in the plan, for whom the savings would be greater. Working in the opposite direction, adverse selection (the second term in Equation (14)) will raise the premium in the generous plan relative to the premium in the moderate plan. Depending on the distribution of medical expenditures, the market differential could thus be above or below the efficient level. The right skewness of medical spending suggests that the adverse selection effect will tend to predominate. This is the situation shown in Figure 9 (by virtue of the fact that the $\Delta P(s)$ line is above the $\Delta P^{\text{marg}}(s)$ line). The premium differential for the generous plan will then be above the efficient differential, and too few people will enroll in the generous plan.

Because of adverse selection, small deviations in price can drive large differences in allocations, and indeed, the generous plan may fail to survive. Starting from $\hat{s}$, suppose the generous plan is priced too high. Marginal enrollees will depart, driving prices up still further, inducing new departures, and so on. The final equilibrium may be quite far from the efficient point. Indeed, Figure 9 also shows the possibility that the entire generous plan is depopulated. If $\Delta P^{\text{alt}}(s)$ described the cost differential, then $V(s)$ would not

[43] If preferences as well as sickness level affect the value of the generous plan, then each individual must pay his personal cost differential, $\Delta P^i(s) = (1-\alpha)s_i$.

intersect that line and the equilibrium would have no enrollment in the generous plan.[44] The disappearance of generous plans as a result of dynamic processes of adverse selection is termed a "death spiral". In such a situation, high risks end up in less generous plans than is optimal, while low risks get their preferred policy.

6.4. Evidence on the importance of biased enrollment

A substantial literature has examined adverse selection in insurance markets. Table 9 summarizes this literature, breaking selection into three categories: traditional insurance versus managed care; overall levels of insurance coverage; and high versus low option coverage.

Most empirical work on adverse selection involves data from employers who allow choices of different health insurance plans of varying generosity; a minority of studies look at the Medicare market, where choices are also given. Within these contexts, adverse selection can be quantified in a variety of fashions. Some authors report the difference in premiums or claims generated by adverse selection after controlling for other relevant factors [for example, Price and Mays (1985), Brown et al. (1993)]. Other papers examine the likelihood of enrollment in a generous plan conditional on expected health status [for example, Cutler and Reber (1998)]. A third group measure the predominance of known risk factors among enrollees of more generous health plans compared to those in less generous plans [for example, Ellis (1989)].

Regardless of the exact measurement strategy, however, the data nearly uniformly suggest that adverse selection is quantitatively large. Adverse selection is present in the choice between fee-for-service and managed care plans (8 out of 12 studies, with 2 findings of favorable selection and 3 studies ambiguous), in the choice between being insured and being uninsured (3 out of 4 studies, with 1 ambiguous finding), and in the choice between high-option and low-option plans within a given type (14 out of 14 studies).

Figure 10 shows a particularly salient example of adverse selection, taken from experience at Harvard University.[45] The Harvard experience is nice because adverse selection was driven by a policy change, and thus one can view the beginning of adverse

[44] Whether a death spiral actually occurs will depend on the distribution of risk levels, and the strength of the risk-preference interaction. The fatter the upper tail, the stronger the interaction, the more threatening is the possibility of a spiral. A numerical example illustrates this possibility. Suppose that the highest cost person has expected spending of $50,000 and that the average costs of the whole population in the moderate policy (with or without this person, if he comprises a small part of the total risk) is $3,000. Suppose further that the high cost person values the generous policy at $20,000 more than the moderate policy, and that he spends only $5,000 less in the moderate policy than with the generous policy (for example, a 10 percent savings if the plans are an indemnity policy and an HMO). Efficiency demands that he should be in the generous policy; the additional value of that policy ($20,000) is greater than the additional cost he imposes there ($5,000). If the high cost person were the only person in the generous policy, however, the cost of that policy would be $47,000 more than the cost of the moderate policy, which would lead him to opt for the moderate policy.

[45] See Cutler and Reber (1998) and Cutler and Zeckhauser (1998).

Table 9
Evidence on adverse selection in health insurance

Paper	Data	Empirical methods	Highlights of results	Selection
Selection between managed care and indemnity plans				
Bice (1975)	East Baltimore public housing residents (random sample)	tests of means of health status variables by Medicaid enrollment	poor health and high expected use of medical services is positively correlated with enrollment in prepaid plans; expected costs are reduced	favorable
Scitovsky, McCall and Benham (1978)	Stanford University employees' enrollment and survey data	least-squares regression of plan choice (note dependent variable is binary)	fee-for-service patients are older and more likely to be single or without children	adverse
Eggers (1980)	Group Health Cooperative (GHC) of Puget Sound's Medicare Risk Contract, 1974–76	comparison of usage statistics with control sample from Medicare 20 percent (Part A) and 5 percent (Part B) Research Discharge Files	Length of stay 25 percent higher for non-GHC patients; inpatient reimbursements per person are 2.11 times higher outside GHC	adverse
Juba, Lave, and Shaddy (1980)	Carnegie-Mellon University employees' health insurance enrollment and survey, 1976	maximum likelihood logit estimates of determinants of plan choice	lower family self-reported health status results in significantly less chance of selecting HMO enrollment	adverse
McGuire (1981)	Yale University employees' health plan enrollment statistics (random sample)	logistic regression of health plan choice given some plan is chosen	women are less likely to join the prepaid health plan than men, but no significant effect is associated with age	adverse
Jackson-Beeck and Kleinman (1983)	11 employee groups from Minneapolis-St. Paul Blue Cross and Blue Shield, 1978-81	comparison of costs and utilization for HMO enrollees and nonenrollees in period before HMO availability	HMO joiners averaged 53 percent fewer inpatient days before joining than those who chose to stay in FFS	adverse
Griffith, Baloff, and Spitznagel (1984)	physician visits in the Medical Care Group of St. Louis	nonlinear regression of frequency of visits	high usage rates at managed care plan's initiation eventually fall to lower steady-state levels	ambiguous

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Merrill, Jackson and Reuter (1985)	state employees' enrollment and utilization data from Salt Lake City and Tallahassee	tests of means in plan populations and logit regression of health plan choice	HMO joiners are younger, more often male, less likely to use psychiatric services, but have more chronic conditions in their family units	ambiguous
Langwell and Hadley (1989)	1980–81 Medicare Capitation Demonstrations	comparison of HMO enrollees and non-enrollees using two-tailed tests of means; comparison of enrollees and disenrollees using surveys	non-enrollees' reimbursements are 44 percent higher than enrollees in two years before capitation; disenrollees have worse past health	adverse
Brown et al. (1993)	Medicare spending for enrollees who stayed in traditional system versus those who moved into managed care	Comparison of spending in the two years prior to HMO enrollment	enrollees who switch to managed care had 10 percent lower spending than enrollees who stayed in traditional system.	adverse
Rodgers and Smith (1996)	summary of 1992 Mathematica Policy Research study of Medicare enrollees	measure cost differences between elderly customers covered by standard Medicare FFS and capitated HMO care	HMO patients are 5.7 percent costlier	favorable
Altman, Cutler and Zeckhauser (1998)	claims and enrollment data from the Massachusetts Group Insurance Commission (GIC)	age- and sex-adjusted analysis of costs among individuals with different plan choice histories	adverse selection accounts for approximately 2 percent of differences between indemnity and HMO plan costs	adverse
SUMMARY				adverse

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Selection of reenrollment versus disenrollment/uninsurance				
Farley and Monheit (1985)	1977 National Medical Care Expenditure Survey	OLS and 2SLS estimation of health insurance purchases	ambulatory care expenditures have an insignificant impact on health insurance purchases	ambiguous
Wrightson, Genuardi, and Stephens (1987)	disenrollees from seven plans offering different types of managed care	comparison of costs and disenrollment rates for insurees	disenrollees have lower inpatient costs and occupy less risky demographic groups than continuing enrollees	adverse
Long, Settle, and Wrightson (1988)	enrollment patterns of subscribers to three Minneapolis-St. Paul HMOs	probit estimation for chance of insuree disenrolling from each of three HMOs	likelihood of disenrollment rises significantly with increases in relative premium of own plan	adverse
Cardon and Hendel (1996)	National Medical Expenditure Survey	Tobit-style model of insurance choice	individuals who are younger, male, or in "excellent" self-reported health are significantly less likely to become insured	adverse
SUMMARY				adverse
Selection of high-option plan within type of plan				
Conrad, Grembowksi, and Milgrom (1985)	1980 random sample of claims and eligibility data for dental health insurance by Pennsylvania Blue Shield	2SLS and 3SLS estimation of demand models for premiums and total expenditures	worse self-perceived dental health corresponds to higher valuation of insurance; experience rating does not always lower premiums	adverse
Ellis (1985)	1982–83 employee health plan enrollment and expense records of a large firm	logit estimates of health plan choice	age and worse previous year's health expenses are associated with choice of more generous health coverage for the next year	adverse
Dowd and Feldman (1985)	survey data from 20 Minneapolis-St. Paul firms	tests of means of characteristics of health plan populations	fee-for-service patients are older and more likely have serious medical conditions or relatives with such conditions	adverse

continued on next page

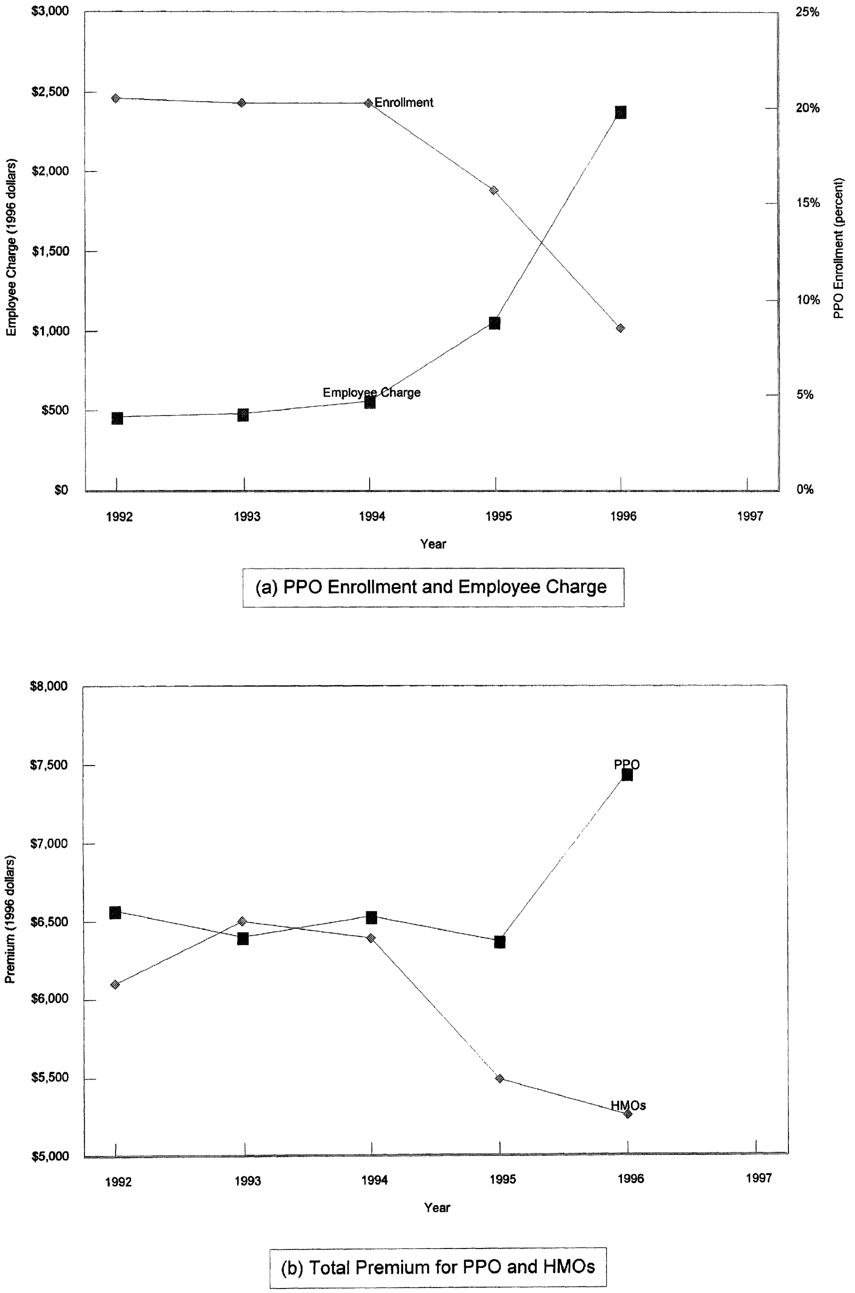
Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Luft, Trauner and Maerki (1985)	California state employees' enrollment and utilization data	comparisons of risk indices across plans and years	patient risk in high option indemnity and fee-for-service plans increases faster than risk in managed care	adverse
Price and Mays (1985)	Federal Employees Health Benefits Program proprietary data	comparison of costs and premiums across plan choices	high option Blue Cross plan undergoes a premium spiral with enrollment cut in half over only three years	adverse
Marquis and Phelps (1987)	Rand Health Insurance Experiment	probit estimation for hypothetical take-up of supplementary insurance	families in highest expenditure quartile were 42 percent more likely to obtain supplementary insurance than those in lowest quartile	adverse
Ellis (1989)	claims and enrollment data from a large financial services firm	analysis of different plans' member characteristics and expenses	employees in high option plan are 1.8 years older, 20.1 percent more likely to be female, and have 8.6 times the costs of the default plan.	adverse
Feldman, Finch, Dowd and Cassou (1989)	survey of employee health insurance programs at 7 Minneapolis firms	nested logit for plan selection	age varies positively with selection of a (relatively generous) IPA or FFS single-coverage health plan	adverse
Welch (1989)	Towers, Perrin, Forster, and Crosby Inc. study of Federal Employees Health Benefits program	comparison of premiums between high and low option Blue Cross plans for government workers	high-option premium is 79 percent higher than low option	adverse

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Marquis (1992)	plan selection of families in Rand Health Insurance Experiment	comparison of plan choices with age/sex adjustments under various group-rating regimes	73 percent more individuals in high risk quartile choose most generous plan than those in low risk quartile, even with age/sex/experience rating	adverse
Van de Ven and Van Vliet (1995)	survey and claims data from 20,000 families insured by largest Dutch insurer, Zilveren Kreis	regression of risk factors on prediction error of difference in costs between members of high- and low-cost plans.	age-and sex-composition of plans explain 40 percent of error in predicted cost differential between plans	adverse
Buchmueller and Feldstein (1997)	University of California Health Benefits Program enrollment figures	historical analysis of enrollment changes and premium increases	two high-option plans suffered fatal premium spirals in a six-year period; a third was transformed from FFS into POS to prevent a spiral	adverse
Cutler and Reber (1998)	claims and enrollment data from Harvard University	calculation of welfare loss and simulation of long-run effects of changes in health plan prices	adverse selection creates a welfare loss equal to 2 percent of baseline health spending; price responses in long run are triple those in short-run	adverse
Cutler and Zeckhauser (1998)	claims and enrollment data from Harvard University and the Massachusetts Group Insurance Commission (GIC)	analysis of different plans' member characteristics and expenses	employees in GIC's FFS plan spend 28 percent more, are older, and have significantly more births and heart attacks than HMO members	adverse
SUMMARY				adverse



(a) PPO Enrollment and Employee Charge

(b) Total Premium for PPO and HMOs

Figure 10. Adverse selection at Harvard University. Note: Dollar figures are for a family policy. Source: Cutler and Reber (1998).

selection and its subsequent effects. In the early 1990s, Harvard University offered its employees two types of health insurance plans: a generous PPO and a number of HMOs. The University paid about 90 percent of each plan's premium; thus, the employee cost of the PPO, shown in Figure 10(a), was a relatively modest $500 per year. To trim costs, Harvard in 1995 moved to a more competitive health insurance system. Under the new system, the University pegged its contribution at a fixed percentage of the lowest cost plan. Employees paid the entire amount above this for the plan of their choice. The hope was that competition among plans would drive down premiums and thus save the University money.

When the new system was introduced, the cost of the PPO rose, and PPO enrollment fell. As Figure 10(a) shows, about one-quarter of PPO enrollees left the plan between 1994 and 1995. These enrollees were disproportionately the younger and healthier employees in the PPO, however. As a result of the biased disenrollment, the PPO lost money in 1995; in 1996, it had to raise its premium by nearly $1,000. This led to a further decline in PPO enrollment; over half the remaining PPO enrollees left the plan after 1996. Again, these employees were disproportionately younger and healthier than those that remained in the PPO. Thus, the PPO premium lost money again in 1997 and would have had to increase premiums substantially in 1998, just to prevent losses. In fact, the required premium increase would have been too large for the insurer and Harvard to bear. The PPO was disbanded before that year. Adverse selection thus produced a death spiral, and did so very quickly. The disappearance of the PPO is a welfare loss to employees who would have chosen it at their individual-specific cost. Cutler and Reber estimate the size of the welfare loss at 2 to 4 percent of baseline premiums.

The importance of adverse selection has had direct impacts on policy. For example, Brown et al. (1993) found that Medicare enrollees who enroll in a managed care plan would have spent 10 percent below average if they had been in the traditional system. Since Medicare paid only 5 percent less to managed care companies for enrolling these people, Medicare lost money as HMO enrollment increased. In 1997, Federal legislation reduced payments to HMOs by an additional 5 percent, to avoid these continuing losses.

6.5. *Evidence on the importance of plan manipulation*

There are substantially fewer empirical studies on plan manipulation than on adverse selection. Plans, of course, differ greatly in their generosity. But it is difficult to know, and plans do not want to reveal, the extent to which the observed variation in plan benefits reflects manipulation by the plans to attract healthy risks as opposed to the self-interested choice of insurance arrangements among people already enrolled in the plans. Adverse selection aside, plans with sicker enrollees probably should be more generous.

Though evidence on plan structures is ambiguous, the marketing of managed care plans shows clear efforts to promote favorable selection. Maibach et al. (1998) document the marketing practices managed care plans use to attract healthy Medicare enrollees, including television ads that show seniors engaged in physical and social activities and marketing seminars held in buildings that were not wheelchair accessible. Whether such practices extend to the types of benefits these plans offer is unknown.

6.6. The tradeoff between competition and selection

In weighting the consequences of competition, losses from adverse selection must be balanced against the gains, if any, from lower premiums that competition induces. The Harvard University study discussed above [Cutler and Reber (1998)] shows such a tradeoff. As Figure 10(b) demonstrates, premiums for the HMOs fell by over $1,000 when the University moved to flat-rate pricing. The savings to Harvard from these lower premiums was estimated at 5 to 8 percent of baseline health spending. This cost savings is greater than the 2 to 4 percent loss from adverse selection noted above. Thus, the net effect of competition in the Harvard circumstance appears to be beneficial, although the adverse selection losses were quite large.

With few exceptions [Wholey et al. (1995), Feldman and Dowd (1993), Baker and Corts (1995)], few studies have examined how competition affects health insurance premiums. It is often difficult to gather data on premiums, since most insurers charge different groups different amounts. In addition, premiums need to be adjusted for differences in the quality of benefits, but the many dimensions of quality are very difficult to control for. Thus, the tradeoff between cost savings and adverse selection in other situations is generally unknown.

6.7. Risk adjustment

The fundamental question about health insurance design is how to achieve the benefits of competition while containing the costs of adverse selection. A natural solution is suggested by the model above. Suppose that individuals were not charged the full difference in premiums between plans, but that instead the employer or government entirely running the insurance system offset some of the difference. For example, if the generous plan has above average risks in the amount $E[s|s > s^*] - E[s]$, the government would give the plan a per capita subsidy equal to this amount. The subsidy would be financed by a tax on the moderate plan, which has below average risks, by the amount $E[s|s < s^*] - E[s]$. The contribution from the plans would just match,[46] so there would be no net cost to the government.

In a competitive market, plans that receive subsidies (or are forced to pay taxes) would pass these subsidies on to consumers. Therefore, the premium for the generous plan would fall to $P_G = s_G - \text{subsidy}_G = E[s]$, and the premium for the moderate plan would rise to $P_M = \alpha s_M + \text{tax}_M = E[s] - (1-\alpha)s_M$. The adjusted premium difference between the plans, which individuals would face, would thus be

$$\Delta P^{\text{adj}} = P_G - P_M = (1-\alpha)s_M. \tag{16}$$

[46] This is because, taking expectations, $(E[s|s > s^*] + E[s|s < s^*])/2 = E[s]$.

This quantity is the savings for the *average* person in the moderate plan. It is closely related to the optimal price difference in Equation (15), which is the savings for the *marginal* person in the moderate plan. Plan choices made on the basis of the price difference in Equation (16), though not optimal, are likely to be more efficient than plan choices made on the basis of unadjusted price differences.

This form of differential payment is termed "risk adjustment" [Van de Ven and Ellis (2000)]. Risk adjustment must be carried out by some entity that can require individuals to insure or convince them to do so through subsidies. Otherwise, low cost individuals would choose not to participate. One possibility would be for the government to impose risk adjustment, whoever is the payer. But employers providing subsidized health insurance can do the job just as well. Employers have an incentive to risk adjust since it promotes efficiency and thus lowers the overall cost of providing health coverage.[47]

Empirically, risk adjustment can be carried out in four ways. Plans can pay or receive payments based on: (1) demographic variables (for example, more for taking on older people); (2) medical conditions (for example, more for people with diabetes); (3) past medical expenditures, which help predict future expenditures; or (4) actual experience in a year (for example, $50,000 extra for each organ transplant patient). The first three approaches attempt to predict experience; the last is after-the-fact reinsurance.

The tradeoffs between these different forms of risk adjustment are related to the ability of health plans to manipulate the risk adjustment system. Information about diagnosis, past claims, and actual use increase the ability to measure differential enrollment, but are susceptible to distortion by the plans. For example, plans may code borderline people as having diabetes if risk adjustment is done on the basis of the number of diabetics. Plans might creatively assign costs to high cost cases, when such cases are largely reimbursed. Even if risk adjustment is done on a prospective basis, plans have an incentive to exaggerate current sickness and expenditure levels, since the vast majority of insureds stick with their plans from year to year. A final, at least theoretical, concern about risk adjustment is that it may diminish plans' incentives to maintain their enrollees' health. Keeping people healthy disqualifies the plan from receiving additional risk adjustment payments, thus reducing the value of the health investment.

Because so few employers or governments have used formal risk adjustment systems, the relative advantages and drawbacks of different risk adjustment methodologies are unknown. New efforts may provide some of this information, however. In January 1999, in a major initiative, the federal government announced its intention to employ risk adjustment on the basis of past diagnoses to pay HMOs that enroll individuals in Medicare. Evaluating the impact of this system is a major research priority.

[47] Some employers have made second-best efforts to implement risk adjustment, at times inadvertently. The heavy subsidy of premiums – many employers pay 85 percent or more – in effect covers 85 percent of cost differentials due to varying mixes of insureds. Alas, heavy subsidies also significantly diminish the incentives of insureds to shop around, hence of health plans to hold down their costs.

7. Person-specific pricing, contract length, and premium uncertainty

Adverse selection is a problem of asymmetric information – individuals know their likely medical care utilization but insurers either do not, or are not allowed to use this information. Increasingly, however, information is becoming equalized. Insurers question individuals or monitor their past utilization to forecast their future costs. Equipped with such knowledge, insurers may know more about expected costs for the groups they are insuring than the members of the groups do themselves.

Insurers can use this information to set premiums. While such "experience rating" is rare at the individual level, it is common at the group level. Most private health insurance in the United States is at least partly experience rated. The bigger the group purchasing insurance, the more likely is experience rating. Hence, older and sicker groups are charged more per capita for the same coverage.

But experience rating creates its own set of problems, particularly when carried out at the individual level. When people face premiums that depend on their sickness, they are denied a form of insurance – the ability to obtain the same insurance premiums as their peers at the same price. The welfare loss can be significant.

Consider, for example, a situation where individuals are insuring themselves, diabetes is the only disease, and both people and plans know who is diabetic. Plans would offer full insurance to everyone but would charge diabetics more than non-diabetics; after all, no one who is not diabetic would be willing to pay extra to insure the diabetics. Given the distribution of diabetics and non-diabetics, the higher premiums charged to diabetics create a distributional issue. Diabetics pay more, and non-diabetics pay less relative to level premiums.

But from an *ex ante* perspective, before anyone knows who will contract diabetes, the distributional issue represents an efficiency loss. Suppose that before an individual knew if she would be diabetic or not – potentially before birth – she was offered insurance against the risk that she would become diabetic and thus face higher insurance premiums in the future. Full insurance would guarantee that if she developed diabetes, the policy would give her sufficient income each year to cover the higher diabetes premium she would then face. The benefits would be financed by payments from non-diabetics. Individuals would be willing to purchase this insurance were it sold at fair odds; they get a reduction in financial risk at no expected cost.

In real-world markets, however, such insurance against falling into a worse risk class is not offered. Some of the insurance would have to be purchased before birth. People obviously cannot do this, and even their parents might be unable to buy it for them, if there is a genetic predisposition towards disease. Other insurance could wait until mid-life for the unpredictable infirmities of old age. The key is to contract for insurance before the risk is resolved. While long-term anticipatory insurance is possible, health insurance in actual markets is rarely sold for over one year. People consequently lose welfare *ex ante*; there is an insurance policy they want but cannot obtain.

This loss at first may seem counterintuitive: everyone has full information and everyone gets full insurance every year. Where is the source of the loss? The welfare loss derives from a missing market for insurance against one's risk type. Risk-averse individuals would like to insure against the possibility of being discovered to be high risk. There is no market where they can do so, however. Given that a market is missing, there is no guarantee that efficient pricing on the basis of known information as opposed to level pricing (as if ignorant) will enhance welfare. This illustrates the theory of the second-best. The market failure might also be thought of as a recontracting failure. We recontract for health insurance annually despite the fact that we learn about expected future health costs during the year. Such periodic recontracting breaks the contractual arrangements that would characterize optimal insurance.

This problem has variously been termed the problem of renewable insurance or the problem of intertemporal insurance [Pauly, Kunreuther, and Hirth (1995), Cochrane (1995), Cutler (1996) and Zeckhauser (1974)]. It is likely to grow in importance in health insurance markets as our ability to predict medical spending rises, as it will, for example, through advances in genetic screening. We note this as the fourth lesson of health insurance:

> *Lesson* 4: *Information and long-term insurance*. More information about individual risk levels allows for more efficient pricing of risk, but portends a welfare loss from incomplete insurance contracts.

Might markets develop to deal with this problem? Some possibilities suggest themselves. People might purchase insurance for their lifetime rather than annually. If insurance choices were made early enough (or high-cost people were compensated when insurance choices were made), people would not suffer from knowledge gained over time. Long-term purchases, such as those associated with whole life insurance, are made in this fashion. Individuals buy level premium life insurance when they are young and healthy; they will wish to retain it, even if relatively healthy, when they grow old and annual risks escalate.

In theory, health insurance could be sold for the long term on a level premium basis. In practice, matters will be more complex. Much health insurance is now bundled with the provision of care. If an individual left a geographic region, he might have to change provider, and no new provider/insurer would want to take him own at his old level rate. Portability is but one problem. Once individuals purchase lifetime medical insurance, why should an insurer strive for efficiency when people are stuck in his plan? This problem is exacerbated since the insurer must agree to pay for or provide a changing level of services. Health insurance policies optimally change from year to year, as medical technology improves and knowledge about optimal treatments expands. Finally, with future medical costs so unpredictable, insurers cannot take on the risk, which would apply to all policies, that costs will escalate beyond expectation. With life insurance, by contrast, portability, changing service mix, and varying costs are not problems.

A second approach to long-term health insurance would be to develop policies offering insurance against learning one is high cost [Cochrane (1995)]. Imagine that people purchase two insurance policies in a year; one to cover their medical costs that year, and a second to cover any increase in premiums they may face in the future. The second policy – the "premium insurance" policy – might look like a standard health insurance policy: people pay in money and if they learn they are likely to have high costs in the future they receive money back. Full premium insurance would give people an amount of money equivalent to the discounted expected increase in their future medical spending they learn about during the year.[48] Why don't we observe premium insurance? Several factors have been identified. Moral hazard (people with premium insurance would take insufficient care of their health) and adverse selection (people expecting declines in health would more likely take up the insurance) are possibilities.

The aggregate risk phenomenon provides still a third explanation [Cutler (1996)]. Full premium insurance would have to insure a person against the risk that the medical policy that a representative individual will need in the future will cost more then than it is forecast to cost today. But future medical costs are not known. For example, a half century ago, the cost of treating cardiovascular disease patients was minimal with little prospect for rapid increase. Bypass surgery, angioplasty, and the like unexpectedly increased the cost of treating cardiovascular disease. Diversifying such a risk of significant cost increases for a common ailment is not possible. It is what is termed an aggregate as opposed to an idiosyncratic risk, where the latter apply to individuals one at a time. Insurers generally eschew aggregate risks. By contrast, insurers accept risks readily when they can lean comfortably on the Law of Large Numbers to spread them, as they can with idiosyncratic risks. They generally refuse to write insurance for risks that are unpredictable or nondiversifiable since they could bankrupt the company. Cost increases associated with future medical care suffer both disqualifications.

The result is that even though improved insurer information may reduce adverse selection over time, problems in insurance markets may grow. If people are increasingly charged on the basis of their individual risk characteristics, the efficiency losses could be severe.

Does employer-based insurance, where individuals choose from a menu of options, help? Under such plans, there is a range of potential costs individuals can face for choosing more generous insurance. At one extreme such plans are fully subsidized; people pay the same amount for each plan. At the other extreme there is no subsidy; people pay the expected cost in each plan on a group or individual basis. A system of risk adjustment lies in between; people pay the average cost of more generous plans assuming the mix of insureds is constant across plans.

We have stressed the efficiency aspects of risk-adjusted premiums, but such a system may not spread risks to a sufficient extent. Even in the perfect risk-adjusted equilibrium, the sick will pay more than the healthy, since they will be more attracted to the generous plan. People would presumably like to insure some of even this efficient price difference. There is, in terms of our earlier discussion, a tradeoff between moral hazard and risk sharing. Risk spreading considerations suggest that people should pay nothing

[48] This is related to the solution in Pauly, Kunreuther, and Hirth (1995). They propose paying a large premium in the first year, which is used to finance additional care for those who become sick in later years.

additional for selecting more generous plans, assuming risk level was the driving factor in their choice. Efficiency dictates that they should pay the expected additional cost they incur by choosing more generous care. The optimal differential lies between the two extremes, at the point where the marginal costs in terms of misallocation of people across plans exactly offsets the marginal benefits of increased risk sharing. Of course, price setting to this level of refinement may not be possible.

8. Insurance and health outcomes

Our empirical analysis to this point has focused on the impact of health insurance on medical spending. Ultimately, people care about health insurance because they are concerned about their health. A central research issue is therefore how alternate insurance arrangements affect health.

Much policy rhetoric expounds on the effects of not having insurance on health. Evidence on this issue shows that the effect of being without insurance can be large. See Weissman and Epstein (1994) for a review. For our purposes, however, we are interested in how variations among the set of insurance plans affect health. One might expect an attenuated version of the same finding – that people carrying less generous insurance, either indemnity insurance with high cost sharing or managed care insurance, would suffer worse health outcomes than people with more generous insurance. This might be particularly expected since medical treatment differs across insurance categories.

But several factors work in the other direction. Some of the additional care provided under more generous insurance may be iatrogenic (harmful to the patient), conceivably provided by physicians to increase their income. Perhaps more important, managed care policies may improve outcomes. One feature of managed care is that it standardizes the care that is received by classes of patients. These standards, if based on sound science and carefully crafted to patient characteristics, may be superior to what physicians conclude on their own. In addition, managed care usually involves less cost sharing for primary care, preventive services, and prescription drugs than does indemnity insurance. Greater use of these services may improve health outcomes.

Evidence on the effect of different insurance arrangements on health outcomes generally suggests very little difference in health produced across plans. The clearest findings on the impact of differing levels of demand-side cost sharing emerge from the Rand Health Insurance Experiment [Newhouse et al. (1993)]. The Rand study measured a broad array of health indicators. For most people, outcomes did not differ across plans. This is true even though spending differed across plans by up to one-third. Insurance did have a small effect on the health of the sick poor: poor people achieved better outcomes in more generous plans with blood pressure control, vision correction, and filling decayed teeth. Of course, the Health Insurance Experiment lasted for only a few years, which may have tilted the test against more generous plans. Increased primary and preventive care, even if strongly beneficial, may not be so important in such a short period of time.

Many studies have examined the impact of supply-side cost sharing on medical outcomes. Such studies must adjust for differing population mixes across plans, which is a difficult challenge. Important evidence comes from the implementation of prospective payment for hospital admissions covered by Medicare. At the time of the change, the critics of the new prospective payment warned that patients would be discharged from hospitals "quicker and sicker." Several papers examined this question, as shown in Table 7. The most detailed studies are the papers grouped under Kahn et al. (1990), which examined patient medical reviews before and after prospective payment was implemented to measure changes in health. That research found no increase in adverse outcomes for the average patient after prospective reimbursement, although it did find that with prospective payment more patients were discharged from the hospital in an unstable condition. The lack of significant adverse effect on quality of care was also found by Desharnais, Chesney, and Fleming (1988).

Some papers have found evidence of adverse outcomes resulting from prospective payment. Fitzgerald et al. (1987, 1988) found that patients admitted to a hospital in the midwest with a hip fracture were discharged sooner after prospective payment but were more likely to be in a nursing home 6 months and 1 year after the hip fracture. In response, many other researchers have examined this question, finding that length of stay for hip fracture patients fell but there was no effect on nursing home residence, functional status, or mortality after 1 year [Gerety et al. (1989), Palmer et al. (1989), Ray, Griffin, and Baugh (1990)].

Two studies have looked at the impact not of the prospective payment system, but of the revenue changes stemming from prospective payment [Cutler (1995) and Staiger and Gaumer (1995)]. These studies compared patients admitted to hospitals that lost revenue with patients admitted to hospitals that gained revenue. The former patients experience a compression of mortality into the period just after the hospital admission in comparison to the latter; some classes of patients that formerly survived several months after being hospitalized did not live as long after revenues fell. The effect diminished over the succeeding year, however. For patients who survived a year or longer, there was no increase in mortality.[49] The authors conclude that price changes have a small adverse effect on the very sick, but little effect on others.

A second set of evidence examines the effect of managed care on health. Miller and Luft (1997) summarize 35 studies comparing medical outcomes in managed care and indemnity insurance. They find no clear difference; some studies find that managed care does worse, while an equally large number find it does better. Many find no difference in outcomes.[50]

One is tempted to conclude from these findings that managed care is superior to traditional insurance – it saves money without substantial adverse effects. Such a conclusion

[49] After a phase-in period, hospital payments in total were not substantially affected by prospective payment, so these results are consistent with the Kahn et al. (1990) finding of no change in health for the average patient.

[50] See also Cutler, McClellan, and Newhouse (1998).

is premature, however, until long-term evidence on the effect of managed care has been obtained. We note the focus on health and lack of conclusive results as the fifth lesson of health insurance.

> *Lesson* 5: *Health insurance and health.* The primary purpose of health insurance and delivery is to improve health. Unfortunately, conclusive results are not in on which insurance and provision arrangements do this most effectively.

9. Conclusions and implications

Health insurance has a complex anatomy. The lens of economics brings many of its critical features – incentives, risk spreading and asymmetric information – into sharp focus. The understanding thus gained, however helpful, does not solve all of the problems. Indeed, the primary message of this chapter is that health insurance design is a challenging exercise in the second-best. On each of a variety of dimensions, goals must be traded off against each another, since first principles are in conflict.

Our lessons about health insurance, highlighted in Table 10, are instructive in this respect. We start with a single insurer. Lesson 1 stresses the tradeoff between efficient risk spreading and excessive utilization. Optimal risk sharing puts all the burden on the risk-neutral insurer, but this induces moral hazard (excess consumption of services) and possibly supplier-induced demand (excessive provision). Lesson 2 finds that integration of insurance and provision of services, which is absent in other insurance contexts, may be desirable to align producer and insurer incentives in the delivery of medical care.

Lessons 3 and 4 highlight second-best problems in the health insurance marketplace. Lesson 3 shows that competitive markets, the traditional lodestar of economics, may have undesirable side effects in health insurance. Most important, competition induces adverse selection, hence the misallocation of people to plans and the incentive for insurers to trim their offerings to deter the sick. In theory at least, risk-adjustment methods, which are just now being tried in practice, can counter these phenomena. Lesson 4 alerts us, however, that even if we slay the dragons of adverse selection and plan manipulation, a fierce risk remains. Since insurance is written on an annual basis, individuals are denied crucial protection against becoming sick and having their premiums escalate substantially in the future.

Lesson 5 reminds us that the ultimate goal of health insurance does not involve the usual economic concepts of prices, incentives and costs. Rather, the central objective of health insurance is to maintain and enhance our health. The payoff question, therefore, is what can we get for alternative levels of expenditure? The contribution of economics is to enable us to sketch the production function.

Health insurance is a service in society, like a haircut or tennis lesson. Why then does health insurance cause so many more problems than the other two? Both the insurance aspect, and its area of application, health, produce problems. In any insurance situation, moral hazard and adverse selection plague outcomes. In the case of health insurance, the problems are magnified, since health-promoting and care-seeking actions are difficult

Table 10
Five central lessons about health insurance

Lesson 1: *Risk spreading versus incentives*	Health insurance involves a fundamental tradeoff between risk spreading and appropriate incentives. Increasing the generosity of insurance spreads risk more broadly but also leads to increased losses because individuals choose more care (moral hazard) and providers supply more care (principal-agent problems).
Lesson 2: *Integration of insurance and provision*	Medical care is unlike other insurance markets in that insurers are often involved in the provision of the good in addition to insuring its cost. The integration of insurance and provision, intended to align incentives, has increased over time. Managed care, where the functions are united, is an extreme version. Under it, doctors have dual loyalties, to the insurer as well as the patient.
Lesson 3: *Competition and consumer identity*	When consumer identity affects costs, competition is a mixed blessing. Allowing individuals to choose among competing health insurance plans can allocate people to appropriate plans and provide incentives for efficient provision. But it can also bring with it adverse selection – the tendency of the sick to prefer the most generous plans. Adverse selection induces people to enroll in less generous plans, so they can be in a healthier pool, and gives plans incentives to distort their offerings to be less generous with care for the sick.
Lesson 4: *Information and long-term insurance*	More information about individual risk levels allows for more efficient pricing of risk, but portends a welfare loss from incomplete insurance contracts.
Lesson 5: *Health insurance and health*	The primary purpose of health insurance and delivery is to improve health. Unfortunately, conclusive results are not in on which insurance and provision arrangements do this most effectively.

to monitor, and it is widely believed to be unfair to charge people more if they contract diseases that are not their fault. Moreover, the payoff from health insurance, unlike say life insurance, is quite variable, and subject to human choice made after the contract is written. In addition, for justifiable reasons, health care is written on an annual basis, though today's chance outcomes often have cost implications that stretch for decades. Finally, health has a privileged position above other goods and services. For a range of philosophical and moral reasons, societies care deeply that their citizens receive health care, even if that is not what they would buy were they given the money.

These fundamental issues surrounding the equitable and efficient provision of health insurance make government involvement inevitable, and in many contexts desirable. The range of government involvement in health care and health insurance is enormous. At one end, many governments provide medical care directly; they raise money through taxes, hire doctors and run public hospitals. Less extreme are countries where the government is the sole insurer, but provision of services remains private. More market-oriented systems such as the United States have most of the population in private insurance and most of the provision of medical care done by private providers. Even there, though, government plays a sizeable role, refereeing the playing field and insuring those who the market would leave behind. Thus, the federal government insures people through Medicare and Medicaid, provides tax subsidies to private insurance, defines permissible structures for supplementary Medicare insurance, and requires insurers to cover people who recently lost or changed jobs. Moreover, many states mandate that particular benefits be part of any health insurance plan.

Discussions of medical care reform in the United States and elsewhere often lead to extreme positions. Advocates at one end believe that the problems with markets in health care are so severe that government control, at least of expenditures, is necessary. The Canadian system – tax-supported, privately provided, but publicly regulated – is held up as an exemplar. The claimed merits are that one insurer eliminates adverse selection, tight supply restrictions manage costs, and tax financing enables everyone to be insured. Of course, in such a system competition between insurers plays no role in promoting efficiency.

At the other extreme are free-market advocates, who believe that market institutions, if guided correctly, would produce a superior outcome. The government should stay out of the insurance business, but implement a risk adjustment system, directly or at arm's length, so that people face efficient prices. Moreover, the government should remove the tax subsidy favoring employer provision of insurance, which would lead to trimmed plan generosity and more cost sharing by employees. Where necessary, the government should give high cost individuals risk-related subsidies that enable them to buy health insurance in the marketplace.

The fundamental difference between the public and private approaches to medical care reform is indicative of the enormous problems in medical care markets and the central role that health plays in our utility. Can risk adjustment work well enough to deter plan manipulation and cream skimming? Without subsidies, would employers provide insurance? If they stopped doing so, how many more people would be uninsured, and how much would their health suffer? These are questions at the heart of health insurance reform.

And beyond the question about organizing the health insurance system, there remain questions of how plans should interact with providers. Should providers be paid by capitation or by fee-for-service, or might there be a happy medium? Will providers respond to a payment schedule by either skimping on patients or driving up costs? Only experience in the future, coupled with a delicately balanced wisdom, will enable us to answer these questions.

Economics does not offer robust conclusions about the virtues and liabilities of markets in second-best situations. Hence, it is not surprising that the debate on who should provide health insurance and how it should best be structured rages on, even among economists. Ultimately, of course, many of the issues cannot be answered on the basis of first principles, much less the dogma that is too often brought to the debate. They require empirical investigations.

An impressive array of data has been brought to bear one-to-one on central issues in health insurance, but the grand synthesis needed for effective prescription awaits us. Which medical system around the world is best, and what would make it even better? Might the best system for Germany or Japan differ significantly from that for the United States? To understand the attractiveness of alternative health insurance structures, not unlike much of medical care itself, many consequences must be weighed, and many side effects considered. This chapter provided an anatomy to help organize those investigations.

Appendix

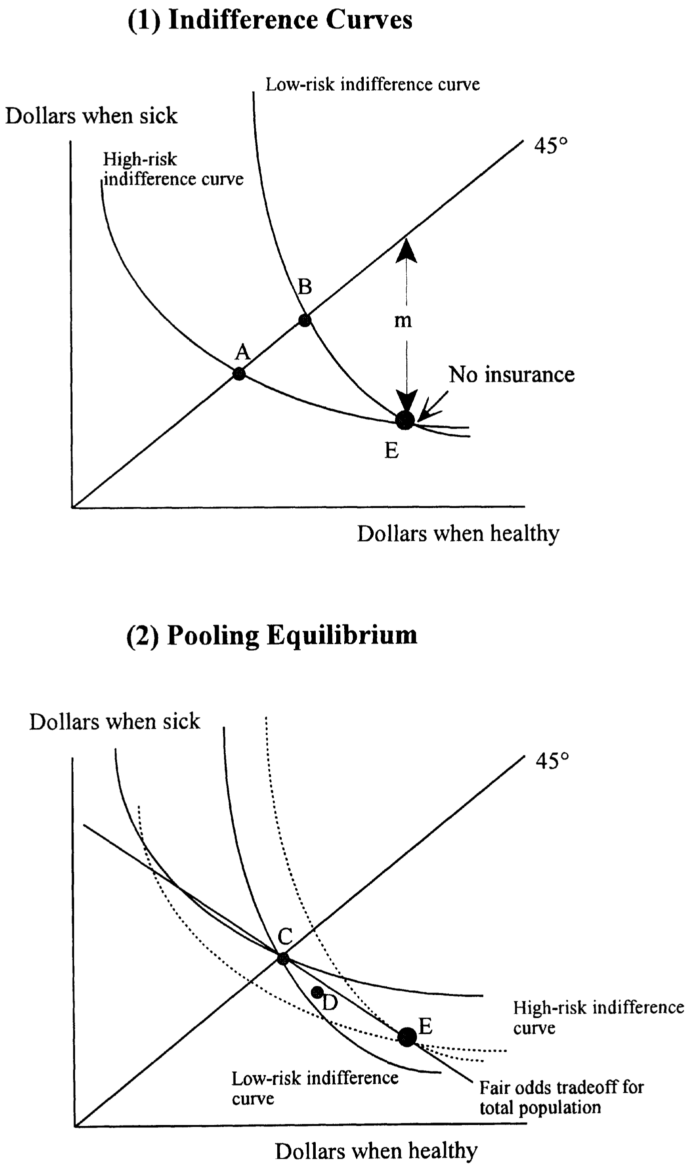
This appendix shows the classic treatment of equilibrium with adverse selection and two individuals, from Rothschild and Stiglitz (1976).

For simplicity, assume that spending when sick, m, is the same for HIGH and LOW, i.e., there is no moral hazard. HIGH is more likely to be sick. Figure A(1) shows the indifference curves for these two people. LOW's indifference curve is steeper than HIGH's, since LOW is not willing to give up as much money when healthy to get a dollar when sick. With no moral hazard, both LOW and HIGH would optimally want full insurance, if charged their fair price for it. Points A and B represent their respective efficient levels of insurance when purchased at actuarially fair rates.

Figure A(2) shows the potential pooling equilibrium. The fair odds line that is shown is the average premium for the two together. At point C, both LOW and HIGH purchase full insurance at this price. But this equilibrium cannot prevail. If an insurer entered the market offering policy D, which has incomplete coverage but a lower premium, he would attract LOW but not HIGH. LOW prefers the policy because he gets the cost savings from not pooling with HIGH, which more than makes up for his loss of full insurance. This is parallel to what happens with the introduction of the basic plan in our numerical analysis, which breaks the pooling equilibrium at moderate.

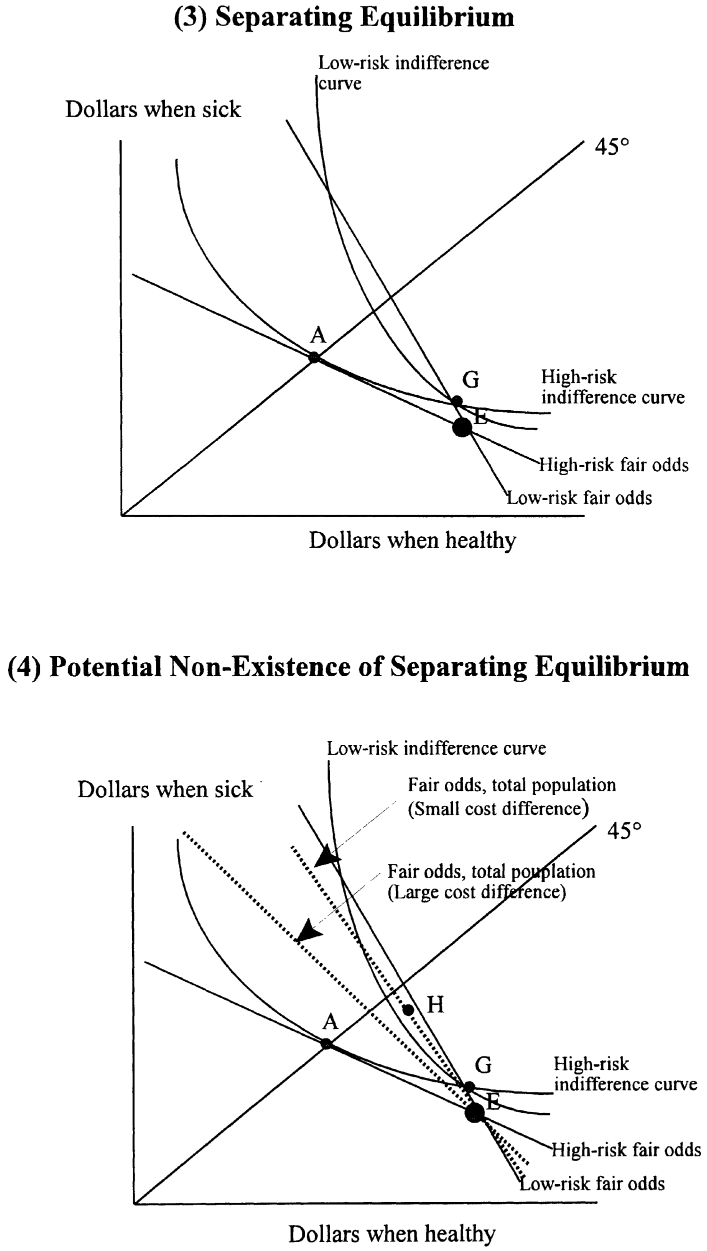
Figure A(3) shows the equilibrium with plan manipulation. HIGH receives full insurance (point A). To separate himself out and thereby reduce his payments, LOW insures incompletely, at point G. Point G makes HIGH just indifferent between staying in the full insurance plan and enrolling in the less generous, but less expensive, policy. Though optimality requires that both groups insure fully, only HIGH does so.

Figure A(4) shows how the separating equilibrium may be broken. We show two fair odds line for the average of HIGH and LOW – one where costs for the two are far apart and one where they are closer together (for simplicity, we show only one indifference curve for HIGH). In the case where HIGH and low have very different costs, the pooled fair odds line will not attract LOW; they do not want to pay the additional amount for more generous coverage because doing so necessitates pooling with HIGH. If the costs are closer together, in contrast, the average fair odds line for the two as a whole


(1) Indifference Curves
Low-risk indifference curve
Dollars when sick
High-risk indifference curve
45°
B
m
A
No insurance
E
Dollars when healthy
(2) Pooling Equilibrium
Dollars when sick
45°
C
D
E
High-risk indifference curve
Low-risk indifference curve
Fair odds tradeoff for total population
Dollars when healthy

Dashed lines are indifference curves through no insurance, point E

Figure A. Adverse selection and plan manipulation.



(3) Separating Equilibrium

Low-risk indifference curve
Dollars when sick
45°
A
G
High-risk indifference curve
E
High-risk fair odds
Low-risk fair odds
Dollars when healthy

(4) Potential Non-Existence of Separating Equilibrium

Low-risk indifference curve
Dollars when sick
Fair odds, total population (Small cost difference)
45°
Fair odds, total population (Large cost difference)
H
A
G
High-risk indifference curve
E
High-risk fair odds
Low-risk fair odds
Dollars when healthy

Figure A. (*Continued.*)

will be close to the fair odds line for LOW. Relative to points A and G, there may be a point such as H that will be preferred by LOW to the separating equilibrium. It will also be preferred by HIGH, who benefits from pooling with the healthier group in the population. It will thus undermine the separating equilibrium. With no stable pooling equilibrium and no stable separating equilibrium, the market will not reach an equilibrium.

References

Akerlof, G. (1970), "The market for 'Lemons': qualitative uncertainty and the market mechanism", Quarterly Journal of Economics 74:488–500.

Altman, D., D.M. Cutler and R.J. Zeckhauser (1998), "Adverse selection and adverse retention", American Economic Review 88(2):122–126.

Arrow, K. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53(5):941–973.

Arrow, K. (1965), Aspects of the Theory of Risk Bearing (Yrjo Jahnssonin Saatio, Helsinki).

Arrow, K. (1985), "The economics of agency", in: J. Pratt and R. Zeckhauser, eds., Principals and Agents: The Structure of Business (Harvard Business School Press, Cambridge, MA) 37–51.

Baker, L.C., and K.S. Corts (1995), "The effects of HMOs on conventional insurance premiums: theory and evidence", NBER Working Paper No. 5356.

Baumgardner, J. (1991), "The interaction between forms of insurance contract and types of technical change in medical care", Rand Journal of Economics 22(1):36–53.

Beck, R.G. (1974), "The effects of co-payment on the poor", Journal of Human Resources 9(1):129–142.

Berk, M.L., and A.C. Monheit (1992), "The concentration of health expenditures: an update", Health Affairs 11(4):145–149.

Bhattacharya, J., W.B. Vogt, A. Yoshikawa and T. Nakahara (1996), "The utilization of outpatient medical services in Japan", Journal of Human Resources 31(2):450–476.

Bice, T.W. (1975), "Risk vulnerability and enrollment in a prepaid group practice", Medical Care 13(8):698–703.

Blomqvist, A.G. (1997), "Optimal non-linear health insurance", Journal of Health Economics 16(3):303–321.

Brown, R.S., et al. (1993), "The Medicare risk program for HMOs – Final summary report on findings from the evaluation", Final Report under HCFA Contract No. 500-88-0006 (Mathematica Policy Research, Inc., Princeton, NJ).

Buchanan, J.L., E.B. Keeler, J.E. Rolph and M.R. Holmer (1991), "Simulating health expenditures under alternative insurance plans", Management Science 37(9):1067–1089.

Buchmueller, T.C., and P.J. Feldstein (1997), "The effect of price on switching among health plans", Journal of Health Economics 16(2):231–247.

Cardon, J., and I. Hendel (1996), "Asymmetric information in health care and health insurance markets: evidence from the National Medical Expenditure Survey", mimeo.

Carroll, N.V., and W.G. Erwin (1987), "Patient shifting as a response to Medicare prospective payment", Medical Care 25(12):1161–1167.

Cave, J. (1985), "Subsidy equilibrium and multiple-option insurance markets", in: R. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research. Biased Selection in Health Care Markets, Vol. 6 (JAI Press, Greenwich, CT) 27–45.

Cherkin, D.C., L. Grothaus and E.H. Wagner (1989), "The effect of office visit copayments on utilization in a health maintenance organization", Medical Care 27(7):669–679.

Cochrane, J. (1995), "Time consistent health insurance", Journal of Political Economy 103(3):445–73.

Colle, A.D., and M. Grossman (1978), “Determinants of pediatric care utilization”, Journal of Human Resources 13(Suppl.):115–153.

Conrad, D.A., D. Grembowski and P. Milgrom (1985), “Adverse selection within dental insurance markets”, in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 171–190.

Congressional Budget Office (1992), “The effects of managed care on use and costs of health services”, mimeo.

Cutler, D.M. (1991), “Estimating the effect of reimbursement policy on medical outcomes”, Doctoral dissertation (Massachusetts Institute of Technology, MA).

Cutler, D.M. (1995), “The incidence of adverse medical outcomes under prospective payment”, Econometrica 63(1):29–50.

Cutler, D.M. (1996), “Why don’t markets insure long-term risk?”, mimeo.

Cutler, D.M., M. McClellan and J.P. Newhouse (1998), “What does managed care do?”, mimeo.

Cutler, D.M., and S.J. Reber (1998), “Paying for health insurance: the tradeoff between competition and adverse selection”, Quarterly Journal of Economics 113(2):433–466.

Cutler, D.M., and R.J. Zeckhauser (1998), “Adverse selection in health insurance”, in: A. Garber, ed., Frontiers in Health Policy Research, Vol. 1 (MIT Press, Cambridge, MA) 1–31.

Davis, K., and L.B. Russell (1972), “The substitution of hospital outpatient care for inpatient care”, Review of Economics and Statistics 54(2):109–120.

De Meza, D. (1983), “Health insurance and the demand for medical care”, Journal of Health Economics 2(1):47–54.

DesHarnais, S.I., J. Chesney and S. Fleming (1988), “Trends and regional variations in hospital utilization and quality during the first two years of the prospective payment system”, Inquiry 25:374–382.

DesHarnais, S.I., R. Wroblewski and D. Schumacher (1990), “How the Medicare prospective payment system affects psychiatric patients treated in short-term general hospitals”, Inquiry 27:382–388.

Diamond, P. (1998), “Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates”, American Economic Review 88(1):83–95.

Dowd, B., and R. Feldman (1985), “Biased selection in twin cities health plans”, in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 253–271.

Eggers, P.W. (1980), “Risk differential between Medicare beneficiaries enrolled and not enrolled in an HMO”, Health Care Financing Review 1:91–99.

Eichner, M.J. (1998), “Incentives, price expectations and medical expenditures: an analysis of claims under employer-provided health insurance”, mimeo.

Eichner, M.J., M. McClellan and D. Wise (1998), “Insurance or self-insurance?: Variation, persistence, and individual health accounts”, in: D. Wise, ed., Inquiries in the Economics of Aging (University of Chicago Press, Chicago, IL) 19–45.

Ellis, R.P. (1985), “The effect of prior-year health expenditures on health coverage plan choice”, in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 127–147.

Ellis, R.P. (1989), “Employee choice of health insurance”, Review of Economics and Statistics 71(2):215–223.

Ellis, R.P. (1998), “Creaming, skimping and dumping: provider competition on the intensive and extensive margins”, Journal of Health Economics 17(5):537–555.

Ellis, R.P., and T.G. McGuire (1986), “Provider behavior under prospective reimbursement: cost sharing and supply”, Journal of Health Economics 5(2):129–152.

Ellis, R.P., and T.G. McGuire (1996), “Hospital response to prospective payment: moral hazard, selection, and practice-style effects”, Journal of Health Economics 15:257–277.

Eze, P., and B. Wolfe (1993), “Is dumping socially inefficient? An analysis of the effect of Medicare’s prospective payment system on the utilization of Veterans Affairs inpatient services”, Journal of Public Economics 52:329–344.

Farley, P.J., and A.C. Monheit (1985), "Selectivity in the demand for health insurance and health care", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, vol. 6 (JAI Press, Greenwich, CT) 231–252.

Feder, J., J. Hadley and S. Zuckerman (1987), "How did Medicare's prospective payment system affect hospitals?", New England Journal of Medicine 317(14):867–873.

Feldman, R., and D. Dowd (1991), "Must adverse selection cause premium spirals?", Journal of Health Economics 10(3):350–357.

Feldman, R., and B. Dowd (1993), "The effectiveness of managed competition in reducing the costs of health insurance", in: R.B. Helms, ed., Health Policy Reform: Competition and Controls (AEI Press, Washington, DC) 176–217.

Feldman, R., M. Finch, B. Dowd and S. Cassou (1989), "The demand for employment-based health insurance plans", Journal of Human Resources 24(1):117–142.

Feldstein, M.S. (1970), "The rising price of physicians' services", Review of Economics and Statistics 52(2):121–133.

Feldstein, M.S. (1971), "Hospital cost inflation: a study of nonprofit price dynamics", American Economic Review 60:853–872.

Feldstein, M.S., and B. Friedman (1977), "Tax subsidies, the rational demand for insurance and the health care crisis", Journal of Public Economics 7(2):155–178.

Feldstein, P.J. (1964), "General report", Report of the Commission on the Cost of Medical Care, Part 1 (American Medical Association, Chicago).

Fisher, C.R. (1992), "Hospital and Medicare financial performance under PPS, 1985–90", Health Care Financing Review 14(1):171–183.

Fitzgerald, J.F., L.F. Fagan, W.M. Tierney and R.S. Dittus (1987), "Changing patterns of hip fracture care before and after implementation of the prospective payment system", Journal of the American Medical Association 258(2):218–221.

Fitzgerald, J.F., P.S. Moore and R.S. Dittus (1988), "The care of elderly patients with hip fracture: changes since implementation of the prospective payment system", New England Journal of Medicine 319(21):1392–1397.

Folland, S., and R. Kleiman (1990), "The effect of prospective payment under DRGs on the market value of hospitals", Quarterly Review of Economics and Business 30(2):50–68.

Frank, R.G., and J.R. Lave (1986), "The effect of benefit design on the length of stay of Medicaid psychiatric patients", Journal of Human Resources 21(3):321–337.

Frank, R.G., and J.R. Lave (1989), "A comparison of hospital responses to reimbursement policies for Medicaid psychiatric patients", RAND Journal of Economics 20(4):588–600.

Frank, R.G., and T. McGuire (1998), "Economic functions of carve-outs", American Journal of Managed Care 4(SP):SP31–SP39.

Frank, R.G., T.G. McGuire and J.P. Newhouse (1995), "Risk contracts in managed mental health care", Health Affairs 14(3):50–64.

Frank, R.G., J. Glazer and T.G. McGuire (1998), "Measuring adverse selection in managed health care", NBER Working Paper no. 6825, December.

Fuchs, V.R., and M.J. Kramer (1972), "Determinants of expenditures for physicians' services in the United States, 1948–68", National Bureau of Economic Research Occasional Paper Series, No. 117.

Gaumer, G.L., E.L. Poggio, C.G. Coelen, C.S. Sennett and R.J. Schmitz (1989), "Effects of state prospective reimbursement programs on hospital mortality", Medical Care 27(7):724–736.

Gerety, M.B., V. Soderholm-Difatte and C.H. Winograd (1989), "Impact of prospective payment and discharge location on the outcome of hip fracture", Journal of General Internal Medicine 4(5):388–391.

Glied, S. (2000), "Managed care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 13.

Goldman, F., and M. Grossman (1978), "The demand for pediatric care: an hedonic approach", Journal of Political Economy 86(2):259–280.

Griffith, M.J., N. Baloff and E.L. Spitznagel (1984), "Utilization patterns of health maintenance organization disenrollees", Medical Care 22(9):827–834.

Guterman, S., S.H. Altman and D.A. Young (1990), "Hospitals' financial performance in the first five years of PPS", Health Affairs 9(1):125–134.

Guterman, S., and A. Dobson (1986), "Impact of the Medicare prospective payment system for hospitals", Health Care Financing Review 7(3):97–114.

Hadley, J., S. Zuckerman and J. Feder (1989), "Profits and fiscal pressure in the prospective payment system: their impacts on hospitals", Inquiry 26:354–365.

Hodgkin, D., and T.G. McGuire (1994), "Payment levels and hospital response to prospective payment", Journal of Health Economics 13:1–29.

Hurley, J. (2000), "An overview of normative economics of the health sector", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 2.

Jackson-Beeck, M., and J.H. Kleinman (1983), "Evidence for self-selection among health maintenance organization enrollees", Journal of the American Medical Association 250(20):2826–2829.

Juba, D.A., J.R. Lave and J. Shaddy (1980), "An analysis of the choice of health benefits plans", Inquiry 17:62–71.

Kahn, K.L., et al. (1990) (series), "The effects of the DRG-based prospective payment system on quality of care for hospitalized Medicare patients", Journal of the American Medical Association 264(15):1953–1994 (eight articles).

Keeler, E.B., J.P. Newhouse and C.E. Phelps (1977), "Deductibles and demand: a theory of the consomer facing a variable price schedule under uncertainty", Econometrica 45:641–655.

Keeler, E.B., G. Carter and J.P. Newhouse (1998), "A model of the impact of reimbursement schemes on health plan choice", Journal of Health Economics 17(3):297–320.

Kotowitz, Y. (1987), "Moral hazard", in: New Palgrave Dictionary of Economics.

Langwell, K.M., and J.P. Hadley (1989), "Evaluation of the Medicare competition demonstrations", Health Care Financing Review 11(2):65–80.

Lave, J.R., R.G. Frank, C. Taube, H. Goldman and A. Rupp (1988), "The early effects of Medicare's prospective payment system on psychiatry", Inquiry 25:354–363.

Long, S.H., R.F. Settle and C.W. Wrightson (1988), "Employee premiums, availability of alternative plans, and HMO disenrollment", Medical Care 26(10):927–938.

Luft, H.S., J.B. Trauner and S.C. Maerki (1985), "Adverse selection in a large, multiple-option health benefits program: a case study of the California Public Employees' Retirement System", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 197–229.

Ma, C.A., and T. McGuire (1997), "Optimal health insurance and provider payment", American Economic Review 87(4):685–704.

Maibach, E., K. Dusenbury, P. Zupp et al. (1998), "Marketing HMOs to Medicare Beneficiaries: An Analysis of Four Medicare Markets" (Kaiser Family Foundation, Menlo Park, CA).

Manning, W.G., and M.S. Marquis (1996), "Health insurance: the tradeoff between risk pooling and moral hazard", Journal of Health Economics 15(5):609–639.

Marquis, M.S. (1992), "Adverse selection with a multiple choice among health insurance plans: a simulation analysis", Journal of Health Economics 11(2):129–151.

Marquis, M.S., and C.E. Phelps (1987), "Price elasticity and adverse selection in the demand for supplemental health insurance", Economic Inquiry 25(2):299–313.

McAvinchey, I.D., and A. Yannopoulos (1993), "Elasticity estimates from a dynamic model of interrelated demands for private and public acute health care", Journal of Health Economics 12(2):171–186.

McGuire, T.G. (1981), "Price and membership in a prepaid group medical practice", Medical Care 19(2):172–183.

McGuire, T.G. (2000), "Physician agency", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 9.

Menke, T. (1990), "Impacts of PPS on Medicare Part B expenditures and utilization for hospital episodes of care", Inquiry 27(2):114–126.

Merrill, J., C. Jackson and J. Reuter (1985), "Factors that affect the HMO enrollment decision: a tale of two cities", Inquiry 22(4):388–395.

Miller, R.H., and H.S. Luft (1997), "Does managed care lead to better or worse quality of care?", Health Affairs 16(5):7–25.

Mirrlees, J.A. (1971), "An exploration in the theory of optimum income taxation", Review of Economic Studies 38:175–208.

Morrisey, M.A., F.A. Sloan and J. Valvona (1988), "Medicare prospective payment and posthospital transfers to subacute care", Medical Care 26(7):685–698.

Newhouse, J.P. (1989), "Do unprofitable patients face access problems?", Health Care Financing Review 11(2):33–42.

Newhouse, J.P. (1996), "Reimbursing health plans and health providers: efficiency in production versus selection", Journal of Economic Literature 34(3):1236–1263.

Newhouse, J.P., and the Insurance Experiment Group (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge, MA).

Newhouse, J.P., and D.J. Byrne (1988), "Did Medicare's prospective payment system cause length of stay to fall?", Journal of Health Economics 7(4):413–416.

Newhouse, J.P., and C.E. Phelps (1974), "Price and income elasticities for medical care services", The Economics of Health and Medical Care (John Wiley & Sons, New York), ch. 9, 140–161.

Newhouse, J.P., and C.E. Phelps (1976), "New estimates of price and income elasticities of medical care services", The Role of Health Insurance in the Health Services Sector (National Bureau of Economic Research, New York), Chapter 7, 261–313.

Nichols, A., and R. Zeckhauser (1982), "Targeting transfers through restrictions on recipients", American Economic Review 72(2):372–377.

Palmer, R.M., R.M. Saywell Jr., T.W. Zollinger, B.K. Erner, A.D. LaBov, D.A. Freund, J.E. Garber, G.W. Misamore and F.B. Throop (1989), "The impact of the prospective payment system on the treatment of hip fractures in the elderly", Archives of Internal Medicine 149(10):2237–2241.

Pauly, M. (1968), "The economics of moral hazard: comment", American Economic Review 58:531–536.

Pauly, M. (1974), "Overinsurance and public provision of insurance: the roles of moral hazard and adverse selection", Quarterly Journal of Economics 88(1):44–54.

Pauly, M. (1986), "Taxation, health insurance and market failure", Journal of Economic Literature 24(2):629–675.

Pauly, M., H. Kunreuther and R. Hirth (1995), "Guaranteed renewability in insurance", Journal of Risk & Uncertainty 10(2):143–156.

Pauly, M., and S. Ramsey (1998), "Would you like suspenders to go with that belt? An analysis of optimal combinations of cost sharing and managed care", mimeo.

Phelps, C.E. (1973), "The demand for health insurance: a theoretical and empirical investigation", RAND Research Paper Series, No. R-1054-OEO.

Phelps, C.E., and J.P. Newhouse (1972a), "Effect of coinsurance: a multivariate analysis", Social Security Bulletin 20–28.

Phelps, C.E., and J.P. Newhouse (1972b), "Effects of coinsurance of demand for physician services", RAND Research Paper Series, No. R-976-OEO.

Phelps, C.E., and J.P. Newhouse (1974), "Coinsurance, the price of time, and the demand for medical services", Review of Economics and Statistics 56(3):334–342.

Plato, The Republic.

Price, J.R., and J.W. Mays (1985), "Biased selection in the Federal Employees Health Benefits Program", Inquiry 22(1):67–77.

Ramsey, S.D., and M. Pauly (1997), "Structural incentives and adoption of medical technologies in HMO and fee-for-service health insurance plans", Inquiry 34(3):228–236.

Ray, W.A., M.R. Griffin and D.K. Baugh (1990), "Mortality following hip fracture before and after implementation of the prospective payment system", Archives of Internal Medicine 150(10):2109–2114.

Rodgers, J., and K.E. Smith (1996), "Is there biased selection in Medicare HMOs?", Health Policy Economics Group Report (Price Waterhouse LLP, Washington, DC).

Roos, N.P., E. Shapiro and R. Tate (1989), "Does a small minority of elderly account for a majority of health care expenditures? A sixteen-year perspective", Milbank Quarterly 67(3-4):347–369.

Rosenthal, G. (1970), "Price elasticity of demand for short-term general hospital services", in: H.E. Klarman, ed., Empirical Studies in Health Economics (Johns Hopkins Press, Baltimore, MD) 101–117.

Rosett, R.N., and L. Huang (1973), "The effect of health insurance on the demand for medical care", Journal of Political Economy 81(March/April):281–305.

Rothschild, M., and J.E. Stiglitz (1976), "Equilibrium in competitive insurance markets: an essay on the economics of imperfect information", Quarterly Journal of Economics 90(4):630–649.

Russell, L.B., and C.L. Manning (1989), "The effect of prospective payment on Medicare expenditures", New England Journal of Medicine 320(7):439–444.

Sager, M.A., D.V. Easterling, D.A. Kindig and O.W. Anderson (1989), "Changes in the location of death after passage of Medicare's prospective payment system", New England Journal of Medicine 320(7):433–439.

Scitovsky, A.A., and N. McCall (1977), "Coinsurance and the demand for physician services: four years later", Social Security Bulletin 19–27.

Scitovsky, A.A., N. McCall and L. Benham (1978), "Factors affecting the choice between two prepaid plans", Medical Care 16(8):660–675.

Scitovsky, A.A., and N.M. Snyder (1972), "Effect of coinsurance on use of physician services", Social Security Bulletin 3–19.

Shaw, G.B. (1911), The Doctors Dilemma.

Sheingold, S.H. (1989), "The first three years of PPS: impact on Medicare costs", Health Affairs 8(3):191–204.

Sheingold, S.H., and T. Buchberger (1986), "Implications of Medicare's prospective payment system for the provision of uncompensated hospital care", Inquiry 23(4):371–381.

Sloan, F.A., M.A. Morrisey and J. Valvona (1988), "Medicare prospective payment and the use of medical technologies in hospitals", Medical Care 26(9):837–850.

Smith, A. (1776), The Wealth of Nations.

Spence, M., and R. Zeckhauser (1971), "Insurance, information, and individual action", American Economic Review 61(2):380–387.

Staiger, D., and G.L. Gaumer (1995), "Price regulation and patient mortality in hospitals", mimeo.

van de Ven, W.P.M.M., and R.P. Ellis (2000), "Risk adjustment in competitive health plan markets", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 14.

van de Ven, W.P.M.M., and R.C.J.A. van Vliet (1995), "Consumer surplus and adverse selection in competitive health insurance markets: an empirical study", Journal of Health Economics 14(2):149–169.

Wagstaff, A., and E.K.A. van Doorslaer (2000), "Equity in health care finance and delivery", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 34.

Welch, W.P. (1989), "Restructuring the Federal Employees Health Benefits Program: the private sector option", Inquiry 26(3):321–334.

Weissman, J., and A. Epstein (1994), Falling Through the Safety Net: Insurance and Access to Medical Care (Johns Hopkins University Press, Baltimore, MD).

Wholey, D., R. Feldman and J.B. Christianson (1995), "The effect of market structure on HMO premiums", Journal of Health Economics 14(1):81–105.

Williams, A., and R. Cookson (2000), "Equity in health", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 35.

Willke, R.J., W.S. Custer, J.W. Moser and R.A. Musacchio (1991), "Collaborative production and resource allocation: the consequences of prospective payment for hospital care", Quarterly Review of Economics and Business 31(1):28–47.

Wilson, C. (1980), “The nature of equilibrium in markets with adverse selection”, Bell Journal of Economics 11(1):108–130.

Wrightson, W., J. Genuardi and S. Stephens (1987), “Demographic and utilization characteristics of HMO disenrollees”, GHAA Journal 23–42.

Zeckhauser, R. (1970), “Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives”, Journal of Economic Theory 2(1):10–26.

Zeckhauser, R. (1974), “Risk spreading and distribution”, in: H.M. Hochman and G.E. Peterson, eds., Redistribution Through Public Choice (Columbia University Press, New York) 206–228.

Zweifel, P., and W.G. Manning (2000), “Moral hazard and consumer incentives in health care”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 8.

This Page Intentionally Left Blank

Chapter 12

HEALTH INSURANCE AND THE LABOR MARKET*

JONATHAN GRUBER

MIT and NBER

Contents

*I am grateful to Brigitte Madrian and Joseph Newhouse for helpful comments, and to the National Institute on Aging, the Sloan Foundation, and the National Science Foundation for financial support.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

A distinctive feature of the health insurance market in the US is the restriction of group insurance availability to the workplace. This has a number of important implications for the functioning of the labor market, through mobility from job-to-job or in and out of the labor force, wage determination, and hiring decisions. This paper reviews the large literature that has emerged in recent years to assess the impact of health insurance on the labor market. I begin with an overview of the institutional details relevant to assessing the interaction of health insurance and the labor market. I then present a theoretical overview of the effects of health insurance on mobility and wage/employment determination. I critically review the empirical literature on these topics, focusing in particular on the methodological issues that have been raised, and highlighting the unanswered questions which can be the focus of future work in this area.

Keywords

health insurance, labor market, wage determination

JEL classification: I18, J33, J68

A distinctive feature of the health insurance market in the US is the restriction of group insurance availability to the workplace, with few pooling mechanisms available for insurance purchase outside of work. As a result, ninety percent of the privately insured population currently obtains their insurance coverage through the workplace, either through their own employment or the employment of a family member [Employee Benefits Research Institute (2000)].

This restriction of health insurance purchase to the workplace setting has potentially quite important implications for the functioning of the US labor market. Counting employer and employee insurance spending, health insurance amounts to 7.1% of compensation in 1996; this share has grown by over 300% over the past 30 years.[1] This large increase in health insurance costs has been derided by some as a drag on hiring and an impediment to our international competitiveness. Others have argued that these costs have been passed onto workers wages, resulting in the lack of wage growth witnessed by the US economy in recent years.

Moreover, workplace pooling has been cited as a cause of potential labor market inefficiencies through reduced mobility. Workers are said to be "locked" into their jobs for fear of losing health insurance, and may be reticent to switch jobs, even if they have opportunities for higher productivity matches. As President Clinton said in motivating his health care reform plan of 1994: "Worker mobility is one of the most important values in an entrepreneurial society, where most jobs are created by small businesses. The present health care system is a big brake on that" [Holtz-Eakin (1994)]. In addition, individuals receiving free public insurance on public assistance programs may be reticent to leave those programs for work since they cannot be assured of finding a job with insurance. As a result, a central feature of Clinton's proposed plan was a universal employer mandate, which would have made it possible for workers to maintain insurance coverage when they switched jobs, and guaranteed health insurance for those moving into the labor force.

Despite these concerns, until the late 1980s there was little research by economists on the effects of health insurance on the labor market. This deficiency has been remedied by a flurry of research activity over the past decade. Large literatures have emerged to investigate the impact of health insurance on mobility, earnings, employment, and hours. This substantial and growing body of work has dramatically increased our knowledge of how health insurance affects the functioning of the US labor market. In addition, this literature has introduced wide variety of innovative techniques for dealing with the selection problem inherent in estimating the effect of health insurance on worker and firm behavior.

This paper critically reviews the literature on health insurance and the labor market, in four steps. First, in Section 1, I provide a brief overview of the relevant institutional details on the US health insurance market, and its interaction with the labor market. Then,

[1] National Income and Product Accounts data on health insurance component of wages and salaries. This share has declined by almost 10% from its peak in 1994 of 7.6% of compensation.

in Section 2, I present a theoretical overview of the effects of health insurance on the labor market, focusing in particular on two areas: mobility, and wage and employment determination. In Section 3, I summarize the evidence on health insurance and job-job mobility. In Section 4, I turn to three other aspects of mobility that are affected by the restriction of health insurance offering to (some) workplaces: mobility from work to retirement; mobility from public assistance programs to work; and mobility by secondary earners into and out of the labor force. In Section 5 I review evidence on the effect of health insurance costs on labor market equilibrium outcomes: wages, employment, and hours. Section 6 concludes by focusing on the priorities for future work in this area.

1. Background on health insurance and the labor market

1.1. Health insurance coverage

The distribution of health insurance coverage in the US in 1998 is presented in the final column of Table 1, from Employee Benefits Research Institute (2000) tabulations of the March 1999 Current Population Survey (CPS).[2] 170 million people, or 71% of the non-elderly population, were covered by private health insurance. Of that total, 90% were covered through employer-provided insurance, roughly one-half in their own name and one-half through others. Another 34.1 million persons, or 14% of the non-elderly, have public coverage. This public coverage is obtained primarily from three sources. The first, and most important for the non-elderly population, is the Medicaid program, the state/federal program of health insurance for low income persons; this accounts for three-quarters of the public coverage of the non-elderly. The others are the Medicare program, which predominantly covers those over age 65 but also covers the disabled below age 65, and CHAMPUS/CHAMPVA, the insurance program for the dependents of military personnel. Finally, over 18% of the non-elderly population has no insurance coverage.

This table also documents the time series trends over the past decade in sources of insurance coverage. Several trends are immediately apparent. There has been a substantial decline in the share of the population with employer-provided health insurance, from 69% in 1987 to 64% in 1996, followed by a slightly rise to 65% in 1998. This decline has been driven by falling employer-provided insurance coverage, with other private coverage rising over this period; and much of the decline of employer-provided coverage has not been declining coverage of workers, but rather of their dependents. There has also been a substantial rise in public coverage; this is completely driven by increases in the size of the Medicaid program, this rise has reversed in recent years,

[2] The subcategories of insurance do not add to the totals, since the CPS asks about insurance coverage at any point during the previous year, so that individuals may have had more than one type of coverage.

Table 1
Nonelderly Americans with selected sources of health insurance coverage, 1987–1998

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997[b]	1998
						(millions)						
Total population	214.4	216.6	218.5	220.6	222.9	225.5	228.0	229.9	231.9	234.0	236.2	238.6
Employement-based coverage	148.5	149.4	149.8	147.7	147.7	145.9	144.9	146.3	147.9	149.8	151.7	154.8
Own name	72.5	73.5	74.0	73.1	73.1	71.7	74.9	75.2	75.9	76.9	77.4	79.1
Dependent coverage	75.9	75.9	75.8	74.7	74.6	74.3	69.9	71.1	72.1	72.9	74.3	75.7
Individually purchased	14.3	13.5	14.5	14.3	13.6	14.6	16.6	16.4	16.0	16.0	15.8	15.5
Public	28.5	28.8	28.7	31.9	34.4	36.0	38.1	38.9	38.4	37.4	34.9	34.2
No health insurance	31.8	33.6	34.3	35.6	36.3	38.3	39.3	39.4	40.3	41.4	43.1	43.9
						(percentage)						
Total population	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Employement-based coverage	69.2	69.0	68.6	67.0	66.3	64.7	63.5	63.6	63.8	64.0	64.2	64.9
Own name	33.8	33.9	33.9	33.1	32.8	31.8	32.9	32.7	32.7	32.9	32.8	33.1
Dependent coverage	35.4	35.0	34.7	33.8	33.5	32.9	30.7	30.9	31.1	31.2	31.5	31.7
Individually purchased	6.7	6.3	6.6	6.5	6.1	6.5	7.3	7.1	6.9	6.8	6.7	6.5
Public	13.3	13.3	13.2	14.5	15.5	16.0	16.7	16.9	16.6	16.0	14.8	14.3
No health insurance	14.8	15.5	15.7	16.1	16.3	17.0	17.3	17.1	17.4	17.7	18.3	18.4

Notes: From EBRI (2000).

however.[3] Finally, over 18% of the non-elderly population has no insurance coverage.

1.2. Features of private health insurance policies

There are several salient features of private insurance policies which are useful for understanding the potential impact of insurance on the functioning of the labor market. Traditionally, there were two types of private insurance plans. Blue Cross/Blue Shield plans, which dominated insurance markets in the pre-war period, charged "community rated" insurance premiums, whereby employers paid only the average expenditure for a broad risk class. Beginning in the 1940s, there was a rapid growth in commercial insurance companies who "experience rated" their customers, charging firms based on their actual (projected and past) cost experience. By the late 1980s, most Blue Cross/Blue Shield plans had also moved to experience rating for all large groups, and even for some smaller groups as well. Experience rating of small firms is particularly detailed; in the extreme, if a particular worker is found to be very costly, he may be "underwritten out" of the policy, or the entire group may be rejected [Congressional Research Service (1988)]. Experience rating has been taken a step further by the growth in self-insurance of medical expenses across firms. In 1993, 19% of all firms self-insured, and 63% of firms with more than 500 employees did so [EBRI (1995)].

As a result of experience rating, there is tremendous dispersion in the cost of health insurance across firms, as documented by Cutler (1994). He finds that, for employer-provided individual insurance plans, the premium at the 90th percentile of the premium distribution is 2.5 times as large as the premium at the 10th percentile. Only a small share of this substantial variation can be explained by plan features, suggesting that most is due to experience rating.

A common feature of traditional insurance plans was unrestricted fee for service medicine: individuals could use the provider of their choice, and that provider was reimbursed based on "usual, customary, and reasonable" costs. The past twenty years, however, have seen a radical reorganization of private insurance towards the "managed care" model. Organizations such as Health Maintenance Organizations (HMOs) and Preferred Provider Organizations (PPO) have both restricted (to varying extents) patient choice or provider, and reimbursed providers on prospective fee schedules, not retrospective costs. Managed care is quickly becoming the dominant type of private insurance coverage; in 1993, 67% of persons covered by employer-sponsored health plans were enrolled in managed care [Health Insurance Association of America (1996)]. This issue is discussed in much more detail in Glied's (2000) chapter in this Handbook.

[3] Recent research highlights one channel through which these trends might be linked: the "crowdout" of private insurance purchases by public insurance eligibility. This research is reviewed in Section 4.

1.3. The role of the workplace

Why is the workplace the predominant source of private health insurance in the US? There are at least two candidate explanations. The first is workplace pooling economies. There are enormous economies of scale in insurance purchase resulting from fixed costs in administration that must be paid for any size group. Large workplace pools also provides a means for individuals to purchase insurance without the adverse selection premium that insurers demand in the individual health insurance marketplace, since the unobservable components of health will average to zero in large groups fored for purposes other than obtaining health insurance. For smaller groups, on the other hand, there is the risk that insurance purchase is driven by the needs of one or two (unobservably) very ill employees, whose costs cannot possibly be covered by the premium payments of healthier workers. As the Congressional Research Service [CRS (1988)] reports, the loading factor on insurance purchases by the very smallest groups (firms with less than 5 employees) is over 40% higher than that on very large groups (more than 10,000 employees), and the loading factor for individual insurance is even higher. Moreover, Cutler (1994) reports that the dispersion in health insurance premiums is much greater for small firms that for larger ones, which is consistent with greater adverse selection problems in the small group market.

The second is the tax deductibility of employer insurance purchases. Employer payments for insurance are not treated as taxable income to employees, unlike wages, which are taxed by both the OASDI payroll tax, and state and federal income taxes. This tax expenditure cost the government $60 billion in lost revenues in 1994 [Gruber and Poterba (1996)]. As a result, there is a large subsidy to the purchase of insurance through the workplace as opposed to through extra-workplace groups. Gruber and Poterba (1996) estimate that the relative price of insurance at the workplace is 27% lower as a result of this tax subsidy.[4]

Despite this subsidy to employer payments, only a minority of employers currently pay all of the cost of health insurance, and employee contributions for insurance have been rising as a share of total insurance payments: in medium and large firms, the share of individual plans that are wholly employer financed has fallen from 74% in 1980 to 37% in 1993; for family plans, the decline has been from 52% to 21% [EBRI (1995)]. Under Section 125 of the Internal Revenue Code, employee payments for insurance can be made tax deductible as well, but only roughly 25% of firms currently make employee premiums deductible.[5] Levy (1997) provides a detailed discussion of the two primary

[4] The correct computation of this subsidy is somewhat subtle, as it involves incorporating the share of premiums paid by employees and the fact that uninsured employees can deduct some of their medical expenses through the income tax; see Gruber and Poterba (1996) for details. There is no work which has explicitly addressed the important question of the role of the tax subsidy in promoting workplace pooling, as opposed to other workplace pooling economies.

[5] Gruber and Poterba (1996). The reasons for such limited takeup of this option are unclear. It may have to do with more extensive regulatory and reporting requirements on Section 125 plans than on traditional insurance plans; alternatively, limited takeup may simply reflect imperfect knowledge about the availability of this option.

motivations for taxable employee contributions: providing an incentive for employees to choose low cost plans within firms that offer several insurance options; and selecting out workers from the insurance pool who do not have a strong demand for insurance, allowing within-workplace sorting by insurance tastes. One source of such heterogeneity could be spousal coverage by insurance, as emphasized by Dranove and Spier (1996). Levy finds evidence to support both models of employee contributions.

Another important restriction on workplace insurance is anti-discrimination regulations, through Section 89 of the Internal Revenue Code [CRS (1988)]. These regulations make it illegal to offer insurance selectively to highly compensated employees in the firm.[6] As a result, it is impossible to selectively offer insurance to only some employees, without making it a workplace wide option.

While insurance is predominantly obtained through the workplace, there is substantial variation across workplaces in insurance offering and employee takeup. This variation is documented in Table 2, which is tabulated from the April 1993 Employee Benefits Supplement to the Current Population Survey. Each cell gives the employee weighted mean of the variable listed in the first column, for the sample denoted in the first row. Overall, 72.5% of employees work in firms that offer health insurance. Of those firms that offer health insurance, 91% offer family coverage as well as individual coverage. Only 57% of workers are covered by insurance, however, for a takeup rate of less than 80%.

The reason for non-takeup is split roughly evenly between employee ineligibility and coverage from other sources. Employee ineligibility typically arises from one of two sources. The first is pre-existing conditions exclusions, which state that the insurance plan will not cover the costs of illnesses existing before enrollment, for some period of time after enrollment.[7] The second is waiting periods (or tenure requirements) for coverage for new employees. As reported in General Accounting Office (1995), 62% of firms with more than 200 employees have a waiting period for coverage, although it is typically quite short (less than 3 months); and 60–70% of plans have a pre-existing conditions exclusion clauses, the majority of which last for 12 months or more.

[6] More specifically, non-highly compensated employees must constitute at least 50 percent of the group of employees eligible to participate in the plan; at least 90 percent of the employer's non-highly compensated employees must be eligible for a benefit that is at least 50 percent as valuable as the benefit made available to the highly compensated employee with the most valuable benefits; and the plan must not contain any provision relating to eligibility to participate that suggests discrimination in favor of highly compensated employees. Alternatively, so long as at least 80% of non-highly compensated employees benefit from the plan, it qualifies as well.

[7] As Gruber and Madrian (1994) report: "A *pre-existing condition* is generally defined as any medical problem that has been treated or diagnosed within the past six months to two years. In some cases it may be more broadly defined as any medical problem for which an individual has *ever* received care. It may also be extended to include medical conditions for which a prudent person would have sought care even if no physician was actually consulted. An insurance company may also require all employees to undergo medical examination, which it then uses to exclude certain medical conditions on an individual basis for the life of the contract. This practice is known as *medical underwriting*."

Table 2
Characteristics of employer-provided health insurance

	All employers	Fewer than 10 employees	10–24 employees	25–49 employees	50–99 employees	100–249 employees	250+ employees
Offer insurance	0.725	0.366	0.686	0.817	0.886	0.918	0.961
Family cover offered (if offered)	0.912	0.822	0.877	0.898	0.909	0.942	0.960
Covered by insurance	0.569	0.274	0.492	0.585	0.683	0.727	0.828
Takeup rate	0.785	0.749	0.717	0.716	0.771	0.792	0.862
Why no insurance?							
Ineligible	0.411	0.333	0.398	0.410	0.434	0.415	0.469
Other coverage	0.413	0.469	0.411	0.388	0.407	0.407	0.397
Firm offers insurance							
Weekly earnings	526.9	470.8	471.4	474.2	513.1	511.3	604.8
Firm offers pension	0.755	0.502	0.588	0.686	0.781	0.834	0.918
Firm offers ST disability	0.711	0.555	0.629	0.675	0.713	0.736	0.819
Firm offers LT disability	0.490	0.380	0.383	0.420	0.481	0.515	0.606
Firm doesn't offer insurance							
Weekly earnings	262.9	265.2	252.6	249.5	278.5	248.9	309.9
Firm offers pension	0.089	0.046	0.080	0.141	0.241	0.290	0.405
Firm offers ST disability	0.106	0.128	0.128	0.118	0.184	0.142	0.238
Firm offers LT disability	0.062	0.048	0.048	0.077	0.081	0.077	0.141

Note: Tabulations by author from April 1993 Current Population Survey Employee Benefits Supplement. Earnings in 1993 dollars.

These findings differ dramatically across firm size categories, however.[8] Coverage rates among the smallest (fewer than 10 employees) firms are only about 37%, while among the largest (greater than 250 employees) it is over 96%. The coverage rate grows rapidly across firm size categories; even among firms with 25–49 employees, over 80% offer insurance. Similarly, among those with insurance, the likelihood of being offered family coverage rises with firm size as well.

Interestingly, however, there is relatively little variation in takeup rates across firm size. The takeup rate is actually higher in the smallest firms that in the next two categories of firm size, and only in the very largest category of firm size is the takeup rate appreciably different than that of smaller firms. There are important differences in the reason for lack of takeup, however. Among small firms, employees are much more likely to not be taking up because of coverage from others, rather than being ineligible. But there is a steady rise in ineligibility, and a fall in other coverage, as firm size grows, so that in the largest firms ineligibility is a much more important barrier to coverage. This pattern is explained by two phenomena. First, insurers offering policies to small firms insist that eligibility be loose and takeup high, to ensure that the policy does not just provide coverage for one or two sick employees [CRS (1988)]. Second, employees at small firms, which traditionally offer less generous insurance plans, are more likely to rely on coverage from spouses than are employees at large firms with better plans.

A natural explanation for low rates of insurance offering at small firms is the much higher loading factors that they face when attempting to purchase insurance. Another explanation, offered by Long and Marquis (1992), is worker demand: they document that the types of workers who work at small firms have characteristics similar to those who work at large firms and decline coverage. This would be consistent with spousal insurance being the predominant cause of non-takeup at small firms.

Finally, it is important to highlight that insurance at even the smallest firms, and those that provide the least generous policies, is cheaper and more comprehensive than the typical individual insurance policy. Individual insurance generally costs at least 50% more than group policies. Moreover, individual policies are much less generous along a number of dimensions. Relative to group policies, non-group policies are only half as likely to have major medical coverage, coverage for physician visits, or coverage of prescription drugs; they are only two-thirds as likely to receive ambulance, mental health, and outpatient diagnostic service coverage. Furthermore, non-group policies generally feature both higher deductibles and higher copayments [Gruber and Madrian (1994)].

2. Health insurance and labor market equilibrium – theory

2.1. Employer-provided health insurance and mobility

One of the potentially most important impacts of health insurance on the labor market is its effects on mobility. Concerns over "job lock", or health insurance-induced reductions

[8] These breakdowns refer to size of establishment, not total firm size.

in worker mobility, were a driving force behind calls for comprehensive health reform, and have motivated recent partial reforms of the individual insurance market. In this section, I outline the theoretical motivation for these concerns.

The very notion that health insurance is responsible for imperfections in the functioning of the US labor market is somewhat curious. After all, health insurance is a voluntarily provided form of employee compensation. There is little discussion of the distortions to the labor market from cash wages. Why is health insurance different?

To see the difficulties introduced by health insurance in reality, it is useful to begin with a very stylized "pure compensating differentials" model [Rosen (1986)]. I construct a highly stylized example in which there is no distortionary effect of health insurance on the labor market. I then relax the very strong assumptions that are required by this example, to illustrate the source of distortions to mobility.

In this example, health insurance coverage consists of a binary, homogeneous good; individuals are either covered or not, and if covered have the exact same insurance plans. Insurance is perfectly experience rated at the worker level. That is, firms essentially purchase insurance on a worker-by-worker basis, and are charged a separate premium for each worker. Jobs that offer health insurance feature a negative compensating wage differential. Moreover, each individual job (worker-firm match) can have its own compensation structure; firms can offer insurance to some workers and not others, and can pay lower wages to those workers whose insurance costs more. Individuals have preferences over wage compensation and health insurance:

$$U_{ij} = U(W_{ij}, H_{ij}), \tag{1}$$

where W_{ij} is the wage level of worker I at firm j, and H_{ij} is a binary indicator for insurance coverage of worker I at firm j ($H_{ij} = 1$ or 0). The (pre-compensating differential) wage rate for each worker/job match is equal to the worker's marginal product at that job.

Given these preferences, individuals will desire health insurance coverage if there is a compensating wage differential ΔW_{ij} such that:

$$U(W_{ij} - \Delta W_{ij}, 1) - U(W_{ij}, 0) = V_{ij} \geqslant 0. \tag{2}$$

Suppose that there are a continuum of jobs in the economy, and that the labor market is perfectly competitive. Firms face identical worker-specific insurance price schedules; a given worker I incurs a cost of insurance $C_{ij} = C_i$ in whatever firm he works. In this world, firms will provide insurance to their workers if:

$$\Delta W_{ij} \geqslant C_i. \tag{3}$$

As a result of perfect competition, firms will bid the compensating differential down to the level C_i. Thus, all workers covered by insurance will earn exactly:

$$W_{ij} - \Delta W_{ij} = W_{ij} - C_i \tag{4}$$

on whatever job they hold.

In this simplified model, there is no real effect of health insurance on the labor market equilibrium. The introduction of health insurance simply leads to lower wages for workers who value that insurance at its cost or more. If individuals wish to change jobs, they can simply ask their new employer to provide them with insurance and lower their wage by C_i. Workers for whom $V > 0$ are earning rents from the fact that they value insurance at above its costs, but firms cannot extract those rents, since workers will be bid away by other employers who charge them the appropriate compensating differential. Most importantly, there is no inefficiency from health insurance: since workers will pay the same compensating differential C_i wherever they work, they will choose the job with the highest level of wages W_{ij}. So workers will find the best job-specific matches, regardless of their tastes for insurance.

This highly stylized model is useful for illustrating the conditions necessary to generate no mobility effects of insurance. But reality departs from this model in at least two important ways. First, employers are unable to set completely employee-specific compensation packages, offering insurance to some workers and not to others. As documented above, the Internal Revenue Code gives favorable tax treatment to employer expenditures on health insurance only if most workers are offered an equivalent benefits package. Moreover, the costs of administering such a complicated benefits system would absorb much of the rents that workers would earn from its existence. And the problems of preference revelation in this context are daunting; it is difficult in reality to see how firms could appropriately set worker-specific compensating differentials. This departure implies that there will be match-specific rents for workers attached to particular jobs.

Second, employers differ dramatically in the underlying costs of providing health insurance. As documented earlier, the loading factors on insurance purchase are substantially higher for small firms than for larger firms, and even conditional on observable factors there is huge variation in insurance premiums [Cutler (1994)]. This variation arises from both unobserved differences in the relationship between firm characteristics and insurance supply prices, and from heterogeneity in the workforce along health dimensions. This implies that workers may be unable to obtain health insurance on comparable terms across jobs.

As a result of these two features, there will be matching of particular workers and firms in labor market equilibrium: those workers who most desire health insurance coverage will work at firms offering insurance, and those firms who can provide that insurance most cheaply will offer it. In the extreme case of a perfectly competitive labor market, there will be a market-wide compensating differential ΔW. Workers will only work at firms offering insurance if their valuation of insurance is at least as great as this compensating differential, $V_{ij} > 0$. Firms will only offer insurance if the cost of insurance to that firm per worker, C_j, is less than the compensating differential, $C_j < \Delta W$. This is the compensating differentials equilibrium described by Rosen (1986). As highlighted by his discussion, in equilibrium all of the workers whose valuation of insurance

V_{ij} is greater than ΔW will be earning rents from working at a job with health insurance; similarly, all firms whose costs of insurance C_j are below ΔW will earn rents.[9]

Adding these complications introduces the possibility of job lock. Suppose that an individual now holds job 0, but would be more productive on job 1 $(W_{i1} > W_{i0})$. The cost of insurance to firm 1 is much higher, however $(C_1 > C_0)$. This high cost might arise from a high loading factor, or from the fact that the firm has a relatively unhealthy workforce and is experience rated. As a result, firm 1 does not offer insurance; even though this insurance would attract worker I, it will cost too much to provide for the rest of the workforce. And, most importantly, the insurance can't just be provided for worker I. As a result, if:

$$U(W_{i0} - \Delta W, 1) - U(W_{i1}, 0) > 0 \tag{5}$$

then the worker will not switch jobs, even though he would be more productive on the new job. This is the welfare loss from job lock: productivity improving switches are not made.

Note that, in theory, firm 0 could extract the surplus from this worker, knowing that he will not move to firm 1. Full extraction of these rents would mean that there was no net "locking" of the employee into his job at firm 0. The key question, of course, is the extent to which firms can pay discriminate on the basis of the value of insurance. In practice, full rent capture on a worker-by-worker basis seems unlikely, due to preference revelation and administrative difficulties. I review some evidence below suggesting that rent capture across relatively broad demographic groups within the workplace is possible. But, as I highlight, the level at which pay discrimination by valuation of insurance occurs is an open question; so long as it doesn't occur on a person-by-person basis, there will be job lock.

It is important to note that this type of lock arises from any employee benefit where there is differential valuation across workers, differential costs of provision across employers, and the inability to set worker-specific compensation packages (i.e. workplace safety, or location of the firm). The key insight is that in this situation, a firm cannot offer the benefit just to the marginal worker that it wishes to attract, leading to job-specific rents and job-lock. In practice, however, this effect is likely to be largest for health insurance, since both the variation in valuation across workers and the variation in costs of provision across firms are much higher than for other workplace amenities.

In theory, this problem only arises for workers considering switches from the sector providing insurance to the sector not providing insurance. But, even within the insurance providing sector, there may be job lock arising from the fact that health insurance coverage is not a homogenous good. For example, pre-existing conditions exclusions

[9] Olson (1993) provides some supportive evidence for this self-selection model. He finds that workers with greater than expected health needs (wives whose husbands do not have health insurance, and who have less healthy children) self-select into firms that provide health benefits.

may leave the worker exposed for large medical costs if he switches to a new plan. There are also probationary periods for new coverage and (in the extreme) medical underwriting and exclusion of costly new employees from insurance coverage. And job changers may lose credit towards deductibles and out-of-pocket payment limits under their old plans, raising the out of pocket costs of medical care on a new job relative to an old one. In addition, health insurance is not a discrete choice but rather a continuum of policy features. The worker's current job may offer a wider range of insurance options that is not available at other jobs which offer insurance, making job switching unattractive, in particular if the worker are restricted (through a managed care plan) from using his traditional medical providers. Finally, the fact that insurance purchased in the individual market is very expensive, less comprehensive, and potentially not even available to very unhealthy applicants, raises the costs of off-the-job search. This further mitigates against leaving a job that currently has insurance even if the next job will have insurance as well.

This last consideration highlights the fact that insurance may inhibit mobility along another dimension: in and out of the labor force. As a result of failures in the individual insurance market, those persons with high valuation of insurance, who will earn rents at insured jobs, will be reluctant to leave the workforce. This means, for example, that less healthy older workers will be unwilling to retire from firms that offer health insurance. This is a form of "lock" because even if the value of leisure is greater than the marginal product of labor for a given worker, the high cost of insurance may prohibit his leaving the job.

2.2. *Health insurance costs and labor market equilibrium*

A pervasive feature of the health care sector over the past several decades has been health care cost increases that have exceeded the rate of inflation, often by large amounts. Health care costs have tripled as a share of GNP over the past 35 years (although cost growth has slowed recently). A natural question is the implications of this dramatic cost growth for the labor market.

To understand these implications, it is useful to draw on the seminal analysis of Summers (1989). Summers' paper, as well as a number of the papers referenced in this section, addressed the question of the effects of a government mandate that employers provide health insurance to their workers, but this can naturally be extended to consider the implications of rising employer insurance costs.[10] Summers' analysis is depicted in Figure 1, which shows supply and demand in the labor market, with an initial equilibrium at (L_0, W_0); for the moment assume that labor supply consists simply of a (1, 0) participation decision. An increase in the costs of providing insurance will raise labor costs, shifting the demand curve inwards, and leading to lower wages (W_1) and employment (L_1).

[10] There are a number of subtleties involved in comparing these cases, a point to which I return below.

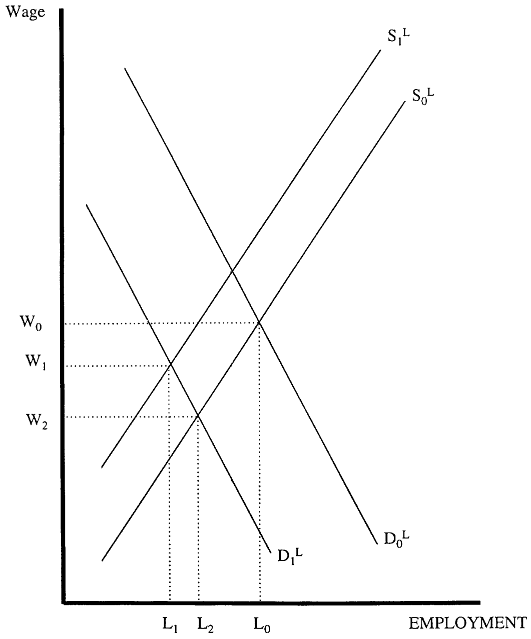


Figure 1.

Summers' key insight, however, was that workers may also value health insurance more now that health care costs have increased, since the costs of being uninsured have risen.[11] As a result, they will increase their desired labor supply in order to obtain employer-provided coverage. This outward shift in supply lowers wages further (to W_2), but mitigates the loss of jobs. In fact, if workers value the increased insurance at its cost, this increase will be fully shifted to wages, with no effect on total employment. In principle, rising health insurance costs could increase employment: if individuals are risk averse, then increasing the size of the risk of being uninsured will raise the desire for insurance.[12]

Gruber and Krueger (1991) provide a formalization of this graphical analysis. Suppose that labor demand (L_d) is given by:

$$L_\mathrm{d} = f_\mathrm{d}(W + C), \tag{6}$$

[11] This effect will be augmented by income effects, since families are now poorer, increasing desired labor supply. For an analysis which incorporates this point, see Feldman (1993).

[12] Obviously this effect depends on how the coefficient of risk aversion changes with income.

where W is wages and C is insurance costs. Further suppose that labor supply is given by:

$$L_s = f_s(W + \alpha C), \tag{7}$$

where αC is the monetary value that employees place on health insurance. For determining the effect of rising costs on the labor market, the relevant concept is marginal α, the valuation of the marginal dollar of health insurance spending. A key determinant of marginal α will be the *source* of the insurance cost increase. If insurance costs are increasing because of an underlying rise in the cost of valuable health care services, marginal α is likely to be high. However, if costs are rising because of increases in the cost of administering insurance, then marginal α will be close to zero, since the value of insurance has not risen relative to the alternative (self-insurance). For the purposes of this discussion, I assume that average and marginal α are equal; that is, that increases in the cost of insurance are valued in the same way as is the existing level of insurance spending.

Using this notation, it can be shown that:

$$\frac{\delta W}{\delta C} = \frac{-\eta^{\mathrm{d}} - \alpha\eta^{\mathrm{s}}}{\eta^{\mathrm{d}} - \eta^{\mathrm{s}}}, \tag{8}$$

where η^{d} and η^{s} are the elasticities of demand and supply for labor, respectively. This equation differs from the standard expression for the incidence of a tax on labor by the term $\alpha\eta^{\mathrm{s}}$ in the numerator, which captures the increase in labor supply due to employee valuation of more expensive insurance. This leads to a change in employment of:

$$\frac{\delta L}{L} = \frac{-\Delta C - \Delta W}{W^0} * \eta^{\mathrm{d}}, \tag{9}$$

where ΔW is the change in wages and W_0 is the initial wage level.[13]

It is clear from Equation (7) that the reduction in wages will be less than the increase in costs if $\alpha < 1$. That is, if employees value the increased insurance at less than its cost to the employer, the costs cannot be fully shifted to wages, leading to a fall in employment. However, if employees value this increase in the health insurance at its full employer cost ($\alpha = 1$), wages will fall by exactly the amount that costs rise, with no effect on employment; in principle, if $\alpha > 1$, employment could even rise. Thus, the

[13] Subsequent models following this formulation have considered in more detail particular aspects of the incidence of increased employer costs. Gruber and Hanratty (1995) develop a model of payroll-tax financed national health insurance, and Anderson and Meyer (1995) illustrate the impact of payroll-tax financed unemployment insurance, for the case of differential employer experience rating (which is clearly appropriate to health insurance markets as well).

implications of this basic model are that rising health care costs should lead to lower wages with an ambiguous effect on employment.

This analysis is obviously simplified along at least eight dimensions. First, labor supply is not simply a discrete choice, but rather a combination of participation and hours of work decisions.[14] Increases in costs will have effects on both the supply of and the demand for work hours conditional on participation.[15] From the employer perspective, increases in health insurance costs are an increase in the fixed cost of employment and are as a result more costly (as a fraction of labor payments) for low-hours employees. If employers are able to lower each worker's wages by the lump-sum increase in costs, then neither hours nor employment should change. However, if (as seems likely) employers are not able to implement a percentage reduction in pay that is inversely proportional to hours worked, then covered low hours workers will become more expensive. Employers will therefore desire increased hours by fewer workers, lowering the cost per hour of the health insurance for a given total labor supply.

Of course, if the wage offset is lower for low-hours workers, workers will demand the opposite outcome: there will be increasing demand for part-time work, with hours falling and employment increasing. Moreover, since part-time workers may be more readily excluded from health insurance coverage, there may also be a countervailing effect on the employer side, as full-time employees are replaced with their (uninsured) part-time counterparts. In this case as well, hours would fall and employment would rise. Thus, the effect on hours of work is uncertain.[16]

Second, employers may react along another dimension: dropping health insurance coverage altogether. Increases in the cost of insurance will reduce the desire of employers to offer insurance, lowering the number of jobs offering insurance and raising the compensating differential. At the same time, as argued above, increases in costs may raise the demand for jobs that offer insurance, raising further the compensating differential and counteracting the decline in the number of jobs with insurance. As earlier, in principle increases in the cost of health care could actually raise the total demand for health insurance. So the net effect on employer insurance offering is ambiguous.

Third, this analysis has ignored existing constraints on compensation design in the labor market. For example, for workers already at the minimum wage, firms will be unable to shift to wages increase in the cost of health insurance.[17] Similarly, union

[14] In fact, as Feldstein (1995) emphasizes, appropriately defined labor supply also includes other features such as choice of job and work effort. It is difficult to assess the impact, either theoretically or empirically, of rising health care costs on these dimensions.

[15] This discussion follows Gruber (1994a).

[16] This is obviously a simplified discussion of the complicated process by which hours is determined, but it captures the basic intuition. For models of health insurance and hours, see Cutler and Madrian (1998) or Hashimoto and Zhao (1996).

[17] As Gruber (1994c) discusses, however, this may not be a very important consideration empirically, since recent research suggests little employment effects in changes in the minimum wage [Card (1992a, 1992b), Katz and Krueger (1992), Card and Krueger (1994)]. This research is consistent either with a monopsony model of the low wage labor market, or with very inelastic demand for low wage labor; in either case, there will be little disemployment effect from increases in health care costs for minimum wage workers.

contract or other workplace pay norms may interfere with the adjustment of wages to reflect higher costs. These institutional features could increase the disemployment effects of rising health costs.

Fourth, this analysis ignores heterogeneity across workers. Increases in the costs of health insurance may not be uniform throughout the workplace; for example, costs may rise more for family insurance than for individual coverage, or they may rise more for older workers than for younger workers. In the limit, with extensive experience rating, costs may rise for particular workers; for example, a worker may be diagnosed with cancer, substantially increasing firm average insurance expenses. Gruber (1992) extends the model of Gruber and Krueger (1991) to the case of two groups of workers, where costs increase for one and not the other. If there is group-specific shifting, then the solution collapses to the one group model. If not, however, the substitutability of these groups will also determine the resulting labor market equilibrium; in general, there will be effects on both the group for which costs increase and the group for which they do not.

In practice, there may be a number of barriers to group, and in particular individual-specific shifting. Most obviously, there are anti-discrimination regulations which prohibit differential pay for the same job across particular demographic groups, or which prevent differential promotion decisions by demographic characteristic.[18] Workplace norms which prohibit different pay across groups or union rules about equality of pay may have similar effects. Thus, a central question for incidence analysis is *how finely* firms can shift increased costs to workers' wages. If there is imperfect group or worker-specific shifting, there may be pressure on employers to discriminate against costly workers in their hiring decisions.

Fifth, this model assumes that the only dimension of compensation offset is wages. In fact, employers may offset rising health insurance costs along other dimensions, such as reducing the generosity of other benefits. Indeed, for fixed cost (per-worker) benefits such as vacation time or other workplace amenities, there will be a natural substitution that will not involve distorting the employment/hours margin.

Sixth, this discussion ignores taxes. As noted earlier, health insurance payments by employers are not treated as taxable income to the employee, while wages are. This means that a dollar of health insurance is worth more than a dollar of wages, increasing the extent to which individuals may be willing to forgo wages as health insurance costs rise.

Seventh, there may be general equilibrium consequences of rising health insurance costs. These considerations will arise from shifts in demand across firms where health care costs rise at different rates, and from substitution between labor and capital. General equilibrium analyses of health care costs include Danzon (1989), Sheiner (1995b), and Ballard and Goodeeris (1993).

[18] See Ehrenberg and Smith (1991) for a discussion of US anti-discrimination legislation.

Finally, one issue that is ignored even by these general equilibrium analyses is changes in mobility. The net effect of increasing health care costs on mobility is ambiguous, and depends on the adjustment of the compensating wage differential and the rate of firm insurance offering. In addition, there may be effects on the insurance market itself which inhibit mobility: firms may find it more advantageous to pay fixed screening costs when insurance becomes more expensive. As emphasized by Triplett (1983), these mobility effects have a feedback implication for wage setting. Firms desire to minimize total costs, including the costs induced by high turnover. If increases in the cost/value of insurance lowers turnover, then firms may be willing to continue to offer insurance even if they are not able to lower wages by a comparable amount. This means that the measured cash wage offset may be lower than dollar for dollar, even if firms are seeing no net rise in labor costs.

2.3. Health insurance mandates

Rising uninsurance in the US is a continuing source of policy concern. One frequently discussed approach to addressing this problem is mandating that employers provide health insurance to their workers. Over one-half of the uninsured are in families where the head is a full-time, full-year worker, and another quarter of the uninsured are in families where there is at least part-year and/or part-time attachment to a job [EBRI (2000)]. Thus, a broad mandate to workers would potentially go a long way towards eradicating the problem of uninsurance. Moreover, in this era of tight fiscal budget constraints, an "off-budget" approach such as a mandate is politically appealing. It is perhaps for this reason that an employer mandate was the centerpiece of the failed Clinton health care reform effort of 1994.

At the most basic level, the effects of an insurance mandate can be modelled using the same framework described above for rising health care costs; indeed, this was the original application of Summers' (1989) analysis. But in reality analyzing a mandate introduces several important complications. First, the value of α may be low when mandating the provision of insurance to firms that have chosen not to provide that insurance. As discussed earlier, these firms may face high costs of insurance and/or low worker demand (willingness to pay compensating differentials). Alternatively, however, Summers argues that α may still be close to one in these firms, since they may not be offering insurance due to adverse selection. Moreover, some part of these high loading factors in the uninsured sector are due to adverse selection, through fixed costs of screening potential enrollees. An employer mandate will substantially reduce the potential for adverse selection by making coverage close to universal; as a result, loading factors might fall, raising α.

Second, Summers' analysis applies to the case of a mandate to *workers only*. In reality, most mandate plans would make some effort to cover non-workers as well. For example, the Clinton reform plan would have offered substantial income-related subsidies for insurance purchase by non-workers. As the generosity of insurance for non-workers rises, it lowers the benefit linkage that causes small efficiency costs of mandates; if there is no need to go to work to obtain insurance, then labor supply will not rise. In the limit, with comparable coverage for workers and non-workers, there will be no linkage, and the mandate will simply operate as a standard tax on employers, with the resultant efficiency cost.

Third, a realistic mandate will have some mechanism for redistributing towards low wage workers for whom the increased compensating differential is particularly burdensome. The structure of the Clinton mandate provides a benchmark for understanding how these subsidies might be structured. There was a cap on employers' health insurance contribution as a share of payroll; firms with average payroll below a given threshold received a subsidy, and firms with average payroll above the threshold purchased their own unsubsidized insurance. As Sheiner (1995b) highlights, this amounts to a tax on firm payroll below the subsidy level, since increases in payroll reduce the subsidy amount. As a result, this type of subsidy structure gives firms the incentive to split into high and low wage components, in order to maximize the subsidies for the low wage component and simply purchase unsubsidized insurance for the high wage component. In practice, such splits may be difficult, leading to incentives for sectoral shifts by workers in order to maximize homogeneity within firms.

Finally, Browning (1994) emphasizes that the efficiency cost of a mandate must be determined with reference to other pre-existing distortions in the labor market. The marginal deadweight loss of a distortionary intervention rises with the distance from the competitive equilibrium. Thus, if there is not full shifting to wages, mandates will have larger efficiency costs if they are imposed in a market which is already relatively far from competitive equilibrium due to labor market regulations and taxation.

3. Evidence on health insurance and job-job mobility

Mobility from job to job and in and out of the labor force is a fundamental feature of the US labor market. Over 20 million Americans change jobs each year. Nearly 12 million of those leave jobs with health insurance, and this group has 7 million dependents. And there are potentially millions more who do not leave jobs with health insurance because of fear of losing that coverage, or facing limitations on coverage at their new jobs.[19]

The key question which has been addressed by a small, but growing, literature is: what is the effect of health insurance on mobility decisions? There is considerable anecdotal evidence that "job lock" is an important phenomenon. Surveys have found that between 11 and 30 percent of individuals report that they or a family member have remained in a job at some time because they did not want to lose health insurance coverage [Government Accounting Office (1995)]. Twenty percent of those who reported job lock in their households cited preexisting conditions as the main reason for not changing jobs. The purpose of the empirical studies in this area is to assess whether these survey responses have real content for mobility decisions in the US.

[19] Facts from GAO (1995).

3.1. Health insurance and job mobility: empirical considerations

In theory, testing for the effects of health insurance on mobility is straightforward: one can simply assess whether individuals are less likely to leave jobs that offer insurance. If so, there is prima facie evidence that "job lock" exists.

This was the approach taken by the early literature on benefits and mobility. This literature was primarily focused on the effects of pensions on mobility, but one article, Mitchell (1982), employed this approach to look at the effect of health insurance benefits.[20] She found that having health insurance on the job resulted in a substantial 22% reduction in the odds of quitting that job for men, but the estimate was not significant; there was no effect for women. This finding highlights an important consideration throughout the literature on mobility: power. A number of studies find sizeable mobility effects that are not significant. Some authors refer to these findings as evidence of no effect, but this is not correct; in fact, estimates such as Mitchell's cannot rule out huge effects. Without sufficient precision, it is difficult to draw useful conclusions about health insurance and mobility.

More recently, Cooper and Monheit (1993) augmented this approach to consider not only whether the worker held insurance on their current job, but the likelihood of finding insurance on the new job. They find very large (and significant) effects on mobility, with health insurance reducing the odds of job leaving by 23–39% across the different demographic groups that they study.

The problem with this approach is that of *selection*, both on the worker and firm side. We have already seen that, on average, the least healthy workers should choose to work at firms that offer health insurance. But underlying health may be correlated with mobility. An obvious dimension along with such a correlation exists is age: older workers are less healthy, and are less likely to change jobs. But this correlation may exist along dimensions unobserved to the econometrician as well. As a result, a finding that workers at firms that offer insurance are less likely to change jobs may simply reflect the fact that these are the least healthy, and therefore least mobile (for other reasons), workers.

Moreover, the firms that offer health insurance are not directly comparable to firms that do not. This point is illustrated in the lower two panels of Table 2, which show four characteristics of workers in firms that do and do not offer insurance: average weekly earnings; likelihood of firm-offered pension; likelihood of firm-offered short-term disability coverage; and likelihood of firm-offered long term disability coverage. The differences across these two types of firms is dramatic. Workers in firms that offer

[20] Results for pensions are not necessarily informative in this context, since the explicit backloaded nature of defined benefit pension plans should increase the mobility-inhibiting feature of this benefit. For work on pensions (or total fringe payments) and mobility, see Bartel and Borjas (1977), Bartel (1982), Mitchell (1982, 1983), and Gustman and Steinmeier (1987). Early work on this topic suggests that pensions significantly reduce quit rates, but Gustman and Steinmeier argue that this result is driven by higher compensation levels at firms that offer pension plans.

insurance have earnings that are over twice as high, and they are roughly *eight times* more likely to be covered by other benefits.

These differentials do not arise simply because of differences in the size of firms that offer and do not offer insurance. The remainder of these panels show wages and benefits offering within firms that do and do not offer insurance, divided by firm size category. Within every size category, wages are much higher and benefits much more generous at firms that offer health insurance relative to those that do not. These findings are consistent with the large labor economics literature on inter-industry wage differentials, which documents persistent pay differences between "good" and "bad" jobs over space and time; this taxonomy could apply equally well to the rate of health insurance offering.[21]

As a result of these differentials, it is difficult to disentangle the effect of insurance per se on the mobility decision. If individuals are reticent to leave these "good" (high wage/generous benefit package) jobs for reasons other than health insurance, then this would be perceived as "job lock". What is needed to disentangle the effect of insurance per se is some way to control for the confounding influence of these other job characteristics.[22]

3.2. Solution: variation in the value of health insurance

While this selection problem was perhaps recognized, it was not seriously addressed by empirical economists until the early 1990s. At that point, a series of articles proposed to address this problem by application of "differences-in-differences" (DD) methodology. The idea of this approach is to find two groups for whom job-lock should operate differentially strongly, for example because the former group has much higher expected medical expenditures than the latter, but for whom the other characteristics of the "good jobs" that offer insurance should be valued equally. Then, one can contrast the effect of employer provided insurance on these two groups. If job lock is important, the reduction in mobility from employer-provided insurance should be much stronger for the group with high insurance valuation than for the comparison group.

This approach is illustrated nicely by Madrian (1994a). Consider the following matrix of mobility rates:

Value of health insurance	Employer-provided health insurance	
	No	Yes
High	M_{00}	M_{01}
Low	M_{10}	M_{11}

[21] See, for example, Katz and Dickens (1987).

[22] Cooper and Monheit do attempt to control for selection on both the worker and firm side by including a number of controls for worker characteristics (including health status) and other firm benefits. The fact that estimates from Madrian (1994a, 1994b), using the same data set but a plausibly more convincing identification strategy, are similar suggests that the approach used by Cooper and Monheit may be sufficient to control for selection.

One approach to measuring job lock would be to compute $(M_{01} - M_{00})$, which would measure the effect of having employer-provided insurance on mobility rates for a group that should be job-locked. But this approach runs into the criticism levied above that the jobs that offer insurance may also offer other amenities which make job leaving unattractive. The advantage of differences-in-differences analysis is that this criticism can be addressed by using those with low value of insurance as a control group. That is, the difference $(M_{11} - M_{10})$ should not reflect job lock, but should reflect the other amenities of jobs that offer health insurance. Thus, by computing the difference of these differences, $(M_{01} - M_{00}) - (M_{11} - M_{10})$, one can measure the pure effect of job lock net of any other amenities. That is, by using the low valuation group as a control, one can hold constant the value of other job attributes, and identify separately the value of insurance.

More precisely, this approach suggests a regression specification of the following form:

$$\text{MOVE} = f\bigl(\alpha + \beta_1 \text{HI} + \beta_2 \text{VALUE} + \beta_3 \text{HI} * \text{VALUE} + X'\delta + \varepsilon\bigr), \tag{10}$$

where MOVE is a dummy variable for job switching, HI is a dummy for having health insurance on one's job, VALUE is an index of the value of that insurance, and X is a set of person and/or job-specific covariates. In this formulation, β_1 captures the other aspects of jobs that affect mobility, and β_2 controls for secular differences in mobility rates between workers who do and do not value insurance. The interaction β_3 measures the differential value of having health insurance for those who value that insurance, relative to those who do not, which proxies for job lock.

Several studies using this approach are reviewed in Table 3. Madrian (1994a) employs three proxies for VALUE using data from the National Medical Care Expenditure Survey (NMES). This survey followed individuals for four quarters in 1987, and collected information on job transitions; it is also the best source of data (since its 1977 counterpart) on health insurance expenditures and health status. Her first proxy is an indicator for whether the spouse does not have insurance coverage: if spouses are insured, it lowers the value of own insurance, mitigating job lock. The other proxies are more direct indicators for potential medical expenditures (and thus the value of having insurance coverage): family size and pregnancy of the spouse. All three DD estimates yield significant and sizeable estimates, suggesting mobility reductions on the order of 30–67%.

Holtz-Eakin (1994) pursues the spousal insurance and health status interactions in the Panel Study of Income Dynamics (PSID) for 1984; this longitudinal survey also collects information on job transitions and health insurance coverage. He finds some evidence of job lock from the spousal insurance interaction over one year, but the effect is small (8.6% mobility reduction from no spousal insurance) and insignificant; and it is wrong-signed over a three year period. He also finds little effect from his health status interactions, although he only reports t-statistics and not coefficients so that it is hard to assess the magnitude of the results.

Table 3
Research on health insurance and mobility

Paper (date)	Data (years) and sample	Empirical strategy	Results
Mitchell (1982)	Quality of employment survey (1973, 1977); 18–65 wage earners	Model of quits as a function of fringe benefits, including health insurance	Insignificant mobility reduction: (a) men: 22% (b) women: 0
Cooper and Monheit (1993)	NMES (1987); 25–54 wage earners	Model of job change as a function of predicted gain or loss of health insurance from change; controls for health status, some other benefits	Signif. mobility reduction: married men: 24.8% single men: 23% married women: 34.7% single women: 38.8%
Madrian (1994a)	NMES (1987); 20–55 married males, non self-employed	DD model of voluntary job leaving – value proxies: (a) spouses HI (b) family size (c) pregnancy	Signif. mobility reduction: (a) 31% (b) 37% (moving 1–5 kids) (c) 67%
Holtz-Eakin (1994)	PSID (1984–1986); 25–55 workers	DD model of job change – value proxies: (a) spousal insurance (b) health status	Insignificant (a) 8.6% 1 yr, neg. 3 yr (b) mixed, insignif.
Holtz-Eakin, Penrod, and Rosen (1996)	SIPP (1984–86), PSID (1984); 16–62 workers	DD model of transition to self-employment – value proxies: (a) spousal Insurance (b) health costs (c) continuation laws	SIPP: insignif., large (a) 15.3% (b) 10/12 coeffs right signed, 9.2% largest (c) 14% PSID: insignif.
Gruber and Madrian (1994)	SIPP (1983–89); 20–54 non self-employed males	Model of job leaving as a function of availability of continuation of coverage laws	1 Year of Coverage = signif. 10% mobility increase
Buchmueller and Valletta (1996)	SIPP (1984–86); 25–55 non-self-employed, non-construction	DD model of job leaving, using spousal insurance, with pension/tenure controls and accounting for endogeneity of dual job change	Men: 25–32%, insignif. women: 34–49%, signif.

continued on next page

Table 3, *continued*

Paper (date)	Data (years) and sample	Empirical strategy	Results
Anderson (1997)	NLSY (1979+); 20–40 males	Hazard DD model of job leaving, using pregnancy, family size, and self-reported health limitation	Signif. mobility reductions: pregnancy: 34% family size: 37% (2 kids) health limit: 0% (insignif.)
Gruber and Madrian (1997)	SIPP (1984–88); 25–54 males	Model effect of continuation mandates on health insurance coverage of non-employed, on transitions to non-employment, on total weeks of non-employment, and on re-employment earnings	One year of continuation cov: (a) increases ins cov by 6.7% (19% for those non-empl. for > 1 year) (b) increases transitions to non-employment by 14% (c) increases weeks non-empl. by 15% (d) increases reemp earnings by 100%
Kapur (1998)	NMES (1987); 20–55 married males, non self-employed	DD model using only those with no spousal insurance, with value proxied by medical conditions	Small and uniformly insignificant effects; often wrong-signed
Madrian and Lefgren (1998)	SIPP (1984–93); 20–64 persons	DD model of transition to self-employment – value proxies: (a) spousal insurance (b) family size (c) continuation laws	Significant and large mobility reductions: (a) 25% (b) 18% (2 kids) (c) 10% (1 year)

Notes: NMES = National Medical Expenditure Survey; PSID = Panel Study of Income Dynamics; SIPP = Survey of Income and Program Participation; NLSY = National Longitudinal Survey of Youth.

Anderson (1997) estimates the effect of job lock in the National Longitudinal Survey of Youth (NLSY), which follows a sample of 14–21 year olds in 1979 over the subsequent years. This provides the longest panel of longitudinal data that has been used to address this issue, allowing Anderson to estimate sophisticated hazard models of mobility. Following Madrian's pregnancy identification strategy, she estimates significant job lock: job lock among men with a pregnant spouse lowers mobility by about 34%. She also draws a potentially important distinction between "job lock" and "job push", where the latter is defined as individuals who leave jobs without health insurance because of a desire for coverage; she finds that roughly half of the total effect estimated in her paper is actually "job push".

One potential problem with the use of the spousal insurance proxy is that spousal insurance is not exogenously assigned. It is therefore plausible that the effects of health insurance on mobility may differ across workers with and without spousal insurance for other reasons.[23] This point is addressed in more detail by Buchmueller and Valletta (1996). They also use spousal insurance as a proxy for VALUE in the Survey of Income and Program Participation (SIPP), which interviews individuals every four months for up to three years, collecting information on job transitions, insurance coverage, and health care utilization. They control for whether the job offers a pension, as well as job tenure, both important correlates of mobility. They also account for the potential endogeneity of spousal insurance through modeling the joint mobility decisions of husbands and wives. Their results are very similar to Madrian's in magnitude, with mobility reductions of 25–32% for dual earning men from no spousal insurance, and of up to 49% for dual earning women; they find little effect of accounting for potential endogeneity. This suggests that the omission of pensions and tenure in Madrian's estimation did not lead to significant bias, which is consistent with the assumption that the spousal insurance interaction proxies for job lock and not other job features. On the other hand, their estimates for men are not significant.[24]

This potential criticism is also levied by Kapur (1998). She addresses this question more directly in the NMES data by examining job lock only among those with insured spouses, using as the measure of value three different indicators of medical demand (such as the presence of chronic illness). She finds that in this sample there is little evidence of job lock, and her estimates are fairly precise. She also argues that Madrian's

[23] For example, among the class of workers with health insurance, husbands with working wives may hold jobs with worse amenities along other dimensions, such as pay, which is why the wife is working. If this is not true for workers without health insurance, the DD estimate would understate the effect of health insurance, since the control group will be more "locked" into their job by the other amenities than will the treatments. Madrian (1994a, 1994b) addresses this point to some extent by conditioning on the wife's labor force status, but she does not control in detail for other amenities of either the husband's or wife's job. This point is made in Slade (1997) as well.

[24] Buchmueller and Valletta also estimate job lock for sole earning men, single men, and single women, estimating mobility reductions on the order of 17% to 45%. But these models are identified only by the effect of insurance, and not by an interaction with VALUE, raising the identification problems noted earlier (particularly given the lack of controls for worker health status in these models).

estimates using pregnancy and family size are biased, and she demonstrates that fixing these biases results in small and insignificant findings.

The other approach that has been taken in this literature is to rely on variation in the availability of government mandated continuation coverage. Over the past twenty years, states and the federal government have passed continuation of coverage laws which mandate that employers sponsoring group health insurance plans offer terminating employees and their families the right to continue their health insurance coverage through the employer's plan for a specified period of time. Although individuals must pay the full average cost of their group insurance, the price may be well below that of a policy purchased in the individual market, especially for individuals with high medical expenditures.[25] Moreover, as documented above, group insurance is typically much more generous than policies purchased in the non-group market along a number of dimensions, including the fact that pre-existing conditions exclusions and underwriting are much more severe in the individual market. Thus, having continuation of coverage benefits available provides a potentially valuable temporary source of portability for the worker who leaves his job. Indeed, Gruber and Madrian (1994, 1996) estimate takeup rates of continuation of coverage benefits to be roughly two-thirds among younger job leavers and retirees.

Continuation of coverage laws generally apply to all separations (except those due to an employee's gross misconduct), although in some states benefits are restricted to those who leave their jobs involuntarily. They often also provide benefits to divorced or widowed spouses and their families. The first such law was implemented in Minnesota in 1974. More than 20 states passed similar laws over the next decade before the federal government, as part of its 1985 Consolidated Omnibus Reconciliation Act (COBRA), mandated such coverage at the national level. The state laws generally provided continuation coverage for 6–12 months, while the federal statute mandated such coverage for 18 months.

The availability of continuation coverage should mitigate job lock, by providing a temporary bridge to those who will be unemployed during their search, who will move to a job without coverage, or who will be at temporarily uncovered on their new job.[26] This suggests that a natural test for job lock is an assessment of whether easing job lock through continuation coverage affects mobility. The advantage of this approach relative to the DD tests denoted above is that the variation across states and over time in the availability of continuation coverage provides clearly exogenous variation in the extent of job lock. The disadvantage is that this is only limited portability, as opposed to the more permanent portability represented by (for example) spousal insurance coverage.

[25] Gruber and Madrian (1994) estimate that the price of a family policy purchased in the non-group insurance market for a family policy for a 40 year old man with a wife and two children was 40% higher than the price of continuing group coverage; Gruber and Madrian (1996) estimate this differential to be 70% for a married couple with a 58 year old head.

[26] In principle, before 1990 COBRA coverage could not be continued if individuals found a new job that offered insurance, even if they were not yet covered by that insurance. In practice, it is difficult to know how well this provision was enforced.

Gruber and Madrian (1994) model transition rates out of jobs as a function of the months of continuation coverage that are available, in the SIPP data for 1983–1989. They find a significant effect of continuation coverage on mobility rates: one year of such coverage raises mobility rates by 12–15%. Given that this is relatively limited portability, this is a sizeable effect, which is consistent with an important role for job lock. Gruber and Madrian (1997) follow on this analysis by considering specifically the impact of continuation of coverage mandates on transitions out of employment, as opposed to job-job movements. They find that (a) almost all of the effect of continuation mandates is on movements out of employment; (b) there appears to be relatively little effect on non-employment durations, conditional on separation; and (c) continuation mandates are important in maintaining the insurance coverage of job leavers, particularly those who are subsequently non-employed for a year or more, where continuation coverage raises the odds of insurance coverage by 19%.

To summarize, the weight of the evidence on job lock suggests that it is a significant phenomenon, with employer-provided insurance reducing mobility by roughly 25–30%. But there remains considerable disagreement. This disagreement revolves around two issues. The first is the validity of spousal insurance as a proxy for value. Virtually all studies that have used spousal insurance as a value proxy have found significant job lock, and the estimates in Madrian (1994a) and Buchmueller and Valetta (1996) that attempt to control for omitted variables correlated with spousal insurance still yield large effects. But Kapur's (1998) criticism that the population with spousal insurance is simply not comparable to the population with such insurance has merit, and estimates that use other value proxies are somewhat more mixed. The second is power considerations, as highlighted by the fact that even the relatively large estimates for married men in Buchmueller and Valetta (1996) are not significant. Definitive resolution of this debate will require further investigation with larger samples in longitudinal databases like SIPP and NLSY, using the variety of identification strategies suggested in this literature.

3.3. Self-employment decisions

A related but distinct question to that of job-job mobility effects is the effect of health insurance on decisions to move into self-employment. The first study to examine this was Holtz-Eakin, Penrod, and Rosen (1996) examined the transition to self-employment. Their estimates of this "employment lock" from the SIPP are quite large, ranging from 9.2% to 15.3%; but they are generally insignificant. Their estimates from the PSID are smaller and also insignificant, but the confidence intervals are once again very large.

A more recent study by Madrian and Lefgren (1998) revisits this issue using a larger number of years of SIPP data. They find somewhat larger effects that are statistically significant, due to the resulting increase in precision. For example, using the presence of spousal health insurance as a proxy for value, they estimate that job lock lowers transition rates to self-employment by 25%; they also find significant effects using family size and continuation of coverage mandates as value proxies.

3.4. Welfare implications

While there is some uncertainty about the empirical importance of job lock, it pales in comparison to the uncertainty about its normative implications. On the one hand, reduced mobility should have negative implications for economic efficiency because workers are not moving to jobs where they are most productive for fear of losing health insurance. On the other hand, reduced mobility has the benefit that it allows firms to reap the benefits of firm-specific human capital investments. That is, by locking workers into their jobs, health insurance may induce firms to invest more in their firm-specific capital. As Madrian (1994a) notes, however, job lock is a particularly inefficient means of reducing employee turnover, since it is the least healthy employees who will stay in their jobs; it would be more efficient to use other mechanisms such as age-wage profiles or pension benefits to achieve this goal.

To the extent that there is inefficiency in job matching, the next question is the empirical magnitude of the efficiency loss. A number of studies document large wage gains from job-job mobility, suggesting important efficiency costs to mobility restrictions through job lock. Bartel and Borjas (1977) estimate that individuals who report leaving their job because they found a better one (presumably the relevant population for computing the benefits of easing job lock) had wage gains that year that were 6% higher per year than those that stayed on their job. On the other hand, they found insignificant effects on future wage growth, suggesting that these gains were short lived. Bartel (1982) also finds wage gains of 3% for young male quitters, but not for mature men. And Topel and Ward (1992) find that job turnover among younger workers is critical to the process by which they settle into lifelong careers. They estimate that there are very large wage increases associated with job changing: quarterly wages rise by 11% for workers who change jobs, as compared to wage gains of roughly 1% for those who remain in their jobs. As a result more than one-third of early career wage growth is associated with job changing.

On the other hand, the literature on wages and mobility does find that the beneficial effects of mobility decline with age. Indeed, Topel and Ward find that the wage change with job changing is only one-third as large at 7.5–10 years of experience as at 0–2.5 years. Thus, the older workers for whom job lock may be most important are the ones for whom the costs of mobility restrictions may be lowest. This suggests that using average wage gains or productivity improvements from better job matching may overstate the benefits of reducing job lock.

Clearly, what is needed here is an empirical investigation of not whether job lock exists, but its implications for productivity. A suggestive piece of evidence on this front is provided by Gruber and Madrian (1997). They model the reemployment earnings of job leavers as a function of whether continuation coverage is available in the worker's state/year; does loosening job lock through providing continuation coverage improve subsequent job matches? They find that one year of continuation coverage availability *doubles* the reemployment earnings of job leavers who take up that coverage. This very large finding suggests that job lock does have very important efficiency consequences;

but the almost implausible magnitude also suggests the value of further investigation of this question.

4. Health insurance and participation in the labor force and public assistance programs

4.1. Health insurance and retirement

As highlighted in Section 2, the existence of rents attached to jobs with health insurance implies that workers will be reluctant to move from these jobs out of the labor force. In particular, this effect might be strongest around the retirement decision, since older workers are the group which are earning the largest rents from within-workplace pooling of insurance purchase.[27] For retirement at age 65 or greater, individuals will have their basic medical needs covered by the Medicare program, so that there should be little net effect of on the job insurance on work decisions.[28] But for individuals contemplating early retirement, the presence of insurance on the job and the lack of insurance off the job may be an important deterrent to job leaving. This is because of the high and variable medical cost exposure for older individuals, as documented in Table 4, which shows a variety of indicators of health status by age.[29]

There is a clear deterioration in health and increase in medical utilization/spending after age 55. Compared to those age 35–44, for example, those age 55–64 are: twice as likely to report themselves in fair health and four times as likely to report themselves in poor health; four times as likely to have had a stroke or have cancer, seven times as likely to have had a heart attack, and five times as likely to have heart disease; twice as likely to be admitted to a hospital (and spending twice as many nights in the hospital if admitted), and 40% more likely to have a prescribed medicine (and having twice as many medicines if they have a prescription). As a result, the medical spending of 55–64 year olds is almost twice as large, and twice as variable, as that of 35–44 year olds.

Despite their higher medical costs, the extent of insurance coverage among 55–64 year olds is similar to that of 25–54 year olds. Overall, 12 percent of 55–64 year olds are

[27] Unless, of course, employers are able to shift the higher costs of experience rated insurance of older workers to their wages. Sheiner (1995a), in a paper discussed below, suggests that this is in fact the case. Even if there is shifting of employer costs to older workers, however, this group may still value employer-provided insurance particularly highly for two reasons. First, the variance of medical expenditures grows with age as well, raising the value of having insurance. Second, the differential cost of individual and group policies rises with age, so that even if they are paying for it, older workers would rather have group coverage than face the individual insurance market.

[28] There may be some remaining effect due to incompleteness in Medicare coverage; along a number of dimensions (high copayments and no prescription drug coverage), Medicare is less generous than existing employer-provided insurance plans.

[29] This table summarizes Tables 1–4 in Gruber and Madrian (1996). Medical expenditures above age 65 is the value for age 65–74 in their Table 4.

Table 4
Health risks by age

	25–34	35–44	45–54	55–64	65+
Self-reported health					
Fair	9.5	11.9	15.6	24.9	36.1
Poor	1.1	1.5	4.1	6.4	11.4
Incidence of specific diseases					
Stroke	0.4	0.8	1.6	3.6	7.4
Cancer	1.6	2.4	4.7	9.7	13.3
Heart attack	0.3	1.1	3.8	7.7	13.3
High blood pressure	10.1	18.2	29.1	41.9	49.8
Emphysema	0.4	1.0	2.6	5.2	8.0
Diabetes	1.7	3.0	5.7	9.8	14.7
Heart disease	0.8	2.2	6.1	11.9	22.2
Health care utilization					
Admitted to hospital?	9.2	6.8	8.7	11.0	20.1
Nights in hospital	5.5	6.8	9.3	11.8	13.8
Prescribed medicines?	52.9	55.6	61.1	71.1	81.9
Number of medicines	5.2	6.6	11.5	14.7	18.5
Visit to doctor?	64.1	67.1	71.1	77.9	85.8
Number of visits	4.6	4.6	5.5	6.0	7.4
Total medical expenditures					
Mean	1176	1135	1395	2144	2877
Standard deviation	4025	3537	4001	6532	7070

Notes: From Gruber and Madrian (1996); originally tabulated from 1987 National Medical Expenditure Survey and (for last two rows) from 1980 National Medical Care Utilization and Expenditure Survey.

uninsured, compared to 15.4 percent of 25–54 year olds [Gruber and Madrian (1995)]. Half of non-working older individuals are covered by employer-provided insurance, either in their own name or a spouse's, which reflects the fact that 45 percent of individuals work in firms that provide retiree health insurance [Madrian (1994b)]. However, 31 percent of older non-workers are either uninsured (14 percent) or purchase insurance in the individual market (17 percent). It is these individuals who potentially find themselves in this situation who might be expected to remain on their (insured) jobs rather than retire, since, as documented above, individual insurance is both very expensive and much less generous than group coverage.[30]

[30] Gruber and Madrian (1996) document that a health insurance policy for a 58 year old man and his wife purchased on the individual market in Massachusetts in 1993 would cost $8640, which was 26% of the average family income of retired individuals age 55–64 in that state and year.

Furthermore, there is also considerable anecdotal evidence that health insurance should be an important determinant of retirement. In a Gallup poll, 63% of working Americans reported that they "would delay retirement until becoming eligible for Medicare [age 65] if their employers were not going to provide health coverage" despite the fact that 50% "said they would prefer to retire early – by age 62" [EBRI (1990)]. Despite these persuasive arguments, and despite the existence of an enormous literature on the effects of health *status* on retirement decisions,[31] it is only over the past five years that researchers have focused on the effect of the availability of retiree health insurance coverage on the retirement decision.

The first approach to answering this question follows the original mobility literature [Mitchell (1982)] by modeling retirement decisions as a function of whether the worker has retiree coverage available. This approach is taken by Gustman and Steinmeyer (1994), Madrian (1994a, 1994b), Headen, Clark, and Ghent (1995), Hurd and McGarry (1996), Blau and Gilleskie (1997), and Rust and Phelan (1997). These studies universally find a very significant effect of retiree health insurance on retirement, particularly if the employer pays the full costs of this insurance. Gustman and Steinmeyer (1994) have the most mixed findings, depending on the concept employed: they find small effects on the average age of retirement, which falls by only 1.3 months, and on the share of the workforce retired at age 62, which rises by only 1 percentage point (2%); but they find large effects on the hazard rate at age 62, which rises by 6 percentage points (47%). The reason for this dichotomy is that part of the effect of retiree insurance in their model is to delay retirement until the age of eligibility, which is assumed to be age 62, so that the effect on the flow at 62 is much larger than the effect on the stock at that age. This large effect on hazard rates is confirmed by Blau and Gilleskie (1997), who find an 80% effect on the hazard rate if insurance is fully paid by the employer, and a 26% effect if it is only partially paid, and by Rust and Phelan (1997). Other studies, such as Madrian (1994a, 1994b), Headen, Clark, and Ghent (1995), and Hurd and McGarry (1996), do find significant effects on the odds of being retired early (a stock measure, as opposed to the flow hazard rate) on the order of 20–50% (with one of Madrian's estimates as high as 80%).

The second approach is adopted by Karoly and Rogowski (1994), who use the SIPP data for 55–64 year olds to examine early retirement. They do not observe in these data whether individuals have retiree insurance available, so they form a proxy based on firm size, industry, and region. They then include this proxy in a reduced form regression for early retirement, so that these excluded variables are in essence serving as instruments for retiree coverage. They also estimate fairly large effects, with retiree coverage associated with an 8 percentage point (47%) rise in the odds of early retirement, and a 100% increase in the hazard at age 60.

[31] See, for example, Bazzoli (1985), Bound (1989), and Stern (1989). See also the recent review in Currie and Madrian (1998).

Table 5
Research on health insurance and retirement

Paper (date)	Data (years), sample	Empirical strategy	Results
Gustman and Steinmeier (1994)	Retirement history survey (1969–79); males 58–63 in 1969	Structural estimation of retirement decision as function of value of retiree HI, controlling for pension value; simulation	1.3 month reduction in retirement age; 1 pp (2%) rise in stock of retired, 6 pp (47%) rise in hazard at age 62
Madrian (1994b)	NMES (1987) SIPP (1983–1986); males age 55–84	Regression of age at retirement on availibility of retiree HI; limited pension controls	Age of retirement reduced by 0.7–1.4 years; 7.5 to 15 pp (44–88%) rise in early retirement
Karoly and Rogowski (1994)	SIPP (1983–1989); 55–64 male wage earners	Model of early retirement on imputed probability of retiree HI coverage (by firm size, industry, and region); limited pension control	8 pp (47%) rise in early retirement; 100% rise in hazard at age 60
Gruber and Madrian (1995, 1996)	Current Population Survey (1980–90), SIPP (1983–89); 55–64 working males	Model of retirement status/rate as function of continuation of coverage availability; limited pension control	2.2 pp (32%) rise in hazard rate for one year of coverage
Lumsdaine, Stock, and Wise (1994, 1996)	Firm data on retirement, pension characteristics, and retiree HI	Structural "option value" model of retirement decision, incorporating valuation of Medicare; contrast of retirement at 65 among firms with and without retiree HI	No evidence of a role for Medicare in explaining "excess" retirement at 65
Headen, Clark, and Ghent (1995)	August 1988 CPS supplement; men and women 55–64	Ordered probit model of retirement and time retired as a function of whether worker has retiree HI	6 pp (35%) increase in odds of being retired; largest effects on being retired 10 years or more

continued on next page

Table 5, *continued*

Paper (date)	Data (years), sample	Empirical strategy	Results
Hurd and McGarry (1996)	HRS (1992); full time men age 51–61 and women age 46–61	Model of intended retirement dates as a function of retiree health insurance availability	Fully employer-paid retiree HI raises odds of early retirement by 11 pp (21%); partially paid raises by 7 pp (15%)
Blau and Gilleskie (1997)	HRS (1992–94); men age 51–62.	Dynamic model of retirement behavior as a function of retiree health insurance availability	Fully employer-paid retiree HI raises odds of exit by 6 pp (80%); partially paid raises odds by 2 pp (26%)
Rust and Phelan (1997)	RHS (1969–79); men age 58–73 who have only SS and not private pension	Dynamic programming model of retirement as a function of retiree health insurance avability and Medicare	Retiree HI is a significant determinant of labor force exit; Medicare can explain "excess" retirement at 65
Madrian and Beaulieu (1998)	US Census (1980 & 1990); married men age 55–69	OLS model of labor force participation as a function of spousal eligibility for Medicare	Significant effect of wife being over age 65 on husband's retirement decision; raises hazard at 60–62 by 25–50%; raises hazard at 63–65 by 10–20%

Notes: NMES = National Medical Expenditure Survey; PSID = Panel Study of Income Dynamics; SIPP = Survey of Income and Program Participation; CPS = Current Population Survey; HRS = Health and Retirement Survey; RHS = Retirement History Survey.

The third approach, used by Gruber and Madrian (1995, 1996), is to model early retirement as a function of the availability of continuation coverage. Continuation coverage acts as partial retiree health insurance coverage, by allowing retirees to buy cheap group coverage to cover at least part of their retirement period. Gruber and Madrian use both CPS and SIPP data to estimate the effect of continuation coverage availability on both the stock of early retirees and flows into early retirement. They find that there are sizeable effects: one year of continuation coverage increased the hazard rate into retirement by 32%.

These three approaches each have potential weaknesses. The first approach suffers from the selection problems discussed under Section 3.1. The potential importance of these problems is illustrated in the results of Blau and Gilleskie (1997), who find that the effects of retiree health insurance are not any larger for those in poor health than for those not in poor health; this suggests that much of the main impact of retiree health insurance may be due to selection. This selection is ideally controlled for in the rich structural modeling of Gustman and Steinmeyer (1994) and Rust and Phelan (1997), but the lack of complete data on retiree health insurance characteristics in the older RHS data (such as data on the exact timing of insurance availability) hamper these efforts.[32] The second approach suffers from the fact that firm size, industry, and region may not be legitimate instruments for retiree coverage, since they may be independently correlated with other determinants of retirement (such as pension coverage).[33] And the third approach, while potentially the cleanest in terms of identification, can only provide a rough indication of the effect of retiree insurance coverage, since continuation benefits are so limited relative to full coverage. Nevertheless, despite these weaknesses, the papers broadly agree that health insurance is an important determinant of retirement decisions, with retiree health insurance raising the odds of early retirement by 20–50%, and the hazard rate into retirement by 50–100%.

A natural implication of these findings is that an explanation for the very high rates of retirement at age 65 is eligibility for the Medicare program. Blau (1994), for example, reports that one-quarter of men who are still working at age 65, the age of Medicare entitlement, retire within three months of their 65th birthday. And, as Lumsdaine, Stock, and Wise (LSW) (1996) note, retirement rates at age 65 are far in excess of what would be predicted based on the incentives inherent in Social Security and private pension plans. But early work by these authors in both this paper and LSW (1994) failed to find an important role for Medicare in retirement decision-making. In LSW (1994), the authors incorporate the valuation of Medicare into a structural retirement model estimated

[32] In addition, studies such as Madrian (1994b) and Headen, Clark, and Ghent (1995) suffer from potential selection bias in the examination of workers that are both already retired (since the point of the paper is that retirement is correlated with retiree insurance) and alive (since retiree insurance may affect the odds that individuals are still living at the survey date).

[33] Karoly and Rogowski do include a variable for pension eligibility in their model, but the fact that it does not enter significantly raises questions about its validity as a control for the effect of pensions.

on data from one firm, and find little effect on retirement behavior. This is perhaps unsurprising, however, since the firm that they use provides retiree health coverage. But in LSW (1996), the contrast the "excess" retirement at age 65 at firms that do and do not have retiree coverage, and they find no major differences, once again belying a causal role for Medicare.

More recent work, however, has begun to uncover evidence of the importance of Medicare which is consistent with the broader literature on health insurance and retirement. Rust and Phelan (1997), using a dynamic programming model, estimate that there is a large role for Medicare, and that it can in fact explain the extent of "excess retirement" at age 65. The major difference between this paper and the LSW work appears to be that in the Retirement History Survey there is much more evidence of differentially large spikes at age 65 for those without employer-provided retiree insurance than for those with this coverage. In addition, Madrian and Beaulieu (1998) find that men are significantly more likely to retire early if their spouse is over age 65, once again suggesting a significant role for Medicare.

Clearly, the next step for research in this area is to build on the strengths of the longitudinal data analysis in Gustman and Steinmeier and LSW, while taking more seriously issues of selection. This should be very feasible given the excellent new data on retirement, pension characteristics, and retiree health insurance in the new Health and Retirement Survey (HRS). Future work using these data, perhaps building on the identification strategies successfully employed in the "job lock" literature, will be useful in pinning down the magnitude of the retirement effect. In particular, an important priority is to further integrate the modeling of employer-provided retiree coverage and Medicare.

As with the mobility literature, there is also an important question here of how to interpret the welfare implications of these findings. For those without retiree coverage, the availability of lower cost group insurance on the job, but only expensive individual insurance after retirement, is a potential source of inefficiency. The fact that workers respond so strongly to retiree coverage suggests that there may be large welfare gains from reducing this inefficiency by increasing the availability of group coverage for early retirees. That is, a policy of continuation coverage which was not limited to 18 months, but which extended until age 65, would increase welfare by "leveling the playing field" between working (where presumably the cost of insurance is paid through lower wages) and retirement (where it would be paid out of pocket).[34] At the same time, there are at least two mitigating factors that reduce the welfare cost of this "retirement lock". First, as noted above, reduced retirement may provide a mechanism for firms to reap the benefits of firm-specific human capital investments. Second, increases in retirement would decrease tax revenues from taxing the high earnings of older workers, which is

[34] Indeed, Gruber and Madrian (1995) infer from the retirement response to continuation availability (relative to the response to pension wealth) that one year of continuation coverage is worth $13,600 to workers, a figure substantially above the $3600 in expected financial savings from having a continuation policy (relative to individual insurance).

not accounted for by workers in making their retirement decision, since they compare their after-tax earnings to the value of leisure.

4.2. Health insurance and public assistance participation

Another margin along which health insurance might affect labor supply is public assistance participation. A key feature of several public assistance plans is that, in addition to cash benefits, individuals qualify for Medicaid coverage of their medical expenses. The major plans that feature this linkage are cash welfare for low income single female-headed families, formerly Aid to Families with Dependent Children (AFDC) and currently Temporary Assistance for Needy Families, and Supplemental Security Income (SSI) for low income disabled persons and elderly. This coverage can amount to quite a valuable benefit, since Medicaid provides first dollar coverage of physician and hospital expenditures, as well as coverage of prescription drugs and other optional benefits (vision, dental care) in many states. In addition, the work opportunities available to potential AFDC and SSI participants are low-wage, low-skilled jobs without health coverage.[35] As a result, the linkage of Medicaid to public assistance participation both encourages non-workers to sign up for the programs, and taxes work among potential recipients. That is, there is a form of "welfare lock": individuals are reticent to leave government programs because they will lose their health insurance.

This effect is illustrated in Figure 2, from Yelowitz (1995). This figure shows the welfare receipt and work decisions of a single woman with children, who can receive AFDC if her income is below $H_{\text{breakeven}}$. This woman trades off utility from leisure and from consumption of goods that is financed from wage income or from welfare payments. The recipient faces a constant post-tax wage w^0. However, she is assumed to be unable to obtain a job with health insurance.[36]

At zero income, this woman receives a certain amount of cash welfare income from AFDC, as well as in-kind benefits, such as Food Stamps and Medicaid. As she earns labor income, her AFDC and non-Medicaid in-kind benefits are taxed away at a high marginal rate, so that her after-tax wage is $w^1 = (1 - \tau_{\text{AFDC}}) * w^0$.[37] Once she works more than $H_{\text{breakeven}}$, the hours of work where the entire welfare benefit is taxed away, she loses her AFDC eligibility, and hence her Medicaid benefits. This creates a dominated part of the budget set, known as the "Medicaid notch". This notch provides a major

[35] I use AFDC to summarize the effects of AFDC/TANF, since all of the work in this area refers to the older program.

[36] Equivalently, she may be able to obtain a job with insurance, but only at a compensating differential which exactly equals her valuation of that insurance. Short, Cantor and Monheit (1988) find that 43% of people who left welfare were covered by private health insurance. Since only those with the best opportunities leave welfare, the likelihood of finding a job with insurance for the average welfare recipient, should they leave the program, is quite low.

[37] This marginal rate is 67% for the first four months, and 100% thereafter (after a basic exemption and some deductions for work and child care expenses).

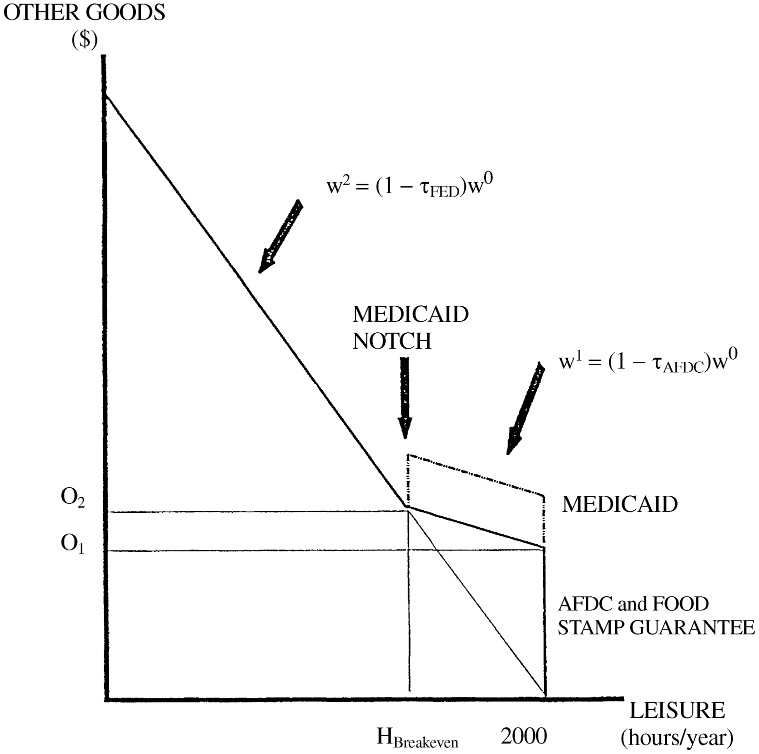


Figure 2.

disincentive to working her way off welfare. As Yelowitz documents, for a mother with 2 children in Pennsylvania in January, 1991, the woman would have to earn more than $5000 additional dollars off welfare to break even with her income on AFDC at point $H_{\text{breakeven}}$.

A number of studies have addressed the welfare lock question in the context of the AFDC program, as reviewed in Table 6. There have been three basic empirical approaches used in this literature. The first is to use differences in individual characteristics to predict who is likely to be "locked" into the AFDC program by Medicaid due to high medical spending, and then to assess differential participation rates by this imputed value of Medicaid. Ellwood and Adams (1990) follow this approach using administrative Medicaid claims data to examine exits from AFDC, and Moffitt and Wolfe (1992) model participation as a function of imputed value in the SIPP. The results are fairly similar, showing sizeable decreases in the likelihood of exiting AFDC as the imputed value of Medicaid rises.

The second approach is to abstract from individual health, and to use variation in the characteristics of state Medicaid programs to identify the value of Medicaid to the

Table 6
Research on health insurance and public assistance participation

Paper (date)	Data (years), sample	Empirical strategy	Results
Ellwood and Adams (1990)	Medicaid claims data for GA & CA (1980–86); women receiving AFDC	Model of leaving AFDC on expected future medical expenses in the next three months based on medical usage of the previous six months	100% increase in expected medical costs lowers exit probability by 6.5–11%
Moffitt and Wolfe (1992)	SIPP (1983–86)	Model of AFDC and labor force participation on a family's predicted Medicaid and private insurance valuation based on family structure	Increasing value of Medicaid coverage by 33% raises AFDC participation by 2% and lowers LFP by 5.5%
Blank (1989)	National medical care utilization & expenditure survey (1980); female heads	Model of AFDC participation on state Medicaid spending per recipient and presence of Medically Needy program	No effect of either program parameter
Winkler (1991)	CPS (1985); female heads	Model of both AFDC and labor force participation on state Medicaid spending per recipient and presence of Medically Needy program	No effects on AFDC participation; 10% increase in Medicaid spending reduces LFP by 0.9 to 1.3 pp
Montgomery and Navin (1992)	CPS (1987–92); female heads	Model of participation on state Medicaid spending per recipient, with and without state fixed effects	No fixed effects: 10% increase in Medicaid spending = 0.36 pp reduction in participation fixed effects: insignificant
Yelowitz (1995)	CPS (1988–91); female heads	Model of AFDC participation and labor force participation on eligibility for Medicaid expansions for children	Increasing income cutoff for eligibility by 25% of the poverty line decreases AFDC participation by 4.6% and increases labor force participation by 3.3%
Decker (1994)	State AFDC caseload data (1964–74); CPS (1966–72) data on single female heads	Model of state caseloads and individual AFDC participation as a function of introduction of Medicaid	Medicaid introduction led to 21% increase in caseloads, 6.4 pp (24%) increase in individual participation, insignificant LFP effects
Yelowitz (1996a)	CPS (1986–1991); age 65 plus	Model of SSI participation as a function of QMB eligibility	QMB program led to a 1.7 pp (40%) reduction in SSI participation
Yelowitz (1998)	CPS (1987–1993); 18–64	Model of SSI participation as function of Medicaid expenditures, instrumented by expenditures on other groups	Rising medical costs explain 0.1 pp rise in participation
Yelowitz (1996b)	SIPP (1986–1994); 18–64	Model of food stamps participation as function of Medicaid eligibility	Making all households eligible for Medicaid would raise FS participation by 0.59 pp (7.5%)

Notes: SIPP = Survey of Income and Program Participation; CPS = Current Population Survey; LFP = Labor Force Participation; QMB = Qualified Medicare Beneficiares; FS = Food Stamps.

potential AFDC participant.[38] Blank (1989) was the first to pursue this approach, estimating models of AFDC participation and hours of work on average state Medicaid expenditures and the presence of a state Medically Needy program, which provides Medicaid to non-AFDC families if their income net of medical expenditures falls below a certain floor. She finds no effect of either policy variable on AFDC participation. Winkler (1991) also finds no effect of average expenditures on AFDC participation, but does find an effect of average expenditures on labor force participation, a finding echoed by Montgomery and Navin (1992) (albeit with a much smaller estimate). But there is no effect of Medicaid expenditures on participation in Montgomery and Navin's work once state fixed effects are included in the regression models.

The third approach that has been taken to this question extends the notion of using state parameters, to exploit the most dramatic change in insurance policy in the US in the past 25 years: expansions of the Medicaid program to children and pregnant women living in non-public assistance receiving households. As described in more detail in Gruber (1996), these Medicaid expansions were phased in across the states since 1984, proceeding first by state option and then by federal mandate. By mid-1991, eligibility was extended to any child under age 6 or any pregnant woman (for the expenses of pregnancy only) in a family living below 133% of the poverty line, as well as to any child born after September 30, 1983 living below the poverty line, regardless of family composition. In addition, states had the option of expanding coverage even higher up the income distribution, an option taken up (in 1996) by over half the states. Currie and Gruber (1996a, 1996b) estimate that as a result of these expansions by 1992 almost one-third of all children in the US and almost one-half of pregnant women are eligible for Medicaid coverage of their medical expenses.

As Yelowitz (1995) notes, these expansions served to decouple Medicaid eligibility from AFDC receipt, thereby providing precisely the variation needed to separately identify the role of Medicaid from that of other factors in determining welfare participation. A key feature of these expansions was variation across the states in the timing and generosity of increased income limits. Indeed, there was even variation within states at a point in time, due to different age cutoffs for eligibility of children across the states. This allows Yelowitz to form plausibly identical groups of families, some of which (the "treatments") were able to leave AFDC and retain their Medicaid coverage, and others of which (the "controls") were not. And he finds significant effects of being in the treatment group on both AFDC participation and labor force participation: he estimates that increasing the income cutoff for eligibility by 25% of the poverty line decreases AFDC participation by 4.6% and increases labor force participation by 3.3%.

A related approach is taken by Decker (1994). She examines the effect of the introduction of the Medicaid program in the late 1960s and early 1970s on AFDC participation in that era. Since the Medicaid program was phased in across the states over a

[38] Features of the state Medicaid program are included in the set of variables used to predict Moffitt and Wolfe's (1992) index, but the papers discussed below use *only* state features for identification.

period of several years, she is able to assess whether states that adopted Medicaid saw a subsequent increase in their AFDC rolls, relative to states that did not. In fact, she finds a very strong effect, with the introduction of Medicaid leading to a 6.4 percentage point (24%) rise in the odds that a single female head participates in AFDC.[39]

As with the retirement literature, these different approaches each have some potential weaknesses. A problem with the individual health valuation approach is it hinges on the assumption that a family's value of Medicaid does not capture other factors that determine AFDC participation. This is unlikely to be true, however, since individual health status (a key predictor of Medicaid valuation) will be independently correlated with desired labor supply and AFDC participation; as noted above, there is a large literature that finds a substantial negative effect of health status on labor force participation. Both studies recognize this potential problem, and attempt to address it by examining separately the effects of the family head's health status and that of the children in the family; Ellwood and Adams find that increases in expected children's spending had similar effects to their main findings, while Moffitt and Wolfe found effects that were only one-third as large for the children's component of their index as for the adult component.[40]

There are potentially more serious problems with using average state Medicaid expenditure as a proxy for the value of the program to the typical family. This is a very noisy proxy for the underlying quality of the Medicaid package; as a result, measurement error will bias downwards the estimated effect of Medicaid. Moreover, much of the variation in this measure comes from variations in the underlying health of the Medicaid population, which will be spuriously correlated with participation decisions. For example, if the marginal persons joining Medicaid is healthier than the average person enrolled, then states with high participation will have low Medicaid costs, once again biasing against a finding of welfare lock.

Finally, while the use of legislative variation in Medicaid in the work of Yelowitz and Decker once again provides potentially the cleanest identification strategy, there is the problem of limited applicability. For example, the Yelowitz findings only apply to the marginal population made newly eligible for the expansions, which may not provide insight for the "harder core" of long-term AFDC enrollees. Nevertheless, the strong findings of this approach, as well as those of the health valuation approach, lead one to the conclusion that welfare lock is an empirically important phenomenon.

In a series of subsequent studies, Yelowitz has explored the effect of Medicaid on participation in other public assistance programs. The first is SSI; as Yelowitz highlights, this program is actually larger in dollar terms than is AFDC, and the same type of welfare-lock problem arises in this context. For elderly SSI recipients, this problem arises because the Medicaid coverage that they receive on SSI pays for their

[39] For this era, however, her results indicate that this increase is primarily due to increased takeup among those already eligible for AFDC, *not* due to reduced labor supply in order to make oneself eligible; but the labor supply effects are imprecisely estimated.

[40] Even this approach has the problem, however, that potential AFDC recipients may be reluctant to go to work if they have a sick child, regardless of Medicaid coverage.

non-covered Medicare expenditures. Using an expansion of Medicaid for the elderly, Yelowitz (1996a) finds a non-trivial welfare lock for this population as well. For the disabled, who get Medicaid if on SSI, Yelowitz (1998) follows the second approach noted above, using variation across states in the Medicaid spending to proxy for the program's generosity. But he addresses the problems with this approach by instrumenting average spending on the disabled by spending on blind recipients, a proxy for program generosity that is uncorrelated with the disabled case mix, and which as a result solves the selection problem inherent in the average expenditures measure. He finds that instrumenting substantially raises his estimates (suggesting that the problems described above are real), and that growth in Medicaid generosity over 1987–1993 can explain almost all of the substantial growth in the SSI disabled caseload. Finally, Yelowitz (1996b) asks whether increased eligibility for Medicaid raises utilization of the food stamps program, both through reducing labor supply and increasing awareness of public assistance programs. Using the same estimation approach as Yelowitz (1995), he finds that Medicaid eligibility does increase food stamp participation, and that this increase occurs through both channels.

Thus, to summarize, this literature suggests that health insurance is a very important determinant of public assistance participation. This has two important welfare implications. First, it suggests that reduced public assistance expenditures may offset a share of the increased costs of expanding health insurance availability. Yelowitz (1995) estimates that expanding eligibility for Medicaid to all women and children with incomes below 185% of the poverty line in 1989 would have saved the government $410 in expenditures per female-headed household per year. Second, there may be non-financial costs to the increase in welfare dependence that results from welfare lock. A number of analysts have suggested a hysteresis-type model of welfare behavior, with exposure to the welfare system increasing future utilization, by both a mother and by her children as adults [Murray (1984)]. Existing evidence on welfare dependence is mixed, with some recent studies concluding that there is little intergenerational transmission of welfare [Zimmerman and Levine (1993)]. But this possibility highlights the benefits of moving welfare recipients off of the public assistance rolls through reducing welfare lock.

Reducing welfare lock through public insurance expansions can also have additional effects on labor market equilibrium, through adjustments of private insurance coverage and wages. As Cutler and Gruber (1996a) note, the typical privately insured family pays for about one-third of its medical costs out of pocket, but Medicaid coverage is comprehensive and free. Moreover, two-thirds of those made eligible by the Medicaid expansions already had private insurance coverage. These facts highlight the possibility that expanded public insurance eligibility could "crowd out" private insurance coverage. Such crowdout could occur through employers dropping insurance coverage if a large share of their workforce is public insurance-eligible, or through employees not taking up somewhat costly employer coverage in the face of eligibility for free Medicaid coverage. Recent evidence suggests that crowdout is quite sizeable. Cutler and Gruber (1996a), who study the Medicaid expansions over the 1987–1992 period, find that for every two persons who joined the Medicaid program one person lost private insurance coverage; although Dubay and Kennedy (1997) find smaller effects.[41]

If there is crowdout, then public insurance expansions will not only reduce welfare lock, but will also potentially reduce job lock as well. By providing extra-workplace insurance coverage for workers or their dependents, Medicaid frees up workers to move to more productive positions. In addition, there may also be effects on wages and hiring, since employer insurance costs have been shifted to the government. As Cutler and Gruber (1996b) note, if the costs of health insurance are fully shifted to wages (as is supported by the literature reviewed below), then the Medicaid expansions provided a transfer of $1523 to the average family made eligible. If they are not shifted to wages, then they provide a subsidy to the hiring of the low wage workers who are likely to be eligible for the program, and who will therefore not take up costly employer-provided insurance. But there is no empirical work to date on the effect of the expansions on job mobility, wages, or employment determination.

4.3. Health insurance and labor force participation and hours worked of prime age workers

Most of the interest in both academic and public policy circles around the labor force participation effects of health insurance has been focused on retirement and public assistance participation. But, in terms of the impacts on aggregate hours worked, the most important effects may well be on the work decisions of prime age workers, and particularly secondary workers. These effects arise because health insurance is generally offered for the entire family, so that having only one spouse with a job offering insurance is enough to provide the opportunity for coverage for the entire family. As a result, the availability and coverage of health insurance for primary workers may be a key determinant of the labor supply decisions of secondary earners in the family.

A small set of recent papers has investigated this question, focusing primarily on the effects of husbands' health insurance on the labor supply decisions of their wives; these studies are described in Table 7. The basic finding of all these papers is clear: wives whose husbands do not have health insurance are much more likely to work, to work more hours, and to be in jobs that offer health insurance. The magnitudes vary somewhat, but the effects are all large, with husband's insurance coverage being associated with a reduction in labor force participation ranging from 11–20%, and an additional reduction in conditional hours on the order of 5–20%. There is also evidence that wives are more likely to choose jobs with health insurance if their husbands are not covered

[41] See Cutler and Gruber (1997) for a response to Dubay and Kennedy (1997). Cutler and Gruber's (1996a) results do not imply that one-half of those joining the Medicaid program came from being privately insured, since some of those losing their private coverage in response to the expansions may become uninsured. For example, a family may drop coverage when the children and wife become Medicaid eligible, with the husband becoming uninsured; alternatively, women may be uninsured when they are not pregnant, gaining Medicaid coverage when they are. See Cutler and Gruber (1996b) for a further discussion.

Table 7
Research on health insurance and prime-age labor force participation

Paper (date)	Data (years), sample	Empirical strategy	Results
Wellington and Cobb-Clark (1997)	CPS (1993); 25–62 year old husbands and wives	Model of hours as a function of being covered by spouse's insurance policy	Husband's insurance = 20% reduction in LFP and 7–15% reduction in hours Wife's insurance = 4–9% reduction in LFP and 0–4% reduction in hours
Schone and Vistnes (1997)	NMES (1987); married women age 25–51	Joint model of hours of work and job choice as function of husband's insurance status	Husband's insurance = 14% reduction in LFP and 30% reduction in job with HI
Olson (1997)	CPS (1993); married women younger than 65	Model of hours and participation as a function of husband's insurance status	Husband's insurance = 13% reduction in total hours; 11% reduction in LFP
Buchmueller and Valetta (1999)	CPS (1993); 25–54 year old married women	Model of wife's hours and participation as a function of husband's insurance status	Husbands' insurance = 36% reduction in total hours and 12% reduction in LFP

Notes: NMES = National Medical Expenditure Survey; SIPP = Survey of Income and Program Participation; CPS = Current Population Survey; LFP = Labor Force Participation; HI = Health Insurance.

[Schone and Vistnes (1997)], and that there is a small effect of the wife's insurance on the husband's labor supply decision [Wellington and Cobb-Clark (1997)].

A potential problem with all of these studies, however, is omitted variables that are correlated with both the husband's insurance coverage and the wife's tastes for work; if husbands who demand "good jobs" are married to women who have preferences against market work, it could cause this result even in the absence of any causal role for health insurance. This issue is not completely satisfactorily addressed in any of these papers, but Buchmueller and Valetta (1999) consider it most carefully. They find that (a) these effects are strongest for those with larger families, which is consistent with the notion that it is health insurance valuation and not tastes for work driving the results; (b) the effects of husband's insurance on wife's hours when the wife is in a job that does not offer insurance are *positive*, suggesting that any unobserved correlation biases against the finding of interest; and (c) husband's insurance is associated only with a reduction in full-time work, and not a reduction in part-time work. There are alternative explanations that one could offer for each of these findings, but taken together they provide fairly strong support for the causal interpretation of their health insurance findings.

These findings have very important implications for the labor market impacts of health insurance policies, particularly policies such as national health insurance; if there is such "wife lock" in practice, it suggests that large scale insurance coverage expansion could cause a non-trivial reduction in the size of the labor force. Once again, however, the welfare implications are unclear. To the extent that health insurance is distorting female labor force behaviour, there are welfare costs for these families; and, if maternal time with children is important for child development, there are potentially even larger long run consequences for child development. On the other hand, the existing US tax code includes several distortions against labor supply by married women, such as the marriage tax penalty against two earner couples and the inframarginality of Social Security tax payments by low earning spouses [Feldstein and Feenberg (1996)]. As a result, this type of "lock" may be appropriately offsetting other distortions against spousal labor supply.

5. Evidence on health insurance and wages, hours, and employment

The discussion in Section 2 highlighted a number of channels through which changes in health care costs, either through inflation in the health sector or government mandates, could affect the functioning of the labor market. In this section, I review the existing evidence on the labor market effects of changing health care costs. In particular, I focus on the key question of whether increases in health care costs are shifted to wages, or whether they are reflected through other channels such as hiring.

5.1. Time series patterns

The notion that there is a tradeoff between fringe benefit costs and wages is suggested by Figure 3, which presents a time series graph of employer-provided health insurance costs and wages. These data are from the Employment Cost Index series, which is based on an establishment-level survey carried out by the Bureau of Labor Statistics. The data cover all private sector employees.

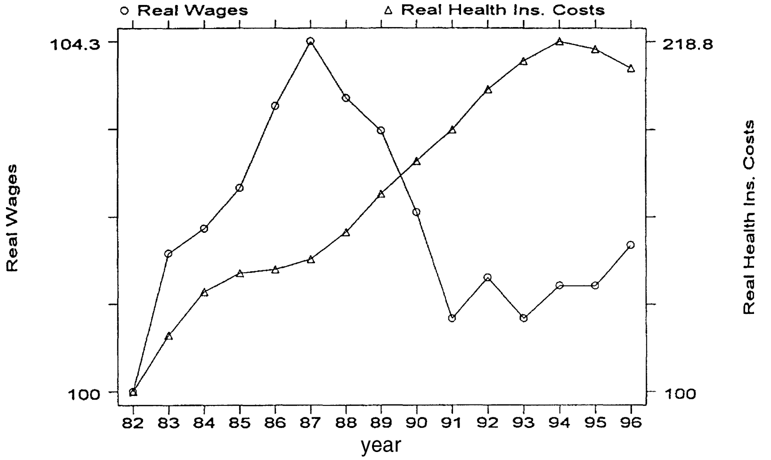


Figure 3. Health insurance costs and wages over time.

For most of this time period, there is a strong negative relationship between the growth in employer health care costs and the growth in wages. In the early 1980s, these series do move together. But then health care cost growth slows in the 1984–87 period, and there is rapid wage growth in these years. Beginning in 1988, however, health care cost growth becomes very rapid, and there is a steep decline in real wages at this same time. Finally, health care cost growth slows in 1992, just as real wages flatten and even rise somewhat. While only suggestive, this time series pattern is certainly consistent with shifting of the costs of health insurance benefits to wages.

5.2. Health insurance and wages

Modeling the effect of health insurance costs on wages is a natural application of the compensating differentials framework described earlier. The standard compensating differentials approach would involve a regression of wages on the existence or cost of health insurance. This is the approach followed by the first two studies described in Table 8, Leibowitz (1983) and Monheit et al. (1985). Both studies, however, find a wrong signed result: health insurance costs, or availability, are *positively*, not negatively, related to wages. A very different approach is taken by Woodbury (1983), who structurally models the substitutability of wages and fringes in firm-based data; he does find a high degree of substitutability between the two.

The finding of a wrong-signed wage offset reflects the difficulty faced by many empirical applications of compensating differentials theory: selection, on both the worker and firm side.[42] High productivity workers may choose to have some share of their compensation in benefits; indeed, given the progressivity of the tax schedule and the deductibility of benefits, the demand for benefits should rise with underlying productivity. And, as highlighted above, the "good jobs" that pay high wages are also the ones that offer generous benefits along a number of dimensions. What is required to identify the effect of health insurance costs on wages is *exogenous* variation in the cost of insurance.

A number of studies over the past decade have attempted to provide such exogenous variation, with results that are supportive of extensive shifting of insurance costs to wages. Eberts and Stone (1985) use variation in the cost of health benefits across school districts in New York from 1972–1977, controlling for unobserved worker and district characteristics by including district fixed effects, and by controlling for other benefits

[42] See Smith (1979), Brown (1980), and Rosen (1986) for general discussions of estimating compensating differentials and reviews of past literature in this area. Triplett (1983) and Smith and Ehrenberg (1983) provide discussions of the estimation problems in the context of worker benefits. There has been more success documenting compensating differentials for job safety [see Viscusi (1992) for a review] and for locational amenities [see Gyourko and Tracy (1989)]. The literature on pensions and wages is much larger than that on health insurance and wages, and has produced mixed results [see Ehrenberg and Smith (1991), Kotlikoff and Wise (1985), Clark and McDermed (1986), Montgomery, Shaw, and Benedict (1990), and Gunderson, Hyatt, and Pesando (1992)].

Table 8
Research on health insurance and wages, employment, and hours

Paper (date)	Data (years)	Empirical strategy	Results
Leibowitz (1983)	RAND Health Insurance Study (1978); full-time workers	Model of wages on health insurance premiums for full-time workers with health insurance	Positive correlation between wages and premiums
Monheit et al. (1985)	National Medical Care Expenditure Survey (1977); workers	Model of wages on indicator for being offered health insurance	Positive correlation between wages and health insurance
Woodbury (1983)	BLS Employee Compensation Survey (1966–74); School Districts (1977)	Structural model of substitutability of wages and fringes	Wages and fringes are highly substitutable – elasticity of substitution greater than one
Eberts and Stone (1985)	New York City public school districts (1972–77); full time teachers	Model of change in wages on change in in cost of health benefits across school districts	83% shifting of increases in health costs to wages
Gruber and Krueger (1991)	CPS, Employment & Earnings (1979–88); workers in 5 high WC cost industries	Model of wages and employment on WC costs/payroll by industry/state/year; CPS: wages on costs; E&E: wages & employment on costs	CPS: 85% shifting to wages E&E: 56–86% shifting to wages. No employment effects
Gruber (1994a)	CPS (1974–82); all 20–64	Model of wages and labor supply on effect of maternity mandates: (a) DDD for 20–40 women (b) Cost of mandate for all	Full shifting to wages No effect on total labor supply: hours up, employment down
Sheiner (1994)	CPS (1990–91); 25–59 workers	Model wages as a function of city-specific costs times: (a) age of worker (b) marital status (c) family vs. indiv. coverage	Men: full shifting to wages from (a)–(c) Women: insignificant for (a) and (b), full shifting for (c)
Olsen (1994)	CPS (1982, 1992); working married women	Husband's insurance as instrument for wife's coverage in wage equation	Insurance for wife = 10% wage reduction

continued on next page

Table 8, *continued*

Paper (date)	Data (years)	Empirical strategy	Results
Miller (1995)	CEX (1988); non self-employed workers age 18+	Model of wage levels and changes as a function of level and changes in insurance status	Positive levels relationship, negative in changes; losing health insurance = 11% wage increase overall; 16% for men vs. 7% for women
Ryan (1997)	SIPP (1988); non-self employed men age 24–64	Model of wage levels and changes as a function of level and changes in insurance status	Positive levels relationship, negative in changes; losing family coverage = $950 gain in wages; losing coverage for singles = $1640 gain
Buchmueller and Lettau (1997)	ECI panel data on jobs (1987–94); private sector jobs of 1500 hours +	Model of changes in wages for job/firm pairs as a function of changes in cost of health insurance	Consistent positive relationship between wage changes and insurance cost changes
Ehrenberg and Schumann (1984)	Establishment data (1976)	Model of log(overtime hours per workers) as function of fringe costs/wage ratio; OLS and instrument by worker characteristics (age, sex, median income)	10% rise in fringe/wage ratio = 4.8–17% rise in overtime/worker in manufacturing; 7.8–12% in non-manufacturing
Ehrenberg, Rosenberg, and Li (1988)	CPS (1983); non self-employed workers	Model of relative part-time work on relative insurance coverage of part-time workers across industries	No effect of relative coverage of part-time workers on use of part-time workers
Montgomery and Cosgrove (1993)	205 Child Care Centers (1989)	Model of part time work as a function of fringe benefits payments and eligibility of part-timers for benefits	1% rise in benefits/wages = share of hours worked by part-timers falls 0.43%; no effect of eligibility for part-timers
Cutler and Madrian (1998)	CPS (1979–92); 25–54 non-self employed men	(a) Time trends in hours by insurance status (b) Differential health cost growth by industry	(a) 0.7 hour/wk increase over 1980s (b) 2.2 hour/wk increase over 1980s
Buchmueller (1998)	Survey of California employers with 3+ employees (1993)	Model part time work as a function of difference in fringe costs between full and part time workers (largely driven by part-time eligibility)	1% rise in relative full-time benefits costs = 1.09% rise in part-time work; elast is 1.19 for HI costs specifically

Notes: NMES = National Medical Expenditure Survey; PSID = Panel Study of Income Dynamics; SIPP = Survey of Income and Program Participation; CPS = Current Population Survey; CEX = Consumer Expenditure Survey; ECI = Establishment Cost Index; WC = Workers Compensation; DDD = Differences-in-Differences-in-Differences model.

costs. They find that 83% of the increases in health costs across districts were reflected in lower wages.

One source of exogenous changes in employer costs is government mandated increases in the cost of insurance.[43] Gruber and Krueger (1991) identify the effect of increased insurance costs on wages by using increases in the employer costs of Workers' Compensation (WC) insurance across industries and states over time. WC provides cash benefits and health coverage to workers injured on the job, and much of the variation in costs in their data comes from increases in the health care component of this program. They focus on workers in five industries for which WC costs are high and rapidly growing; in some industries and states, these costs amounted to over 25% of payroll by 1987, the end of their sample period. They use both micro-data on wages (from the CPS) and aggregate data on employment and wages by state/industry (from administrative data on firm payrolls). They include state and industry fixed effects in their models, so that they are controlling for general differences in pay across industries and places, and estimating only how that pay changed when the costs of WC rose. In both datasets, they find that for these set of industries 85% of increases in workers compensation costs were shifted to wages; for a broader set of industries in the aggregate data, they estimate shifting of 56%.

Gruber (1994a) extends this approach to a group-specific health insurance mandate, mandated comprehensive health insurance coverage for childbirth. Before the mid-1970s, coverage for the expenses of childbirth in health insurance plans was much less generous than coverage of other services, but a series of state laws after 1974, as well as a federal law in 1978, outlawed this practice. This substantially increased the cost of insuring a particular group of workers, women of child-bearing age (and their husbands, who may have covered these women on their health insurance plans). Gruber examines whether this exogenous increase in insurance costs was reflected in the relative wages earned by these affected groups. He does so by extending the DD approach discussed earlier to a "differences-in-differences-in-differences" approach, comparing the change in *relative* wages of these affected groups (relative to unaffected groups such as older workers and single men), in states with mandates relative to those without. Doing so, he finds that there is a significant relative decline in wages for married 20–40 year old

[43] This discussion focuses only on articles that pertain to health insurance mandates (or recent workers compensation mandates, where much of the variation comes from changes in health costs). There are a number of closely related studies which focus on the incidence of government mandates or payroll taxes that do not finance health benefits. Fishback and Kantor (1995) study the introduction of the workers' compensation program in the early 1900s, and find that most of the costs of this new insurance program were reflected in lower wages. Anderson and Meyer (1995) find that the incidence of the payroll tax used to finance unemployment insurance is mostly on wages. Holmlund (1983) uses time-series data on payroll taxes in Sweden to examine wage growth in a period when the payroll tax increased from 14 to 40%, and he estimates that 50% of this increase was shifted to wages in the short run. Hamermesh (1979) uses the variation in payroll tax rates due to the social security payroll tax limit to estimate wage offsets; his estimates indicate that from 0 to 35% of the social security tax is shifted to wages. Finally, Gruber (1997) estimates that the incidence of payroll taxation to finance social insurance programs in Chile was fully on wages, with little effect on employment.

women, whose costs rose most under this mandate. Using data on insurance costs to parameterize the cost of the mandate across the full sample of workers, he finds full shifting of these costs to wages.

Sheiner (1995a) also considers the question of group-specific shifting. She notes that groups with higher baseline insurance costs, such as older workers (relative to younger workers) and workers with family insurance coverage (relative to those with individual coverage), should see the greatest rise in insurance costs when there is a general rise in area medical prices. Using data on changes in insurance costs across cities, interacted with indicators for being in a high cost group, she finds that there is a relative decline in wages for high costs groups when area cost rise; her results indicate full shifting to wages for men, with mixed results for women. Olson (1994) focuses explicitly on women, and uses as an instrument for their health insurance coverage the coverage of their husbands; women whose husbands are uninsured are more likely to demand insurance, and may accept lower wages as a result. Indeed, using this instrument, Olsen finds a weakly significant 10% wage reduction associated with insurance coverage; this is roughly the ratio of health insurance costs to wages for this group.

More recent work in this area has attempted to control for heterogeneity by using fixed effects for persons or jobs. Miller (1995) and Ryan (1997) pursue similar approaches in the Consumer Expenditure Survey (CEX) and the SIPP, respectively, first identifying a wrong-signed (positive) relationship between wage levels and health insurance offering, then showing that the relationship has the expected negative (and highly significant) sign in changes. Miller's estimated effect of an 11% wage effect seem somewhat large, given that his sample is a mix of married and single policies; and his estimate for men only is a very large 16%. Ryan finds much smaller effects for her full sample, with an offset of only $950 that is significantly smaller than average insurance costs; she also finds a much larger offset for single workers, which is counterintuitive given that their policies should be less expensive. These mixed results may reflect the fact that the studies control to some extent for worker characteristics, but not job characteristics; moreover, there may be changes in worker characteristics (e.g., productivity shocks, positive or negative) which are correlated with the change in jobs.

Buchmueller and Lettau (1997) take a different approach, using jobs as the unit of observation in a unique data set with job-specific information on wages and insurance costs over time. This allows them to control for good vs. bad jobs, although potential problems with worker selection into these jobs remains. Unlike Eberts and Stone (1985), however, they do not find the expected negative relationship between wages and health insurance costs.

The primary lesson from this literature is that estimating compensating differentials of this variety is very difficult, and requires sophisticated identification strategies for clean results. But the results that attempt to control for worker selection, firm selection, or (ideally) both, have produced a fairly uniform result: the costs of health insurance are fully shifted to wages. As with the mobility literatures reviewed earlier, each of these approaches has its limitations. The evidence from mandated benefits relies on the exogeneity of the law changes with respect to labor market conditions, and only provides information for the marginal changes that are embodied by the mandates, and not average differences across employers in health insurance costs.[44] And the evidence using cross-city medical prices faces the problem that these prices may be determined by the city-specific labor market conditions that determine wages, due to the wage component of health care costs. Nevertheless, the uniformity of the conclusions across these very different strategies is striking.

5.3. Health insurance, employment, and hours

A natural implication of the full shifting of the costs of insurance to wages is that there should be no effect on the equilibrium level of labor utilization. This contention is supported by two of the studies reviewed above. Gruber and Krueger (1991), using aggregate state/industry data, find no effect of changes in workers' compensation costs on employment levels. And Gruber (1994a) finds no effect of the "maternity mandates" on total hours of work. Thus, the result from the full valuation case of Summers (1989) is supported by the evidence: full shifting to wages with no effect on labor utilization.

As noted earlier, however, even if there is no effect on average, rising health insurance costs may change the compositional mix of employment and hours. There is a large literature on fringe benefits costs (and other fixed labor costs) and use of overtime labor, and the firm or industry level. This literature is reviewed by Ehrenberg and Schumann (1984). They update previous models of hours of work and fringe benefit costs using establishment data for 1976. They acknowledge the endogeneity of fringe costs to hours of work (since non-fixed fringe costs such as pension contributions are themselves a function of earnings), and instrument by employee characteristics which are correlated with fringe demand. They find very large effects, indicating a 5–17% rise in overtime hours/worker in response to a 10% rise in fringe benefits costs.[45]

More recent research assesses the effect of fringe provision on use of part-time workers. Montgomery and Cosgrove (1993) find that increases in benefits costs decrease the share of hours at their sample of child care centers that are worked by part-time workers, which is consistent with employer preferences. But neither they nor Ehrenberg, Rosenberg, and Li (1988) find any effect of variations in the eligibility of part-time workers

[44] Neither of these counterarguments is likely to explain the findings of these papers, however. For example, it may be that governments tend to mandate benefits when the economy is doing poorly, causing a negative correlation between wages and mandates; but this explanation would predict a negative association between mandates and employment as well, which is not supported by the evidence discussed below. And, it seems likely that the increase in costs through mandates are valued less than the general cost differences across employers and over time (marginal α is smaller than average α), so that if anything these case studies understate average shifting to wages.

[45] This instrumental variables strategy may not be valid, however, if the employee characteristics that are correlated with fringe costs are also correlated with tastes for work hours; for example, older or higher wage workers may prefer more generous fringes and shorter work hours. But most stories of this type would suggest downward bias to their estimates, strengthening these results.

for benefits on the use of part-time workers. On the other hand, a recent paper by Buchmueller (forthcoming) finds in a sample of California employers that an increase in the cost of fringes that are provided to full time workers, relative to those provided to part-time workers, increases the share of part-time workers employed. Part of the reason for this change in results may be that the Buchmueller paper takes the additional step of attempting to account for the potential endogeneity of the eligibility determination. Overall, the literature in this area suggests strongly that employers are adjusting to increases in fixed employment costs by increasing hours, with somewhat more mixed evidence that employers are also responding by increasing the share of the workforce that is ineligible for benefits.

Several recent papers investigate more specifically the effect of health insurance costs on hours of work. Gruber (1994a) finds that mandated maternity health insurance led to an increase in hours and a decrease in employment, with total labor input held constant. This is consistent with the argument that the costs of the mandate were shifted to wages on average, but that employers responded along this compositional margin. Cutler and Madrian (1998) estimate time trends in hours of work by insurance status, as health care costs have risen over the 1980s, and find that hours of work have been rising much more rapidly for insured workers than for uninsured workers. They also find that hours rose the most in those industries where health care costs grew the most.[46] Thus, there appears to be strong evidence for a compositional shift towards more hours/worker as health care costs increase.

5.4. Unanswered questions

While this literature has convincingly addressed the effect of insurance on wages, employment, and hours, there are a series of more detailed, yet very important, questions that have been largely ignored. First, what about other margins of response to increases in health care costs? One such margin is reduced insurance coverage. This reduction can occur along the coverage margin, as firms drop insurance altogether, or through changes in the structure of insurance plans, as firms increase employee cost-sharing or drop particularly expensive benefits.

There is a huge literature on the price elasticity of demand of both insurance coverage and total insurance expenditure; see Gruber and Poterba (1996) for a review. Unfortunately, this literature has not produced a consensus on the elasticity of demand for insurance at the firm level, with recent estimates ranging from -0.16 in Thorpe et al. (1992) to greater than -2 in Woodbury and Hamermesh (1992). Gruber (1994b) addresses this question in a particular context, by studying the effect of state-level laws in the US that mandate employers who offer insurance to include certain benefits in their health insurance plans, such as coverage for alcoholism treatment or chiropractic visits. It has been

[46] Their results for overall time trends appear to be driven by increases in overtime, which is consistent with the earlier literature on fringes and overtime. Their results are also reduced by about one-half when pensions are controlled for.

claimed that such "state mandated benefits", by forcing employers who would otherwise offer "barebones" insurance coverage to offer "Cadillac" coverage, have led these employers to drop their insurance altogether; obviously, this would only be a problem if employees did not value the expanded benefits. Gruber studies the effect of the five highest cost state mandates on employer provision of health insurance, and finds that there was no significant effect of mandates on employer insurance coverage. This is consistent either with full employee valuation, or with a low elasticity of demand at the firm level. Gruber offers evidence to support the former view; even in the absence of state mandates, most firms voluntarily offer these mandated benefits.

Another important question is how finely the costs of health insurance can be shifted to wages. Gruber's (1994a) and Sheiner's (1995a) results confirm that group-specific shifting is possible, but do not offer much insight into how finely that shifting can occur. In particular, can firms go beyond broad demographic categorizations and actually reduce the wages of individuals workers who are particularly costly? If not, is there hiring discrimination against particularly costly workers?

There is also considerable uncertainty about the mechanisms of shifting to wages. How quickly does shifting to wages occur? Much of the debate over health care reform surrounded the immediate job impact of the Clinton mandate, not the five to ten year impacts, but no work in this literature separates the long-run and short-run effects. Is there actual scope for nominal wage cuts when benefits rise, or does it occur only through the erosion of real wages (due to money illusion on the part of workers)? If it is the latter, then the underlying macroeconomic environment could have important implications for the efficiency of government intervention; mandates in inflationary periods may have smaller efficiency costs than mandates in non-inflationary periods.

An additional question of importance is the underlying structural mechanism behind a finding of full shifting to wages. In the simple labor market framework above, there are two reasons why increased costs might be shifted to wages: because individuals value the benefits that they are getting fully; or because labor supply is perfectly inelastic.[47] Disentangling these alternatives is very important for future policy analysis. Consider the example of national health insurance, which is financed by a mandate, with an additional payroll tax to cover non-workers. If the full shifting documented earlier is due to full employee valuation with somewhat elastic labor supply, then national health insurance will have important disemployment effects, since labor supply will not increase in response to a benefit that is not restricted to workers. If full shifting is due to inelastic supply, however, then the population which is receiving benefits is irrelevant; in any case the costs will be passed onto workers' wages, so national health insurance will not cause disemployment.

What is needed to convincingly disentangle these views is some variation in one or the other of these dimensions only. For example, is the incidence of employer man-

[47] A third alternative for full shifting to wages would be perfectly elastic demand, but this would imply much larger disemployment effects than those found by Gruber and Krueger (1991) or Gruber (1994a).

dates/payroll taxes significantly different across groups with plausibly different elasticities of labor supply, such as married men and married women? Past evidence is mixed here: Gruber (1994a) finds full shifting to married women, who have been estimated to have much more elastic labor supply than men, while Sheiner (1995a) finds less shifting to women than to men. Alternatively, is there differential incidence with respect to elements of a policy which are likely to be valuable, such as cash benefits for work injury, as opposed to elements which are less likely to be valued, such as insurance administrative loading factors?

Two recent studies of actual policy changes highlight the limitations of the literature reviewed here. Gruber and Hanratty (1995) study the implementation of National Health Insurance (NHI) in Canada in the late 1960s. NHI provided coverage to the entire population, financed through both income and payroll taxation. In addition, NHI was phased in over time across the Canadian provinces, allowing the authors to assess the effect on the labor market in a difference-in-difference framework, comparing outcomes in provinces that converted to NHI to outcomes in provinces that did not. In fact, they find that the implementation of NHI *raised* employment and wages. Similarly, Thurston (1997) examines the impact of an employer mandate on wages in Hawaii, and he finds that the most affected industries actually had *faster* wage growth than their counterparts in the continental US, although slower wage growth than less affected industries within Hawaii. One possible explanation for these findings is that there were unobserved labor demand shocks which offset the effects of these policy interventions. This is certainly supported by Thurston's within-Hawaii estimates, but given the consistent effects across Canadian provinces that implemented NHI at different times it is hard to see how it could be driving the Gruber and Hanratty (1995) results. An alternative explanation is that the benefits of dramatic increases in health insurance availability for the functioning of the labor market (i.e. through reducing job lock, since insurance was employment-based in Canada before NHI) outweigh any costs in terms of disemployment.

Finally, this discussion has focused exclusively on efficiency, and ignored the equity implications of interventions such as mandated employer-provided health insurance. If the government is intervening to correct an insurance market failure, and the mandate is simply a means of financing that intervention, then shifting to wages can be viewed as the "price" that is being paid by workers for government provision of insurance. In the case of full valuation, perhaps due to adverse selection in the private insurance market, government mandates will be an efficient and equitable policy; the mandate is a perfect "benefits tax".

If the goal of a mandate is not to correct a market failure, however, but rather to provide benefits to some specific group in society, then full shifting to wages may not be viewed as a desirable outcome. Rather, this may be viewed as the mandate being "undone" by the adjustment of wages. In this case, the additional deadweight loss from broad-based financing which does not have tax/benefit linkages may be a price that society is willing to pay in order to direct more resources towards one group. Thus, it is important to understand the goal of government mandate policy: is it to correct a market failure, or to redirect resources across groups? [48]

6. Conclusions

While still in its infancy, the literature reviewed here has made enormous strides in increasing our understanding of the interaction between health insurance and the labor market. We have some evidence that non-universal employment-based health insurance limits job-job mobility and the ability of secondary earners to leave the labor force, with a stronger consensus that it limits retirement and movements off of public assistance programs. Moreover, increases in health insurance costs appear to be fully reflected in worker wages, with little net effect on labor supply, although with some shift in the composition of hours and employment. These findings have emerged from a variety of studies that have introduced an exciting new set of empirical techniques.

Nevertheless, while much has been learned, there remain important holes in this literature that need to be filled. These can be classified into four categories:

Replication. While there does appear to be a broad consensus on the basic effects of health insurance on the labor market, there is still disagreement about a number of particulars and magnitudes. For example, there remains considerable uncertainty about the importance of job lock, and estimates of the effect of health insurance on retirement vary substantially. This disagreement often stems from very different methodological approaches applied to very different data sets. An important priority for future work is to reconcile these differences, using a broader range of approaches simultaneously on a number of different data sources.

Extension. Along some dimensions, this literature has raised more questions than it has answered. In particular, the focus has been on the *effects* of health insurance on the labor market, and not on the *process* by which those effects occur; for example, how do employers shift health care costs to wages? Also, there has been very little exploration of *heterogeneity* of responses; for which groups are the various forms of “lock” described above the most sizeable?

Theory. To some extent, the previous point reflects the atheoretical nature of this literature. While the empirical innovations in this area have been impressive, the theoretical advances have been much more modest. If this literature is to move beyond its infancy to a richer understanding of the process by which health insurance influences the labor

[48] Vergara (1990) shows that, if the social welfare function values poor individuals more highly, it will in general be optimal to have some degree of public provision financed by income taxation instead of having all of the intervention financed by a mandate.

market, a firmer theoretical underpinning will be necessary. Moreover, without an underlying theoretical framework, it is difficult to understand the welfare implications of these findings.

Policy. Finally, a central question for such an empirically-based literature is the policy implications of the findings. Despite the failure of sweeping health care reform, government intervention in the health insurance market is alive and well. This is witnessed by the recent passage of the Health Insurance Portability and Accountability Act of 1996 (H.R. 3103), which limits insurance companies' ability to discriminate against children and adults with health problems.[49] But there is little work by economists that is devoted to simulating the effects of policies such as this one, building on the empirical results reviewed here. Moreover, there has been little attempt to contrast the costs and benefits of alternative policy approaches, such as insurance market reform versus expanded public health insurance coverage.

This laundry list should not be taken as a criticism of this literature, which has come a long way in a short time. Rather, it is a suggestion that there is still much work to be done in this exciting and extremely policy-relevant area.

References

Acemoglu, D., and J. Pischke (1996), "Why do firms train? Theories and evidence", NBER Working Paper #5605.

Addison, J., C. Barrett and W. Siebert (1995), "Mandated benefits, welfare, and heterogeneous firms", mimeo (University of Hull).

Anderson, K., and R. Burkhauser (1985), "The retirement-health nexus: a new measure of an old puzzle", The Journal of Human Resources 20:315–330.

Anderson, P.M. (1997), "The effect of employer-provided health insurance on job mobility: job-lock or job-push?", unpublished paper (Darmouth University).

Anderson, P.M., and B.D. Meyer (1995), "The incidence of a firm-varying payroll tax: the case of unemployment insurance", NBER Working Paper #5201.

Ballard, C., and J. Goddeeris (1993), "Financing universal health care in the United States: a general equilibrium analysis of efficiency and distributional effects", National Tax Journal 52(1):31–51.

Bartel, A. (1982), "Wages, nonwage job characteristics, and labor mobility", Industrial and Labor Relations Review 35:578–589.

Bartel, A., and G. Borjas (1977), "Middle-age job mobility: its determinants and consequences", in S. Wolfbein, ed., Men in the Pre-Retirement Years (Temple University Press, Philadelphia).

Bazzoli, G. (1985), "The early retirement decision: new empirical evidence on the influence of health", The Journal of Human Resources 20:214–234.

Berkovec, J., and S. Stern (1991), "Job exit behavior of older men", Econometrica 59:189–210.

[49] In particular, group health plans are not allowed to exclude pre-existing conditions for more than 12 months, and this period is reduced by periods of prior, continuous coverage. In addition, insurers must allow individuals the right to convert from group to individual insurance coverage if they lose their group coverage and have exhausted their COBRA entitlement.

Blank, R. (1989), “The effect of medical need and Medicaid on AFDC participation”, The Journal of Human Resources 24:54–87.

Blau, D. (1994), “Labor force dynamics of older men”, Econometrica 67:117–156.

Blau, D.M., and D.B. Gilleskie (1997), “Retiree health insurance and the labor force behavior of older men in the 1990s”, NBER Working Paper No. 5948 (National Bureau of Economic Research, Cambridge, MA).

Bound, J. (1989), “Self-reported vs. objective measures of health in retirement models”, Journal of Human Resources 26(1):106–138.

Bound, J. (1991), “Self-reported vs. objective measures of health in retirement models”, The Journal of Human Resources 26:106–138.

Brown, C. (1980), “Equalizing differences in the labor market”, Quarterly Journal of Economics 94:133–134.

Browning, E. (1994), “The non-tax wedge”, Journal of Public Economics 53:419–433.

Buchmueller, T. (forthcoming), “Fringe benefits and the demand for part-time workers”, Applied Economics.

Buchmueller, T., and R. Valletta (1996), “The effects of employer-provided health insurance on worker mobility”, Industrial and Labor Relations Review 49:439–455.

Buchmueller, T., and R. Valletta (1999), “The effect of health insurance on married female labor supply”, Journal of Human Resources 34:42–70.

Buchmueller, T., and M. Lettau (1997), “Estimating the wage-health insurance tradeoff: more data problems?” mimeo (UC-Irvine).

Burkhauser, R., J. Butler and Y. Kim (1995), “The importance of employer accommodation on the job duration of workers with disabilities: a hazard model approach”, Labour Economics 2:109–130.

Butler, J., R. Burkhauser, J. Mitchell and T. Pincus (1987), “Measurement error in self-reported health variables”, Review of Economics and Statistics 69:644–650.

Card, D. (1992a), “Using regional variation in wages to measure the effects of the federal minimum wage”, Industrial and Labor Relations Review 46:22–37.

Card, D. (1992b), “Do minimum wages reduce employment? A case study of California, 1987–89”, Industrial and Labor Relations Review 46:38–54.

Card, D., and A.B. Krueger (1994), “Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania”, American Economic Review 84:772–793.

Chirikos, T., and G. Nestel (1981), “Impairment and labor market outcomes: a cross-sectional and longitudinal analysis”, in: H.S. Parnes, ed., Work and Retirement: A Longitudinal Study of Men (The MIT Press, Cambridge).

Clark, R., and A.A. McDermed (1986), “Earnings and pension compensation: the effect of eligibility”, Quarterly Journal of Economics 101:341–361.

Congressional Research Service (1988), Costs and Effects of Extending Health Insurance Coverage (US Government Printing Office, Washington, DC).

Cooper, P., and A. Monheit (1993), “Does employment-related health insurance inhibit job mobility?”, Inquiry 30:400–416.

Costa, D. (1996), “Health and labor force participation of older men, 1900–1991”, Journal of Economic History 56(1):62–89.

Currie, J., and J. Gruber (1996a), “Saving babies: the efficacy and cost of recent expansions of Medicaid eligibility for pregnant women”, Journal of Political Economy 104:1263–1296.

Currie, J., and J. Gruber (1996b), “Health insurance eligibility, utilization of medical care, and child health”, Quarterly Journal of Economics 111:431–466.

Currie, J., and B. Madrian (1998), “Health, health insurance and the labor market”, mimeo (UCLA).

Cutler, D. (1994), “Market failure in small group health insurance”, NBER Working Paper #4879.

Cutler, D., and J. Gruber (1996a), “Does public insurance crowd out private insurance?”, Quarterly Journal of Economics 111:391–430.

Cutler, D., and J. Gruber (1996b), “The effect of expanding the Medicaid program on public insurance, private insurance, and redistribution”, American Economic Review 86:368–373.

Cutler, D., and J. Gruber (1997), “Medicaid and private insurance: evidence and implications”, Health Affairs 16:194–200.

Cutler, D., and B. Madrian (1998), "Labor market responses to rising health insurance costs: evidence on hours worked", RAND Journal of Economics 29(3):509–530.

Danzon, P. (1989), "Mandated employment-based health insurance", mimeo (University of Pennsylvania).

Decker, S. (1994), "The effect of Medicaid on participation in the AFDC program: evidence from the initial introduction of Medicaid", mimeo (New York University).

Diamond, P. (1992), "Organizing the health insurance market", Econometrica 60:1233–1254.

Dickens, W., and L. Katz (1987), "Inter-industry wage differences and industry characteristics", in: K. Lang and J.S. Leonard, eds., Unemployment and the Structure of Labor Markets.

Dranove, D., and K. Spier (1996), "Competition among employers offering health insurance", mimeo (Northwestern University).

Dubay, L., and G. Kennedy (1997), "Did the Medicaid expansions for pregnant women crowdout private coverage?", Health Affairs 16:185–193.

Dwyer, D., and O. Mitchell (1996), "Health problems as determinants of retirement: are self-rated measures endogenous?", mimeo (Social Security Administration).

Eberts, R., and J. Stone (1985), "Wages, fringe benefits, and working conditions: an analysis of compensating differentials", Southern Economic Journal 52:274–280.

Ehrenberg, R., P. Rosenberg and J. Li (1988), "Part-time employment in the United States", in: R. Hart, ed., Employment, Unemployment, and Labor Utilization (Unwin Hyman, Boston).

Ehrenberg, R., and P. Schumann (1984), Longer Hours or More Jobs? (ILR Press, Ithaca).

Ehrenberg, R., and R. Smith (1991), Modern Labor Economics: Theory and Public Policy (Harper Collins, New York).

Ellwood, D., and K. Adams (1990), "Medicaid mysteries: transitional benefits, Medicaid coverage, and welfare exits", Health Care Financing Review (1990 Annual Suppl.):119–131.

Employee Benefit Research Institute (1990), Employee Benefit Notes, November (EBRI, Washington, DC).

Employee Benefit Research Institute (1995), EBRI Databook on Employee Benefits (EBRI, Washington).

Employee Benefit Research Institute (2000), Sources of Health Insurance and Characteristics of the Uninsured, Issue Brief #217 (EBRI, Washington).

Feldman, R. (1993), "Who pays for mandated health insurance benefits?", Journal of Health Economics 11:341–348.

Feldstein, M. (1995), "The effect of marginal tax rates on taxable income: a panel study of the 1986 Tax Reform Act", Journal of Political Economy 103:551–572.

Feldstein, M., and D. Feenberg (1996), "The taxation of two-earner families", in: M. Feldstein and J. Poterba, eds., Empirical Foundations of Household Taxation (University of Chicago Press, Chicago) 39–76.

Fishback, P.V., and S.E. Kantor (1995), "Did workers pay for the passage of workers' compensation laws?", Quarterly Journal of Economics 110:713–742.

General Accounting Office (1995), Health Insurance Portability: Reform Could Ensure Continued Coverage for up to 25 Million Americans (GAO, Washington).

Glied, S. (2000), "Managed care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economic (Elsevier, Amsterdam) Chapter 13.

Gruber, J. (1992), "The efficiency of a group-specific mandated benefit: evidence from health insurance benefits for maternity", NBER Working Paper #4157, September.

Gruber, J. (1994a), "The incidence of mandated maternity benefits", American Economic Review 84:622–641.

Gruber, J. (1994b), "State mandated benefits and employer provided insurance", Journal of Public Economics 55:433–464.

Gruber, J. (1994c), "Payroll taxation, employer mandates, and the labor market: theory, evidence, and unanswered questions", forthcoming in Upjohn Foundation Volume on Fringe Benefits in the US and Canada.

Gruber, J. (1996), "Health insurance for poor women and children in the US: lessons from the past decade", forthcoming in: J. Poterba, ed., Tax Policy and the Economy 11 (MIT Press, Cambridge, MA).

Gruber, J. (1997), "The incidence of payroll taxation: evidence from Chile", Journal of Labor Economics, 15(3):S72–S101.

Gruber, J., and M. Hanratty (1995), "The labor market effects of introducing national health insurance: evidence from Canada", Journal of Business and Economics Statistics 13:163–174.

Gruber, J., and H.B. Kruger (1991), "The incidence of mandated employer-provided insurance: evidence from workers compensation", in: D. Bradford, ed., Tax Policy and the Economy, Vol. 5 (MIT Press, Cambridge, MA) 111–143.

Gruber, J., and B. Madrian (1994), "Health insurance and job mobility: the effects of public policy on job-lock", Industrial and Labor Relations Review 48:86–102.

Gruber, J., and B. Madrian (1995), "Health insurance availability and the retirement decision", American Economic Review 85:938–948.

Gruber, J., and B. Madrian (1996), "Health insurance and early retirement: evidence from the availability of continuation coverage", in: D. Wise, ed., Advances in the Economics of Aging (University of Chicago, Chicago).

Gruber, J., and B. Madrian (1997), "Employment separation and health insurance coverage", Journal of Public Economics 66(3):349–382.

Gruber, J., and J. Poterba (1996), "Fundamental tax reform and employer-provided health insurance", in: H.J. Aaron and W.G. Gale, eds., Economic Effects of Fundamental Tax Reform, 125–170.

Gunderson, M., D. Hyatt and J. Pesando (1992), "Wage-pension trade-offs in collective agreements", Industrial and Labor Relations Review 46:146–160.

Gustman, A., and T. Steinmeier (1987), "Pensions, efficiency, wages, and job mobility", NBER Working Paper #2426.

Gustman, A., and T. Steinmeier (1990), "Pension portability and labor mobility: evidence from the survey of income and program participation", NBER Working Paper #3525.

Gustman, A., and T. Steinmeier (1994), "Employer-provided health insurance and retirement behavior", Industrial and Labor Relations Review 48:124–140.

Gyourko, J., and J. Tracy (1989), "The importance of local fiscal conditions in analyzing local labor markets", Journal of Political Economy 97:1208–1231.

Hamermesh, D. (1979), "New estimates of the incidence of payroll tax", Southern Economic Journal 45:1208–1219.

Harhoff, D., and T. Kane (1994), "Financing apprenticeship training: evidence from Germany", mimeo (University of Mannheim).

Hashimoto, M., and J. Zhao (1996), "Non-wage compensations, employment, and hours", mimeo (Ohio State University).

Headen, A.E., R.L. Clark and L.S. Ghert (1995), "Retiree health insurance and the retirement timing of older workers", mimeo (North Carolina State University).

Health Insurance Association of America (1996), Source Book of Health Insurance Data (HIAA, Washington).

Holmlund, B. (1983), "Payroll taxes and wage inflation: the Swedish experience", Scandinavian Journal of Economics 85:1–15.

Holtz-Eakin, D. (1994), "Health insurance provision and labor market efficiency in the United States and Germany", in: R. Blank and R. Freeman, eds., Social Protection Versus Economic Flexibility: Is There a Tradeoff? (University of Chicago Press, Chicago).

Holtz-Eakin, D., J. Penrod and H. Rosen (1996), "Health insurance and the supply of entrepreneurs", Journal of Public Economic 62(1–2):209–235.

Hurd, M., and K. McGarry (1996), "Prospective retirement: effects of job characteristics, pensions, and health insurance", mimeo (University of California at Los Angeles).

Jovanovic, B. (1979), "Job matching and the theory of turnover", Journal of Political Economy 87:972–990.

Jovanovic, B., and R. Moffitt (1990), "An estimate of a sectoral model of labor mobility", Journal of Political Economy 98:827–852.

Karoly, L., and J. Rogowski (1994), "The effect of access to post-retirement health insurance on the decision to retire early", Industrial and Labor Relations Review 48:103–123.

Katz, L.F., and A.B. Krueger (1992), "The effects of the minimum wage on the fast food industry", Industrial and Labor Relations Review 46:6–21.

Kotlikoff, L., and D. Wise (1985), "Labor compensation and the structure of private pension plans: evidence for contractual versus spot labor markets", in: D. Wise, ed., Pensions, Labor, and Individual Choice (University of Chicago Press, Chicago).

Kapur, K. (1998), "The impact of health on job mobility: a measure of job lock", Industrial and Labor Relations Review 51:282–297.

Leibowitz, A. (1983), "Fringe benefits in employee compensation", in: J.E. Triplett, ed., The Measurement of Labor Cost (The University of Chicago Press, Chicago).

Levy, H. (1997), "Who pays for health insurance? Employee contributions to health insurance premiums", mimeo (Princeton University).

Long, S., and M.S. Marquis (1992), "Gaps in employment-based health insurance: lack of supply or lack of demand?", in: Health Benefits and the Workforce (Pension and Welfare Benefits Administration, Department of Labor, Washington DC).

Lumsdaine, R., J. Stock and D. Wise (1994), "Pension plan provisions and retirement: men and women, Medicare, and models", in: D. Wise, ed., Studies in the Economics of Aging (University of Chicago Press, Chicago).

Lumsdaine, R., J. Stock and D. Wise (1996), "Why are retirement rates so high at age 65?", in: D. Wise, ed., Advances in the Economics of Aging (University of Chicago Press, Chicago).

Madrian, B. (1994a), "The effect of health insurance on retirement", Brookings Papers on Economic Activity 1:181–232.

Madrian, B. (1994b), "Employment-based health insurance and job mobility: is there evidence of job lock?", Quarterly Journal of Economics 109:27–54.

Madrian, B.C., and N. Beaulieu (1998), "Does Medicare eligiblity affect retirement?", forthcoming in: D.A. Wise, ed., Inquiries in the Economics of Aging (University of Chicago Press, Chicago).

Madrian, B.C., and L.J. Lefgren (1998), "The effect of health insurance on transitions to self employment", mimeo (University of Chicago).

Miller, R.D. (1995), "Estimating compensating differentials for employer-provided health insurance benefits", mimeo (University of California at Santa Barbara).

Mitchell, O. (1982), "Fringe benefits and labor mobility", The Journal of Human Resources 17:286–298.

Mitchell, O. (1983), "Fringe benefits and the cost of changing jobs", Industrial and Labor Relations Review 37:70–78.

Moffitt, R., and B. Wolfe (1992), "The effect of the Medicaid program on welfare participation and labor supply", The Review of Economics and Statistics, 615–626.

Monheit, A., and P. Cooper (1994), "Health insurance and job mobility: theory and evidence", Industrial and Labor Relations Review 48:68–85.

Monheit, A., M. Hagan, M. Berk and P. Farley (1985), "The employed and uninsured and the role of public policy", Inquiry 22:348–364.

Montgomery, E., and J. Navin (1992), "Cross-state variation in Medicaid program and female labor supply", International Economic Review 33(1):111–128.

Montgomery, M. (1988), "Notes on the determinants of employer demand for part-time workers", The Review of Economics and Statistics, 112–117.

Montgomery, M., and J. Cosgrove (1993), "The effect of employee benefits on the demand for part-time workers", Industrial and Labor Relations Review 47:87–98.

Montgomery, E., K. Shaw and M.E. Benedict (1990), "Pensions and wages: an hedonic price theory approach", NBER Working Paper #3458.

Murray, C. (1984), Losing Ground (Basic Books, New York).

Newhouse, J. (1992), "Medical care costs: how much welfare loss?", Journal of Economic Perspectives 6:3–21.

Olson, C. (1993), "Health insurance and adverse selection in the labor market", mimeo (University of Wisconsin-Madison).

Olson, C. (1994), "Parttime work, health insurance coverage and the wages of married women", mimeo (University of Wisconsin-Madison).

Olson, C. (1997), "Health insurance coverage and weekly hours worked by wives", mimeo (University of Wisconsin-Madison).

Penrod, J. (1994), "Health care costs, health insurance, and job mobility", mimeo (Princeton University).

Rosen, S. (1986), "The theory of equalizing differences", in: Handbook of Labor Economics, Vol. 1, 641–692.

Rust, J., and C. Phelan (1997), "How social security and Medicare affect retirement behavior in a world of incomplete markets", Econometrica 66:781–831.

Ryan, S. (1997), "Employer-provided health insurance and compensating wage differentials: evidence from the survey of income and program participation", mimeo (University of Missouri-Columbia).

Sammartino, F. (1987), "The effect of health on retirement", Social Security Bulletin 50:31–47.

Sheiner, L. (1994), "Health care costs, wages, and aging: assessing the impact of community rating", mimeo (Federal Reserve Board).

Sheiner, L. (1995a), "Health costs, aging, and wages", mimeo (Federal Reserve Board).

Sheiner, L. (1995b), "Mandates with subsidies: efficiency and distributional consequences", mimeo (Federal Reserve Board).

Schone, B., and J. Vistnes (1997), "The relationship between health insurance and labor force decisions: an analysis of married women", Mimeo.

Short, P., J. Cantor and A. Monheit (1988), "The dynamics of Medicaid enrollment", Inquiry 25:504–516.

Slade, E.P. (1997), "The effect of the propensity to change jobs on estimates of 'job-lock' ", mimeo (Johns Hopkins University).

Smith, R. (1979), "Compensating wage differentials and public policy", Industrial and Labor Relations Review, 339–352.

Smith, R., and R. Ehrenberg (1983), "Estimating wage-fringe trade-offs: some data problems", in: J.E. Triplett, ed., The Measurement of Labor Cost (The University of Chicago Press, Chicago).

Steinberg, B.S., and J.S. Vistnes (1997), "The relationship between health insurance and labor force decisions: an analysis of married women", mimeo (Agency of Health Care Policy and Research).

Stern, S. (1989), "Measuring the effect of disability on labor force participation", The Journal of Human Resources 24:361–395,

Summers, L. (1989), "Some simple economics of mandated benefits", American Economic Association Papers and Proceedings 79:177–183.

Thorpe, K., et al. (1992), "Reducing the number of uninsured by subsidizing employment based health insurance: results from a pilot study", Journal of the American Medical Association 267:945–48.

Thurston, N.K. (1997), "Labor market effects of Hawaii's mandatory employer-provided health insurance", Industrial and Labor Relations Review 51:117–135.

Topel, R., and M. Ward (1992), "Job mobility and the careers of young men", Quarterly Journal of Economics 107:439–479.

Trejo, S. (1991), "The effects of overtime pay regulation on worker compensation", American Economic Review 81:719–740.

Triplett, J. (1983), "Introduction: an essay on labor cost", in: J.E. Triplett, ed., The Measurement of Labor Cost (The University of Chicago Press, Chicago).

US Congress, Committee on Ways and Means (1996), 1996 Green Book (Government Printing Office, Washington, DC, US).

US Department of Labor (1992), Health Benefits and The Workforce (Government Printing Office, Washington).

Vergara, R. (1990), "The economics of mandatory benefits programs", mimeo (Harvard University, 1989).

Viscusi, W.K.P. (1992), Fatal Tradeoffs (Oxford University Press, New York).

Wellington, A.J., and D.A. Cobb-Clark (1997), "The labor-supply effects of universal health coverage: what can we learn from individuals with spousal coverage?", mimeo (Australian National University).

Winkler, A. (1991), “The incentive effects of Medicaid on women’s labor supply”, The Journal of Human Resources 26:308–337.

Woodbury, S. (1983), “Substitute between wage and non-wage benefits”, American Economic Review 73:166–187.

Woodbury, S., and D. Hamermesh (1992), “Taxes, fringe benefits, and faculty”, Review of Economics and Statistics 74:287–296.

Yelowitz, A. (1995), “The Medicaid notch, labor supply and welfare participation: evidence from eligibility expansions”, Quarterly Journal of Economics 105:909–940.

Yelowitz, A. (1996a), “Using the Medicare buy-in program to estimate the effect of Medicaid on SSI participation”, mimeo (UCLA).

Yelowitz, A. (1996b), “Did recent Medicaid reforms cause the caseload explosion in the food stamp program”, mimeo (UCLA).

Yelowitz, A. (1998), “Why did the SSI-disabled program grow so much? Disentangling the effect of Medicaid”, Journal of Health Economics 17(3):321–349.

Zimmerman, D., and P. Levine (1993), “The intergenerational correlation in AFDC participation: welfare trap or poverty trap?”, Working Paper #93-07 (Wellesley College Department of Economics).

Chapter 13

MANAGED CARE*

SHERRY GLIED

Mailman School of Public Health, Columbia University

Contents

*I would like to thank Lauren Baker, Martin Gaynor, Joseph Newhouse, Edward Norton, Mark Pauly, Tomas Philipson, Dahlia Remler, and participants at the Chicago Handbook Conference for very helpful suggestions. Lisa Bramham, Boris Peresechensky, and Yael Rockoff provided useful research assistance.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

By 1993, over 70% of all Americans with health insurance were enrolled in some form of managed care plan. The term managed care encompasses a diverse array of institutional arrangements, which combine various sets of mechanisms, that, in turn, have changed over time. The chapter reviews these mechanisms, which, in addition to the methods employed by traditional insurance plans, include the selection and organization of providers, the choice of payment methods (including capitation and salary payment), and the monitoring of service utilization.

Managed care has a long history. For an extended period, this form of organization was discouraged by a hostile regulatory environment. Since the early 1980s, however, managed care has grown dramatically. Neither theoretical nor empirical research has yet provided an explanation for this pattern of growth. The growth of managed care may be due to this organizational form's relative success in responding to underlying market failures in the health care system – asymmetric information about health risks, moral hazard, limited information on quality, and limited industry competitiveness. The chapter next explores managed care's response to each of these problems.

The chapter then turns to empirical research on managed care. Managed care plans appear to attract a population that is somewhat lower cost than that enrolled in conventional insurance. This complicates analysis of the effect of managed care on utilization. Nonetheless, many studies suggest that managed care plans reduce the rate of health care utilization somewhat. Less evidence exists on their effect on overall health care costs and cost growth.

Keywords

adverse selection, any willing provider, capitation, competition, cost growth, cost-sharing, empirical research, fee-for-service, gatekeeper, growth of managed care, history, HMO – health maintenance organization, malpractice, Medicaid, Medicare, monitoring service utilization (see utilization review), moral hazard, PPO – preferred provider organization, preventive services, quality information, RAND health insurance experiment, risk adjustment, search, selective contracting (selection of providers), technological innovation, total cost of health care, UR – utilization review, utilization

JEL classification: I11, L10

1. Introduction

Managed care dominates the United States health insurance marketplace. By 1993, over 70% of all Americans with health insurance were enrolled in some form of managed care plan [Quinn (1998)]. The term managed care encompasses a diverse array of institutional arrangements. There is no single broadly accepted definition of the term nor do any existing definitions persuasively distinguish managed care from other types of health insurance. Many definitions of managed care focus on the nature of the contract, arguing, in effect, that managed care arrangements are more complete contingent claims contracts than traditional health insurance contracts. For example, managed care organizations may intervene in the relationship between the provider and the insured individual, limiting service use in particular circumstances, or they may selectively contract with a defined set of providers, limiting choice of provider. This broad definition of managed care includes arrangements in which insurance and service delivery are fully integrated, such as staff and group model health maintenance organizations (HMOs); arrangements in which insured people are restricted to a defined set of providers, such as independent practice associations (IPAs); and arrangements in which the choice of providers is unrestricted but insurers provide incentives to use selected providers and monitor the care provided, such as preferred provider organizations (PPOs) that conduct utilization review of costly services (UR).

Managed care is often viewed as a particularly American phenomenon associated with voluntary insurance purchase in a private market. The public sector in the United States, however, has also made increasing use of managed care. Furthermore, many systems with compulsory national insurance have always used or have begun to adopt the same mechanisms used by American managed care plans. Since 1980, several countries, including Great Britain, the Netherlands, Germany, and Israel, have formally incorporated elements of managed care into their national health systems and other countries, such as France, are contemplating such changes [Brown (1998)]. In this discussion, I focus on the US experience with managed care plans, but much of the analysis is equally relevant when the same mechanisms are used in other contexts.

Most of the health economics literature on managed care is an empirical literature. This literature seeks to answer the question: How do managed care arrangements perform relative to other types of insurance arrangements? Economic theory offers an equivocal answer to this question. As discussed below, managed care arrangements are one set of responses to the range of informational asymmetries and other market failures that characterize health care delivery. Other institutional arrangements address the same problems in other ways. There is no theoretical reason to expect managed care arrangements always to perform better or worse across dimensions of performance than should other arrangements [Ramsey and Pauly (1997)]. This theoretical indeterminacy is consistent with both the highly varied nature of managed care in practice and the rather mixed results of the extensive empirical literature along most (though not all) dimensions of performance of managed care plans relative to conventional insurance arrangements (discussed below).

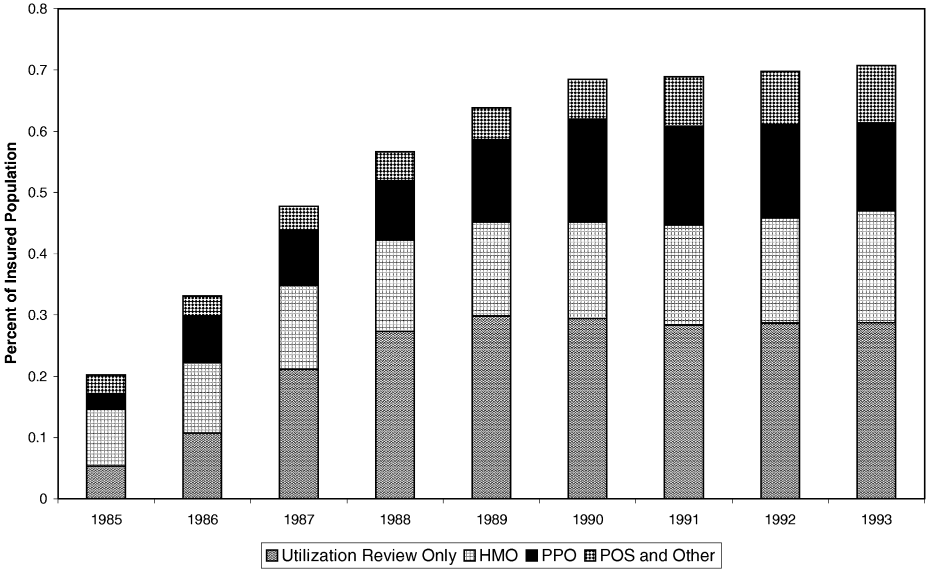


Figure 1. Growth of managed care 1985–1993. Source: Quinn (1998).

One of the most striking features about managed care – and one that is hardly addressed in the existing economic literature – has been its remarkably rapid growth as a share of the health care marketplace. Beginning in the mid-1980s, enrollment in managed care plans in the US grew very rapidly, more than 10% per year [American Association of Health Plans (1998)]. By the end of 1995, over 91 million privately insured Americans were enrolled in HMO, PPO and hybrid managed care plans and almost all conventional insurers incorporated some managed care practices into their plans [Managed Care (1997)] (see Figure 1). An increasing proportion of the publicly insured population is also enrolled in managed care. By 1996, 12% of Medicare beneficiaries and 39% of Medicaid beneficiaries belonged to managed care plans [Physician Payment Review Commission (1997)]. The theory of managed care should provide some answers to the question of why managed care has grown so quickly. If managed care is understood as a response to particular problems of market failure, the growth of managed care should be understood as a response to exacerbations of these particular problems [see, for example, Baumgardner (1991)]. In the discussion of the theoretical literature on managed care below, I assess the potential strengths of existing theory in explaining the growth of managed care.

While market failures are undoubtedly important, the development, early stagnation, and later growth of managed care, in the United States and elsewhere, are not only a product of economic efficiency but also a consequence of the regulatory and institutional environment. In the past, the regulatory and institutional environment has at times discouraged the growth of managed care (for example, through anti-selective contracting legislation) and encouraged the growth of managed care (for example, through passage of the 1973 HMO Act). Furthermore, the future of managed care will depend substantially on the regulatory environment in which it must operate. Both the theoretical and empirical literature on managed care can only be understood within this historical context. I begin this chapter by defining managed care. Section 3 describes the origins of managed care and the regulatory and institutional environment in which it came to exist. Section 4 is a discussion of how managed care addresses imperfections in health care markets. Section 5 presents empirical evidence on the effects of managed care. Section 6 describes some economic problems created by the rise of managed care. Section 7 concludes.[1]

2. What is managed care?

As the broad definition above suggests, the nature of managed care plans varies tremendously across plans and the degree of variation has been increasing over time [Feldman, Kralewski and Dowd (1989)]. As one writer puts it: "If you've seen one managed care plan, you've seen one managed care plan." This tremendous variation makes it difficult to assess the economics of managed care either theoretically or empirically. It makes more sense to think of managed care plans as combining various sets of mechanisms, although these mechanisms, too, have changed over time. In theory and in practice, different combinations of mechanisms may generate different outcomes and some combinations may work together better than others [Robinson (1993)].

In traditional health insurance, a contract can be defined along three dimensions: a premium, a set of covered benefits (such as inpatient hospitalization), and a set of cost-sharing provisions that apply to these benefits (possibly including an out-of-pocket payment limit and limits on annual or lifetime payments). In addition to these, the mechanisms at the disposal of managed care plans consist of the selection and organization of providers, the methods used for paying providers (in addition to the levels of payment), and the methods used for monitoring service utilization. Several authors have developed taxonomies of these plans that describe how they combine these mechanisms [Robinson (1993), Weiner and deLissovoy (1993), Miller and Luft (1994)]. While these taxonomies are helpful, the observed combinations of these mechanisms are constantly changing. There is, as yet, no single clearly superior combination of mechanisms.

The variation in combinations of mechanisms makes it difficult to characterize managed care. It also makes it difficult to assess the effectiveness of any single mechanism. Few plans incorporate just one managed care mechanism. Furthermore, managed care mechanisms differ in their stringency and design in ways that may be hard for researchers to observe. Two plans may cover similar benefits, but limit access in different

[1] Managed care has been particularly important in mental health care. This literature is described in the Handbook chapter on mental health economics [Frank and McGuire (2000)].

ways. They may incorporate cost-sharing, but at very different rates. They may contract with the same providers and hospitals, but one may pay discounted fee-for-service and the other may use capitation payments. They may use utilization review, but differ in how stringently they review claims.

2.1. *Covered benefits*

Managed care plans' contracts often cover a broader scope of benefits than do indemnity plans (in part, as a consequence of Federal regulations described below). In particular, managed care plans, especially the more integrated forms, offer more generous preventive services than do traditional health insurance plans [Weiner and deLissovoy (1993)]. Prior to the passage of the Pregnancy Discrimination Act of 1979 managed care plans also offered better coverage for maternity care. This better coverage for preventive and maternity services is sometimes explained as a natural outgrowth of the fact that managed care plans take on a larger share of the financial risk of health care than do indemnity plans [Pauly (1970)]. If plans prevent disease, proponents argue, overall health care costs to the plan will be reduced [Duston (1978)]. While this argument is appealing in principle, relatively few preventive health services are medical care cost saving [Russell (1986)]. The investment value of preventive services from the perspective of the managed care plan is even more limited because members can, and frequently do, change plans before the payoffs would become evident [Doherty (1979)].

Others have argued that providing better coverage for preventive (and maternity) services helps plans (managed care or traditional) attract a healthier than average population [Frank, McGuire and Glazer (1998)]. If the correlation between the demand for these services and total health care expenditures is negative, then the plan may benefit from expanding coverage.

In other areas, the scope of benefits formally covered by managed care plans is also more generous than under indemnity plans. For example, managed care plans are less likely to incorporate lifetime coverage limits [Jensen et al. (1997)]. This difference in formal definition, however, may be less meaningful than it appears. In indemnity plans, the scope of services is normally defined by service type (e.g., all inpatient hospitalization costs, a specific number of psychiatrist visits). This type of specification describes both the upper and lower bounds of coverage when patients themselves choose services. If providers or plans decide whether or not to authorize an admission or service, formal terms of this type may define only the upper bound of services available under the contract [Glied (1998)].

2.2. *Consumer cost-sharing*

Managed care plans generally rely less on cost-sharing than do conventional indemnity plans. They use cost-sharing in two ways. First, like indemnity insurers, they use cost-sharing to control the use of services within their restricted networks of providers.

Historically, group and staff model HMOs eschewed such consumer cost-sharing altogether. Empirical evidence, however, suggests that, as with conventional insurers, cost-sharing can reduce the use of services in managed care plans [see, for example, Cherkin, Gothaus and Wagner (1989)]. Nominal cost-sharing requirements in managed care plans quadrupled between 1987–1993 [Gabel (1997)]. Today, most plans, even group and staff model plans, have adopted small copayments for routine, non-preventive physician visits.

The second way that cost-sharing is used by managed care plans is as a financial incentive to encourage members to use services provided by the plan's own network of providers. Preferred provider organizations, and looser HMOs (such as point-of-service plans), offer members the choice of network services with low co-pays or out-of-network services with high co-pays.

2.3. Provider selection and organization

The relatively low use of cost-sharing in managed care plans means that plans (or the providers with whom they contract) bear a higher share of the financial risks of medical care use. This risk, borne across all types of medical services, gives the plans (or providers) an incentive to encourage the optimal use of a range of services and to substitute less costly for more costly services (as well as to select healthier patients). One way that the plans can do this is through the selection and organization of participating providers.

Managed care plans may require or encourage patients to use selected providers. Several of the earliest managed care plans were almost fully vertically integrated organizations, in which a limited number of hospitals and physicians were employees of organizations that took on insurance risk. These plans are often referred to as "staff model" HMOs. Closely related to these plans are those (often referred to as "group model" HMOs) in which a fixed group of physicians (and sometimes hospitals) contracts exclusively with an organization that takes on insurance risk.

Much of the early literature on HMOs illustrated the advantages of these vertically integrated delivery systems [Luft (1981)]. Nonetheless, these forms have shrunk in importance, suggesting that the advantages of formal vertical integration have declined over time (or that consumer preferences for choice have increased). Staff and group model HMOs dominated the managed care marketplace through 1983, but their market share has since declined considerably [Feldman, Kralewski and Dowd (1989)]. In 1995, only 25% of those enrolled in HMOs reported that they belonged to a group or staff model HMO [Managed Care (1997)]. New forms of vertical integration, such as hospital- or physician-sponsored networks and plans have begun to develop, but the economic literature has not yet evaluated the efficiency of these organizations or the extent to which these forms of vertical integration behave differently from traditional staff and group model HMOs.

An alternative form of organization is through contractual arrangements with independent providers. Several early HMOs, known as "independent practice associations" operated through non-exclusive contracts with providers who also treated indemnity patients. These IPAs now dominate the HMO segment of the managed care market. Many other managed care forms also use non-exclusive contracts with providers, but do not share all the features of IPA HMOs. The largest of these forms are the preferred provider organizations (PPOs), which negotiate discounted rates with a defined panel of providers. In addition to selecting providers, plans may also restrict access to pharmaceuticals through the use of formularies. Under formulary arrangements, insurers cover the cost of pharmaceuticals only if they are selected from among those on a predetermined (usually discounted) list.

Managed care plans can select the physicians, non-physician providers, and hospitals with whom they contract. Manipulating the composition of provider panels to reduce costs and improve quality could be a valuable tool for managed care plans, but there is very little evidence that they do so systematically. A limited body of research has examined the characteristics of managed care providers. Physicians participating in managed care plans are more likely to be board-certified than average [Brown (1983)]. Early studies suggested that their specialty composition resembled that of the US physician population [Luft (1981)]. Some subsequent studies found that managed care plans were likely to employ fewer physicians per patient and a lower proportion of specialist physicians than the US average [Weiner (1994)]. More recent evidence suggests that as the populations in plans more closely resemble the US population, the physician composition also more closely resembles US averages (although the US average is itself affected by the spread of managed care) [Hart et al. (1997)]. Group and staff model HMOs employ more non-physician providers than the US average [Hart et al. (1997)]. Some evidence suggests that managed care plans choose providers with low-cost practice styles [Robinson (1993)]. Some studies find that managed care plans contract with higher volume hospitals than do other plans [Chernew, Hayward and Scanlon (1996)]; other studies find the opposite [Escarce, Shea and Chen (1997)].

2.4. *Paying providers*

Managed care plans use a wide range of methods to pay physicians and a somewhat narrower range (similar to those used by traditional plans) for paying hospitals. The three basic methods of physician payment are salaries, fee-for-service, and capitation. Plans may also combine these mechanisms, as well as bonuses, withholds, and other incentives, into tailored incentive schemes. Each mechanism generates a set of incentives and a distribution of financial risk [Gaynor (1994)]. Under pure salary payment, physicians have no incentive to see more patients or to provide more services of any particular type. Under fee-for-service payment, providers collect more revenue the more services they provide and, if fees exceed costs, earn more as they provide more services [Pauly (1970)]. Particularly in combination with limited consumer cost-sharing, fee-for-service payment (at fees that exceed costs) can generate excessive service utilization. Nonetheless, many managed care plans continue to pay physicians on a (discounted) fee-for-service basis [Gold et al. (1995)].

Under capitation payment, providers receive a fixed periodic payment for each patient they enroll and can earn more by enrolling more patients (if the capitation fee exceeds expected costs). Capitation makes providers face the full financial cost of their patients' service use, which gives them an incentive to reduce utilization [Pauly (1970), Gaynor and Gertler (1995)]. To the extent that they are also responsible for patients' future service use (which depends on the expected duration of the provider–patient relationship), capitation payment can also encourage the provision of preventive services that reduce the total costs of health care. Capitation arrangements vary according to the scope of services covered within the capitation contract. If the scope of services is very narrow, providers paid a capitation fee have incentives to refer patients to other providers whose services are not included in the capitation fee. Such contracts typically incorporate additional mechanisms to restrict such referrals. Under broad capitation arrangements, providers may also be financially responsible for the costs of services obtained through referral or hospitalization.

Capitation arrangements require providers to share in the financial risk of illness. Thus, they can be thought of as a form of supply-side cost-sharing [Ellis and McGuire (1993)]. Supply-side cost-sharing has several advantages over demand-side cost-sharing as a means of using financial risk to control the use of services. Providers, especially if they form groups, are better able to bear financial risk than are consumers (though risk averse providers also experience disutility from risk bearing). Furthermore, providers generally have more information about risks and benefits than do consumers and are better able to make efficient tradeoffs [Ellis and McGuire (1993)]. Nonetheless, capitation, like other forms of supply-side cost-sharing, poses two serious problems. First, if patients are ill-informed, capitation can lead to underprovision of necessary services [Blomqvist (1991)]. Capitation also gives providers strong incentives to avoid costly cases [Newhouse (1996), Ellis and McGuire (1993), Selden (1990)].

The methods used for paying physicians vary widely, and depend, to some extent, on the extent of vertical integration within the plans. In fully vertically integrated plans, physicians are often paid using salaries [Gold et al. (1995) report that 28% of group and staff model plans pay primary care physicians salaries without further financial incentives]. Where groups of physicians contract with managed care providers, the group may be paid on a capitation basis, per member enrolled with the group. Within these groups, individual physicians may be paid using capitation or salaries. This three-tier system makes it particularly difficult to assess the incentives facing a particular provider [Hillman, Welch and Pauly (1992)]. When individual physicians contract with managed care plans, they may be paid using capitation, discounted fee-for-service, or on an incentive basis. In less integrated arrangements, such as PPOs, discounted fee-for-service is the usual (though not exclusive) payment mechanism. These arrangements can be combined with bonuses, withholds, and other incentive arrangements [see Gold et al. (1995) for examples].

There is very little empirical evidence on the behavior of physicians paid using different payment arrangements. In a study of partnerships, Gaynor and Gertler (1995) find that systems that reward physicians for effort (such as fee-for-service payment)

Table 1
Physician payment arrangements in 1995

	All physicians	Generalists	Medical specialists	Surgeons
Mean % of patients for whom capitation is paid to physician practice	13	18	10	10
Mean % of patients for whom capitation is paid to physician	8	9	5	7
Physicians paid salary	34	43	36	22

Source: Remler et al. (1997).

induce substantially more effort than salary or capitation mechanisms. Hillman, Pauly, and Kerstein (1989) find mixed evidence on the effect of financial incentives. Physicians paid capitation or salary used hospitalization less frequently than did those paid fee-for-service, but other measures were inconsistent with theory. Stearns, Wolfe, and Kindig (1992) find evidence that the same physicians, when paid on a capitation rather than a fee-for-service basis, used significantly fewer hospital admissions in treating patients.

Plans can also combine these payment mechanisms. For example, plans may pay fee-for-service rates but withhold a portion of the payment if utilization exceeds a predetermined level [Hillman (1987)]. Table 1 describes the distribution of physician payment arrangements in 1995 [Remler et al. (1997)].

Managed care plans typically rely less on complex financial incentives for hospitals than for physicians [Luft (1981)], and pay hospitals in much the same way that traditional plans do. Many plans pay hospitals on a per-diem basis based on negotiated rates. Some plans pay using prospective payment mechanisms [Zelman (1996)]. Those vertically-integrated managed care plans that own their own hospitals use internal pricing mechanisms to pay them [Newhouse and the Insurance Experiment Group (1993)]. There are no existing surveys of managed care hospital payment arrangements.

2.5. Monitoring service utilization

In addition to altering the financial incentives affecting providers, managed care plans also directly monitor service utilization. They do this by placing limits on which providers an enrollee may see and by placing limits on what those providers can do. Plans with strong ownership and contractual ties over providers focus on the former type of restriction, while looser plans emphasize the latter. Under capitation or salary payment, physicians may have incentives to underservice patients relative to the health plan's optimum. Plans may also monitor utilization to ensure that it meets minimum quality standards. Finally, plans use a range of management techniques, such as feedback mechanisms and continuous quality improvement programs, that provide information to physicians and assist them in improving quality and reducing costs.

More strongly integrated plans limit enrollee choice by restricting reimbursement to the services of those providers who belong to or contract with the plan. All managed care plans may further restrict choice through the use of “gatekeeper” arrangements. Gatekeeper arrangements require enrollees to obtain a referral from a specified primary care physician before consulting a specialist. In some specialized health plans, such as managed mental health plans, the referral source may be a specialized referral screener, rather than a primary care doctor. Gatekeeper arrangements permit plans to hold primary care physicians financially responsible for the magnitude of referrals, and so strengthen the power of existing financial incentives. Furthermore, to the extent that specialist treatment is more costly than generalist treatment, gatekeepers may reduce total treatment costs, even if they face no financial incentives to limit referrals.

In addition to limiting enrollee choice of provider, most managed care plans also monitor utilization directly. Utilization review is particularly common for high cost services, such as hospitalizations and surgical procedures. About 80% of insurers in 1990 required that enrollees (or their physicians) obtain pre-admission insurer authorization for hospitalization [Sullivan and Rice (1991)]. Many plans also directly limit the number of days that patients spend in hospital. More recently (and particularly for mental health services), plans have begun applying guidelines for the outpatient treatment of particular conditions. In plans with contractual relations with providers, financial incentives may be tied to compliance with these guidelines. Some plans also require patients who seek surgery to obtain a second opinion.

Early studies of utilization review suggested that it had little effect on utilization [IOM (1976)]. Some more recent research suggests that utilization review can reduce hospital expenses by about 7–10% [Wheeler and Wickizer (1990), IOM (1989), Wickizer (1992), Wickizer, Wheeler, and Feldstein (1989), Khandker and Manning (1992)]. Again, however, the results are not unequivocal [Ermann (1988)]. One controlled trial of the use of utilization review in a fee-for-service context found that it had no effect whatsoever on utilization [Rosenberg et al. (1995)]. Even in studies where utilization review is shown to reduce utilization, the source of this reduction in expenses differs across studies. Some studies find that utilization review reduces admissions [Wickizer (1992), Feldstein, Wickizer and Wheeler (1988), Wheeler and Wickizer (1990)]. Other studies find that the effects occur mainly through reductions in length of stay [Khandker and Manning (1992)].

A similar lack of concrete evidence characterizes the literature on second surgical opinion programs. The empirical effectiveness of these programs is unknown [Lindsey and Newhouse (1990)]. Furthermore, as Newhouse and Lindsey (1988) point out, if those who provide second opinions are as likely to make mistakes as the initial physician, these programs may actually worsen outcomes.

3. History of managed care

Managed care has a long history. Arrangements where individuals (often employers) contract with a number of physicians to provide services for a preset fee to a defined population have been noted since 1849 [Friedman (1996)]. Large prepaid group practices, such as the Kaiser health plan, date back to the 1930s [Starr (1981)]. Nonetheless, these plans did not grow quickly until quite recently.

Many physicians and physician associations disapproved of these "contract medicine" plans, and beginning in the 1920s, they pursued both informal and regulatory efforts to ban the practice of contract medicine. For example, in some states physicians who participated in prepaid plans were excluded from medical associations and were denied hospital admitting privileges [Friedman (1996)]. Over 1/2 the states at some point banned consumer-controlled medical plans and 17 required free choice of physician, effectively eliminating most forms of managed care [IOM (1993)]. Indeed, efforts to thwart the growth of prepaid, consumer-controlled group practice plans even led to the formation of other types of managed care plans. These "foundation plans" consisted of physicians in private, independent practice, and were the precursors of today's highly successful independent practice associations [IOM (1993), Starr (1981)]. Together, these efforts to limit the growth of prepaid practice were largely successful, preventing the establishment of more than a handful of prepaid practices (fewer than 40) through the 1960s [Gruber, Shadle and Polich (1988), IOM (1993)]. In the 1950s and 1960s, court and legislative decisions gradually relaxed these restrictions on physician practice, and most studies find no evidence that remaining state legislation limited HMO formation subsequently [Goldberg and Greenberg (1981), Morrisey and Ashby (1982), but see Welch (1985) for some contrary evidence]. Nonetheless, between 1930 and 1970, enrollment in these plans in the United States remained small as a proportion of the insured population. As late as 1980, just 5% of Americans were enrolled in managed care plans [Weiner and deLissovoy (1993)].

Medical reformers as early as the 1930s had pointed to prepaid practices as an ideal model of medical practice [IOM (1993)]. After the passage of Medicare and Medicaid in 1965, as the Federal government became more directly affected by the rising cost of health care, political interest in this model grew. In 1973, the Federal government passed the HMO Act. The Act signaled a substantial change in the regulatory environment. Rather than discouraging (or tolerating) managed care, the Act provided start-up funds to encourage the development of HMOs, overrode State anti-managed care laws, and required large firms to offer an HMO choice to their employees [Brown (1983)]. At the same time, it placed restrictions on the HMOs that were permitted to use these new funds and privileges (these were relaxed somewhat by amendments in 1976). Qualified HMOs were required to offer open enrollment, community rating of health insurance premiums, and comprehensive benefit packages [Brown (1983)]. The HMO Act was somewhat successful in encouraging the growth of HMOs. Between 1970 and 1975, the number of HMOs increased from 37 to 183 and HMO membership doubled [Gruber, Shadle and Polich (1988)].

Despite these advances, HMO enrollment remained small as a fraction of the insured population. HMO custom, Federal rules, and employer practices contributed to this stagnation. In an effort to gain employer acceptance of its prepaid group practice, the Kaiser health plan had insisted that employers who offered it also offer a conventional insurance alternative [Starr (1981)]. This policy was entrenched in the Federal HMO Act, which required that employers who offered a Federally-qualified HMO plan also offer their employees a conventional insurance alternative [Feldman, Kralewski and Dowd (1989)]. When employers did offer multiple competing plans, they typically contributed a fixed share of the premium (often 100%) to both types of plans, regardless of plan cost [Enthoven (1980)]. This practice continues today, with only 28% of employers contributing an equal dollar amount to all health plans in 1997 [Center for Studying Health System Change (1998)]. This structure meant that HMO plans had a limited incentive to control the cost of care relative to competing indemnity insurers. Since employees bore little of the incremental cost of more expensive health plans, they showed little inclination to switch to HMOs. Estimates of the elasticity of employee demand with respect to price were quite low (-0.2 to -0.5) [Cutler and Reber (1998)]. Instead, plans competed principally by offering lower out-of-pocket costs than their indemnity competitors.

Looser selective contracting arrangements between plans and providers, such as PPOs, are a more recent phenomenon than HMOs, emerging in the early 1980s. They too faced legal restrictions. Many states restricted the ability of insurers to selectively contract with physicians and hospitals, and several required all insurers to offer individuals a free choice of qualified providers. In 1980, the regulatory structure in most states effectively prohibited such selective contracting. In 1982, California relaxed selective contracting limits and between 1981–1984, 15 other states passed laws encouraging the growth of PPOs [Gabel et al. (1986)]. Almost immediately, growth in PPO plans escalated rapidly. While data on PPO membership are notoriously unreliable, in 1983, physicians reported that 5% of their patients contacts were governed by a PPO contract; just two years later, they reported that PPO patients accounted for 1/4 of their contacts [Gabel et al. (1986)].

The growth of PPOs also led to changes in the more traditional HMO market. The popularity of PPOs encouraged the growth of independent practice association model HMOs. IPA, group, and staff model plans began to allow "point-of-service" options, which provide partial reimbursement for services that enrollees receive from providers outside the plans.

Through 1990, managed care participation was almost exclusively confined to the private sector. Medicare permitted enrollment in HMOs from its inception, but plans had few incentives to participate [Adamache and Rossiter (1986)]. Reimbursement was cost-based and retrospective and HMOs provided physician (Part B) services only [Gruber, Shadle and Polich (1988)]. In 1983, only 1.5% of Medicare beneficiaries belonged to HMOs [Bonnano and Wetle (1984)]. From 1982 on, changes in Medicare legislation began to authorize prospective contracts with Federally-qualified HMOs. Prospective reimbursement was set based on the age-sex adjusted average per capita cost of Medicare's fee-for-service program in each county, a practice that generated wide variation in Medicare's HMO reimbursement across the country [Physician Payment Review Commission (1997)]. This legislation encouraged some HMOs to join, but requirements remained relatively onerous. Only Federally-qualified plans could participate and hybrid plans, such as point-of-service plans, were generally not permitted. Furthermore, most Medicare beneficiaries held supplementary coverage that effectively eliminated Medicare cost-sharing. As long as costs of supplementary coverage remained relatively low, Medicare beneficiaries given the choice between traditional Medicare with limited cost-sharing and restricted managed care proved understandably reluctant to switch to managed care plans. As late as 1990, only 5.4% of Medicare beneficiaries belonged to HMOs [Physician Payment Review Commission (1997)]. As premiums for supplementary insurance increased, however, managed care became a more attractive option for Medicare beneficiaries. By 1996, one in eight Medicare beneficiaries belonged to a managed care plan [Physician Payment Review Commission (1997)]. Under the Balanced Budget Act of 1997, forms of managed care other than traditional HMOs (such as some point-of-service plans and provider-sponsored plans) are permitted to participate in Medicare.

Under Medicaid, a joint state-federal program, states have always been permitted to contract with managed care plans who could provide services to those who voluntarily enrolled [Brown (1983)]. Through the early 1980s, only a few states pursued such contacts (16 had contracts in 1980), and several of the early efforts were poorly managed [Brown (1983)]. These voluntary plans attracted very few beneficiaries (only 1.3% of all beneficiaries in 1980) both because of difficulties in administering the plans and because Medicaid fee-for-service beneficiaries already received comprehensive services and had little cost-sharing [Brown (1983), Luft (1981)].[2] Legislation in 1981 created the possibility of waivers for mandatory HMO enrollment [Gruber, Shadle and Polich (1988)]. In 1982, Arizona entered the Medicaid program with an all-HMO plan and enrollment in managed care elsewhere grew somewhat during the 1980s. By 1991, nearly 10% of Medicaid beneficiaries were enrolled in managed care plans. Since then, states have been increasingly turning to managed care. By 1996, all states except Utah and Alaska used managed care as a component of their Medicaid programs, and nearly 40% of Medicaid beneficiaries were enrolled in managed care [Physician Payment Review Commission (1997), Holahan et al. (1998)]. The 1997 Balanced Budget Act eliminated the requirement that states seek a Federal waiver to begin mandatory Medicaid managed care programs. While HMOs dominate the Medicaid managed care business, other forms of managed care are also in use. For example, California implemented a system of selective contracting for its Medicaid fee-for-service program in 1982.

Efforts to manage care within traditional health insurance directly were encouraged from the 1950s on [IOM (1993)]. By the early 1960s, many Blue Cross plans reviewed hospital claims [IOM (1993)]. The initial Medicare legislation incorporated a requirement of hospital utilization review. These requirements have been amended several times, but continue in the form of peer review organizations, which examine both quality and hospital costs [Ermann (1988)]. Second surgical opinion programs were attempted in the mid-1950s but were not successfully implemented until the mid-1970s.

[2] Low payment levels, however, may have made it difficult for fee-for-service Medicaid beneficiaries to gain access to services.

By 1984, 76% of conventional insurers had implemented second surgical opinion programs.

Today, managed care is well established in the US health care market, yet the legal requirements that limited the initial growth of these contracts have by no means disappeared. A managed care backlash has led to the passage of new requirements that may (or may not) have desirable effects on the quality of care, but are also likely to inhibit the formation or operation of these arrangements. In 1995, 27 states required state-regulated insurers to permit "any willing provider" to participate in a health plan, although often these requirements only apply to pharmacists [Zelman (1996)]. Some states require managed care plans to permit those holding coverage a free choice of provider or mandate that plans must offer a point-of-service option, sometimes at a defined premium level [Marsteller et al. (1997), Hellinger (1996)]. Overall, in 1996, nearly $1/3$ of the states had strong or medium-strong restrictions on the operations of state-regulated managed care plans [Marsteller et al. (1997)]. At this writing, the Federal government is considering similar legislation that would apply to coverage exempt from state-regulation.[3]

4. Managed care and market failure

Through the use of the mechanisms described above, managed care organizations can respond differently than did traditional health insurers to the underlying characteristics of the health care system. This section considers four well-known features of the health care system and describes how managed care plans respond to them: asymmetric information about health risks (leading to adverse selection), moral hazard, information about health care quality, and industry competitiveness. The growth of managed care may be due to this organizational form's relative success in responding to these underlying features of the health care system. If so, recent changes either in underlying economic problems or in the technology available to address them, should favor managed care. In each case, I assess this possibility and discuss its implications.

4.1. Asymmetric information about health risks

A fundamental problem in the health care market is that individuals have more information about their propensity to use services than do insurers [Arrow (1963)]. This informational asymmetry can lead to adverse selection, and adverse selection can lead to segmentation of the health insurance market. Managed care may be a response to these informational asymmetries and managed care plans may have an advantage over traditional insurers in segmenting the market according to risk (and utilization preferences). Managed care changes the way health care services are rationed. Since people

[3] Self-insured health plans are exempt from state regulation under the Federal ERISA statute.

are heterogeneous (both in their preferences and in their health-related characteristics), these changes are more desirable to some consumers than to others. Patients with longstanding ties to providers do not want to switch doctors, while those who are newly arrived in communities may prefer to choose from a pre-selected list of physicians. Patients who expect to use routine preventive care may prefer organizations that cover such care and do not require consumer cost-sharing while those who require specialty care may prefer organizations that do not require gatekeeper authorization for such care. These differences imply that the populations enrolled in managed care organizations will differ from those enrolled in traditional health insurance plans.

By designing packages that appeal to some consumers and not others, managed care organizations can make consumers reveal information about their expected use of health services and encourage consumers with lower expected use to choose different plans than consumers with higher expected use. Differences in cost-sharing rules under indemnity insurance can have the same effect, but the multiplicity of managed care mechanisms may lead to more market segmentation than under indemnity insurance. Managed care plans can use both explicit prices (consumer cost-sharing rules) and implicit prices (provider selection and incentives) to set different shadow prices for different services [Frank, Glazer and McGuire (1998)].

Segmentation of the health care market through adverse selection means that consumers with high expected use pay high prices while consumers with low expected use pay less. The normative consequences of this risk segmentation are controversial. By generating separating equilibria, risk segmentation may preserve otherwise unstable insurance markets and increase coverage among healthy populations [Pauly (1985)]. At the same time, risk segmentation limits the amount of risk spreading that goes on in health insurance markets. Since risk averse people want insurance against the possibility that they will develop an adverse health condition, segmentation of this type can lead to inefficiency [Cochrane (1995)]. In practice, risk segmentation can also generate welfare losses by leading generous plans that are preferred by some segment of the population to leave the market [Cutler and Reber (1998)].

Managed care plans may (or may not) attract lower utilizers than traditional insurance plans at a point in time (an empirical question addressed in Section 5 below), but can the superior ability of managed care plans to sort people according to expected utilization explain the growth of managed care? If consumers have more private information about health risks than they did previously, managed care plans' advantage in segmenting risk may have become more valuable. In practice, it is unclear that there have been such improvements in private information, so there is little reason to expect the advantage of managed care plans in risk segmentation to have become more important over time. Furthermore, this advantage should have led managed care plans to increase overall coverage among low risk populations. To date, there is no evidence suggesting that the growth of managed care has increased total health insurance coverage rates among these populations.

To the extent that managed care plans do operate by segmenting the market and selecting good risks, they are likely to drive up the costs of their competitors. Overall health care costs will not fall as a consequence of the introduction of managed care [Luft (1981)].[4] Instead, health care costs will simply be distributed differently.

4.2. Moral hazard

Under moral hazard, people with insurance may use more services than they otherwise would [Arrow (1963)]. They may also use more costly services than do those without insurance. Finally, they may prefer relatively more quality-enhancing but cost-increasing technologies than would those without insurance [Goddeeris (1984)]. This last effect may lead to higher rates of growth of health care costs.

Traditional health insurance responds to moral hazard through demand-side cost-sharing – co-payments and deductibles that require consumers to bear a share of the cost of their health care consumption. By contrast, managed care combines cost-sharing with a range of provider-side mechanisms and direct supply constraints to control moral hazard.[5]

One set of mechanisms consists of supply-side cost-sharing arrangements [Ellis and McGuire (1990), Ellis and McGuire (1993)]. Under these arrangements, which include capitation payment and financial penalties for the use of services, providers bear part of the risk of increased utilization. A second set of arrangements, which include provider guidelines and utilization review procedures, uses administrative regulations, rather than financial incentives, to control use of health care resources. These arrangements correspond closely to the theoretical concept of monitoring utilization to control moral hazard. A final set of rationing arrangements focuses on the choice of provider for a given service. These arrangements, which include gatekeepers, closed panels, and preferred provider organizations, seek to address that aspect of moral hazard associated with the use of more costly care under health insurance.

As the discussion above suggests, consumer-side cost-sharing can perform exactly the same functions as managed care in controlling moral hazard. The results of the RAND health insurance experiment, which, in part, compared a staff model health maintenance organization that used no cost-sharing with a series of indemnity plans that used different rates of cost-sharing, suggest that this particular managed care form led to utilization rates equal to those under a 95% cost-sharing plan with an out-of-pocket cap of $1,000 (in late 1970s dollars).

The optimal choice between these mechanisms depends on the distribution of decision making, on risk bearing abilities, and on administrative costs [Ellis and McGuire (1993)]. No single study examines the efficiency of using these three sets of managed care mechanisms together; but the theoretical literature taken as a whole points to the

[4] If risk segmentation allows previously uninsured healthy people to obtain health insurance, managed care may slightly increase total health care costs. If risk segmentation encourages people with poor health habits to improve their behavior, managed care could decrease total health care costs.

[5] Note that conventional insurers may also use direct supply constraints to limit access to technology [Ramsey and Pauly (1997)].

result that neither consumer cost-sharing, nor producer cost-sharing, nor quantity restrictions alone is likely to be optimal [Blomqvist (1991), Newhouse (1996), Ellis and McGuire (1993), Ramsey and Pauly (1997), Selden (1990)].

Mechanisms that control moral hazard at a point in time can also directly affect the choice of technologies and may change the nature and extent of technological innovation in health care. High cost-sharing provisions in indemnity insurance will encourage patients to choose less costly technologies, just as high supply-side cost-sharing arrangements will encourage providers to recommend less costly technologies. Managed care arrangements that directly control the providers and technologies used by patients can also reduce the use of costly technologies [Baumgardner (1991)], a result that can also be obtained through coverage restrictions in conventional contracts [Ramsey and Pauly (1997)]. Managed care responses to moral hazard, such as supply-side cost-sharing and utilization monitoring, may have become more valuable over time, helping to explain the growth in this organizational form. As health care costs rise, the disutility associated with the financial risks of a given consumer-side cost-sharing rule also increases. The RAND Health Insurance Experiment incorporated an out-of-pocket limit of $1,000 in the late 1970s, roughly equal to mean expenditures in the free care plan and under 5% of median family income in 1980 [Newhouse and the Insurance Experiment Group (1993), Bureau of the Census (1996)]. Given medical care cost inflation since then, a comparable out-of-pocket limit in 1996 would exceed 9% of median family income. To the extent that providers are better able than consumers to pool these risks, we would expect the growth in medical costs to lead to a shift toward provider-side cost-sharing. Similarly, if the costs of administering a utilization monitoring system rise more slowly than consumer financial risks, we would expect to see a shift toward this approach to the management of moral hazard. Consistent with this hypothesis, the strongest effects of managed care occur in circumstances where services are very high cost (e.g., hospital care) and where the price elasticity of demand for services is very high (e.g., mental health care). In both these circumstances, the out-of-pocket expenditures necessary to reduce moral hazard may impose greater financial risk costs on consumers than the costs of directly monitoring services.

In functioning as a control on moral hazard, managed care can reduce the cost of health insurance for its members. Indeed, if managed care leads to a proliferation of less intensive practice styles, or reduces the returns to investments in the development of new technology, it might also reduce the cost of health care provided in the non-managed care sector.

Similarly, the growth of managed care may lead conventional insurers to adjust their cost-sharing or utilization management procedures to keep costs low [Enthoven (1978)]. If health care providers induce demand for their services (or raise prices), however, managed care may lead to an increase in the cost of health care in the non-managed sector. Under managed care, providers no longer have an incentive (under capitation) or opportunity (under gatekeeping and utilization review) to induce demand from their patients. If this reduction in moral hazard also reduces provider incomes, they might respond by increasing demand inducement among their non-managed care patients [Enthoven

(1978), McGuire and Pauly (1991)]. Finally, some argue that the growth of managed care may lead to intensified competition among managed care plans [Enthoven (1978)].

4.3. Information

It is difficult for consumers to assess the quality of the health care that they purchase. Several mechanisms in the health care system serve to improve consumers' knowledge of the quality of health care. Patients may rely on general practitioners whose quality they can judge to recommend specialty care [Pauly (1978)]. Physicians may affiliate with hospitals that promise to screen doctors. Hospitals and physician groups may develop brand names that are associated with quality. Managed care plans, particularly those that use restricted provider panels, may act as effective agents, offering another set of mechanisms for assessing the quality of health care.

In order to operate, managed care plans must have the capacity to collect and transfer administrative data within an internal market. This information collection capacity means that plans can collect information on the processes and outcomes of care offered by many different providers to a defined population of enrollees [Miller and Luft (1994), Luft (1981)]. If firms disseminate this information, consumers can use it to compare performance across competing managed care plans.

The type of information generated under managed care is distinct from the type of information available under traditional health insurance. While information about the quality of services provided by specific physicians and hospitals could be generated under indemnity insurance, the use of restrictive panels and defined populations allows managed care plans to generate information both about the processes of care and about the outcomes experienced by those enrollees who did and did not receive specific services (e.g., population level hospitalization rates). Plans, in turn, can use their control over provider patterns and practice guidelines to improve their performance on these quality measures, although it is not yet clear to what extent they actually do this.

The growth of managed care has coincided with renewed efforts to measure the quality of medical services. In part, this information collection and dissemination responds to direct consumer demands, including requirements of regulatory agencies. In addition to this consumer- and regulator-mandated dissemination, most managed care plans routinely collect some data related to quality, particularly data on consumer satisfaction [McGlynn (1997)]. Quality report cards developed by private groups and public payers, are increasingly used to measure the output of managed care plans. In 1997, about 1/4 of large employers disseminated information about plan quality to their employees [Center for Studying Health System Change (1998)]. There is less evidence that firms or their employees actually make use of this information in making health plan choices [Gabel et al. (1998)].

Managed care plans can also generate information about quality through the development of brand names [Klein and Leffler (1981)]. While health care delivery is inherently local, managed care plans may be able to develop national reputations based on the quality of their provider panels, the nature of their incentive systems, and the types of guidelines and utilization mechanisms they use. The development of brand names in health care is consistent with the growing predominance of national firms in the managed care marketplace [Zelman (1996)].

One element in the rise of managed care may be cost-reducing and quality-improving changes in the technology of administration, such as the development of computer systems, which make it possible to monitor transactions and processes across a range of providers. The advantage of managed care over indemnity insurance in generating information about quality in health care markets depends on the extent to which the information generated through these measures meaningfully describes the quality of health care. This question is the focus of considerable research. The answer will help economists understand whether managed care can offer an increase in the efficiency of the health care market through improvements in consumer information.

4.4. Industry competitiveness

Several features of health care have historically limited the extent of price competition in the industry [Arrow (1963)]. First, the industry has maintained formal barriers to competition. As noted above, for many years, the growth of many forms of managed care was stymied by barriers such as prohibitions on contracting and on prepaid practice. Second, the rules of professional practice have also limited competition. In most states, advertising by professionals, particularly price advertising, was until recently prohibited and professional organizations have combined to limit price competition among their members. While these regulatory barriers to competition have been struck down, incentives for price competition were – and continue to be – muted by the provision of public subsidies (including the tax treatment of employer-sponsored health insurance and public programs), which protect consumers from the full cost of their health insurance and health service decisions. Finally, in some areas of the country, small numbers of providers still share considerable market power. Managed care may provide a means to overcome some of these formal and informal barriers to competition.

In a perfectly competitive marketplace where search is costless, price-sensitive consumers should efficiently seek out low-cost producers. In practice, search is costly, especially where provider advertising is prohibited. Furthermore, under indemnity coverage with limited copayments, individual consumers gain only a small fraction of the total benefits of search for lower prices [Newhouse (1978)]. Finally, providers may collude to keep prices uniformly high, limiting the benefit of search.

Certain managed care techniques, particularly selective contracting, can allow consumers to act in combination and exert countervailing pressure against the price setting power of health care providers [Dranove, Shanley and White (1993)] (although note that managed care may lead to new inefficiencies if managed care firms become monopsonist purchasers). Furthermore, managed care plans that selectively contract with providers and sell services to large numbers of consumers can reduce the cost of search and seek out low cost producers. Since they bear (almost) the full cost of services used by their enrollees, they benefit fully from search. Finally, they gain a further advantage in generating price competition because they can promise producers a large volume of service in exchange for lower prices. This last point means that managed care is most likely to be effective in obtaining discounts from prevailing health care prices when producers have substantial excess capacity [see, for example, Kralewski et al. (1992), Morrisey and Ashby (1982)].

Can the advantages of selective contracting explain the rise in managed care? There is little empirical evidence on this point. Nonetheless, the steep reductions in inpatient occupancy rates in the early 1980s may have generated this type of excess capacity, encouraging the growth of plans that were able to negotiate substantial price discounts.

Even if only a few managed care plans are able to search more effectively in the health care marketplace, they may (under restrictive assumptions) lower costs to themselves and to competing plans [Salop (1976)]. If managed care plans are able to obtain discounts by offering health care providers a steady flow of business, they may lower their own costs without affecting the costs faced by their competitors. If, however, health care providers offset reduced prices paid by managed care providers by raising prices or inducing demand among those with traditional health insurance, total health care costs may be unaffected by the growth of managed care [see Mathewson and Winter (1997) for a theoretical discussion of this point; for some evidence consistent with this hypothesis, see Feldman et al. (1986)].

5. Empirical research on managed care

The theoretical structure above suggests that managed care might be expected to affect the utilization of health care services, the quality of health care services, the total cost of health care, and the rate of growth of health care costs. The magnitude of these effects has been the subject of a considerable body of empirical research.

Empirical research on managed care is complicated by two factors. First, as discussed above, the term managed care incorporates many different combinations of mechanisms. Even plans that apparently share common mechanisms may vary in the specifics of their provider or consumer cost-sharing arrangements or in the stringency of their utilization review procedures. Plans often will not release detailed information about these arrangements to researchers, citing competitive concerns. Conventional insurance plans used as comparisons in these studies also vary in their cost-sharing arrangements. By the mid-1980s, many apparently conventional insurance arrangements incorporated some managed care features, particularly utilization review, so that organizational complexity can obscure both sides of the managed care-conventional insurance comparison.

In addition to their use of these specific mechanisms, plans also vary in their organization in ways that might be expected to affect their performance, although the direction of the effects may be unclear. Some plans are for-profit, others are not-for-profit. Some plans have existed for a long time, others are brand new. Some plans are insurer-based, others are provider-based. This substantial, and often unobservable, heterogeneity means that it is very difficult to generalize from the results of managed care studies.

Second, risk segmentation through managed care substantially complicates the analysis of the effects of managed care. If managed care enrollees differ from enrollees of conventional insurance plans, differences in observed utilization at a point in time, growth in utilization over time, and outcomes may be a consequence of the underlying characteristics of the enrolled population, rather than the management of care itself. Furthermore, if managed care is correlated with overall insurance coverage, even measures of costs that combine information from the conventional insurance and managed care sectors may be misleading [Glied, Sparer and Brown (1995)]. A small study in St. Louis in the early 1970s [Perkoff, Kahn and Haas (1976)] and the RAND Health Insurance Experiment [Manning, Leibowitz, Goldberg, Rogers and Newhouse (1984)] are the only studies in which people were randomly assigned to a managed care plan. A few other studies are able to exploit natural experiments that minimize the effects of self-selection [Buchanan, Leibowitz, Keesey, Mann and Damberg (1992), Cutler and Reber (1998)]. Most studies rely on multivariate controls to attempt to remove the effects of selection on the results.

5.1. Selection

A considerable empirical literature has documented differences between managed care and conventional insurance enrollees [Hellinger (1995), Physician Payment Review Commission (1996)]. This literature is summarized in Table 2. Differences across plans are complex and vary across studies. Some managed care plans attract more young families [Berki et al. (1977)]. Some plans attract fewer chronically ill people [Hill and Brown (1990)]. Managed care plans often attract new migrants and do not attract people with long-standing ties to physicians [Luft (1981)]. Many studies find differences in rates of prior health service utilization. The results, however, are not uniform. Several studies find reverse selection, especially with respect to maternity care [e.g., Hudes et al. (1980), Robinson, Gardner, and Luft (1993)]. Some authors have speculated that managed care plan members might differ from conventional insurance enrollees in terms of their health attitudes and behaviors, but there is little evidence to support this conjecture [Feldman, Finch and Dowd (1989), Lairson and Herd (1987)]. The RAND Health Insurance Experiment found no statistically significant differences in the expenditures of those randomly assigned to an HMO and those who had voluntarily chosen the plan [Manning et al. (1984)].

The results of selection studies depend on how selection is measured. Some studies measure selection according to particular conditions (such as maternity or chronic disease). Since disease-specific patterns of care differ by system of care, it is possible for both types of plans to have unfavorable selection of this type simultaneously [Frank, Glazer and McGuire (1998)]. Access to hospital care is easier under conventional insurance, so those who expect to use high levels of inpatient care may select conventional coverage. Access to general practitioners is easier under managed care, so those who expect to use high levels of outpatient services may select managed care coverage. Families who expect to need maternity care may choose HMOs, while those

Table 2
Selection studies

Study	Finding	Sample	Notes
Berki, Ashcraft, Penchanski and Fortus (1977)	reverse selection	one employer; no premium differences across plans	
Goldman (1995)	reverse selection	military enrollees	
Hudes, Young, Sher and Trinh (1980)	reverse selection due to maternity benefits	Kaiser Southern California	
Robinson, Gardner and Luft (1993)	reverse selection due to maternity benefits	Large employer 1981–1984	
Buchanan, Leibowitz, Keesey, Mann and Damberg (1992)	favorable selection in New York, not in Florida	Medicaid	
Luft (1981)	mixed results	Survey of studies	
Feldman, Finch and Dowd (1989)	no difference in health habits	17 Minneapolis firms	
Gordon and Kaplan (1991)	similar health profiles and rates of screening procedures	California residents who either did or did not belong to Kaiser Permanente	
Lairson and Herd (1987)	no difference in health habits	1 large company	
Manning, Leibowitz, Goldberg, Rogers and Newhouse (1984)	no difference	controlled experiment, private population	no premium
Hosek, Marquis and Wells (1990)	no evidence of selection wrt PPO, slight favorable selection wrt HMO	study of 5 employers	
Robinson and Gardner (1995)	differs by plan, not consistent by type	private population	HMO and FFS weights give different results
Billi, Wise, Sher, Duran-Arenas and Shapiro (1993)	19% difference in prior use favoring PPO (relative to traditional coverage)	private population	
Buchanan and Cretin (1986)	lower prior utilization among families who joined HMOs	large firm	
Cutler and Reber (1998)	selection effect about 20% favoring HMOs	private population	Switchers 20% cheaper. Stayers 11% more costly substantial premium difference

continued on next page

Table 2, *continued*

Study	Finding	Sample	Notes
Eggers and Prihoda (1982)	favorable selection into PGPs (20%); no selection in IPA	Medicare enrollment by 3 HMOs	
Brown (1988)	21% lower prior use among HMO enrollees; 54% higher expenditures for disenrollees	Medicare	
Hill and Brown (1990)	23% lower prior spending among HMO enrollees	Medicare	no controls for supplemental coverage
Jackson-Beeck and Kleinman (1983)	lower prior year hospital use	11 employee groups in Minneapolis	
Luft, Trauner and Maerki (1985)	HMO risk profile 17–25% less expensive than BC/BS	California Public Employee system – state payment based on weighted average premium	
Kasper, Riley, McCombs and Stevenson (1988)	24–42% lower prior spending among HMO enrollees	Medicare	
Strumwasser et al. (1989)	Managed care risk profile 30% lower than conventional	Large Midwest Firm	
Zwanziger and Auerbach (1991)	Managed care risk profile 27% lower than conventional	Large Midwest Firm	
Eggers (1980)	Prior use among HMO enrollees 52–62% lower	Medicare	

with heart disease may choose conventional insurance. Consistent with this possibility, Robinson and Gardner (1995) find that the pattern of selection on health characteristics differs according to whether the costs of these characteristics are assessed based on HMO practice patterns or conventional insurance practice patterns.

Prior utilization measures of selection more accurately capture the effect of sets of health characteristics on costs. These measures, however, may overstate selection (especially if they focus on plan switchers). This will occur if prior utilization includes both transitory and permanent components and there is regression to the mean in overall expenditures [Welch (1985)]. As discussed further below, the growing literature on risk adjustment attempts to provide better estimates of differences in expected health care utilization among populations.

Overall, the results of selection studies suggest that managed care plans in the private sector tend to enjoy a 20–30% prior utilization advantage over conventional indemnity plans while Medicare plans enjoy a similar advantage over traditional Medicare. The degree to which managed care plans attract healthier people will depend, of course, on the generosity of the conventional insurance alternative and the stringency of managed care limitations on use. Selection may be more severe (or less severe) as the price differential faced by consumers increases. In practice, the financial implications to the consumer of choosing managed care rather than an alternative depend on employer practices. Since many employers continue to pay a fixed proportion of costs, the cost advantage to an employee of selecting a managed care plan may be relatively small. While less clear, the selection studies also suggest that differences in health outcomes between managed care and conventional insurance enrollees may also depend on the underlying characteristics of these populations. The wide range of estimates and the complicated nature of selection between managed care and non-managed care suggests caution in interpreting the results of non-randomized studies of managed care utilization and quality [Newhouse (1996)].

5.2. Utilization

Analyses of the effects of managed care on utilization examine its effects on inpatient, outpatient and total utilization. Comprehensive reviews of this literature are provided in Luft (1981), Miller and Luft (1994, 1997). Luft (1981) reviewed studies of managed care utilization conducted between 1959 and 1975. Most of these studies compared people in group or staff model managed care plans with those in conventional insurance arrangements. Since conventional arrangements in this period rarely incorporated utilization review, while managed care plans rarely incorporated cost-sharing, the results are somewhat easier to generalize than those from studies conducted after 1980. The managed care plans in Luft's survey include plans that manage only outpatient care, IPA plans, and group and staff plans. The characteristics of the comparison group of conventional insurance plans are rarely specified in detail.

The study of utilization effects is further complicated by the problem of measuring costs within managed care. Managed care plans often do not collect cost information that is comparable to traditional insurance claims costs. Mechanisms such as capitation and salary payment make it especially difficult to measure costs at the level of the individual visit. Instead, many studies impute costs based on observed patterns of utilization measured at traditional insurance claim rates. To the extent that these rates do not accurately reflect costs within a managed care setting (whether because of production efficiencies or volume discounts), estimates of the cost of service use within managed care may be misleading.

In general, Luft finds that managed care plans reduced inpatient admission rates, had mixed effects on length of inpatient stays, and reduced total inpatient costs. The overall effect on inpatient days was a reduction of 5–25% for IPA plans and 35% for group and staff model plans. Results were generally more robust for group and staff plans. Managed care plans, especially IPAs, tended to have higher outpatient visit rates, especially for patient-initiated visits. Overall costs were 10–40% lower for group and staff model plans, but IPA plans did not appear to be less costly than conventional arrangements.

In 1984, the RAND Health Insurance Experiment Group published the results of its randomized study of the effects of managed care [Manning, Leibowitz, Goldberg, Rogers and Newhouse (1984)]. The study assigned 1149 people to Group Health of Puget Sound, a staff model HMO in Seattle, Washington. It also observed the behavior of 733 people who were already enrolled in the plan. In addition to randomizing enrollees, the RAND experiment was unusual in capturing the characteristics of both the managed care plan (which used no consumer cost-sharing), and of the comparison conventional insurance arrangements. The results of the RAND randomized experiment study are broadly consistent with the non-randomized studies summarized in Luft (1981). Enrollees randomized to the managed care plan had inpatient admission levels 40% lower than those randomized to the conventional insurance plan with no cost-sharing. Outpatient spending was slightly, but not significantly higher, than under free care. Total imputed costs were 28% lower than under free care.

Since 1981, many studies have been conducted comparing utilization in managed care and non-managed care plans. These studies, mainly collected in Miller and Luft (1994) and Miller and Luft (1997) are summarized in Table 3. Miller and Luft limited their analyses to studies included in peer-reviewed publications that made some effort to control for differences in the characteristics of managed care and non-managed care enrollees.

There are several major problems in interpreting the results of the studies. First, while all of the studies use some form of statistical control for differences in characteristics (such as health status), only a few use random assignment to managed care. Some of the studies examine patients with a particular condition, but there may be difficult-to-observe differences in the health status of patients with similar conditions. As the selection studies above suggest, differences between managed care and non-managed care enrollees can take a wide variety of forms (and operate in both directions). Many of the characteristics associated with selection, such as preferences over intensity of treatment, are unlikely to be measurable by the researcher. Few of the non-randomized studies describe the terms of the choice faced by potential enrollees, which may also affect the extent and nature of selection.

Second, most of these studies do not fully describe the characteristics of either managed care plans or comparison traditional insurance arrangements. While many studies separate group and staff model plans from network or IPA model plans, there is no empirical or theoretical reason to believe that this is the most important distinction among plans. Some studies compare conventional Medicaid or Medicare to managed care, and in this case, the characteristics of the non-managed care plan are well known. Others, however, simply compare an HMO or PPO to a poorly defined conventional alternative.

Third, many of the studies rely on information from a small number of plans, providers, or employers. Since few details about the contents of plans are provided, it is difficult to generalize from these results.

Table 3
Utilization studies since 1980

Study	Year(s) of data collection and population	Comparison groups (detail – e.g., UR, capitation)	How control for differences in patient characteristics?	Managed care vs. comparison			
				Total charges	Length of stay	Visits	Admits
Angus et al. (1996)	1992 Adults in ICU in Mass	Commercial or Medicare FFS/Commercial or Medicare HMO	Age, sex, severity of illness, co-morbidities, diagnosis, discharge status		<65: −15%* >65: +1.5%		
Arnould, Debrock and Pollard (1984)	1980–1982 1 of 4 surgical procedures in Illinois	Prepaid Network/FFS	Demographic	# −35%–+2%	# −10%–+10%		
Bradbury, Golec, and Stearns (1991)	1988–1989 <65, 10 DRGs in 10 hospitals	IPA/FFS	Age; sex; admissions severity; case mix; hospital; year of admission		−14%*		
Braveman et al. (1991)	1987 Newborns, CA	Medicaid, uninsured, indemnity and prepaid	Demographics; diagnoses; hospital characteristics	−3%*	−1%*		
Buchanan et al. (1992)	1987 Medicaid AFDC, NY, FLA	Prepaid Managed Health Care/FFS	Randomization, sociodemographics, prior use	−30% ψ		−47% NY 1% FLA	−15% ψ
Buchanan, Leibowitz, and Keesey (1996)	1986 Medicaid AFDC, Florida	Staff model HMO/FFS	Age; family size; education; self-reported health status; avg. prior Mcaid expenditures and MD visits	−29%			
Carey et al. (1995)	1992–1993 (North Carolina Back Pain Project) Acute Low back pain	Group model HMO vs. FFS	Demographics, health services use, functional health status, provider type (primary care, specialty), rural/urban	P.C. −11% Spec. −37% ψ		P.C. −31%* Spec. −62%* ψ	

continued on next page

Table 3, *continued*

Study	Year(s) of data collection and population	Comparison groups (detail – e.g., UR, capitation)	How control for differences in patient characteristics?	Managed care vs. comparison			
				Total charges	Length of stay	Visits	Admits
Cole et al. (1994)	Early 1990s Mental health capitation	FFS/Capitation	Baseline differences		−1.28 days*		
Experton et al. (1996)	Early 1990s Medicare home care users	Medicare HMO/FFS/Medicaid	Socioeconomic, health status, functional status, clinical needs	0%	−42%* $	+29%* $	
Fitzgerald, Moore and Dittus (1988)	1981–1986 Medicare hip fracture, 1 hospital	Medicare FFS/HMO	Age; previous hip problems; PPS status			−47%*	
Garnick et al. (1990)	1984 Selected conditions, 1 insurer	PPO/Indemnity	Age, gender, comorbidities, hospitalizations	+3%– +56%* * #		+10%– +50%* * #	
Greenfield et al. (1992)	1986 Random sample >18 in Boston, Chicago, LA	1: Staff Model HMO 2: Prepaid Multi-specialty Group Practice (MGP) 3: FFS MGP 4: small/solo provider pre-paid group practice 5: small/solo FFS group practice	Patient mix, functional health status, sociodemographics, mortality, co-morbidities, history of MI			1/2: +16% 1/3: −12% 1/4: 0% 1/5: −29%*	1/2: −1% 1/3: +12% 1/4: +8% 1/5: +9%
Greenfield et al. (1995)	1986–1994 diabetics, hyptertensives	HMO: staff model IPA: prepaid MSGs and solo or single specialty practices FFS: MSG and solo or single specialty groups	Socio-demographics and health status			HMO-FFS: +6% IPA-FFS: −9% HMO-IPA: +20% ψ	

continued on next page

Table 3, *continued*

Study	Year(s) of data collection and population	Comparison groups (detail – e.g., UR, capitation)	How control for differences in patient characteristics?	Managed care vs. comparison			
				Total charges	Length of stay	Visits	Admits
Hosek, Marquis and Wells (1990)	1985/6 5 employers	FFS/5 PPO plans, cost-sharing specified	Socio-demographics, health status	−11%–+9% δ	−14%–17%* δ	+4%–+75%* δ	
Johnson et al. (1989)	1982–1984 1 of 10 diagnoses in Minneapolis	Group/Staff (GS) IPA/FFS	Demographic; Medical condition		GS −60%* IPA −10%		
Lubeck, Brown, and Holman (1985)	Early 1980s osteoarthritis	Staff model HMO/FFS	Demographics; pain; disability; disease duration	−13%		−22%*	
Lurie et al. (1994)	1980s Non-Institutionalized Medicaid elderly	FFS vs capitated Medicaid organized as 1: closed panel HMO, 2: County-sponsored Network HMO 3: 5 IPA plans	Randomization, health status indicators, sociodemographics	+27%	−38%	−7%	−20%*
Manning et al. (1984)	1976–1981 <62 Seattle in 1976	• Group model HMO • FFS by cost sharing (25%)	Randomization, age, sex	Vs. 25% FFS −16%		Vs. 25% +22%*	Vs. 25% −43%
Mark and Mueller (1996)	1993 National Health Interview Survey	HMO(IPA)/PPO/FFS	Age, sex, family income, health status, limitations on daily activity			HMO-PPO: +7% HMO-FFS: +20%* PPO-FFS: +12%	

continued on next page

Table 3, *continued*

Study	Year(s) of data collection and population	Comparison groups (detail – e.g., UR, capitation)	How control for differences in patient characteristics?	Managed care vs. comparison			
				Total charges	Length of stay	Visits	Admits
Martin et al. (1989)	1979–1982 New enrollees in Seattle HMO	IPA with Gatekeeper vs. IPA w/o gatekeeper	Randomized trial; demographics, perceived health status; other health insurance coverage	−6%	−26%	−1%	−13%
Mauldon et al. (1994)	1984 Medicaid Children in 1 hospital	Primary Care Case Management/FFS	Sex, race, # of health problems, random or self selected			−48%	
McCombs, Kasper, and Riley (1990)	1980–1982 Medicare	Group Model HMO/IPA/FFS Followed over 2 years	Socio-demographics, pre-enrollment charges	IPA: +27%* HMO −39%* ψ			
McCusker, Stoddard and Sorensen (1988)	1976–1982 200 Terminal cancer patients <65 Monroe City, NY	Multispecialty prepaid group practice and multiple-site group practice organization	Age; cancer site; months from diagnosis to death	−10%	−5%		−4%
Newcomer et al. (1995)	6/86–9/89 Medicare 4 sites	2 types social HMOs/FFS	Health status; case mix scores	Healthy +18% Very frail +23% ψ			
Norquist and Wells (1991)	1985 Mental health patients in Los Angeles	Medicare, FFS, Medicaid, uninsured, HMO	Age, sex, ethnicity, physical health, employment			Spec. MH −84%* PC +80%	

continued on next page

Table 3, *continued*

Study	Year(s) of data collection and population	Comparison groups (detail – e.g., UR, capitation)	How control for differences in patient characteristics?	Managed care vs. comparison			
				Total charges	Length of stay	Visits	Admits
Pearson et al. (1994)	1987–1989 Acute chest pain, 1 hospital	Staff model HMO/Commercial Ins. (indemnity + prepaid)/ Medicare/Medicaid/ Self-Pay/Other	Age, history of MI, clinical characteristics, risk category				+3%–+250% * ψ, δ
Rapoport et al. (1992)	1989–1990 ICU patients, 1 hospital	Staff-model HMO, PPO, IPA/FFS	Severity of illness; case mix; mortality	−25%	−28%*		
Reed et al. (1994)	1992 Mental health	FFS/Capitation		−14%			
Sisk et al. (1996)	1994 Medicaid New York City	5 plans vs. FFS	Health status and socio-demographic indicators, Medicaid aid category			Odds of any visits +1.10	Odds of admit −0.88
Stern et al. (1989)	1983–1985 1 of 13 DRGs 1 hospital	Staff model HMO/FFS	DRG, sex, age, similar admission dates	−4%	−14% *		
Sturm et al. (1995)	1986 Depressed patients	Prepaid group plans and FFS	Socio-demographics and health status			+35–40%*	
Szilagyi et al. (1990)	1981–1985 Pediatric ambulatory care Rochester NY	BCBS FFS/2 IPAs Switching study	socioeconomic, family size, health status			Acute: +42%* Well: +22%*	
Udvarhelyi et al. (1991)	1985–1987 Hypertension and preventive services	Network Model HMO (Capitation, UR)	Beseline demographic and clinical characteristics, medical history		+7% ψ		+15%* ψ
Welch (1985)	Late 1970s 2 national surveys	Group/Staff	Demographic characteristics		−32% δ	−25%	−2% δ

continued on next page

Table 3, *continued*

Study	Year(s) of data collection and population	Comparison groups (detail – e.g., UR, capitation)	How control for differences in patient characteristics?	Managed care vs. comparison			
				Total charges	Length of stay	Visits	Admits
Wells, Hosek and Marquis (1992)	1983–1986 Employees mental health use	PPO (2 in FL, 1 in CA)/FFS switching study	Mental health status, level of prior care for mental health, age gender, education	−3%*		−5%*	
Wouters (1990)	1982–1985 California residents in 1 plan	PPO/Non PPO Switching study	Sociodemographics, health status, expected health care utilization			−6%	
Yelin, Criswell and Feigenbaum (1996)	1982–1994 Rheumatoid arthritis	FFS/Prepaid Group Practice Over 11 years	Demographic and clinical characteristics, co-morbid conditions, medical utilization history			−2%	+17%
Yelin, Shern and Epstein (1986)	1982–1986 Rheumatoid arthritis in California	Prepaid Group Practice/FFS	Medical condition; socio-demographic characteristics		+1%	−2%*	+10%
Zwanziger and Auerbach (1991)	1985–1987 Employees Mental health use	PPO/ FFS	Demographics, prior health expenditures		MH: 7% Non-MH: 34% $	MH: 7% Non-MH: 2% $	

Source: Articles identified based on Miller and Luft (1994, 1997).

[#] Depending on condition.

[ψ] Midpoint of range.

[*] Statistically significant $p < 0.05$.

[$] Charges.

[δ] Depending on comparison.

Switching studies are those that compare people who switch from conventional to managed care coverage.

Finally, there is no consistent metric for measuring the effects of managed care. Some studies examine utilization differences in detail, while others report only differences in some measures of utilization.

In general, the results of earlier studies continue to hold in the more recent research, but there is enormous variation in the results. HMO-type managed care plans reduce hospital utilization, primarily through reductions in length of stay and admissions, and tend to increase outpatient utilization. Overall, total charges tend to be about 10–15% lower under these plans than under conventional insurance. One important difference between the more recent results and the earlier findings is that the form of HMO appears to be less important in generating the results. Plans that contract with dispersed providers (such as IPAs) appear to be as successful in controlling costs as more tightly integrated plans.

Some studies since 1982 compare utilization in preferred provider organizations with that in conventional insurance plans. The results for these plans are less clear. Some studies find reductions in unit costs under preferred provider plans [e.g., Smith (1997)], but others find that PPO plans, which often offer lower cost-sharing than conventional insurance, actually have higher costs than other arrangements [Hosek, Marquis and Wells (1990)].

5.3. Quality

Managed care may be a means of generating contracts that offer lower quality at lower cost. Alternatively, managed care may be a means of producing care of equivalent or better quality at lower cost. The literature on outcome differences for enrollees in managed care plans relative to conventional insurance arrangements, summarized in Luft (1981), Miller and Luft (1994), and Miller and Luft (1997), suggests that there are few consistent differences between the quality of care provided in managed care plans and conventional insurance arrangements. Similarly, the results of the RAND experiment found generally equivalent outcomes among HMO and conventional insurance enrollees [Ware et al. (1987)]. Both the Miller and Luft reviews and the RAND study, however, suggest that managed care plans may perform less well than conventional insurance arrangements for groups with serious health conditions, particularly those who also had low incomes.

Subjective measures of quality, such as consumer satisfaction with care, tend to favor conventional insurance arrangements over managed care for most (but not all) populations [Miller and Luft (1997)]. This result is consistent with the nature of rationing in managed care plans. While enrollees in conventional insurance arrangements self-ration through consumer cost-sharing, managed care enrollees are more likely to face a situation where they are willing to pay the (low) cost-sharing to gain access to a service, but the insurer or provider denies such access. Furthermore, enrollees who prefer restrictions on access to high premiums ex ante may be dissatisfied with their choice afterwards. Restrictions on access to providers, limitations on length of stay, and other barriers to care in managed care plans have provoked the widespread regulatory efforts (described above) that would limit the ability of managed care to ration care through such restrictions.

5.4. Spillover effects of managed care

Costs of care in managed care may be low relative to conventional insurance, but if these cost reductions occur as a consequence of selection, or if they lead to demand inducement, apparent savings may be illusory. Total health care costs may rise (or not fall) through the entry of managed care. The potential effects of managed care on the conventional insurance market make it important to look at total costs as a measure of the effectiveness of managed care. Table 4 summarizes the results of these studies.

Managed care effects on the total cost of health care in a market are less likely to be affected by selection problems at the level of the individual (as long as there is no change in the size or characteristics of the overall insured population). Selection may, however, occur at the level of the health plan. Managed care plans may be more likely to enter markets where overall costs are low or are likely to decelerate [Welch (1985)]. Some early studies acknowledge this problem [for example, McLaughlin, Merrill and Freed (1983) and Hay and Leahy (1984)], but it is difficult to correct. More recent studies sometimes use instrumental variable methods to adjust for the entry decisions of managed care firms. Unfortunately, it is difficult to identify factors that should affect the entry of managed care plans while not affecting total costs.

Early studies of the effects of managed care on total costs were generally case studies, and most found no effect. As Frank and Welch (1985) point out, few of these studies address problems of selection bias at the individual level. Most also do not consider selection at the health plan level. More recent studies focus on the rate of cost growth in areas with high managed care penetration. Most, but not all, of the more recent studies find that increases in managed care penetration are associated with reductions in the rate of growth of total costs. While these studies mainly support the hypothesis that managed care can reduce total costs, they do not yet conclude the issue. Indeed, one study found that the entry of managed care plans drove total employer health insurance costs up [Feldman, Dowd and Gifford (1993)]. Furthermore, most of the results are identified mainly from managed care penetration in California (four of the recent studies rely exclusively on data from California). To the extent that managed care takes different, and perhaps less effective, forms in other parts of the country [see, for example, Remler et al. (1997)], or that California's health care climate differs for other reasons, these results may not be generalizable.

5.5. Cost growth

A few studies have examined the rate of growth of costs within managed care plans. This research addresses the question of whether managed care plans are a superior way of addressing problems of dynamic moral hazard in health insurance. Again, the results may be contaminated by selection problems. In particular, if managed care plans

Table 4
Managed care and total health care costs

Study	Result	Sample	Notes
Managed care raises total costs			
Feldman, Dowd and Gifford (1993)	offering an HMO raises total employer costs	Minneapolis area employers	
Hay and Leahy (1984)	increased HMO share increases hospital utilization costs	202 hospital service areas	
McLaughlin, Merrill and Freed (1983)	increased HMO penetration increases hospital utilization costs	25 SMSAs	
Managed care does not reduce total costs			
Baker and Corts (1996)	above 10% HMO, conventional insurance premiums rise	Data on 3000 firms	
Feldman, Dowd, McCann, Johnson (1986)	market share and discounts have no effect on profits		
Johnson and Aquilina (1986)	no overall effect	case study of Minneapolis	
Krueger and Levy (1997)	HMO premiums only slightly below FFS, cannot explain savings		
Luft, Maerki and Trauner (1986)	no consistent effect	case studies of Hawaii, Rochester, and Minneapolis	
McLaughlin (1987)	no effect on average hospital expenses per capita	25 SMSAs 1972–1982	
McLaughlin (1988)	no significant effect of HMOs on per capita, per day, or per admission hospital expenses	283 SMSAs in 1980	
Merrill and McLaughlin (1986)	lower hospital admits and higher expenses per day in high HMO areas	25 SMSAs over 10 years; insurers respond by trying to control own costs	
Managed care reduces total costs			
Baker (1997)	above 18% market share, HMO penetration reduces total Medicare costs		later results suggest may have increased over time
Cutler and Sheiner (1998)	10% increase in HMO enrollment reduces total cost growth about 4%	diffusion of new interventions, lower tech growth in high penetration markets	results control for whether state is a "high-diffuser" or not
Feldstein and Wickizer (1995)	HMO market share reduces growth of insurance premiums (elasticity −0.65)	1985–1992 data – 95 insured groups	

continued on next page

Table 4, *continued*

Study	Result	Sample	Notes
Gaskin and Hadley (1997)	hospital expenses grew 8.3% in high HMO and 11.2% annually in low HMO regions, effects stronger over time	1985–1993	
Goldberg and Greenberg (1979)	increased HMO share reduces overall hospital utilization	insurers respond by trying to control own costs	
Melnick and Zwanziger (1995)	managed care reduces hospital costs relative to nation and rate regulating states	California vs. national average	
Robinson (1991)	hospital costs per admission grew 9.4% slower in high HMO penetration markets than in low penetration markets	California hospitals 1982–1988	
Robinson (1996)	hospital expenditures grew 44% slower in high HMO penetration markets	California hospitals 1983–1993	
Zwanziger and Melnick (1989)	highly competitive markets had lower cost growth	California data	

benefit from positive selection, adverse selection could lead premiums in conventional insurance plans to grow very rapidly as managed care plans enter the market. This rapid growth could mistakenly suggest that managed care plans were better at controlling cost growth.

Studies of cost growth using data through the early 1980s generally find equivalent or very slightly slower rates of growth in managed care plans [Christianson and McClure (1979), Luft (1981), Newhouse et al. (1985)]. More recent studies find that managed care rates of growth are slightly slower, as much as 1 percentage point per year slower than traditional insurance premium growth [Miller and Luft (1997)].

Another way of examining cost growth is by looking at the effects of managed care on choices about the use of technology. Several studies examine how managed care affects technological diffusion. Higher managed care penetration appears to reduce the number of facilities and increase the volume per facility of mammography equipment [Baker and Brown (1997)]; and reduce the rate of Cesarean sections [Tussing and Wojtowycz (1994)]. Not all studies point in this direction, however. Chernew, Fendrick and Hirth (1997) finds that HMOs have had as much difficulty in controlling the diffusion of laparoscopic cholecystectomy as have other plans.

Lower rates of technological diffusion may lead to lower costs at a point in time (or over a brief period). If managed care is able to reduce dynamic moral hazard, it should do so by changing the rate of adoption of new technologies. Only one study to date examines this question, and it finds that the growth of managed care reduced the rate of adoption of new technologies [Cutler and Sheiner (1998)]. In general, the finding that managed care may have led to a lower overall rate of cost growth is still tentative, but it is buttressed by evidence of lower rates of technological adoption and diffusion in areas dominated by managed care.

6. Economic issues related to the growth of managed care

Managed care operates quite differently from conventional insurance policies. These differences imply that the institutional structures established to address concerns in the insurance market may not be equally appropriate in response to problems in the managed care marketplace. Theory and empirical research suggest three areas where the advent of managed care may alter economic research in broader areas: competition policy, malpractice litigation, and public program design.

6.1. Competition among managed care plans

Conventional insurers have relatively few dimensions of performance on which to compete. Under conventional insurance, competition in the health care market occurs mainly at the level of the health care provider. Correspondingly, antitrust scrutiny has focused on health care provider behavior. Managed care, by contrast, is characterized by relationships between insurers and health care providers. The conventional insurance model of competition may not apply in managed care markets. This literature is summarized in the Handbook chapters on Antitrust [Gaynor and Vogt (2000)] and Industrial Organization [Dranove and Satterthwaite (2000)].

As in other arenas, the competitiveness of managed care markets will depend on the underlying extent of economies of scope and scale in managed care operations and on the extent to which managed care markets are contestable. There may be scale economies in the performance of key managed care functions, such as utilization review or guideline formation. Plans may be able to achieve economies of scope (across markets or market segments), by transferring expertise gained in one area; or by developing a brand name that has value across markets.

Empirical research has begun to investigate the extent of economies of scope and scale across managed care plans. Two studies using data from the late 1970s and early 1980s find some evidence of managed care economies of scale in outpatient visits [Bothwell and Cooley (1982), Schlesinger, Blumenthal and Schlesinger (1986)]. More recent studies that examine overall economies of scale find that such economies are present, but at relatively low levels. Given (1996) finds that economies of scale occur up to about

115,000 enrollees; while Wholey et al. (1996) find similar results up to about 50,000 enrollees. Most managed care plan enrollees are members of much bigger plans. In 1997, the median HMO had 40,000 members [HCIA (1997)].

Other analyses suggest that managed care plans do compete with one another, so that premiums fall as the HMO market share rises [Wholey, Feldman and Christianson (1995)]. Together with minimal evidence of scale economies, these results suggest that mergers in the managed care industry might be expected to have anti-competitive effects [Feldman (1994)]. The only empirical analysis of mergers, however, suggests that they have had little effect on health care costs [Christianson, Feldman and Wholey (1997)]. In the past, competition in the health care sector focused on quality, not costs. Economic research to date has not investigated the role of quality competition in the managed care marketplace.

6.2. Malpractice

The malpractice litigation system, like other tort systems, is intended to encourage providers (and patients) to minimize the cost of potential negligent injuries [see Handbook chapter by Danzon (2000)]. The existing model of malpractice in medicine separates decisions about the quality of care received or not received (suits against health care providers) from decisions about coverage (contract cases against insurers). This model may have less applicability when providers bear financial risk for coverage decisions and insurers provide guidelines for treatment. Furthermore, the standard analysis is predicated on the assumption that providers generally have incentives to provide too many services. To the extent that the incentives in managed care operate in the opposite direction, new analyses of the design of malpractice insurance systems are needed [Blomqvist (1991)].

6.3. Risk adjustment

Risk segmentation complicates the evaluation of the effectiveness of managed care and has potentially undesirable normative consequences (as discussed above). Furthermore, risk segmentation makes it difficult to design managed care policy. Consider a payer, such as the Medicare program, that operates its own indemnity plan and contracts with managed care plans. If the payer sets managed care payment rates based on the indemnity population, while the managed care plans enroll healthier-than-average enrollees, total costs under the program may increase. If risk segmentation is important, payers must ensure that the rates they pay to managed care plans accurately reflect the risk profile of the population these plans enroll.

For all of these reasons, the increased diversity of insurance plans that has characterized the growth of managed care has encouraged the development of methods that capture differences in the characteristics of enrollees in different plans. These techniques, or risk adjustment methodologies, are summarized in the Handbook chapter on risk adjustment [Van de Ven and Ellis (2000)].

7. Conclusions

The nature of health insurance in the United States has become much more complex over the past 20 years. Economic theory and empirical research have not entirely kept pace with these changes. Very little theory explores the relative efficiency of consumer cost-sharing, provider cost-sharing, and direct monitoring of service utilization. In consequence, economic theory has little to say about the reasons for the recent growth in managed care arrangements. Empirical research on managed care is hampered by the extraordinary variety of plans that fall into the general category. Research is needed to identify which characteristics of managed care generate economically meaningful differences in outcomes and which are only superficial. The regulation of managed care practice, antitrust and malpractice law concerning managed care, and the integration of managed care into public programs are proceeding rapidly. Theoretical and empirical research in this area are of critical public policy importance.

References

Adamache, K.W., and L.F. Rossiter (1986), "The entry of HMOs into the Medicare market: implications for TEFRA's mandate", Inquiry 23:349–364.

American Association of Health Plans (1998), National Directory of Health Plans and Utilization Review Organizations (American Association of Health Plans, Washington, DC).

Angus, D.C., W.T. Linde-Zwible, C.A. Sirio, A.J. Rotondi, L. Chelluri, R.C. Newbold, J.R. Lane and M.R. Pinsky (1996), "The effect of managed care on ICU length of stay: implications for Medicare", JAMA 276:1075–1082.

Arnould, R.J., L.W. Debrock and J.W. Pollard (1984), "Do HMOs produce specific services more efficiently?", Inquiry 21:243–253.

Arrow, K.J. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53:941–973.

Baker, L., and K. Corts (1996), "HMO penetration and the cost of health care: market discipline or market segmentation", American Economic Review 86(2):389–94.

Baker, L.C., and M.L. Brown (1997), "The effect of managed care on health care providers", Working Paper 5987 (National Bureau of Economic Research).

Baker, L. (1997), "The effect of HMOs on fee-for-service health expenditures: evidence from Medicare", Journal of Health Economics 16(4):453–81.

Baumgardner, J. (1991), "The interaction between forms of insurance contract and types of technical change in medical care", RAND Journal of Economics 22(1):36–53.

Berki, S.E., M. Ashcraft, R. Penchanski and R. Fortus (1977), "Enrollment choice in a multi-HMO setting: the roles of health risk, financial vulnerability, and access to care", Medical Care 15(2):95–114.

Billi, J.E., C.G. Wise, S.I. Sher, L. Duran-Arenas and L. Shapiro (1993), "Selection in a preferred provider organization enrollment", Health Services Research 28(5):563–75.

Blomqvist, A. (1991), "The doctor as double agent: information asymmetry, health insurance, and medical care", Journal of Health Economics 10(4):411–432.

Bonnano, J.B., and T. Wetle (1984), "HMO enrollment of Medicare recipients: an analysis of incentives and barriers", Journal of Health Politics, Policy and Law 9(1):41–62.

Bothwell, J.L., and T.F. Cooley (1982), "Efficiency in the provision of health care: an analysis of health maintenance organizations", Southern Economic Journal 47:970–984.

Bradbury, R.C., J.H. Golec and F.E. Stearns (1991), "Comparing hospital length of stay in independent practice association HMOs and traditional insurance programs", Inquiry 28:87–93.

Braveman, P.A., S. Egerter, T. Bennett and J. Showstack (1991), "Differences in hospital resource allocation among sick newborns according to insurance coverage", JAMA 266:3300–3308.

Brown, L.D. (1983), Politics and Health Care Organization: HMOs as Federal Policy (Brookings Institution, Washington, DC).

Brown, L.D. (1998), "Exceptionalism as the rule? US health policy innovation and cross-national learning", Journal of Health Politics, Policy, and Law 23(1):35–51.

Brown, R.S. (1988), "Biased selection in the Medicare competition demonstrations", Report to the Health Care Financing Administration (Mathematica Policy Research Inc., Princeton).

Brown, R.S., and J. Hill (1993), "The Medicare risk program for HMOs: final summary report on findings from the evaluation" (Mathematical Policy Research Inc., Princeton, NJ).

Buchanan, J.L., A. Leibowitz and J. Keesey (1996), "Medicaid health maintenance organizations: can they reduce program spending?", Medical Care 34:249–263.

Buchanan, J., and S. Cretin (1986), "Risk selection of families electing HMO membership", Medical Care 24(1):39–51.

Buchanan, J.L., A. Leibowitz, J. Keesey, J. Mann and C. Damberg (1992), "Cost and use of capitated medical services: evaluation of the program for prepaid managed health care", RAND R-4225-HCFA (RAND Corporation, Santa Monica, CA).

Bureau of the Census (1996), "Money income and poverty status of families and persons in the United States", Current population reports, Series P-60 (US Dept. of Commerce, Bureau of the Census, Washington, DC).

Carey, T.S., J. Garrett, A. Jackman, C. McLaughlin, J. Fryer and D.R. Smucker (1995), "The outcomes and costs of care for acute low back pain among patients seen by primary care practitioners, chirpractors, and orthopedic surgeons", The New England Journal of Medicine 333:913–917.

Center for Studying Health System Change (1998), Data Bulletin 12, Summer.

Cherkin, D., L. Grothaus and E. Wagner (1989), "The effect of office visit copayments on utilization in a health maintenance organization", Medical Care 27:1036–45.

Chernew, M. (1995), "The impact of non-IPA HMOs on the number of hospitals and hospital capacity", Inquiry 32:143–154.

Chernew, M., M.A. Fendrick and R.A. Hirth (1997), "Managed care and medical technology", Health Affairs 16(2):196–206.

Chernew, M., R. Hayward and D. Scanlon (1996), "Managed care and open-heart surgery facilities in California", Health Affairs 15(1):191–201.

Christianson, J.B., R.D. Feldman and D.R. Wholey (1997), "HMO mergers: estimating impact and costs", Health Affairs 16(6):133–41.

Christianson, J.B., and W. McClure (1979), "Competition in the delivery of medical care", New England Journal of Medicine 301(15):812–18.

Cochrane, J.H. (1995), "Time consistent health insurance", Journal of Political Economy 103(3):445–473.

Cole, R.E., S.K. Reed, H.M. Babigian, S.W. Brown and J. Fray (1994), "A mental health capitation program: I. Patient outcomes", Hospital and Community Psychiatry 145:1090–1096.

Cutler, D., and S. Reber (1998), "Paying for health insurance: the tradeoff between competition and adverse selection", Quarterly Journal of Economics 113(2): 433–466.

Cutler, D.M., and L. Sheiner (1998), Managed Care and the Growth of Medical Expenditures in Frontiers in Health Policy Research, A.M. Garber, ed., Vol. 1 (MIT Press for the National Bureau of Economics Research, Cambridge and London) 77–116.

Danzon, P.M. (2000), "Liability for medical malpractice", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 26.

Doherty, N.A. (1979), "Insurer and provider as the same firm: HMO's and moral hazard-comment", Journal of Risk and Insurance 46(3):550–553.

Dranove, D., and M. Satterthwaite (2000), "Industrial organization of health care markets", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 20.

Dranove, D., M. Shanley and W.D. White (1993), "Price and concentration in hospital markets: the switch from patient-driven to payer-driven competition", Journal of Law and Economics 36(1; Part 1):179–204.

Duston, T.E. (1978), "Insurer and provider as the same firm: HMO's and moral hazard", Journal of Risk and Insurance 45(1):141–147.

Eggers, P., and R. Prihoda (1982), "Pre-enrollment reimbursement patterns of Medicare beneficiaries enrolled in 'at-risk' HMOs", Health Care Financing Review 4(1):55–73.

Eggers, P. (1980), "Risk differentials between Medicare beneficiaries enrolled and not enrolled in an HMO", Health Care Financing Review 1(3):91–99.

Ellis, R.P., and T.G. McGuire (1993), "Supply-side and demand-side cost sharing in health care", Journal of Economic Perspectives 7(4):135–152.

Ellis, R.P., and T.G. McGuire (1990), "Optimal payment systems for health services", Journal of Health Economics 9(4):375–396.

Enthoven, A. (1978), "Competition of alternative delivery systems", in: W. Greenberg, ed., Competition in the Health Care Sector: Past, Present, and Future. Proceedings of a Conference Sponsored by the Bureau of Economics, Federal Trade Commision, March 1978 (Aspen Systems Corporation, Germantown, MA) 225–278.

Enthoven, A. (1980), Health Plan: The Only Practical Solution to the Soaring Cost of Medical Care (Addison-Wesley, Reading, MA).

Ermann, D. (1988), "Hospital utilization review: past experience, future directions", Journal of Health Politics, Policy and Law 13:683–704.

Escarce, J.J., J.A. Shea and W. Chen (1997), "Segmentation of hospital markets: where do HMO enrollees get care?", Health Affairs 16(6):181–92.

Experton, B., R.J. Ozminkowski, L.G. Branch and Z. Li (1996), "A comparison by payor/provider type of the cost of dying among frail older adults", Journal of the American Geriatrics Society 44:1098–1107.

Feldman, R. (1994), "The welfare economics of a health plan merger", Journal of Regulatory Economics 6(1):67–86.

Feldman, R., B. Dowd and G. Gifford (1993), "The effect of HMOs on premiums in employment-based health plans", Health Services Research 27(6):779–781.

Feldman, R., B. Dowd, D. McCann and A. Johnson (1986), "The competitive impact of health maintenance organizations on hospital finances: an exploratory study", Journal of Health Politics, Policy and Law 10(4):675–97.

Feldman, R., M. Finch and B. Dowd (1989), "The role of health practices in HMO selection bias: a confirmatory study", Inquiry 26(3):381–87.

Feldman, R., J. Kralewski and B. Dowd (1989), "Health maintenance organizations: the beginning or the end", Health Services Research 24(2):191–211.

Feldstein, P.J., and T.M. Wickizer (1995), "Analysis of private health insurance premium growth rates: 1985–1992", Medical Care 33(10):1035–50.

Feldstein, P.J., T.M. Wickizer and J.R. Wheeler (1988), "Private cost containment. The effects of utilization review programs on health care use and expenditures", New England Journal of Medicine 318(20):1310–1314.

Fitzgerald, J.F., P.S. Moore and R.S. Dittus (1988), "The care of elderly patients with hip fractures: changes since implementation of the prospective payment system", New England Journal of Medicine 319:1392–1397.

Frank, R.G., J. Glazer and T.G. McGuire (1998), "Measuring adverse selection in managed health care", Working paper.

Frank, R.G., and T.G. McGuire (2000), "Economics and mental health", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 16.

Frank, R.G., and W.P. Welch (1985), "The competitive effects of HMOs: a review of the evidence", Inquiry 22(2):148–61.

Friedman, E.S. (1996), "Capitation, integration, and managed care: lessons from early experiments", JAMA 275(12):957–62.

Gabel, J., D. Ermann, T. Rice and G. de Lissovoy (1986), “The emergence and future of PPOs”, Journal of Health Politics, Policy and Law 11(2):305–22.

Gabel, J.A. (1997), “Ten ways HMOs have changed during the 1990s”, Health Affairs 16(3):134–145.

Gabel, J.A., K.A. Hunt and K.M. Hurst (1998), “When employers choose health plans: do NCQA accreditation and HEDIS data count?”, Commonwealth Fund Paper, August.

Garnick, D.W., H.S. Luft, L.B. Gardner, E.M. Morrison, M. Barrett, A. O’Neil and B. Harvey (1990), “Services and charges by PPO physicians for PPO and indemnity patients: an episode of care comparison”, Medical Care 28:894–906.

Gaskin, D., and J. Hadley (1997), “The impact of HMO penetration on the rate of hospital cost inflation, 1985 and 1993”, Inquiry 34(3):205–216.

Gaynor, M. (1994), “Issues in the industrial organization of the market for physician services”, Journal of Economics and Management Strategy 3(1):211–255.

Gaynor, M., and P. Gertler (1995), “Moral hazard and risk spreading in partnerships”, Rand Journal of Economics 26(4):591–613.

Gaynor, M., and W. Vogt (2000), “Antitrust and competition in health care markets”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 27.

Given, R. (1996), “Economies of scale and scopes as an explanation of merger and output diversification”, Journal of Health Economics 15(6):685–713.

Glied, S., M. Sparer and L. Brown (1995), “Comment: containing state health expenditures”, American Journal of Public Health 85(10):1347–1349.

Glied, S. (1998), “Getting the incentives right for children”, Health Services Research 33(4; Part II):1143–1160.

Goddeeris, J.H. (1984), “Medical insurance, technological change, and welfare”, Economic Inquiry 22:56–67.

Gold, M.R., R. Hurley, T. Lake, T. Ensor and R. Berenson (1995), “A national survey of the arrangements managed care plans make with physicians”, New England Journal of Medicine 333(25):1678–1683.

Goldberg, L.G., and W. Greenberg (1979), “The competitive response of Blue Cross and Blue Shield to the growth of health maintenance organizations in Northern California and Hawaii”, Medical Care 17(10):1019–1028.

Goldberg, L.G., and W. Greenberg (1981), “The determinants of HMO enrollment and growth”, Health Services Research 16:421–438.

Goldman, D.P. (1995), “Managed care as a public cost-containment mechanism”, Rand Journal of Economics 26(2):277–95.

Gordon, N., and G. Kaplan (1991), “Some evidence refuting the HMO “favorable selection” hypothesis: the case of Kaiser Permanente”, Advances in Health Economics and Health Services Research 12:19–39.

Greenfield, S., E.C. Nelson, M. Zubkoff, W. Manning, W. Rogers, R.L. Kravitz, A. Keller and A.R. Tarlov (1992), “Variations in resource utilization among medical specialties and systems of care: results from the Medical Outcomes Study”, JAMA 267:1624–1630.

Greenfield, S., W. Rogers, M. Magotich, M.F. Carneyand and A.R. Tarlov (1995), “Outcomes of patients with hypertension and non-insulin dependent diabetes mellitus treated by different systems and specialties: results from the Medical Outcomes Study”, JAMA 274:1436–1444.

Gruber, L.R., M. Shadle and C.L. Polich (1988), “From movement to industry: the growth of HMOs”, Health Affairs 7(3):197–208.

Hart, L.G., E. Wagner, S. Pirzada, A.F. Nelson and R.A. Rosenblatt (1997), “Physician staffing ratios in staff-model HMOs: a cautionary tale”, Health Affairs 16(1):55–70.

Hay, J.W., and M.J. Leahy (1984), “Competition among health plans: some preliminary evidence”, Southern Economic Journal 50(3):831–845.

HCIA (1997), The Guide to the Managed Care Industry (HCIA Inc., Baltimore).

Hellinger, F.J. (1995), “Selection bias in HMOs and PPOs: a review of the evidence”, Inquiry 32(Summer):135–143.

Hellinger, F.J. (1996), “The expanding scope of state legislation”, JAMA 276(13):1065–1070.

Hill, J., and R. Brown (1990), "Biased selection in the TEFRA HMO/CMP program", Final Report to the Department of Health and Human Services, September (Mathematica Policy Research, Princeton).

Hill, J., R. Brown, D. Chu and J. Bergeron (1992), The Impact of the Medicare Risk Program on the Use of Services and Costs to Medicare (Mathematica Policy Research, Princeton).

Hill, S., and B. Wolfe (1997), "Testing the HMO competitive strategy: an analysis of its impact on medical care resources", Journal of Health Economics 16(3):261–86.

Hillman, A.L. (1987), "Financial incentives for physicians in HMOs: is there a conflict of interest?", New England Journal of Medicine 317:1743–1748.

Hillman, A.L., M.V. Pauly and J.J. Kerstein (1989), "How do financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations?", New England Journal of Medicine 321:86–92.

Hillman, A.L., W.P. Welch and M. Pauly (1992), "Contractual arrangements between HMOs and primary care physicians: three-tiered HMOs and risk pools", Medical Care 30(2):136–48.

Holahan, J., S. Zuckerman, A. Evans and S. Rangarajan (1998), "Medicaid managed care in thirteen states", Health Affairs 17(3):43–63.

Hosek, S.D., M.S. Marquis and K.B. Wells (1990), "Health care utilization in employer plans with preferred provider organization options", RAND R-3800-HHS/NIMH.

Hudes, J., C.A. Young, L. Sohrab and C.N. Trinh (1980), "Are HMO enrollees being attracted by a liberal maternity benefit?", Medical Care 18(6):635–48.

Institute of Medicine (1976), Assessing Quality in Health Care (National Academy of Sciences, Washington, DC).

Institute of Medicine (1989), B. Gray and M. Field, eds., Controlling Costs and Changing Patient Care? The Role of Utilization Management (National Academy Press, Washington, DC).

Institute of Medicine (1993), Employment and Health Benefits: A Connection at Risk (National Academy Press, Washington, DC).

Jackson-Beeck, M., and J.H. Kleinman (1983), "Evidence for self-selection among health maintenance organization enrollees", JAMA 250(20):2826–29.

Jensen, G.A., M.A. Morrisey, S. Gaffney and D.K. Liston (1997), "The new dominance of managed care: insurance trends in the 1990s", Health Affairs 16(1):125–136.

Johnson, A.N., and D. Aquilina (1986), "The impact of health maintenance organizations and competition on hospitals in Minneapolis/St. Paul", Journal of Health Politics Policy Law 10(4):659–74.

Johnson, A.N., B. Dowd, N.E. Morris and N. Lurie (1989), "Differences in inpatient resource use by type of health plan", Inquiry 26:388–398.

Kasper, J.D., G.F. Riley, J.S. McCombs and M.A. Stevenson (1988), "Beneficiary selection, use, and charges in two Medicare capitation demonstrations", Health Care Financing Review 10(1):7–49.

Khandker, R., and W.G. Manning (1992), "The impact of utilization review on costs and utilization", Health Economics Worldwide. Developments in Health Economics and Public Policy series, Vol. 1, 47–62.

Klein, B., and K.B. Leffler (1981), "The role of market forces in assuring contractual performance", Journal of Political Economy 89(4):615–641.

Kralewski, J.E., T.D. Wingert, R. Feldman, G.J. Rahn and T.H. Klassen (1992), "Factors related to the provision of hospital discounts for HMO inpatients", Health Services Research 27(2):133–53.

Krueger, A.B., and H. Levy (1997), "Accounting for the slowdown in employer health care costs", Working Paper 5891 (National Bureau of Economic Research).

Lairson, D., and A. Herd (1987), "The role of health practices, health status, and prior health care claims in HMO selection bias", Inquiry 24(3):276–84.

Lindsey, P.A., and J.P. Newhouse (1990), "The cost and value of second surgical opinion programs: a critical review of the literature", Journal of Health Politics Policy and Law 15(3):543–70.

Lubeck, D.P., B.W. Brown and H.R. Holman (1985), "Chronic disease and health system performance: care of osteoarthritis across three health services", Medical Care 23:266–277.

Luft, H.S. (1981), Health Maintenance Organizations: Dimensions of Performance (John Wiley and Sons, New York).

Luft, H.S., S.C. Maerki and J.B. Trauner (1986), "The competitive effects of health maintenance organizations: another look at the evidence from Hawaii, Rochester, and Minneapolis/St. Paul", Journal of Health Politics, Policy, and Law 10(Winter):625–658.

Luft, H.S., J.B. Trauner and S.C. Maerki (1985), "Adverse selection in a large, multiple-option health benefits program: a case study of the California Public Employees' Retirement System", Advances in Health Economics and Health Services Research 6:197–229.

Lurie, N., J. Christianson, M. Finch and I. Moscovice (1994), "The effects of capitation on health and functional status of the medicaid elderly", Annals of Internal Medicine 120:506–511.

Managed Care: Facts, Trends and Data: 1997–98, 2nd edn. (1997) (Atlantic Information Services, Inc., Washington, DC).

Manning, W.G., A. Leibowitz, G.A. Goldberg, W.H. Rogers and J. Newhouse (1984), "A controlled trial of the effect of a prepaid group practice on the use of services", New England Journal of Medicine 310(23):1505–1510.

Mark, T., and C. Mueller (1996), "Access to care in HMOs and traditional insurance plans", Health Affairs 15:81–87.

Marsteller, J.A., R.R. Bovbjerg, L.M. Nichols and D.K. Verrilli (1997), "The resurgence of selective contracting restrictions", Journal of Health Politics Policy and Law 22(5):1133–89.

Martin, D.P., P. Diehr, K.F. Price and W.C. Richardson (1989), "Effect of a gatekeeper plan on health services use and charges: a randomized trial", American Journal of Public Health 79:1628–1632.

Mathewson, F.J., and R.A. Winter (1997), "Buyer groups", International Journal of Industrial Organization 15(2):137–164.

Mauldon, T., A. Leibowitz, J.L. Buchanan, C. Damberg and K.A. McGuigan (1994), "Rationing or rationalizing children's medical care: comparison of a medicaid HMO with fee-for-service care", American Journal of Public Health 84:899–904.

McCombs, J.S., J.D. Kasper and G.F. Riley (1990), "Do HMOs reduce health care costs? A multivariate analysis of two Medicare HMO demonstration projects", Health Services Research 25:593–613.

McCusker, J., A.M. Stoddard and A.A. Sorensen (1988), "Do HMOs reduce hospitalization of terminal cancer patients?", Inquiry 25:263–270.

McGlynn, E.A. (1997), "Six challenges in measuring the quality of health care", Health Affairs 16(3):7–21.

McGuire, T.G., and M.V. Pauly (1991), "Physician responses to fee changes with multiple payers", Journal of Health Economics 10(4):385–410.

McLaughlin, C.G. (1987), "HMO growth and hospital expenses and use: a simultaneous equation approach", Health Services Research 22(2):183–205.

McLaughlin, C.G. (1988), "Market responses to HMOs: price competition or rivalry?", Inquiry 25:207–218.

McLaughlin, C.G., J.C. Merrill and A.J. Freed (1983), "The impact of HMO growth on hospital costs and utilization", Advances in Health Economics and Health Services Research 5:57–93.

Melnick, G.A., and J. Zwanziger (1995), "State health care expenditures under competition and regulation, 1980 through 1991", American Journal of Public Health 85(10):1391–6.

Merrill, J., and C. McLaughlin (1986), "Competition versus regulation: some empirical evidence", Journal of Health Politics, Policy, and Law 10(4):613–624.

Miller, R.H., and H.S. Luft (1997), "Does managed care lead to better or worse quality of care?", Health Affairs 16(5):7–25.

Miller, R.H., and H.S. Luft (1994), "Managed care plan performance since 1980: a literature analysis", JAMA 271(May 18):1512–1519.

Morrisey, M., and C. Ashby (1982), "An empirical analysis of HMO market share", Inquiry 19(2):136–49.

Morrison, E., and H. Luft (1991), "Alternative delivery systems", in: E. Ginzberg, ed., Health Services Research.

Newcomer, R. (1995), "Case mix controlled service use and expenditures in the Social/Health Maintenance organization", Journal of Gerontology, Medical Sciences 50A:M35–M44.

Newhouse, J.P. (1978), “The structure of health insurance and the erosion of competition in the medical marketplace”, in: W. Greenberg, ed., Competition in the Health Care Sector: Past, Present, and Future (Aspen Systems, Germantown, MD).

Newhouse, J.P., et al. (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge, MA).

Newhouse, J.P., and P.A. Lindsey (1988), “Do second opinion programs improve outcomes?”, Journal of Health Economics 7:285–288.

Newhouse, J.P., and the Insurance Experiment Group (1996), “Reimbursing health plans and health providers: efficiency in production versus selection”, Journal of Economic Literature 34(3):1236–63.

Newhouse, J.P., W. Schwartz, A. Williams and C. Witsberger (1985), “Are fee-for-service costs increasing faster than HMO costs?”, Medical Care 23(8):960–66.

Norquist, G.S., and K.B. Wells (1991), “How do HMOs reduce outpatient mental health costs?”, American Journal of Psychiatry 148:96–101.

Pauly, M.V. (1970), “Efficiency, incentives and reimbursement for health care”, Inquiry 7(1):115–131.

Pauly, M.V. (1978), “Is medical care really different?”, in: W. Greenberg, ed., Competition in the Health Care Sector (Aspen Systems, Germantown, MD).

Pauly, M.V. (1985), “What is adverse about adverse selection?”, Advances in Health Economics and Health Services Research 6:281–286.

Pearson, S.D., T.H. Lee, E. Lindsey, S.T. Hawkings, E.F. Cook and L. Goldman (1994), “The impact of membership in a health maintenance organization on hospital admission rates for acute chest pain”, Health Services Research (April):59–74.

Perkoff, G.T., L. Kahn and P.J. Haas (1976), “The effects of an experimental prepaid group practice on medical care utilization and cost”, Medical Care 14(5):432–449.

Physician Payment Review Commission (1996), Annual Report to Congress (The Commision, Washington).

Physician Payment Review Commission (1997), Annual Report to Congress (The Commision, Washington).

Quinn, K. (1998), The Sources and Types of Health Insurance (Abt Associates Inc., Cambridge).

Ramsey, S., and M. Pauly (1997), “Structural incentives and adoption of medical technologies in HMO and fee-for-service health insurance plans”, Inquiry 34(3):228–36.

Rapoport, J., S. Gehlbach, S. Lemeshow and D. Teres (1992), “Resource utilization among intensive care patients: managed care vs. traditional insurance”, Archives of Internal Medicine 152:2207–2212.

Reed, S.K., K.D. Hennessey, O.S. Mitchell and H.M. Babigian (1994), “A mental health capitation program: II. Cost benefit analysis”, Hospital and Community Psychiatry 45:1097–1103.

Remler, D.K., et al. (1997), “What do managed care plans do to affect care? Results from a survey of physicians”, Inquiry 34(3):196–204.

Robinson, J.C. (1991), “HMO market penetration and hospital cost inflation in California”, JAMA 266(19):2719–2723.

Robinson, J.C. (1993), “Payment mechanisms, nonprice incentives, and organizational innovation in health care”, Inquiry 30(Fall):328–333.

Robinson, J.C. (1996), “Decline in hospital utilization and cost inflation under managed care in California”, JAMA 276(13):1060–4.

Robinson, J.C., and L.B. Gardner (1995), “Adverse selection among multiple competing health maintenance organizations”, Medical Care 33(12):1161–1175.

Robinson, J.C., L.B. Gardner and H.S. Luft (1993), “Health plan switching in anticipation of increased medical care utilization”, Medical Care 31(1):43–51.

Rosenberg, S., D. Allen, J. Handte, T. Jackson, L. Leto, B. Rodstein, S. Stratton, G. Westfall and R. Yasser (1995), “Effect of utilization review in a fee-for-service health insurance plan”, The New England Journal of Medicine 333(20):1326–30.

Russell, L. (1986), Is Prevention Better than Cure? (Brookings Institution, Washington, DC).

Salop, S.C. (1976), “Information and monopolistic competition”, American Economic Review 66:240–245.

Schlesinger, M., D. Blumenthal and E. Schlesinger (1986), “Profits under pressure. The economic performance of investor-owned and nonprofit health maintenance organizations”, Medical Care 24(7):615–27.

Selden, T. (1990), “A model of capitation”, Journal of Health Economics 9(4):397–409.

Sisk, J.E., S.A. Gorman, A. Lenhard-Reisinger, S.A. Glied, W.H. DuMouchel and M.M. Hynes (1996), “Evaluation of Medicaid managed care: satisfaction, access and use”, JAMA 276:50–55.

Smith, D.G. (1997), “The effects of preferred provider organizations on health care use and costs”, Inquiry 34(Winter):278–287.

Starr, P. (1981), The Social Transformation of American Medicine (Basic Books, New York).

Stearns, S., B. Wolfe and D. Kindig (1992), “Physician responses to fee-for-service and capitation payment”, Inquiry 29(4):416–25.

Stern, R.S., P.I. Juhn, P.J. Gertler and A.M. Epstein (1989), “A comparison of length of stay and costs for health maintenance organizations and fee-for-service patients”, Archives of Internal Medicine 149:1185–1188.

Strumwasser, I., N.V. Paranjpe, D.L. Ronis, J. McGinnis and D.W. Kee (1989), “The triple option choice: self-selection bias in traditional coverage, HMOs, and PPOs”, Inquiry 26(4):432–41.

Sturm, R., C.A. Jackson, L.S. Meredith, W. Yip, W.G. Manning, W.H. Rogers and K.A. Wells (1995), “Mental health care utilization in prepaid and fee-for-service plans among depressed patients in the Medical Outcomes Study”, Health Services Research 25:319–340.

Sullivan, C., and T. Rice (1991), “The health insurance picture in 1990”, Health Affairs 10(2):104–115.

Szilagyi, P.G., K.J. Roghmann, H.R. Foye, C. Parks, J. MacWhinney, R. Miller, L. Nazarian, T. McInerny and S. Klein (1990), “The effect of independent practice association plans on use of pediatric ambulatory medical care in one group practice”, JAMA 263:2198–2203.

Tussing, A.D., and M.A. Wojtowycz (1994), “Health maintenance organizations, independent practice associations, and Cesarean section rates”, Health Services Research 29(1):75–93.

Udvarhelyi, I.S., K. Jennison, R.S. Phillips and A.M. Epstein (1991), “Comparison of the quality of ambulatory care for fee-for-service and prepaid patients”, Annals of Internal Medicine 115:394–400.

van de Ven, W.P.M.M., and R.P. Ellis (2000), “Risk adjustment in competitive health plan markets”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 14.

Ware Jr., J.E., R.H. Brook, W.H. Rogers, E.B. Keeler, A.R. Davies, C.D. Sherbourne, G.A. Goldberg, P. Camp and J.P. Newhouse (1987), “Health outcomes for adults in prepaid and fee-for-service systems of care: results from the health insurance experiment”, RAND R-3459-HHS (RAND Corporation, Santa Monica, CA).

Weiner, J.P. (1994), “Forecasting the effects of health reform on US physician workforce requirement. Evidence from HMO staffing patterns”, JAMA 272(3):222–30.

Weiner, J.P., and G. de Lissovoy (1993), “Razing a tower of babel: a taxonomy for managed care and health insurance plans”, Journal of Health Politics Policy and Law 18(1):75–103.

Welch, W.P. (1984), “HMO enrollment: a study of market forces and regulations”, Journal of Health Politics, Policy, and Law 8(4):743–758.

Welch, W.P. (1985), “Health care utilization in HMOs: results from two national samples”, Journal of Health Economics 4(4):293–308.

Wells, K.B., S.D. Hosek and M.S. Marquis (1992), “The effects of preferred provider options in fee-for-service plans on use of outpatient mental health services by three employee groups”, Medical Care 30:412–427.

Wheeler, J.R.C., and T.M. Wickizer (1990), “Relating health care market characteristics to the effectiveness of utilization review”, Inquiry 27(4):344–351.

Wholey, D., R. Feldman and J. Christianson (1995), “The effect of market structure on HMO premiums”, Journal of Health Economics 14(1):81–105.

Wholey, D., R. Feldman, J. Christianson and J. Engberg (1996), “Scale and scope economies among health maintenance organizations”, Journal of Health Economics 15(6):657–684.

Wickizer, T.M. (1992), “The effect of utilization review on hospital use and expenditures: a covariance analysis”, Health Services Research 27(1):103–21.

Wickizer, T.M., R.C.J. Wheeler and P.J. Feldstein (1989), “Does utilization review reduce unnecessary hospital care and contain costs?”, Medical Care 27:632–647.

Wouters, A.V. (1990), "The cost of acute outpatient primary care in a preferred provider organization", Medical Care 28:573–585.

Yelin, E.H., L.A. Criswell and P.G. Feigenbaum (1996), "Health care utilization and outcomes among persons with rheumatoid arthritis in fee-for-service and prepaid group practice settings", JAMA 276:1048–1053.

Yelin, E.H., M.A. Shern and W.V. Epstein (1986), "Health outcomes for a chronic disease in prepaid group practice and fee-for-service settings", Medical Care 24:236–246.

Zelman, W.A. (1996), The Changing Health Care Marketplace (Jossey-Bass, San Francisco).

Zwanziger, J., and R. Auerbach (1991), "Evaluating PPO performance using prior expenditure data", Medical Care 29(2):142–51.

Zwanziger, J., and G.A. Melnick (1989), "The effects of hospital competition and the Medicare PPS program on hospital cost behavior in California", Journal of Health Economics 7:301–320.

This Page Intentionally Left Blank

Chapter 14

RISK ADJUSTMENT IN COMPETITIVE HEALTH PLAN MARKETS*

WYNAND P.M.M. VAN DE VEN

Erasmus University Rotterdam

RANDALL P. ELLIS

Boston University

Contents

*The authors gratefully acknowledge very useful comments from: Arlene Ash, Ernie Berndt, Richard Frank, Jeremiah Hurley, Leida Lamers, Tom McGuire, Joe Newhouse, Erik Schokkaert, Matthias von der Schulenburg, Erik van Barneveld, René van Vliet and Peter Zweifel.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

In the 1990s many countries have chosen to use prospective payment arrangements for health plans (e.g., health insurers, sickness funds or HMOs) together with health plan competition, as a means of creating incentives to be cost conscious, while preserving quality, innovation and responsiveness to consumer preferences. Risk adjustment is an important mechanism for attenuating problems that threaten the effectiveness of this strategy for resource allocation in health care. Without adequate risk adjustment, competing health plans have incentives to avoid individuals with predictable losses and to select predictably profitable members. This selection and the resulting risk segmentation can have adverse effects in terms of access to care, quality of care and efficiency in the production of care.

This chapter first provides a conceptual framework for thinking about risk adjustment. Second, it gives an overview of the progress developing risk adjustment models in recent years. Third, several forms of risk sharing are discussed, which can be used as a tool for reducing selection in case of imperfect risk adjustment. Fourth, an overview is given of the current practice of risk adjustment and risk sharing in 11 countries. Finally some directions for future research are discussed.

JEL classification: C10, D82, G22, I10, I11, I18

1. Introduction

More than any other good routinely used by consumers, health care expenditures are characterized both by large *random* variation as well as large *predictable* variation across individuals. Such differences create the potential for large efficiency gains due to risk reduction from insurance and raise important concerns about fairness across individuals with different expected needs for services. In this chapter, we examine the principles and practice of risk adjustment and how it may contribute to both efficiency and equity in competitive health plan markets.

Because the term "risk adjustment" is used in different contexts to mean different things, we begin by defining how we shall use the term. Throughout this chapter we use *risk adjustment* to mean the use of information to calculate the expected health expenditures of individual consumers over a fixed interval of time (e.g., a month, quarter, or year) and set subsidies to consumers or health plans to improve efficiency and equity. By this definition we intend to exclude the use of risk modeling for profiling, or measuring resources defined over episodes of treatment or episodes of illness [see Iezzoni (1994) for discussion of this practice], which is also known as severity adjustment. We also exclude the adjustment of expected expenditures at the family, group, or plan level, such as is commonly done by actuaries using occupational and demographic averages. Although risk adjusters may be used by insurers for risk-rating their premiums, we do not focus on this application. Risk adjusters may also be used for monitoring, or for internal financing decisions within managed care organizations (e.g., risk-adjusted capitation payments and shared risk pools), or included as control variables in prediction models with other objectives, but these uses are not the focus of our chapter.

As our title indicates, we focus our discussion on risk adjustment in the context of competitive health plan markets. By *competitive*, we mean markets in which individual consumers have a periodic choice of health plan and health plans may take actions, such as designing, pricing and marketing their products, to attract or repel enrollees. By *health plan* we mean a risk-bearing entity that performs at least some insurance function – i.e. it bears some or all of the financial risk associated with the random variation in health expenditures across individuals. Health plans may also manage or provide health care, and this can influence how risk-adjusted payments should be made; however we focus primarily on plan-level rather than provider-level incentives. Examples of health plans are: private health insurance companies, sickness funds (Israel, Netherlands), managed care organizations like Health Maintenance Organizations (US) and capitated provider groups like general practitioner-fundholders (UK).

1.1. Efficiency and fairness

Imperfect information is a serious problem in health plan markets. Yet efficiency and equity issues would need to be addressed even in a world with *perfect information*, since plans will face large differences in expected health costs due to heterogeneity in demographics and the incidence of illness. A competitive market forces health plans to break even, in expectation, on each insurance contract offered. If a health plan does not adjust its premium for a risk factor that is known either to individuals or to plans, then low-risk individuals will tend to choose a competing plan that offers a lower premium or a contract specifically designed to attract low risk individuals. Consequently the first plan, left with only high-risk individuals, will have to increase its premium. In this way, in the absence of any restrictions on premium rates, a competitive health plan market will tend to result in plans' charging risk-adjusted premiums that differentiate according to the individual consumer's risk. This is called the equivalence principle.[1] Risk-adjusted premiums are the norm, not the exception, in competitive markets, and in the absence of regulation, health plans will tend to charge premiums that differ across both observable risk factors and benefit packages designed to attract specific risk types.

This raises the equity question: is this fair? As we document below, risk-adjusted premiums can easily differ by a factor of ten or more for demographic risk factors such as age, and factors of 100 or more once health status is also taken into account. Almost universally, people agree that premiums which reflect such large differences are not fair, and that cross subsidies are needed.

In addition to the equity concerns there are also efficiency problems: consumers are not permitted to equalize the marginal utility of income across different annual or lifetime health profiles. Risk averse consumers would like to buy insurance against the risk of becoming a bad risk in the future. However, in practice there is no market for such insurance. The welfare losses resulting from this inefficiency[2] of a competitive health plan market are discussed in Cutler and Zeckhauser (2000) [see also Newhouse (1984), Pauly (1992), Diamond (1992), Cochrane (1995)].

Problems are exacerbated if there is *asymmetric information*, with consumers knowing more than health plans. This asymmetry can create moral hazard and adverse selection inefficiencies. Consider the moral hazard problem that arises when consumers have private information about their health care needs which is not known to the health plan. If consumers are fully insured against financial risks, then they will tend to over consume health services because of the moral hazard problem [Arrow (1963), Zweifel and Manning (2000) and Cutler and Zeckhauser (2000)]. To reduce this problem, health plans typically try to constrain the use of services through demand-side incentives (such as user fees, deductibles, copayments, waiting time, etc.) or supply incentives (supply-side cost sharing, case management, selection of providers, etc.). Unfortunately, the same tools that health plans use to offset the patient-level moral hazard problem can

[1] We assume that, up to a sophisticated level of risk-rating, the costs of risk-rating are not prohibitively high. If risk-rating becomes too costly, technically infeasible, or politically unacceptable, the equivalence principle may force health plans to exclude from coverage the costs related to some preexisting medical conditions or to refuse to contract with high-risk individuals altogether.

[2] The fairness issue discussed above can alternatively be thought of as an inefficiency because there is no market for buying insurance against a bad draw from the gene pool (i.e. lifetime insurance).

also be used to compete for profitable enrollees: competing health plans will design their plans so as to attract a favorable selection of enrollees.[3]

1.2. The supply price and demand price of insurance

The payment received by a health plan for an individual enrollee need not be the same as the payment made by that same enrollee: the supply price and the demand price for health insurance can differ. This important distinction is often missed. Note that we refer here to the health plan price, commonly called the insurance premium, not the price paid at the time health services are received. Subscribers rarely pay the full insurance premium. Instead, with only a few exceptions, a substantial part of the insurance premiums tend to be paid by a *sponsor*.[4] The sponsor acts as a broker in structuring coverage, contracting with and regulating health plans, and managing enrollment. The sponsor also reallocates the burden of health plan premiums across consumers, and enters into risk-sharing arrangements with health plans [cf. Enthoven (1988)]. The demand price and the supply price will differ only if a sponsor redistributes the financial burden.

The sponsor can be of many types – an employer, a coalition of employers, a government agency, a nonprofit organization, or a distinct insurance entity empowered to use coercion to redistribute risk. Examples of sponsors are the Health Care Financing Administration in the US which negotiates "at-risk" contracts with HMOs for Medicare beneficiaries and the government agencies which regulate and even pay the competitive sickness funds in several European countries. In many countries, the sponsor role is fulfilled by the government agency that regulates access to individual (or small group) private health insurance coverage in a competitive market.[5] In the US, the role of sponsor is also fulfilled by (large) employers who offer group health insurance to their employees.[6]

There is no widely used terminology for distinguishing the demand and supply prices for health plans, so we define our own. On the demand side, we call payments made by the consumer *contributions*, the two most important of which are *premium contributions* – the contribution of a consumer towards his own health insurance coverage – and *solidarity contributions*, which are made toward all consumers covered by the sponsor (see Figure 1). The term solidarity contribution derives from a substantial literature in Europe on the 'solidarity principle', which holds that high-risk individuals should

[3] A point we develop more fully below is that this selection can arise either because of asymmetric information, or because of regulation-induced pooling of people with different known risks.

[4] Newhouse (1996) calls the sponsor the regulator. We use sponsor to highlight the redistribution role, not just the fact that the sponsor may also regulate the characteristics of health plans that are offered.

[5] For example, in Australia, Chile, Ireland, the Netherlands, Portugal, Spain and the US.

[6] In a later section of this chapter, we discuss the implications of the sponsor being voluntarily chosen by consumers, but for the most part we focus on the common case in which there is no consumer choice of sponsor.

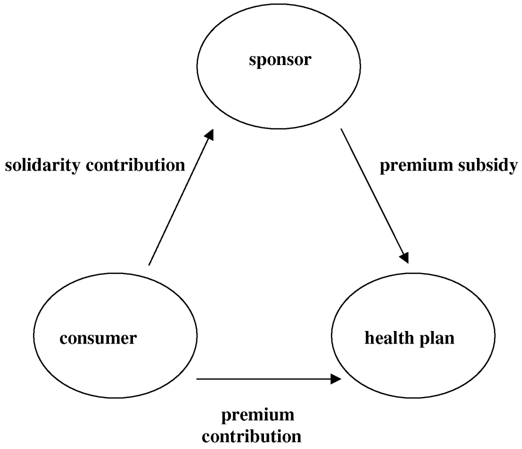


Figure 1. Risk adjustment system (Modality A).

receive a subsidy to increase their access to health insurance coverage [see, e.g., Hamilton (1997), and Chinitz et al. (1998)].[7] On the supply side, we call the payments made by the sponsor *subsidies*.[8] The most important type of sponsor subsidy is the *premium subsidy*, an ex-ante subsidy mostly paid directly to the health plan.[9] The sum total of ex-ante payments received by the health plan for one consumer, i.e. the premium contribution plus the premium subsidy, is the (supply side) *health plan premium*, or simply the *premium*. As discussed below, a wide variety of mechanisms are used for calculating the consumer contributions and the premium subsidies, as well as for organizing the actual payment flows in practice.

1.3. The role of the sponsor

The sponsor plays a crucial role in enabling health plan premiums to be risk-adjusted (reflecting the expected health cost of the plans' enrollees) while not insisting that payments by individuals reflect each person's own expected cost. One mechanism for doing this is to *risk adjust* the premium subsidies to competing health plans while charging consumers a solidarity contribution that does not reflect the person's own expected cost.

[7] In the US there does not appear to be any widely-used terminology for describing the normative concept that high-risk individuals should receive a cross-subsidy from low-risk individuals.

[8] Researchers in the US are more used to thinking of the employer as being the sponsor, and focusing on employee and employer contributions toward the premium. This terminology ignores the fact that the sponsor need not to be an employer, and that the consumers may make other payments besides the premium contributions.

[9] Another type of sponsor subsidy is the ex-post payments made by the sponsor to the health plans because of the risk-sharing arrangements between the sponsor and the health plans (see Section 4).

Another mechanism to reduce the variation in contributions across consumers is to *regulate* the rate classes, plan features, and premium contributions that health plans are allowed to charge. As we highlight below, it is difficult for the sponsor to fully risk adjust health plans subsidies, but it is also difficult to fully regulate all of the dimensions in which health plans will try to differentiate their plan features.

Sponsors have many mechanisms for allocating financial burdens among consumers through the contribution side of the market, as well as great flexibility in redistributing financial revenues among health plans on the supply side of the market. Note that once the linkage between the contribution and expected health care use is broken by the sponsor, then solidarity contributions can be based on information that may have little relation to future health costs, such as income. It is common for solidarity contributions to be income-based, or to be a flat payment that does not vary across plans with different benefit designs. Throughout this chapter, we focus primarily on so-called "risk-solidarity", that is solidarity between high- and low-risk individuals. Solidarity between high- and low-income individuals, so-called "income-solidarity", is a redistribution concept that varies across countries and is relatively independent of the incentive issues and fairness across risk types that is the primary focus here.

1.4. Policy relevance

The policy relevance of an adequate risk adjustment mechanism has increased during the 1990s as many countries make their individual health insurance market more competitive or reform their already competitive markets in order to increase access to coverage for high-risk individuals.[10] Many countries have chosen to use prospective payment arrangements (pure or otherwise) for health plans as a means for creating incentives to be cost conscious, together with competition among health plans as a tool for preserving quality, innovation and responsiveness to consumer preferences. Risk adjustment is a key strategy for attenuating problems that threaten the effectiveness of this strategy for resource allocation in health care. Without adequate risk adjustment it is hard, if not impossible, to achieve both efficiency and fairness objectives in a competitive health plan market.[11]

Despite its increasing relevance, the practical application of risk adjustment is still at early stages. For reasons that are not clear to us, most sponsors around the world do not use risk adjustment. Instead, they regulate the dimensions along which health plans are allowed to compete. They force plans to pool consumers into a relatively small number

[10] For example, Belgium, Colombia, the Czech Republic, Germany, Ireland, Israel, The Netherlands, Poland, Russia, Switzerland and the US.

[11] Risk adjustment is also relevant for a competitive provider market where risk-adjusted payments are used – often by a large monopsonistic insurer (e.g., a governmental agency) – to push financial risks all the way down to providers. For example, in the GP fundholder system of the UK, primary care physicians receive a risk-adjusted capitation payment for some or all of the follow-up care of their patients. In the terminology of this chapter, such a GP fund holder is considered a health plan.

of rate categories and regulate the characteristics of contracts offered to each of these categories.[12]

Whereas a system of risk-adjusted subsidies attempts to provide *explicit* subsidies to high-risk individuals, the effect of regulating plan design and restricting the variation of premium contributions is to create *implicit* cross-subsidies from low-risk to high-risk individuals. Although this risk pooling may foster the solidarity principle, it creates *predictable* losses for health plans on their high-risk individuals. In so doing, it creates incentives for health plans to avoid individuals with predictable losses and to select predictably profitable insureds.[13] This selection and the resulting risk segmentation can adversely affect access to care, quality of care and efficiency (see Section 2.5).

If premium subsidies cannot be adequately risk adjusted or if loosening the restrictions on the variation of the premium contributions is not socially acceptable, the adverse effects of selection may also be reduced by various forms of *ex post* risk sharing between the sponsor and the health plans. *Risk sharing* implies that the health plans are retrospectively reimbursed by the sponsor for some of their costs. Although risk sharing effectively reduces the health plans' incentives for selection, it also reduces their incentives for efficiency [Newhouse (1996)].

The conclusion is that in competitive health plan markets – given that risk-adjusted subsidies will always be imperfect – there will always be selection incentives. Because the effects of selection have consequences for both efficiency and fairness, we are confronted with a complicated *tradeoff between efficiency and fairness objectives*. The relevance of an adequate risk adjustment mechanism is that the better the explicit subsidies are adjusted for relevant risk factors, the less severe is the tradeoff. In theory, perfect risk adjustment can eliminate this tradeoff entirely.

1.5. Outline

This chapter gives an overview of all aspects of risk adjustment in competitive health plan markets. We also discuss at length the major mechanisms that can be either a complement or an alternative to risk adjustment, namely plan regulation, carveouts, and ex post risk sharing. The chapter is relevant for voluntary health plan markets as well as for mandatory health plan membership.

The organization of this chapter is as follows. Section 2 presents a conceptual framework of risk adjustment. Section 3 extensively discusses the state of the art of empirical

[12] Although these regulations reduce the ability of plans to select profitable enrollees, they increase the incentive for health plans to try to do so.

[13] If the sponsor (e.g., an employer) contracts with only one health plan, risk adjustment is not needed to prevent selection by the plan. However, if the single health plan offers its beneficiaries a menu of several options to choose among, selection may occur *within* the health plan. Even if there is only one plan and no choice by enrollees, risk adjustment may still be used within the health plan to allocate payments among providers. We do not focus attention on how risk adjustment may be used for these internal financing decisions within health plans.

risk adjusters, i.e. the predictors used in risk adjustment. Section 4 discusses several forms of risk sharing, which can be used as a tool for reducing selection. The practice of risk adjustment and risk sharing in several countries is discussed in Section 5. Finally some directions for future research are discussed in Section 6.

2. Conceptual aspects of risk adjustment

In Subsection 2.1 we briefly consider each of the three payment flows identified in the preceding section: risk-adjusted premium subsidies, solidarity contributions, and premium contributions. In Section 2.2 we discuss some conceptual aspects of how to calculate the risk-adjusted subsidies. In Subsections 2.3–2.6 we discuss the consequences of regulations that sponsors may implement as a substitute for, or as a complement to, risk adjustment.[14]

2.1. Payment flows

2.1.1. Risk-adjusted premium subsidies

The central feature of any risk adjustment system is a risk-adjusted premium subsidy (or voucher) from the sponsor to each consumer or to high-risk consumers only. In most countries the sponsor pays the subsidy directly to the consumer's health plan and thereby lowers the consumer's premium contribution (see Figure 1). The risk-adjusted premium subsidy has several general properties that are worth highlighting. The subsidy is generally worth a specified amount of money, dependent only on the individual's relevant risk characteristics.[15] We assume that the subsidy does not depend on the premium that the consumer pays or the specific health plan chosen by the consumer. The subsidy may be earmarked for the purchase of a specified health plan with specified coverage features, or may be portable across plans.[16] The risk-adjusted subsidy is not transferable. The information that may be used by the sponsor to calculate the risk-adjusted subsidy is discussed in Section 2.2 and Section 3.

[14] Section 2 is partly based on Van de Ven et al. (1997).

[15] That is the risk factors for which solidarity is desired (see Section 2.2).

[16] The sponsor may define a *minimum* benefits package which health plans may extend with additional benefits (like, e.g., the US Medicare risk contracts) or the sponsor may require all health plans to offer a fully *standardized* benefits package (as in the Dutch sickness fund system). The advantage of a minimum package is that health plans can be responsive to consumer preferences (no one-size-fits-all coverage). Disadvantages of a minimum package are that (1) the benefits package can be used as a tool for cream skimming; (2) it reduces the transparency of health plan products; (3) it reduces the price competition because of segmentation of the market.

2.1.2. Solidarity contributions

Solidarity contributions are payments made by consumers toward the health needs of everyone covered by the sponsor, not payments made for a consumer's own health care. Such payments may reflect information that is largely unrelated to the individual's health care needs (income, or wealth). Solidarity contributions are mandatory payments by enrollees, made independently of the plan or benefit features selected.

Although in Figure 1 for simplicity we show the premium subsidies as financed entirely by mandatory solidarity contributions from enrollees, the sponsor's outlay may also include financing from other sources. In the US the risk-adjusted subsidies to HMOs with Medicare risk contracts (based on an Adjusted Average Per Capita Cost calculation) are financed primarily out of federal payroll taxes. In the Netherlands the risk-adjusted subsidies are supported from a combination of earmarked income-related enrollees' contributions, general taxes, and a mandatory levy on the premium of each private health insurance contract.

In some countries, such as the US, some individuals get to choose their sponsor when they change employment (e.g., their employer, or a sponsor for the unemployed). When solidarity contributions or premium subsidies differ across sponsors for identical plans, then individuals have an incentive to select a sponsor that contributes more generously. Such differences can also make enrollees reluctant to leave a sponsor with a favorable solidarity contribution. For example, in the US, unemployed persons often have more generous coverage through the Medicaid programs than do low-wage workers. See Gruber (2000) for a discussion of distortion in labor markets resulting from concerns about loss of sponsorship.

2.1.3. Premium contributions

A premium contribution by an enrollee is a payment for his or her own health plan. A consumer's premium contribution equals the health plan's premium minus the premium subsidy. Differences in expected costs across individuals may be reflected either in differences in premium subsidies or in differences in premium contributions. If the premium subsidies are adjusted for differences in health status across individuals, the premium contribution will be unrelated to an enrollee's health status. If the premium subsidies are not adjusted for differences in plan benefit features or efficiency of provision, these differences in expected costs will typically be reflected in the premium contributions.

In the sickness fund system in the Netherlands the risk-adjusted subsidy equals the risk-adjusted predicted per capita costs at the national level minus a fixed amount that is identical for all persons. In the US Medicare system, risk-contracting HMOs are paid 95 per cent of the risk-adjusted predicted per capita costs. In both countries the health plans are allowed to make up for any potential shortfall in this premium subsidy by charging a *community-rated* (i.e. the same) premium contribution to all enrollees who

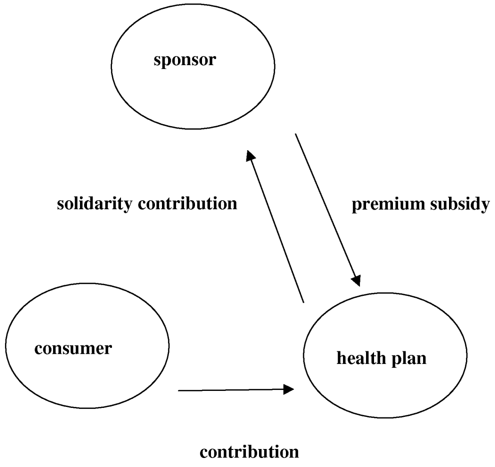


Figure 2. Risk adjustment system (Modality B).

choose the same plan.[17] Each health plan is free to set its own premium contribution. In the Netherlands in 1999 the premium contributions varied between 345 and 441 Dutch guilders per enrollee per year.[18] In the US in 1996 63 per cent of Medicare risk-contract enrollees were quoted a zero premium contribution. The other 37 per cent of enrollees paid an average premium contribution of 162 US$ per enrollee per year [Lamphere et al. (1997)]. In other countries, (e.g., Israel and Colombia) the sponsor requires the premium contribution to be zero for all enrollees. That is, the health plan premiums equal the risk-adjusted premium subsidies.

2.1.4. *Different modalities of payment flows*

Figure 1 shows schematically how the risk adjustment system is applied in Medicare in the US and the sickness fund system in the Netherlands. We refer to such an implementation as modality A. However, actual payment flows in a risk adjustment system need not follow this pattern. One alternative is that the premium subsidies go to the consumer, who then pays the total premium directly to the health plan (a so-called "voucher model"). A second alternative is that the sponsor also collects the premium contributions and transfers them to the health plans. This alternative is applied by some employer purchasing coalitions in the US that use risk adjustment. A third alternative, depicted in Figure 2 which we call modality B, is that the consumer pays the total contribution, i.e. solidarity contribution plus premium contribution, to the health plan and

[17] In the specific case of community-rated premium contributions the premium subsidy is often referred to as "*capitation*".

[18] I.e. between 172 and 220 US$ per year (1999 exchange rate).

that the health plan transfers the solidarity contributions to the sponsor. To reduce the actual flows of money, each health plan and the sponsor net the difference of all the solidarity contributions and premium subsidies for all members of a health plan. This way of organizing the payment flows in a risk adjustment system is being applied in the mandatory sickness fund insurance in Germany and Switzerland and in the voluntary health insurance in Ireland. This modality of organizing the payment flows was also proposed by the White House Task Force on Health Risk Pooling (1993). In Germany the contribution is a certain percentage of the consumer's income. The sponsor requires this percentage to be the same for all members of a sickness fund, but allows it to differ across sickness funds. In Switzerland and Ireland the contribution must be community-rated per health plan (in Switzerland: per region).

As the figures suggest, the direct payment from the consumer to the health plan in Modality A is considerably less than in Modality B. Hence, cost savings by health plans will have a much larger proportional effect on the level of direct payments in Modality A than in Modality B. Both the difference in proportional change and in absolute level of direct payments may result in different responses by consumers [Buchmueller and Feldstein (1997)].

2.2. *Subsidy formula*

The formula to calculate the risk-adjusted premium subsidies and solidarity contributions can in principle be independent of how the actual payment flows are organized. In practice, however, there is often a relation. Assume, for example, that age is the only risk adjuster. In countries that use modality A (US, the Netherlands, Israel) the health plans receive an age-related subsidy for each consumer, while in countries that use modality B (Germany, Switzerland and Ireland) only health plans with an overrepresentation of elderly receive a subsidy and only health plans with an underrepresentation of elderly pay a risk-adjusted solidarity contribution.

In this chapter, for convenience, we assume that risk solidarity is fully reflected in the risk-adjusted premium subsidy, and that the solidarity contribution is not risk adjusted. Broadly speaking, this assumption is not restrictive and sacrifices little generality.[19]

For the calculation of the risk-adjusted premium subsidies a central question is: on what costs should the subsidies be based? We shall call these costs the *acceptable costs*. Acceptable costs can be conceptualized as those generated in delivering a "specified basic benefit package" containing only medically necessary and cost-effective care. In principle, the cost of hospitalizations could be excluded when only day surgery is medically indicated; as could the cost of psychiatric care when care by a psychologist is

[19] Assume, for example, that age is the only risk adjuster, and that E_i is the average expenditures in age group i, with E the grand average. Assume that the solidarity contribution is $E - E_i$ and has to be paid only by individuals belonging to an age group i with $E > E_i$; and the risk-adjusted subsidy is $E_i - E$ and is received only by individuals belonging to age group i with $E < E_i$. This situation is identical to the situation that each individual pays a non risk-adjusted solidarity contribution E and receives a risk-adjusted subsidy E_i.

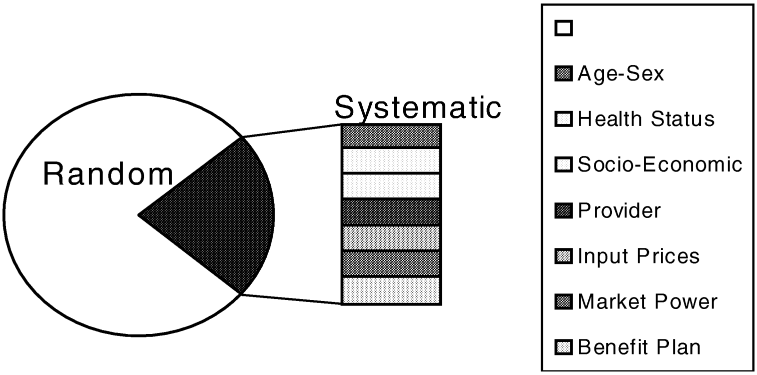


Figure 3. Factors explaining variation in health spending.

appropriate. Because the cost level of such a benefits package is hard to determine, in practice subsidies are based on observed expenses rather than needs-based costs. This is true of social health insurance programs such as Medicare in the US or the sickness fund systems in Germany, Israel and the Netherlands.

Observed expenses are determined by many factors, not all of which need to be used for calculating the risk-adjusted subsidies. Ideally, subsidies should only be adjusted for those risk factors for which solidarity is desired. Society has to decide for which risk factors, and to what extent, it seeks solidarity. Figure 3 summarizes seven classes of risk factors that explain variations in health spending across individuals. The first three groups are characteristics of individuals: age and sex; health status;[20] and socio-economic factors such as lifestyle, taste, purchasing power, religion, race, ethnicity, and population density. The fourth group includes all provider characteristics, such as practice style and whether there is an oversupply of providers or facilities. Input prices are a characteristic of the region in which the providers are located, and are largely exogenous to the patient or provider. The final two groups are characteristics of the health plan. By market power, we mean to indicate the health plan's ability to negotiate price discounts. Benefit plan features include conventional demand side features such as deductibles, copayments and decisions about covered services, but also include supply side features such as utilization review, various health management strategies, and characteristics of the contracts and financial incentives between plans and providers. Even after controlling for these seven systematic factors that affect costs, considerable variation in spending across individuals will remain, which ex ante is random and will be

[20] In this chapter we will use the term health status without going into details concerning either the difference between health status and need, or the various concepts of need, such as normative need, felt need, expressed need and comparative need [Bradshaw (1972)]. For a discussion of the concepts morbidity, need and demand, see, e.g., Ashley and McLachlan (1985).

averaged out by health plans by risk pooling. We recognize that not all of these factors are independent, and indeed some are reasonably thought of as partially endogenous to others (we return to this in Section 3.4). We use X to denote the full set of risk factors that predict variations in health spending across individuals.

Should all risk factors X that are observed by the sponsor be used to calculate risk-adjusted subsidies? The answer may vary with the sub-population, context and country. In the US, on the one hand, the widespread practice of experience rating health premiums at the employer level is consistent with the view that health premium subsidies by the sponsor (employer) should reflect just about any information that explains variation in spending. On the other hand, individual premium contributions differ greatly across sponsors, and sponsors differ dramatically in how they calculate their subsidy payments to health plans. In Europe, national solidarity is more prized, and there is greater standardization of benefits and sponsor subsidy formulas within each country. Europe is characterized by narrower ranges of individual premium contributions than in the US.

Despite differences in the specifics, most systems implicitly seek to achieve solidarity along some specified dimensions. We divide the risk factors X into two subsets: those factors for which solidarity is desired, the S-type; and those factors for which solidarity is not desired, the N-type. In most societies age, sex and health status are S-type risk factors. Differences in input prices are also likely to be considered S-type risk factors. It may be argued that differences in costs caused by the other risk factors can be influenced by the insurer or by the insured, and should be reflected in the premium contribution. To the extent that the division into S-type and N-type factors is not clear, society should make an explicit choice. For example, hospitalizations for lung cancer, AIDS, obesity, and skiing accidents are all health-related as well as life-style related risk factors. To the extent that consumers and health plans cannot be held responsible for cost differences or to the extent that society decides that solidarity is desired, the subsidies could be adjusted for these factors.

Assume that $E(X)$ is the best estimate of the expected expenses for a person with risk characteristics X in the next contract period. An estimate of the acceptable cost level $A(X)$, which serves as the basis for setting the sponsor subsidies, could then be $E(X)$ with the values of the N-type risk factors set at an acceptable level (e.g., the acceptable level of the price or supply of health care or the acceptable practice style).[21] The risk-adjusted premium subsidy could then be a function of $A(X)$, e.g., it could be $A(X)$, or $A(X)$ minus a fixed amount (as in the Netherlands), or a certain percentage of $A(X)$ (as in the US Medicare). The calculation of $A(X)$ will be discussed in Section 3.

2.3. Regulation

If health plans were fully free to set their risk-adjusted premiums, the set of rating-factors and the resulting range of premiums could be substantial. For example, the

[21] For example, in Belgium the weights of the subsidy formula are estimated based on the relevant risk factors including indicators of the supply of health care facilities. However, when calculating the subsidies the differences in supply, an N-type factor, are ignored [Schokkaert and Van de Voorde (1998)].

premium for private health insurance in the Netherlands may be related to age, gender, family size, region, occupation, length of contract period, individual or group contract, the level of deductible, health status at time of enrollment, health habits (smoking, drinking, exercising) and – via differentiated bonuses for multi-year no-claim – of prior costs. Also in the US the premiums for individual health insurance are substantially risk-rated. Insurers commonly use age, gender, geographic area, tobacco use and family size as risk adjusters to determine standard premiums; and dependent on the applicant's health status insurers may charge premiums up to seven times the standard rates [US General Accounting Office (1996, 1998)]. In a competitive health plan market with unregulated premiums, the maximum premium for full health plan coverage (i.e. without cost-sharing) could be expected to exceed the average premium for the same product by a factor 10 or more, with a minimum premium of around 10 per cent of the average.[22]

To what extent is a system of risk-adjusted premium subsidies able to reduce such a range of consumer payments? In most countries that have implemented a system of risk-adjusted subsidies in a competitive health plan market, age and gender are used as risk adjusters, sometimes supplemented with an indication of disability (the Netherlands) and institutional and welfare status (US). Region often is a controversial candidate for being a risk adjuster, since it can either reflect input cost variation (usually a solidarity factor) or practice style variation (which many may consider undesirable). Risk adjustment models that use only these variables routinely do a poor job. For example, in a simulation based on a simple premium model and subsidy formula, the range of premium contributions was 14,297 Dutch guilders without any risk adjustment versus 11,571 guilders using age and sex to risk adjust [Van de Ven et al. (1997, Table 7)]. Using age and gender for risk adjustment reduced the range of total individual payments only by 20 per cent.

If the resulting range of individual payments is considered to be too large, the sponsor may combine the system of risk-adjusted subsidies with restrictions related to premium contributions and with a periodic open enrollment for a specified basic health plan coverage. A *periodic open enrollment* requirement implies that during the open enrollment period, for example one month every year, consumers are allowed to change plans and each health plan must accept anyone who wants to join.

Restrictions related to the premium contributions can take several forms: community rating, a ban on certain rating factors (for example health status, genetic information, duration of coverage, or claim experience) or rate-banding (i.e. a minimum and maximum premium contribution).[23] Community rating implies that a health plan must ask the

[22] In a simulation based on a simple premium model, the minimum premium, the average premium and the maximum premium were respectively 199; 1,500; and 14,496 Dutch guilders [Van de Ven et al. (1997, Table 3)].

[23] Ideally restrictions related to premium contributions should only relate to the S-type risk factors and not the N-type factors. In practice this may be hard to effectuate, especially when S-type and N-type risk factors are correlated.

same premium contribution from each individual, independent of the individual's additional risk characteristics. A variant is adjusted community rating, that is, adjustments in the community rate are allowed for various factors (for example, claim experience) with various limits imposed on the extent to which rates, after adjustment, may vary. Rate banding can take several forms: per health plan or nation-wide; and may specify either an absolute or a relative difference between maximum and minimum premium contribution. An extreme form of restriction on premium contributions is that health plans are required to accept the individual's risk-adjusted premium subsidy, which is determined by the sponsor, as the full premium. This is the case in the competitive social health insurance systems in, e.g., Colombia, Israel and Russia.

The goal of restrictions related to the variation of premium contributions is to fulfil the solidarity principle by creating *implicit* cross-subsidies from low-risk to high-risk individuals (whereas a system of risk-adjusted subsidies implies *explicit* cross-subsidies). However, restrictions on premium contributions also imply predictable profits on low-risk consumers and predictable losses on high-risk consumers. If the premium contributions must be community-rated and if the premium subsidies depend on age and gender only, the health plans will incur substantial predictable losses on their high-risk members. For example, Van Barneveld et al. (1998, Table 2) show that if a health plan were to use information on prior hospitalizations and prior costs in the three preceding years, it could identify a subgroup of 4 per cent of its members whose predicted costs are threefold their average age/gender-adjusted expenses. Another example is that the five per cent of the individuals with the highest health care expenditures in any year can be predicted to have per capita expenditures over (at least) the next four years that are twice their average age/gender-adjusted expenses [Van Vliet and van de Ven (1992, Table 3)]. Ideally, for each health plan the predictable losses on its high-risk members should be compensated by the predictable profits on its low-risk members. However, this ideal situation may not be achieved because of *selection*, i.e. actions[24] by consumers and health plans to exploit unpriced risk heterogeneity and break pooling arrangements [Newhouse (1996)]. Often the term selection is also used to refer to the outcome of these actions. The literature identifies two forms of selection: adverse selection and cream skimming.[25] Because these forms of selection may differ in the consumers' or health plans' actions as well as in their effects on efficiency and fairness, we will discuss each of them.

2.4. Selection

2.4.1. Adverse selection

Adverse selection is the selection that occurs because high-risk consumers have an incentive to buy more coverage than low-risk consumers within the same premium risk

[24] Not including risk-rated pricing by health plans.

[25] For the relevance of the distinction between these two forms of selection, see, e.g., Pauly (1984).

group. A necessary condition for adverse selection to occur is that the consumers themselves know whether they are a high- or low-risk *within* their premium risk group, i.e. consumers must have more information about their future risks than the information health plans use for premium differentiation. As Wilson (1977, pp. 167–168) highlighted, this consumer information surplus vis à vis the health plan may be caused either by regulation or by a limitation of the health plans' knowledge. That is, either restrictions on premium rates or asymmetric information between health plans and consumers may result in similar adverse selection problems. In the case of asymmetric information the health plans may know that consumers vary in the level of risk, but they cannot discern who are the high- and low-risk individuals within a premium risk group. Pauly (1984) referred to this as "true adverse selection". In the case of regulatory restrictions on health plans' abilities to differentiate premiums, the health plans may know the consumer's level of risk, but are not allowed to use this information to set premiums.[26]

Rothschild and Stiglitz (1976) showed that in a market with asymmetric information a competitive equilibrium may not exist. This would be the case if there are relatively few high-risk individuals, which seems a quite realistic assumption for the health plan market.[27] As a result of adverse selection a competitive health plan market may be unstable. Low-risk individuals will persistently (try to) separate themselves from the high-risk individuals by buying new products that are especially designed to lure them from the more heterogeneous risk pool. Premium for the old products will have to rise as they come to be predominantly bought by high-risk individuals. As the low-risk individuals avoid the generous health plans, these plans may be confronted with a fatal spiral of ever rising premiums. Rothschild and Stiglitz showed that if equilibrium exists, high-risk individuals buy full coverage and low-risk individuals buy incomplete coverage (i.e. a separating equilibrium or an "adverse selection equilibrium"). In their model a pooling equilibrium cannot exist.

The strong predictions of the Rothschild–Stiglitz model have subsequently been softened by Wilson (1977), Schut (1995) and Newhouse (1996), among others, who show that pooling equilibria are at least possible. Wilson (1977) shows that if the losses to low-risk individuals from separating themselves from high-risk individuals are greater than the cross-subsidy implied by a pooled equilibrium, then a pooling equilibrium can result. Schut (1995, Chapter 3) shows that costly risk classification may stabilize a competitive health plan market and may result in a Pareto-type welfare improvement. Newhouse (1996) shows that the presence of sufficiently large contracting costs can result in a pooling equilibrium with the low-risk group at its most preferred point and the high-risk group at its most preferred feasible point.

[26] For a discussion of regulation-induced adverse selection see, e.g., Newhouse (1984, p. 99), Pauly (1984) and Keeler et al. (1998).

[27] When applying the Rothschild–Stiglitz theory in our case, we have to interpret "high-risk and low-risk individuals" as "high-risk and low-risk individuals *within* their premium-risk-group" (e.g., an age/gender-group). The Rothschild–Stiglitz theory then applies to the submarket for each premium-risk-group [see footnote 5 in Rothschild and Stiglitz (1976)].

Empirical simulation results by Marquis (1992) suggest that adverse selection is sufficient to eliminate high-option benefit plans in multiple choice markets if health plans charge a single, experience-rated premium. Similar results are found by Keeler et al. (1998). Cutler and Reber (1998) analyzed the health insurance pricing reform by Harvard University in the mid-1990s. Harvard had historically subsidized the most generous plan quite generously at the margin. Under the new policy, Harvard contributes an equal amount per individual/family to each plan regardless of which plan an employee chooses. The plans' premiums are only differentiated for individual/family. Because of adverse selection, the most generous policy could not be sustained under an equal contribution rule (i.e. without risk adjustment). In three years the adverse selection "death spiral" was completed at Harvard. Price et al. (1983) analyzed the instability of the Federal Employees Health Benefits Programme (FEHBP), which offers comprehensive benefits to federal workers and retired employees in the US. All FEHBP-plans are subject to annual open enrollment and the premiums are differentiated only according to single/family (that is community rating by single/family class). Price and Mays (1985) found substantial adverse selection within the FEHBP-market. Price et al. (1983) concluded that the FEHBP's lack of stability raises important questions about the viability of some pro-competition proposals involving multiple-insurer systems.

2.4.2. *Cream skimming*

Cream skimming (or preferred risk selection or cherry picking) is the selection that occurs because health plans prefer low-risk consumers to high-risk consumers within the same premium-risk-group. A necessary condition for cream skimming to occur is that the health plans know that there are high- and low-risk individuals within the premium-risk-groups. Such a situation may be caused by regulation or by transaction costs related to (further) premium differentiation. Even if there is an open enrollment requirement cream skimming can take place in several ways. Health plans may actively cream the preferred consumers and dump nonpreferred consumers [Ellis (1998)]. The precise form of the selection that may occur, depends on the additional information that health plans have. We distinguish three situations.

First, if health plans only know that there are high- and low-risk individuals within the premium-risk-groups, but they cannot ex-ante identify who are the high-risk individuals and they also don't know what the relevant omitted risk factors are, they may structure their coverage such that the plan is unattractive for the high-risk individuals [Newhouse (1996), Glazer and McGuire (forthcoming)]. For example, plans may exclude prescription drugs from coverage or may offer a low-option plan with a high deductible and other cost-sharing. In this way health plans use adverse selection as a tool for cream skimming. They stimulate the different risk groups to reveal themselves. Even if the benefits package and the cost-sharing structure are fully specified, health plans may differentiate their coverage conditions by contracting with different panels of providers. For example, a health plan may contract with a selected panel of providers who work according to strict protocols, or it may apply strict utilization management techniques or contract with managed care firms that do so. Such a health plan is more attractive for the low-risk individuals than for the high-risk individuals within each premium-risk-group. Health plans may also share financial risk with the contracted providers in a way that encourages providers to cream skim. Health plans may also try to attract the low-risks by offering a package deal of health insurance and other forms of insurance or services bought mostly by relatively healthy people, including fitness club memberships.

Second, if health plans know that some omitted risk factors are relevant (e.g., AIDS, disability, prior utilization or hypochondria), but they cannot ex-ante identify the individuals with these characteristics, they may deter the high-risk consumers by selectively not contracting with physicians who have the best reputation of treating patients with such problems. Health plans also could contract with providers who have no interpreters, or whose facilities have no disabled access [Luft (1987)]. They may also select by the design of their supplementary health insurance (no coverage for mental health care, prescription drugs and reconstructive breast surgery) or by putting the brochures of competing health plans on the counter in places where sicker people are likely to be, such as in pharmacies and hospitals.

Third, if health plans can ex-ante identify predictably unprofitable individuals based on certain risk characteristics, they can focus their selection strategy directly on those identifiable individuals, e.g., by providing the high risks with poor quality of care or poor services (such as delayed payments of reimbursement and delayed answers to letters); by not working to coordinate the multiple visits that people with many problems may need; by selective advertising and direct mailing; by contracting with providers who practice in "healthy districts"; by providing the insurance agent with incentives to advise relatively unhealthy persons to buy health insurance from another company; or by a golden handshake for unhealthy members at disenrollment, such as offering an AIDS patient a large sum of money to choose a different plan during the next open enrollment.

2.5. *Effects of selection*

The primary rationale for regulating a competitive health plan market is to provide financial access to health plan coverage for the high-risk individuals. Because regulation induces selection, we have to understand the effects of selection to evaluate the overall effects of regulation.[28]

As stated above, depending on the relative proportion of high-risk individuals within each premium-risk-group and contracting costs, *adverse selection* may either cause a competitive health plan market to be unstable or it may result in a pooling equilibrium or it may result in a separating equilibrium. In the last case high-risk individuals pay a high premium for generous coverage and low-risk individuals pay a low premium for stingy

[28] Another effect of regulation is that it limits health plans in designing and pricing their products (e.g., managed care and no-claim bonuses) such that they reduce undesirable moral hazard.

coverage. So, adverse selection may decrease access to coverage for non-affluent high-risk individuals. The inefficiency that arises in an adverse selection equilibrium is that, depending on the contracting costs, either the low-risks or the high-risks cannot obtain as much coverage as they wish.[29] Another inefficiency arising from adverse selection is the welfare loss due to the potential non-existence of a competitive equilibrium. The continuous exit (bankruptcy) and re-entry of health plans has real social costs.

Even with a periodic open enrollment requirement (to prevent health plans' refusing relatively high-risk individuals) there may be *cream skimming*. First, the larger the predictable profits resulting from cream skimming, the greater the disincentive for health plans to respond to the preferences of high-risk consumers. Health plans may give poor service to the chronically ill and choose not to contract with providers who have the best reputations for treating chronic illnesses. This in turn can discourage physicians and hospitals from acquiring such a reputation. To the extent that a health plan and its contracted providers of care share financial risk, the providers share the incentive to attract profitable patients and to deter patients who generate predictable losses. As Newhouse (1982) highlighted in his famous "mother with an asthmatic child" example, providers of care have subtle tools to encouraging high cost patients to seek care elsewhere, such as keeping the patient in uncertainty about the correct diagnosis, making the patient wait for an appointment, making the patient wait in the office, being discourteous to the patient, or advising chronically ill patients to consult another physician who is "more specialized in treating their disease". Health plans who specialize in care for high-risk patients, have to ask a high premium (because of adverse selection).[30] So, as a result of selection, high-risk patients may either receive poor care and poor service or pay a very high premium for good care and good service. If the regulation implies a nation-wide maximum premium instead of a maximum per health plan, health plans that experience adverse selection cannot raise their premium and will go bankrupt. In that case, it is suicidal for a plan to become known for providing the best care for chronically ill, because it will be flooded by individuals who predictably generate more costs than revenues.

Second, the larger the predictable profits resulting from cream skimming, the greater the chance that cream skimming will be more profitable than improving efficiency. At least in the short run, when a health plan has limited resources available to invest in cost-reducing activities, it may prefer to invest in cream skimming rather than in improving efficiency. In the long run, improving efficiency may be rewarding, independent of the level of cream skimming, as long as these improvements are perceived as desirable by

[29] Of course, the *desired* level of health plan coverage depends on the tradeoff between moral hazard and risk aversion [Zeckhauser (1970), Manning and Marquis (1996)].

[30] In the short run, a small health plan which specializes in care for high-risk patients may be confronted with financial problems if, *after* it has determined its premium for the next contract period, it is flooded by a group of high-risk members.

Table 1
Effects of selection

Effects of *adverse selection*:

- high premiums for high-risk individuals;
- dependent upon the level of the contracting costs either the low-risk individuals or the high-risk individuals cannot obtain as much health plan coverage as they wish;
- welfare loss in the case of an unstable market (including bankruptcy of adversely selected health plans).

Effects of *cream skimming*:

- disincentive for the health plans to respond to the preferences of high-risk consumers;
- incentive to provide poor quality of care and poor service to high-risk individuals;
- disincentives for providers and health plans to acquire the best reputation for treating chronic illness;
- dependent upon the form of premium rate restrictions (per health plan or nation-wide): high premiums for high-risk patients or bankruptcy of non-skimming selected health plans;
- investments in cream skimming have higher returns than investments in improving efficiency;
- investments in cream skimming (e.g., resources to identify and attract high-risk consumers) are a welfare loss.

consumers.[31] Efficient health plans who do not cream skim applicants, may lose market share to inefficient health plans who do, resulting in a welfare loss to society.

Third, while an individual health plan can gain by cream skimming, for society as a whole, cream skimming produces no gains. Thus, any resources used for cream skimming represent a welfare loss.[32]

In sum, regulations that are intended to increase access to coverage for high-risk individuals may instead induce selection efforts with the following unintended effects (see Table 1): problems with financial access to coverage for high-risk individuals, reductions of the quality of certain kinds of care, or reduction of allocative efficiency and efficiency in the production of care. So, given a system of imperfectly risk-adjusted subsidies, there is a *tradeoff between access to coverage and the adverse effects of selection*. A relevant question therefore is: How can we prevent selection?

2.6. *How can we prevent selection?*

Theoretically, the best strategy to reduce selection is *good risk adjustment* (see Section 3), so that the heterogeneity of the subsidy-risk-groups is small and the expected cost of cream skimming exceeds its expected profitability. The more homogeneous costs are within a rate category the harder it will be for health plans to attract only enrollees

[31] Cost-reducing efficiency gains need not always be desired by consumers that receive subsidized insurance premiums. It may fall upon the sponsor to decide what the acceptable costs are.

[32] Resources used by health plans for product innovation or for designing contracts which provide consumers an incentive to become/remain in good health, but which may also attract low-risk individuals, are not considered a welfare loss [Beck and Zweifel (1998)].

whose average expected profit is high.[33] Whether feasible levels of risk adjustment still allow serious adverse selection remains an empirical question [see, e.g., Pauly (1996)]. In case of perfect risk adjustment, there is no selection.

As perfect risk adjustment is still a long way off, a second strategy to reduce selection is *risk sharing* between the sponsor and the health plan, which we discuss in Section 4. However, risk sharing reduces a health plan's incentive for efficiency, causing a tradeoff between selection and efficiency.

A third strategy to reduce selection is to allow health plans to *risk rate the consumer's premium contribution* within a certain range. Consequently, any information surplus the health plans might have over the sponsor would be focused on premium differences rather than on cream skimming. This could potentially worsen access for the high-risk individuals, yielding a tradeoff between access and selection. If health plans are required to identify any risk factors they use for premium differentiation, the sponsor could try to include these risk factors in the subsidy formula in subsequent years, thereby reducing the potential for cream skimming. Potentially, market-driven improvements of the risk adjustment mechanism may be more effective and more workable than research-driven improvements.

Several *additional measures* may be adopted to reduce selection. One straightforward way to prevent an extreme form of adverse selection – that is, one in which low-risk individuals do not buy health plan coverage at all and thereby do not cross-subsidize the high-risk individuals – is to mandate everyone to buy some minimum basic health plan coverage. Mandating a minimum health plan reduces but does not eliminate the possibility that health plans may differentiate their insurance plans to try to enroll profitable individuals. Forbidding selective contracting, such as by imposing an "any-willing-provider" mandate is a related tactic. Given the many subtle ways health plans can differentiate the coverage of their benefits package, this type of regulation may be hard to enforce. However, even if a sponsor could successfully implement mandatory health plan membership with uniform conditions, it could have several adverse effects.[34] First, it would impede health plans from selectively contracting with only cost-effective providers. This reduces the potential for managed care activities by the health plans, implying a loss of efficiency in production. Second, a "one-size-fits-all" plan reduces the consumer's choice and yields a welfare loss[35] because it reduces the health plans' responsiveness to consumer preferences. Third, a standardized plan reduces the health plan's initiatives to design insurance contracts that reduce undesired

[33] Although a refinement of the subsidy formula *on average* lowers the profits of cream skimming, for some individuals it might *increase* the profits [see, e.g., Beebe et al. (1985)]. Therefore a detailed exploration of the distribution of the potential profits and losses per individual insured may be necessary.

[34] In contrast to these adverse effects, a certain degree of standardization may have the advantage of making the market more transparent and reducing the consumers' search costs.

[35] For an estimate of this welfare loss, see Keeler et al. (1998).

moral hazard. Fourth depending on the generosity of the fully standardized benefits package, a mandatory health plan membership may increase moral hazard problems.[36]

A second and closely related measure for reducing selection incentives is to "carve out" or separately cover services on which health plans may potentially have the greatest incentive to select. Pharmaceuticals, mental health treatment, and dental care are frequently not included in the standard benefit package, but are either not covered or covered separately. The classic rationale is concern about demand-side moral hazard response since these services appear to be more price responsive to insurance coverage [Morrisey (1992)]. More recently Frank et al. (1997) and Ettner et al. (1998) have examined the rationale behind carving out these services, which includes the fact that these services are more predictable, and hence more prone to selection activities.

A third additional measure to reduce selection might be to increase plan level entry or exit barriers. The qualification or certification of health plan contracts by the sponsor or by an independent organization will make it more expensive for plans to enter so as to cream skim, or exit so as to avoid adverse selection. Sponsor subsidies can be earmarked for the purchase of qualified or certified health plan contracts only. The requirements for qualification of health plan contracts may relate to the design of the benefit package, the copayment structure, the quality of the contracted specialty-mix, the forms of risk sharing between the health plan and the contracted providers, the location and accessibility of the contracted facilities, etc. The pricing and selling of qualified health plans should not be tied-in with other products and services.

Fourth, regulations of the enrollment procedure may influence selection activities. Enthoven (1978, 1986) proposed that there be no direct interaction between a health plan's sales representative and a potential member in the enrollment process. The potential members should deal with an independent agency (or the sponsor itself) that notifies the health plans of those who have enrolled for the coming contract-period. Every family would receive a booklet, published by the administrative agency, containing meaningful, useful information on the features and merits of the presented alternatives. Furthermore, the contract period should not be too short. The shorter the contract period, the higher is the proportion of predictable episodes of costly illnesses (predictable by both the health plans and the consumers) during the next contract period(s). An example is the potential dumping of some patients at high risk of death [Newhouse (1986)]. Switching plans to take advantage of better pregnancy and birth benefits is another important example. The short (one month) lock-in period for a Medicare insured who chooses an HMO provides many opportunities for selection. The one-year lock-in period as applied in the Dutch sickness fund market may be a good compromise between sufficient consumer choice and not too much selection. Pauly (1988) proposed requiring

[36] In addition, even if the implementation of mandatory health plan membership with uniform conditions could successfully prevent adverse selection in a competitive health plan market, health plans would still be left with other tools for cream skimming, such as tie-in sales, selective advertising, design of supplemental health insurance, providing poor services to high-risk individuals, selective advice by insurance agents and a golden handshake.

consumers to choose their health plan option a long time before the renewal date of the contract. This lowers the predictability of future costs during the new contract period and thereby reduces the potential profits of selection.

Fifth, improved consumer information may mitigate selection, particularly monitoring and publicizing of information on plan quality. Luft (1982) suggested monitoring systems in which people who change plans are asked about any problems they experienced and whether they felt pushed out. Such information and data from more broadly targeted consumer satisfaction surveys could be very worthwhile for the consumer. The sponsor also could examine the health care needs and costs of those consumers who switch plans as a way of monitoring health plans' (and the contracted providers') behavior. In theory, the sponsor could raise the cost of cream skimming by dissemination of such information to consumers. In practice, this methodology probably has even further to go than risk adjustment or risk sharing.

Sixth, ethical codes for health plans might be designed to reduce incentives to select. Codes could be developed either by the sponsor or by professional organizations; violation of these codes could be a punishable offence. The ethical codes could relate to things such as the quality of the contracted providers, procedures for making and handling complaints, selective advertising, golden handshakes, etc.

Seventh, the sponsor will need to evaluate and periodically adjust and improve the risk-adjusted subsidy formula over time. Risk adjustment should not be done once and left alone. The sponsor will need to update the risk adjustment formula in light of technological change or behavioral responses to risk adjustment by the health plans and by consumers. The credible announcement by the sponsor of its intention to improve on the accuracy of its risk adjustment methodology periodically will reduce the expected profitability of certain cream skimming activities, and may lessen their use.

The extent of the success of these measures to prevent selection largely depends on the size of the predictable profits and losses that result from the regulation, as well as on the costs of selection, including the cost to a health plan of losing its good reputation.

3. Risk adjustment models

In this section we concentrate on the theoretically most preferred strategy to reduce selection, i.e. risk adjustment. We examine the specific risk factors and models that can be used for calculating the best estimate of acceptable costs. By *acceptable costs* we mean the cost of the set of services and intensity of treatment that the sponsor has chosen to subsidize, as defined in Section 2.2. We begin with a discussion of criteria that can be used for assessing risk adjustment models, and apply these criteria to issues related to designing, evaluating, and choosing a model. Specifically, we consider: criteria for selecting a risk adjustment model; choice of prediction period; choice of explanatory variables to use for risk adjustment; selection of a functional form; and use of summary statistics to assess and compare alternatives. The section ends with a review of selected state-of-the-art risk adjustment models that are compared in terms of their ability to achieve the objectives set out at the beginning of this section.

3.1. *Criteria for choosing among risk adjustment models*

A number of very useful surveys of risk adjustment models have proposed criteria for comparing different risk adjustment models [Thomas et al. (1983), Newhouse (1986), Epstein and Cummella (1988), Van de Ven and van Vliet (1992), US General Accounting Office (1994), Ingber (1998)]. Although more than a dozen criteria can be listed, they can usefully be grouped into three broad criteria, which may be mutually related:

- Appropriateness of incentives;
- Fairness;
- Feasibility.

In addition to the "appropriateness of incentives", an efficiency concept, and "fairness", which have already been emphasized, we see here the new concept of "feasibility". The feasibility of risk adjustment models imposes constraints on the key tradeoff between efficiency and fairness discussed in previous sections. Although a perfect risk adjustment model might be able to eliminate this tradeoff, such a model might not be feasible to implement.

3.1.1. *Appropriateness of incentives*

Correcting for selection and moral hazard problems are the primary reasons for implementing risk adjustment. Thus, the most important criterion for evaluating risk adjustment models is by the extent to which they create *appropriate incentives*. There are many possible distortions or undesirable responses to risk adjustment, in particular when combined with restrictions related to the premium contributions and with open enrollment. Since it is an area of keen interest and research by economists, Table 2 provides an extensive list of the ways that provider and health plan behaviour may respond to incentives created by the risk adjustment and the various regulations.

It is beyond the scope of this chapter to fully review the literature on each of these topics. The literature has traditionally focused on how benefit design and premiums influence plan selection by enrollees who differ in expected health [Morrisey (1992), Jensen and Morrisey (1990)]. Differences in expected costs that result from cost sharing differences should in most cases be taken into account when developing and implementing risk adjustment models [Van de Ven and Van Vliet (1995)].

Plans can use a range of strategies for attracting profitable enrollees and avoiding unprofitable ones, such as by denying coverage, exclusions for preexisting conditions, and selective enrollment or disenrollment counseling. Many of these strategies are regulated or prohibited in some, but not all, countries and settings. These strategies should probably be addressed through regulation rather than asking risk adjustment models to solve all of the problems of creating an effective health care market (see Section 2.6).

In the presence of government regulations prohibiting explicit selection, health plans have incentives to manipulate the specific services that they offer to enrollees. This topic has been the focus of a recent flurry of research, perhaps reflecting growing concerns about its potential importance. Ellis (1998) develops a framework in which health

Table 2
Health plan response to incentives created by the way that health plans are reimbursed

- Choice of plan benefit features
  - Deductibles or copayments for selected conditions
  - Coverage limits (lifetime or annual)
  - Coverage of pharmaceuticals or other specific services
  - Exclusions for preexisting conditions
- Responses to regulated rate classes
  - Efforts to attract more profitable rate classes such as:
    - family or individual contracts
    - employee or retiree
    - specific geographic area
  - Selection of relative premiums by rate classes
- Plan level efforts to attract profitable/avoid unprofitable enrollees
  - Denying coverage ("medical underwriting")
  - Canceling coverage
  - Selective advertising
  - Pre-enrollment screening
  - Selective enrollment and disenrollment counseling
- Changes in service offerings
  - Selection of specialists to include or exclude from plan network
  - Overprovision of services that attract profitable enrollees
  - Underprovision of services that attract unprofitable enrollees
  - Change of place of service to increase payments
  - Unnecessary provision of services to code a diagnosis
  - Change in timing of services to increase payment
- Changes in diagnostic coding or other claims information
  - Upcoding of diagnoses to more serious conditions
  - Proliferation of diagnoses
  - Fraudulent diagnostic coding
  - Coding of "rule out" diagnoses
- Attempts to influence survey-based health measures
  - Enrollee coaching
  - Nonrandom enrollee sampling
  - Biased corrections for nonresponse

plans have incentives to oversupply services to profitable patients ("cream skim") and undersupply ("skimp on") or "dump" (avoid treating) patients that are unprofitable. Improved risk adjustment reduces the incentive for plans to engage in these activities, but also changes the particular enrollees that plans will compete to attract. For example, increasing payments for individuals expected to cost more than the average can result in plans competing to attract such individuals, a reversal of the incentives with unadjusted capitation payments.

Although premium subsidies that are fully adjusted for the consumer's health status make selection unimportant, these payments may be criticized because they create inappropriate incentives for health-improving activities. One could argue that a health plan that improves its members' health status by good quality care and effective prevention is penalized by lower future revenues [McClure (1984), Luft (1996)]. A counter argument, however, is that improved health status not only reduces future revenues, but also future expected expenditures. Furthermore, if a health plan effectively reduces the incidence of lung cancer or heart diseases, it fully benefits from not having the high first-year expenses related to these diseases.[37] In addition the plan fully benefits from not having expenses related to preventable transitory health problems for which the subsidies are not adjusted (e.g., fever and flu). Nevertheless, it is true that a health plan bears the full costs of health-improving activities and preventive services such as smoking cessation, weight loss, and nutritional guidance, while it may lack a part of the future returns. In other words, from the point of view of the health plan health adjustment may reduce the cost-effectiveness of some prevention programs. Whether in practice these incentives override the professional ethics of the providers and the consumer preferences, remains an empirical question.

McClure (1984) suggested the following two solutions. The first is to make bonus adjustments based on change in health status over time. With care and ingenuity, it may be possible to devise subsidies that reward health improvement but that cannot be gamed by the plans. Secondly, McClure suggested making public to beneficiaries any change in overall health status levels in each health plan, so beneficiaries might shop for health plans on the basic of health status improvement figures. Plans would thus gain a reward for improving health status by attracting new enrollees. A third solution is to provide health plans with earmarked payments for effective prevention programs.

Several studies have discussed the incentive for health plans to distort information reported to the sponsor if that information is used for payment purposes [e.g., Epstein and Cumella (1988)]. This may occur either with diagnosis- or survey-based risk adjustment. Carter, Newhouse and Relles (1990) examined changes in diagnostic coding in the United States in response to the Medicare Program's payment system based on Diagnosis Related Groups (DRGs), which they termed "DRG Creep". They also suggest that such changes in diagnostic coding appear to be one-time. It seems plausible that similar responses might occur from risk-adjusted capitated payments, but we are not aware of any studies documenting this result empirically to date.

The predictive accuracy of different models is by far the most common criterion on which risk adjustment models are compared. Yet the goal of risk adjustment is not accuracy per se, but rather improved incentives and fairness. Using prior information that is known to the individual or plan to adjust payments is important because it should lessen the danger of cream skimming or dumping. Specific measures of predictive power are discussed below, along with a consideration of whether it is individual or group resources that should be predicted.

Although greater predictive power is generally desirable, it is important to emphasize that higher predictive power is not necessarily preferred to less. For example, actual expenditures are perfectly correlated with actual expenditures, and are an excellent

[37] This argument does not hold in case of retrospective risk adjustment (see Section 3.2.2).

"predictor" of the health care use in that same year. Yet such fee-for-service reimbursement is a very imperfect basis for payment since it creates undesirable disincentives for efficiency, and "costs" are difficult to measure and monitor. Similarly, models that base their predictions upon the type of service provided, the use of specific procedures, or concurrent year diagnoses, can be more accurate, but may be create inappropriate incentives.

Finally one may argue that mortality as a risk adjuster provides health plans with inappropriate incentives ("mortal hazard").

3.1.2. Fairness

We discussed fairness within the framework of the solidarity principle of Sections 1 and 2. While the fairness of the method of collecting premiums and calculating risk adjustment subsidies has been the topic of considerable discussion in many European countries, it has received considerably less attention in the United States. For example, the *fairness* of the risk adjustment model does not enter explicitly into the list of criteria used to compare across different models in the reviews of Epstein and Cumella (1988), the US Government Accounting Office (1994), or Ingber (1998).

Decisions about fairness and about what risk factors should be labeled an S-type or N-type factor, reflect value judgments that differ across countries and among individuals. There appears to be a consensus that factors that reflect purely tastes (e.g., religion or a preference for cost-ineffective care) may have predictive power but do not belong in a risk adjustment model based on commonly held fairness principles. Lifestyle is a more problematic risk factor. On the hand one could argue that health care expenditures that are purely related to smoking or sexual behavior should not be subsidized because these expenses can be influenced by the individual. On the other hand, many people will argue that these expenditures should be subsidized because it is unfair if people with lung cancer or AIDS cannot receive an appropriate medical treatment.[38]

Another discussible factor is average distance between patients and providers or density. Should the premium subsidies be lower for geographically dispersed regions with poor access to health services? In the United States, the Medicare program's formula for reimbursing HMOs (in 1998) fully reflects county level geographic variation in average health costs, but it is not clear that it should do so.[39] Some of the geographic variation in health costs is due to differences in cost of living between different regions. Many people consider it fair that risk-adjusted subsidies for persons living in high wage cost regions, where medical care is more expensive, should be higher than those for people living in low wage cost regions. This argument holds in particular if the solidarity contribution in the high wage cost regions is higher than in the low wage cost regions. But

[38] An alternative is to let the solidarity contribution partly depend on lifestyle factors. E.g., a surcharge on tobacco could go to the sponsor.

[39] The US Balanced Budget Act of 1997 seeks to reduce differences among county level averages used for risk adjustment in the Medicare program.

if the variation is due to practice style variation, taste differences, over- or undersupply, or differences in access, geographic adjustment may be viewed as unfair. The same argument may hold for factors that primarily reflect differences in access, such as race, minority group or ability to pay (income). By not adjusting the subsidy for these access indicators, individuals with poor access will either become preferred risks, which may increase their access, or they will pay a lower premium contribution.

A different type of equity argument is that individuals who are sicker should have risk-adjusted subsidies that are higher than for those who are less sick. This implies that evidence of a new disease or chronic condition for a person should never result in a reduction in the risk-adjusted subsidy for that individual if there exists a cost-effective medical treatment for the person's health problem. This equity argument (monotonicity) does not always hold in empirically derived risk adjustment models. For example, in the empirical risk adjustment models described in Ellis et al. (1996b), in many specifications it was found that among US Medicare enrollees, individuals classified as having dementia (e.g., Alzheimer's disease) have lower predicted medical costs than persons with otherwise identical demographic and diagnostic information. If this reflects underutilization, it seems unfair to reduce payments for this group, even if it is predictive of lower costs. Similarly, Ash et al. (1998) find that in some samples those with profound and severe mental retardation have lower predicted costs than those with mild retardation. If this lower utilization reflects underutilization, one may argue that it should not be reflected in the subsidies for fairness reasons. In this way the underserved become the preferred risks, which may reduce their underutilization.

As we suggested in the preceding subsection, a risk adjustment system will often be considered fairer if it predicts a larger proportion of the variation in health spending. If health plans are fully compensated for the higher expected costs of enrollees with chronic conditions, then it is more likely that they will enroll them, thereby increasing the access of these high cost people. In addition, health plans will bear less risk. Yet as the above examples highlight, improved accuracy that comes from using information for which solidarity is not desired or from risk factors indicating poor access or underutilization may worsen rather than improve fairness.

3.1.3. Feasibility

Administrative *feasibility*, closely related to the criteria of obtainability discussed in Van de Ven and Van Vliet (1992), is the requirement that the measures are feasible to obtain for all potential enrollees without undue expenditures of time or money. Information that is routinely collected, standardized and comparable across different health plans, and measures that are easily validated have greater feasibility than measures that require separate data collection, validation and processing.

A further dimension of feasibility is that large, representative samples exist on which risk adjustment models can be developed and parameterized prior to implementation, or used for recalibrating subsequent to adoption. This weakness is particularly serious for survey-based predictors. Another dimension of feasibility is length of the time lag required between the collection of data and its feasible use for payments. Long lags between the date when a health service is provided and the date on which a claim is submitted and processed, can constrain the feasibility of diagnosis- or other claims-based risk adjusters.

Risk adjustment will be feasible only if it is accepted by consumers, providers, health plans, and sponsors. Although a considerable amount of academic research has gone into improving the predictive power and incentives of risk adjustment, relatively little has been published on making risk adjustment acceptable to all parties involved.

One dimension of acceptability is that a risk adjustment model should not compromise the right to privacy of consumers and providers [Epstein and Cumella (1988) and Van de Ven and van Vliet (1992)]. For example, a risk adjustment approach that requires individuals or providers to identify specific individuals who are HIV positive or who suffer from mental illness may be unacceptable to consumers, regardless of other merits.

Race, ethnic background, and religion are examples of demographic variables that may not be acceptable for risk adjustment primarily due to concerns about fairness. Paying more to a plan during the year in which an enrollee dies may be an actuarially good way to recompense it for the known high costs incurred in the last months of life. However many are repelled by the idea of paying more to health plans because their mortality rates are higher.

Clinical credibility is another dimension of acceptability, since doctors and clinically trained health administrators are important decision-makers. Regardless of whether it affects the predictive accuracy of the risk adjustment model, if clinicians see large differences in payments based on apparently trivial classification differences, then this will undermine acceptability to clinicians.

One last group for whom acceptability is central is actuaries, who typically work for sponsors or health plans and traditionally calculate premiums and provider payments based on demographics and prior experience measures. An important criterion for them is that risk adjustment models are actuarially fair. In the United States, actuaries have been slow to accept health-based risk adjustment, despite its greater accuracy.

3.2. Preliminary issues in designing or implementing risk adjustment

3.2.1. Individual versus contract level risk adjustment

As stated earlier, we take it as given that it is desirable to calculate health-based payments at the level of individuals rather than contracts, such as families or employers. Actuaries in the United States and elsewhere often focus attention on calculating expected payments at the contract level, with the employee and all dependents counting as one unit of analysis. However this approach, focusing solely on the number of people and their relationship to the enrollee without regard to the age and sex breakdown, ignores obviously important information. According to our approach, the expected payment at the contract level (family or employer) can be calculated as the sum of the expected payments for the covered individuals. Although we understand the actuaries' argument that in a competitive market an insurer has to break even on each insurance contract and not on each insured person, the advantage of our approach is that when one individual (e.g., an HIV patient) goes from contract unit A to B, we can easily recalculate the expected payments at the contract level.

3.2.2. Prospective versus retrospective use of risk adjustment information

In developing or implementing risk adjustment, important choices must be made about how information will be used. One alternative is that payments are calculated prospectively, at the beginning of the prediction period using only prior information. A second alternative is to calculate payments retrospectively, at the end of the period. Retrospective payments can reflect information that becomes known during the period being predicted. As Ellis and McGuire (1986) and Newhouse (1986) have highlighted, these two extremes are not the only ones possible: one can also make payments that are a mixture of the two. We focus here on the two pure cases, and defer to section 4 the discussion of risk sharing arrangements that are implied by taking combinations of the two.

The per cent of the variance in health spending at the individual level that is predicted using a retrospective framework is considerably greater than what can be predicted prospectively. However, a retrospective framework may not be preferable in practice. While there are estimates of the maximum potential variance predictable by prospective risk adjustment models (see Section 3.2.6), we do not have a standard for how much variance a good retrospective adjuster should predict [Newhouse et al. (1997)]. The incentive and fairness properties of retrospective adjusters are not inherently superior, and the feasibility of using retrospective models is probably worse. Dunn et al. (1996) compared the predictive accuracy of prospective and retrospective frameworks on groups of enrollees and found surprisingly small differences in predictive power for groups when the samples were reasonably large. Ellis et al. (1996b) and Ash et al. (1998) likewise find that prospective models do nearly as well as retrospective models when nonrandom groups of individuals are formed using only prior-year information. Chapman (1997) finds a greater advantage of retrospective models over prospective models in his plan level analysis, but he focuses primarily on group level predictions rather than individual level predictions.

Conceptually, an argument for preferring the use of prospective information for risk adjustment is that only prospective information is potentially known to health plans and individuals at the time that they are making enrollment decisions, and hence used for risk selection. Prospective models attach relatively more weight to information related to chronic conditions that persist over time, while retrospective models attach more weight to information that signals the presence of acute problems. If two persons are ex ante observationally identical, but ex post only one of them turns out to have a heart attack, then under a wide range of assumption, it should not matter for incentives on the plan whether they are compensated ex post for the actual cost of the one getting the heart attack, or ex ante for the expected cost of the likelihood that one of the two will have a heart attack. Newhouse et al. (1997) highlight that explaining truly random events is unimportant when the risk is averaged over many conditions and many individuals. On the other hand, if there is moral hazard on the probability of having the heart attack, or discretion in the treatment of and recording the acute diagnosis of heart attack, then the two systems are not the same. In the US, for example, many believe that there is too little prevention and too much treatment. In such an environment, paying prospectively rather than retrospectively will create superior incentives to avoid and not over-diagnose heart attacks. This moral hazard problem is potentially quite important for the many health conditions for which treatment or prevention activities are discretionary.[40]

Although one may give a high weight to the above argument, it ignores that a retrospective framework protects health plans against adverse selection by individuals with a diagnosis that yields high costs in the period (e.g., a year) in which the diagnosis is set and from which moment it can be used as a risk adjuster. If this argument is relevant, which still is an empirical question, sponsors may consider to extend a prospective risk adjustment model with selected one-year retrospective elements.

Prospective models tend to be more feasible than retrospective models. As a practical consideration, prospective frameworks have the advantage that the information is available sooner, and health plans have more predictable revenues at the beginning of each prediction period. This predictability is attractive both for plans and for sponsors.[41] A second practical consideration is administrative feasibility of available data. Developing a retrospective model has the advantage of only requiring data from a single period, versus two for prospective modeling. Implementing each model imposes similar data collection burdens.

Although the arguments are not all unambiguously in favor of a prospective setting, our interpretation weights the arguments in favor of a prospective framework as relatively more important. Therefore, we focus our attention in this chapter primarily on prospective risk adjustment models. For clarity of presentation, we describe the various models as if a prospective setting is the only intended use. We return at the end of this section to compare various prospective and retrospective models, and in Section 4 we compare various risk sharing strategies that share much in common with retrospective adjustment models.

3.2.3. Functional form

There is a considerable literature in statistics, econometrics, and health economics that examines and assesses alternative functional forms for estimating models of health spending. Although these models often include many N-factors, and not just S-factors that policy makers and researchers are interested in using for potential risk adjustment

[40] A contrary view is that prospective payment may overpay for persons with high-blood pressure who don't use any medicine.

[41] On the other hand, the predictability of a health plan's margin is higher under a retrospective model than a prospective model.

models, this literature has an important bearing on the selection of models. The classic article in this literature is that of Duan et al. (1983), which developed the so called "two-part model" of health spending. This model decomposes the expected level of spending (Y) given a vector of explanatory variables (X) into the two parts using the identity:

$$E(Y) = \Pr(Y > 0 \mid X)E(Y \mid Y > 0, X).$$

Several specifications have been used for each part of this model, including Probit, logit, and linear probability models for the first part, and linear, log-linear and square root models for the second part, which is conditioned only on observations that are strictly positive.[42] Both the use of two-part models and nonlinear transformations of the second part are used to improve consistency of the ordinary least squares (OLS) model given the highly heteroskedastic errors. Conventionally, both parts of the two-part model are estimated independently and a smearing transformation [see, Duan et al. (1983)] is used to generate unbiased estimates of the second part of the model in the common situation in which nonlinear transformations of the dependent variable are used. The classic article using this approach is Manning et al. (1987).

Since this issue is already examined at length in Jones (2000) (which examines econometric issues), we highlight here only two observations based on the recent literature relevant for applications of risk adjustment in practice. The first observation, made by Mullahy (1998), is that for the two-part models to yield unbiased estimates of both partial effects and conditional means, it is important that the error structure strictly satisfy the homoskedastic error assumption, or else a nonlinear smearing correction can lead to seriously biased estimates. This point is reinforced by the companion article by Manning (1998) which demonstrates that predicted means can be seriously underpredicted (e.g., 20%) if heteroskedasticity is not taken into account. Manning makes the important point that the use of the simple transformation $\log(Y + 1)$, motivated by its convenience, has very poor statistical properties for use in risk adjustment.

The second point is that rather than using nonlinear, two-part models of health spending, the problem with health spending having a thick upper tail can be dealt with by using extremely large samples, and correcting standard errors for heteroskedasticity using the Huber/White formula. Mullahy (1998) notes in a footnote that when sample sizes are large, using simple nonparametric techniques such as cell means or linear regressions may be sufficient, an argument that we find convincing. Monte Carlo simulations presented in Ellis and Azzone (1998) suggest that the attractiveness of simple linear models relative to two-part models increases as the predictive power of the risk adjustment models increase. With only a few exceptions, the major risk adjustment models

[42] For useful references and discussion of this literature see Mullahy (1998). Also relevant are the debates of the 1980s, most notably that of Hay and Olsen (1984), which examined the desirability of estimating two part models while assuming that $E[Y \mid Y > 0, X]$ can be estimated consistently using only observations where $Y > 0$.

discussed below have used simple linear models, for which there is no retransformation problem. Another argument for the use of simple linear models is to stay as close as possible to the cell-based approach, i.e. the calculation of the average expenditures per risk group, which is mostly used by sponsors for risk adjustment and by health plans for premium rating. For the remainder of this chapter, we focus on simple linear models, comparing them to nonlinear models only to make a point about the effect of nonlinear transformations on measures of predictive power.

3.2.4. Adjustments for partial years of eligibility

It is very common for people to be eligible for health coverage for fractions of a year. This happens automatically with births and deaths, and may also occur due to enrollment or disenrollment. This presents a problem for risk adjustment models both in terms of efficient estimation and in terms of prediction.

It is clearly undesirable to simply exclude those with partial years of eligibility when the goal is unbiased prediction, since partial year eligibles tend to be systematically different from average. Simply including observations of those with partial years of spending without any recognition of the partial year eligibility is also undesirable, since the resulting models will tend to underpredict spending if the model is used to make predictions that are used for partial year rather than full year payments.

Consider the following example involving only two persons. Suppose person A is eligible for all 12 months and costs \$6,000, while person B is eligible for only 6 months, but costs \$12,000. Total spending on these two persons is \$18,000, and total eligible months are 18, so the correct monthly average is \$1,000 per month, or \$12,000 per year.

Two corrections for partial year eligibility have been made in the literature, one focusing on unbiasedness, the other focusing on maximizing statistical efficiency. Ellis and Ash (1995) argue that spending for partial year eligibles should be annualized, and then each observation should be weighted by the fraction of the year that the person is eligible. Hence in the above example, person B has \$24,000 in annualized expenditures, with weight 0.5. The weighted average is then $(\$6{,}000 * 1 + \$24,000 * 0.5)/1.5 = \$12{,}000$, which gives the correct annual average. An alternative approach is developed in Hornbrook et al. (1998), who assume that person A reflects 12 draws of monthly spending, while person B reflects only six draws. If the monthly draws are independent and homoskedastic, then efficient weighting reverts to the formula used by Ellis and Ash. However if the monthly draws are correlated (which empirically they are), then the efficient weights are to place relatively less weight on person A relative to person B than the ratio 2 : 1. Alternatively, once heteroskedasticity rather than correlation is modeled, empirically it is generally true that the monthly draws for people with shorter eligibility have a higher monthly variance than those with a full year of information. Hence efficient weighting would place relatively less weight on person B. It is easy to show that predictions based on either of these two weighted least squares models, in general, will generate biased estimates of the sample means. Whether they are more accurate predictors empirically does not yet appear to have been answered.

3.2.5. Determinants of R^2

A common measure of the predictive power of different risk adjustment models, but by no means the only one, is the conventional R^2, which measures the proportion of the variance in individual expenditures that is explained by a set of risk adjusters. Nearly all empirical studies on risk adjustment present R^2-values. Ideally, in order to prevent overfitting R^2-values should be reported which are based on out-of-sample predictions. In that case Efron's (1978) R^2 should be used. Some studies have dealt with the question what the maximum R^2 is that can be achieved by a set of prospective risk adjusters (see Section 3.2.6). For the interpretation of R^2-values presented in the literature it is important that the R^2 (as well as the total variation) may depend on: (1) the type of services under analysis; (2) the (sub)population under analysis; (3) the variation in explanatory factors; (4) the level of medical technology; (5) the year of the data analyzed; and (6) the length of the time period being predicted. We discuss each of these determinants, which may be mutually related.

The relation between R^2 and *type of service* can be illustrated as follows. Newhouse et al. (1989) found an R^2 of 0.05 for inpatient care and an R^2 of 0.25 for outpatient care, using the same comprehensive set of risk adjusters for the same population (14–64 years old); the R^2 for total acute care was 0.09. Wouters (1991) also found much higher R^2-values for outpatient expenditures than for inpatient expenditures, using the same set of adjusters. In addition, she found that among the various types of outpatient services there is a wide variability in out-of-sample prediction R^2-values, using the same set of adjusters. Drugs ranks first ($R^2 = 0.40$), followed by visits, diagnostics, procedures, and surgery ($R^2 = 0.005$). Van Barneveld et al. (1997) analyzed expenses for several forms of expensive long-term care, like institutional care for mentally handicapped persons, nursing home care and institutional psychiatric care. Using 2 year prior costs as a risk adjuster they found an R^2 of 0.56. This figure is much higher than the comparable R^2-values for acute care, which typically are below 0.15.

The relation between R^2 and *subpopulation* can be illustrated by the results of Kronick et al. (1995). Analyzing US Medicaid claims they concluded that expenditures are much more predictable among persons with Medicaid entitlement based upon disabilities than for other populations. Using prior year expenditures as a risk adjuster they found R^2-values on four different data sets ranging from 0.29 to 0.51. In explaining these relatively high R^2 Kronick et al. suggest that among persons with disabilities a much greater portion of resource utilization results from chronic problems and their complications which persist from year to year, and a smaller portion from acute episodes that lead to short-term spikes in resource use but are not followed by long-term needs.

A third determinant of the R^2-value is the *variation in explanatory variables*. A greater variation in the factors explaining variation in health spending (see Figure 3) ceteris paribus increases total variation. Whether the proportion of predictable variation (R^2) then increases or decreases depends on whether the variation in explaining factors is known ex-ante or can be accurately predicted. For instance, greater variation in practice style, supply or input prices which are stable over time, enlarges both total variation in expenditures and the R^2-value. However, greater variation in input prices resulting from unpredictable changes in market power or government regulation, increases total variation but decreases R^2.

Fourth we hypothesize a positive relation between R^2 and the level of *medical technology*. This level may change from country to country and, within a country, it may change over time. An increase of the level of diagnostic technology may result in a better prediction of future (genetically determined) diseases and expenditures. In addition it may result in more protocolized treatments and thereby reduce random variation in treatments. An increase of the level of effective therapeutic medical technologies may keep alive at-risk patients who otherwise would have died, e.g., cancer patients, heart patients and patients with a transplantation. As a result the proportion of chronically ill persons may increase. As Kronick et al. (1995) stated, the expenses of the chronically ill are relatively more predictable because they pertain to chronic problems and their complications which persist from year to year.

A fifth determinant of the R^2-values that are presented in the literature is the *year of the data*. Many studies analyze data from the 1970s, while others use data from the 1990s. For the following reasons, which may be mutually related, we expect an increase in R^2 over time. First, medical technology has increased. Second, we have seen a substitution of outpatient care for inpatient care over the last decades, with outpatient care being more predictive than inpatient care. Third, the proportion of expenditures spent on prescribed drugs has increased over the last decades, with prescription drugs costs being relatively more predictable (see Section 3.2.6). Fourth, the proportion of elderly and chronically ill persons, whose expenditures are more predictable, has increased. Fifth, the predictive power of age has increased over time. For example, Schut (1995) calculated that in the Netherlands from 1979 to 1986 the average hospital costs of men over 80 years old increased from 4.9 to 7.6 times the average hospital cost of men in the 45–49 years age group. So, over time more variation in expenditures can be explained by age. Based on these arguments we hypothesize an increase of R^2-values over time.

A sixth and final determinant of the R^2 values that are presented in the literature is the *length of the time period being predicted.* Using longer periods averages out some of the randomness, and tends to improve predictive power. Ellis and Ash (1989) developed models that predict a one month prediction period that achieved an R^2 of only 0.0089 on the monthly observations versus 0.04 using the same information with an annual prediction period. Garber, MaCurdy, and McClellan (1998) examined the predictability of health spending over multiple years, and demonstrate the effects of smoothing out of random variation.

3.2.6. Maximum R^2

In an important set of articles Newhouse et al.[43] and Van Vliet (1992) ask the question: what is the maximum potential variance predictable by prospective risk adjustment models, i.e. models using only information from a past period or periods? The

[43] Newhouse (1996) and Newhouse et al. (1989, 1993, 1997).

literature usually tries to answer this question by dividing the variance in actual spending into different components. The component indicating the between-person variance was estimated by McCall and Wai (1983) to be 0.15 and by Newhouse et al. (1989) at 0.145. Additionally, some within-person variance is predictable because of the autoregressive error-component [Newhouse (1996)]. As an upper bound for this component, exclusive of time-varying covariates, 0.04 could be used,[44] making the predictable proportion around 0.20. This corresponds with the 0.174 estimated by Van Vliet (1992), who used an autoregressive moving averages (ARMA) model. However, the "around 20 per cent" is a lower bound on the ability to predict future spending because other predictive factors may be observed that are not reflected in past spending. (So it is a "lower bound on the upper bound", rather than a true upper bound.) Examples of such predictive factors are a pregnancy, a recent diagnosis of cancer, a terminal illness, or being on the waiting list for an expensive treatment [Newhouse et al. (1989, 1997), Van Vliet (1992)]. Plans and individuals could potentially predict more than the 20 per cent of actual variance, but how much more is unclear.

Results about the maximum R^2 as presented in the literature are consistent with the above mentioned determinants of R^2. Newhouse et al. (1989) estimated the maximum R^2 for inpatient care to be 0.08 and for outpatient care, 0.48. Similar results were found by Van Vliet (1992), who also concluded that the expenditures for prescription drugs together with GP consultations are extremely predictable (maximum R^2 of 0.80). This finding has serious implications for comparing R^2-values from a setting where expenditures for prescription drugs are not included (e.g., US Medicare data) with a setting where they are included (e.g., the Netherlands sickness fund data).

With respect to the relation between R^2 and subgroups Van Vliet (1992) found evidence supporting the hypothesis that predictability increases with age and that differences in health expenditures for older individuals are more predictable than those for young people. This hypothesis is consistent with the findings by Newhouse et al. (1989, 1993) that the maximum R^2 for outpatient expenditures are higher for the age-group 14–64 years (maximum $R^2 = 0.48$) than for the age-group 3–13 years (maximum $R^2 = 0.37$).

Because of the relation between R^2 and both medical technology and the variation in factors explaining the variation in expenditures (such as input prices, supply, practice style and benefit plan features) it is important to note that these determinants may strongly vary from country to country. So one should be careful to apply in one setting the maximum R^2 estimated in another setting. Ideally, to have a benchmark researchers should estimate the maximum R^2 on the same (longitudinal) data base that is used for analyzing risk adjusters.

The relation between R^2 and year of data analyzed is relevant for the interpretation of the above mentioned lower bound of maximum R^2 (around 20 per cent). This estimate

[44] This value is based on Newhouse (1996). We consider 0.04 as an upper bound because Table 3 of Newhouse et al. (1989) contains correlations between expenditures and not between residual spending, as stated in footnote 62 of Newhouse (1996).

is based on different data sets from the 1970s and early 1980s. Based on the above arguments we are not surprised to see higher lower bounds to be estimated on data of more recent years. For example, Lamers (1999b) analyzed acute care expenditures, including prescription drugs, for Dutch sickness fund members[45] for the years 1992–1996. Using the ARMA model [see Van Vliet (1992)] she found a lower bound of the maximum R^2 of 0.33.[46]

We started with the question: what proportion of variance in expenditures is potentially predictable by a health plan? We may conclude that the maximum is, in any event, much less than 100 per cent because many health expenditures cannot be foreseen by either the individual or the health plan [Newhouse et al. (1989)]. Furthermore a lower bound of the maximum percentage can be estimated, which depends on the type of care, the (sub)population, and the specific setting and year. However, we do not know how much more variation is predictable than indicated by this lower bound.

3.2.7. How successful can risk adjustment be?

With respect to the success of risk adjustment two types of concern can be discerned: (1) can risk adjustment be *sufficiently* successful?; and (2) can risk adjustment be *too* successful? We discuss both.

Newhouse et al. (1997) raise the question of how close to perfect the formula must be to make plans' incentive and ability to seek favorable risks a *de minimus* problem. We share their view that a workable formula need not achieve the ideal, but that it is unknown how far from perfection will be sufficient.[47] As stated in Section 2.6 an adequate risk adjustment formula should be such that health plans expect the transaction costs of cream skimming (including the loss of good reputation) to exceed its profits. A second reason why the variation in the risk-adjusted premium subsidies will not equal the maximum potential variation in predicted expenditures, is that the sponsor ideally will only compensate for variation in S-factors and not in N-factors [Van Vliet (1992)]. A third reason why an adequate risk adjustment formula need not be perfect, is that cream skimming strategies based on one-year savings may have longer-run opportunity costs. Beck and Zweifel (1998) present an example in which 50 per cent of 'bad' risks turn out to be good risks in the long run, while 20 per cent of the 'good' risks become bad ones [because of regression towards the mean; see also Welch (1985)]. Fourth, the

[45] Disabled persons and chronically ill are over-represented among the Dutch sickness fund members.

[46] Van Vliet and Lamers (1998), analyzing the same data base, found an R^2 of 0.19 for a model with the risk-adjusters 3-year DCGs and three years of prior expenses.

[47] Under the assumption of lognormally distributed expenditures Newhouse et al. (1989) provided evidence that as explained variation improves, incentives to select do not diminish proportionately. From this finding Newhouse (1996) concludes that the risk adjustment formula must be close to perfect to reduce greatly the incentives to select. However, as Van de Ven et al. (1994) put forward, after correcting for the overestimation of the nonlinearity in Newhouse et al. (1989), the relation between the square root of explained variance (which, just a profits, is a linear function of predicted expenditures, rather than a quadratic function) and profits appears to be linear.

sponsor could periodically adjust and improve the formula, thereby lessening the attractiveness of some selection strategies in the long-term. Fifth, when the sponsor improves the formula, not only will a health plan's potential profits from selection decrease, but also the standard deviation of its profits will increase [up to a factor of three; Van de Ven et al. (1994)], thereby reducing the attractiveness of selection strategies. So, an imperfect formula may be sufficient to make selection unimportant. However, how much imperfection a sponsor can permit, is an unanswered empirical question.

A second concern about the potential success of risk adjustment is the question: can risk adjustment be *too* successful? A formula based on age and region, each divided into two groups, clearly creates large incentives for selection, by pooling heterogeneous people in the same groups. Assume that, in order to reduce these incentives, a sponsor refines the subgroups and replaces the two age-groups by 40,000 birthday-groups and replaces the two region-groups by 10,000 zipcode-groups. Assume further that each of these 400 million subgroups contains at most one individual. Although this birthday-zipcode formula largely reduces the incentives for selection (except for those who change plans, for newborns and for those who are expected to die), it also reduces the health plans' incentives for efficiency because of the large extent of cost-based reimbursement with a one period delay. Most sponsors will reject the birthday-zipcode formula because of inappropriate incentives. The birthday-zipcode formula also lacks robustness in the sense of stability of the weights over time, and it suffers from overfitting in the estimation model.[48]

In the discussion of alternate risk adjustment models, we mention the conventional R^2 measures in a few instances in order to convey an initial picture of the explanatory power of different sets of information. We present other reasons why R^2 can be misleading and difficult to compare as well as alternative measures of predictive power useful for assessing different modeling frameworks in a subsequent section.

3.3. Alternative risk adjustment models

Considerable research has been conducted on alternative risk adjustment models in many countries, using a wide range of information. We discuss these models in groups defined by the kind of data used for prediction: demographics only, prior year expenditures, diagnoses, information derived from prescription drugs, self-reported health and functional health status measures, mortality, and other types of information.

3.3.1. Demographic models

The most basic type of information used to adjust payments to health plans (or providers) are age and sex. Figures 4A through 4C illustrate that there are pronounced

[48] The birthday-zipcode formula illustrates the need to make a distinction between the R^2 for explanation (in this case: $R^2 = 1.0$) and prediction (with Efron's R^2 being negative).

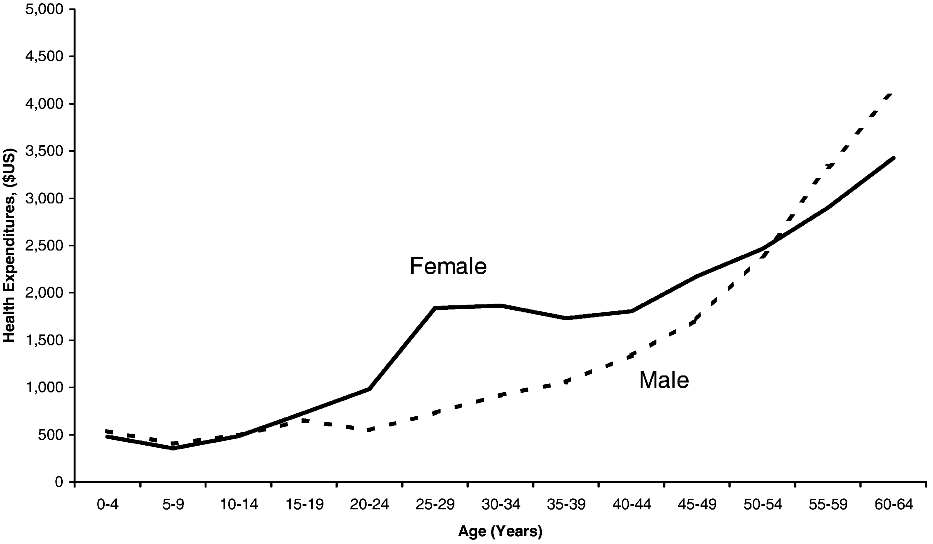


Figure 4A. Health spending by gender and age in the USA. 1.0 Million Privately Insured Individuals, 1992–93.

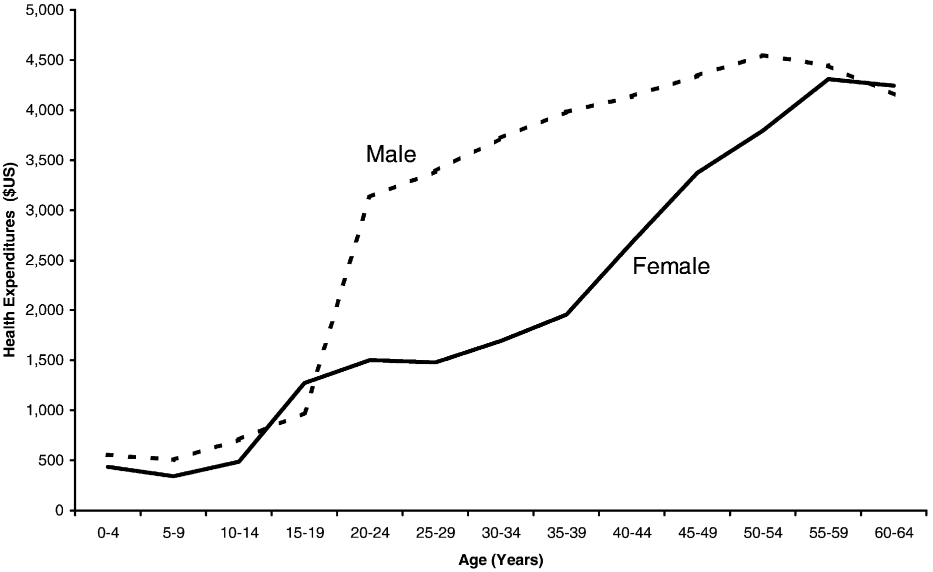


Figure 4B. Health spending by gender and age in the USA. 1.3 Million Medicaid Eligibles Under age 65, Michigan, 1991–92.

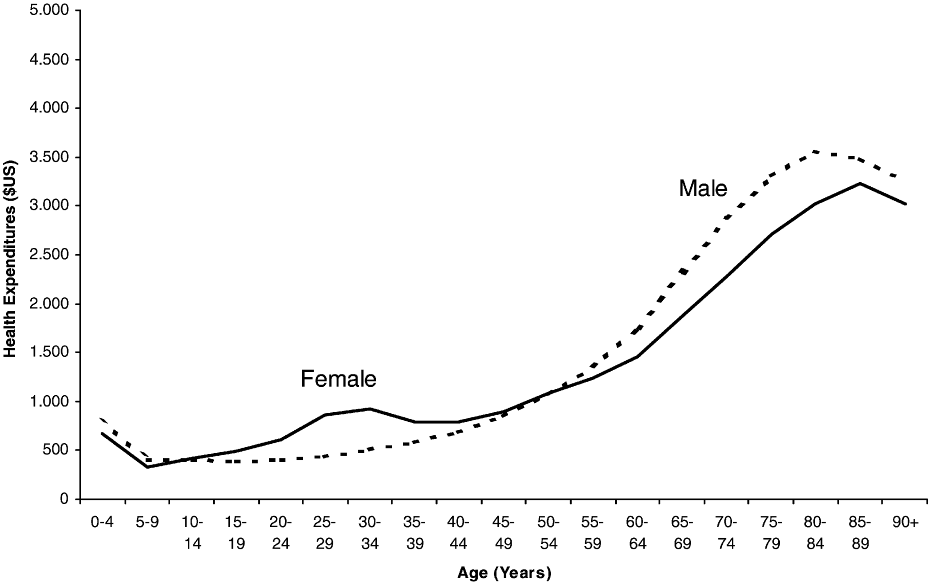


Figure 4C. Health spending by gender and age in the Netherlands. 9.6 Million Enrollees in Sickness Fund Basic Benefits Package, 1995.

differences in expenditures across individuals by age and sex, that these patterns differ according to the country and sub-population studied. Among privately insured enrollees in one large cross section of US firms from 1992–93 (Figure 4A), average health expenditures on men aged 60–64 are $4,100 versus only $350 for females aged 5–9, an 11-fold difference. The corresponding numbers for Medicaid (Figure 4B) are $4,160 and $340, a 12-fold difference. As shown in Figure 4C, the distribution of health spending by age and sex in the Netherlands for a large cross-section of people with the same insurance plan, there is also a more than tenfold difference in average costs between the highest and lowest expenditures.[49]

Age and sex are easy to document and use for risk adjustment, are fair, and generally accepted by all parties involved.[50] Because the information is independent of medical care, and not readily gamed, it appears attractive in terms of incentives. The most serious

[49] Spending in the Netherlands was converted to US currency using the 1998 exchange rate of 2 Dutch Guilders per US dollar. Over the age range from 0 to 64, there is somewhat less variation in health spending in the Netherlands than in the United States.

[50] Separate calculation of health insurance premiums by sex is generally considered acceptable in the US and elsewhere, while separate calculation by race or religion is generally not considered acceptable. Interestingly, charging different insurance premiums by gender for automobile, life and liability insurance in the USA is generally NOT considered acceptable, even though gender-based differences in expected costs may be at least as large in these other insurance markets as in health care.

drawback of age and sex as risk adjusters is simply that they are weak predictors of individual expenditures.[51]

3.3.2. Prior year expenditures

Because expenditures in one year are correlated with expenditure the following year – the correlation coefficient for total health expenditures is on the order of 0.2 to 0.3 – a simple proposal has often been made to regress expenditures in year two on year one expenditures (together with other demographic variables) and use this model for calculating risk-adjusted payments. Newhouse et al. (1989), Van de Ven and van Vliet (1992) and Ash et al. (1998) have all estimated such models and typically find that spending an extra dollar on health care in year one "predicts" spending of $0.20 to $0.30 in year two. The R^2 from a regression that includes age, sex and prior year expenditures, is generally estimated to be in the range of 0.06 to 0.10, with two recent estimates being 0.073 [Van Vliet and van de Ven (1992)] and 0.098 [Ash et al. (1998)]. These measures are a substantial improvement over demographic only models, and comparable to the predictive power achieved by diagnosis-based models or models that use self-reported health status measures.

Although the accuracy of prior year expenditures is reasonable compared to many alternatives, this approach is inferior to others according to several of the above criteria. The feasibility of using such a model is often a concern. In some cases it can be seen as requiring the sponsor to "assume the can opener", since a major reason why a risk-adjusted capitation payment rather than cost-based payment is used is precisely because health expenditures are difficult to measure or monitor. In the USA in particular, a growing number of health plans do not collect individual level cost information that can be used for calculating payment for specific conditions. Instead, many plans have subcontracts with provider groups that do not even require that service and cost information be shared with the plan. The absence of cost or charge information undermines the feasibility of its use for payment in some settings, such as HMOs in the USA.[52] However in other settings, such as the Netherlands, the feasibility requirement is met, since prior year expenditures are routinely available in the administration of the sickness funds.

Although prior year expenditures or utilization appears to be the best single predictor of an individual's future health expenditures, some argue that using it as a risk adjuster creates inappropriate incentives. Firstly, some differences in prior use among individuals

[51] Altman et al. (1998) highlight the important nonlinear relationship between age and health spending, as shown here in Figures 4A to 4C. They note that if a health plan has an above average age, then even without enrollment changes the average health spending of the health plan's enrollees will increase faster than the average because of this nonlinearity. They define this concept as adverse retention, and demonstrate in one sample that as much as biased enrollment and disenrollment, adverse retention may explain why costs of health plans diverge.

[52] An alternative may be to use imputed spending based on a price or fee schedule applied to observed utilization.

could reflect differences in physician discretionary practice patterns (an N-type factor). Premium subsidies based on prior utilization would pay health plans without regard to the appropriateness of the care [McClure (1984)]. Secondly, the premium subsidies would be based on an average relationship between prior use and subsequent medical expenditures. The expected future costs, however, may differ widely for persons with high prior use associated with chronic medical conditions in contrast to those with one-off acute conditions. This might lead to inappropriate provider incentives or to new selection problems [Beebe et al. (1985)].

Some providers and researchers also challenge the fairness of using prior expenditures to calculate payments [e.g., Lubitz (1987) and Porell and Turner (1990)]. The usual argument is that payments based on prior year expenditures reward plans for spending more on individual patients, and punish "well managed" plans that conserve on spending. However, this argument misses the fact that plans are still only compensated for a proportion of their spending on health services. Ellis and McGuire (1993) and Newhouse (1996, 1998) have argued that this may be a desirable practice to soften the incentives of a fully prospective system (a thought that we develop further below). One last argument against using prior utilization/expenditures as a risk adjuster is that it does not provide higher subsidies to individuals with medical problems who have not sought care.

3.3.3. Diagnosis-based risk adjustment

The potential equity and inefficiency problem of inappropriate incentives related to prior utilization as a risk adjuster may be reduced by combining prior utilization with diagnostic information. Since the early 1980's a considerable amount of research has developed risk adjustment models that use diagnoses from insurance claims to calculate risk-adjusted payments. The three most widely known classification systems are the Ambulatory Care Group (ACG) system developed at Johns Hopkins by Jonathan Weiner and colleagues [Weiner et al. (1991, 1996)], the Diagnostic Cost Group (DCG) family of models developed at Boston University and Health Economics Research by Arlene Ash, Randall Ellis, Gregory Pope and colleagues [Ash et al. (1989, 1998), Ellis et al. (1996a, 1996b), Pope et al. (1998a, 1998b, 1999)], and the Disability Payment System (DPS) developed by Richard Kronick and Anthony Dreyfus [Kronick et al. (1996)] primarily for US Medicaid disabled enrollees.[53] Although the above authors have led the development of these classification systems, the models themselves have also been applied by numerous other researchers, notably Van Vliet and Van de Ven (1993) and Lamers (1998) in the Netherlands. Although each of these systems has its own unique features, they share several characteristics that are worth highlighting.

[53] Diagnosis-based risk adjustment models have also been developed by Hornbrook et al. (1991), Clark et al. (1995), and Carter et al. (1997).

The starting point for all diagnosis-based risk adjustment models is the concept that certain diagnoses predict of health care expenditures. Each of the three major diagnosis-based models begins by identifying a subset of all diagnoses that predict current or subsequent year resource use. Although the three models differ in how they choose their subset of diagnoses, each attempts to identify codes that are assigned only for encounters involving a professionally trained clinician. In particular, diagnoses appearing on laboratory, diagnostic testing, and medical supplies claims are uniformly not used in classifying individuals for prediction, on the grounds that they are less reliable than those assigned by clinicians.

Since there are approximately 15,000 valid International Classification of Diseases (ICD9) codes, it is intractable to classify individuals at this level of detail because in most cases there will be too few people with a diagnosis to properly calibrate a model. Each of the models therefore begins by grouping ICD-9 codes into more aggregated groups based on clinical, cost, and incentive considerations. The most refined versions of the ACG, DCG, and DPS systems begin by classifying diagnoses into a large number of diagnostic-based groups, then use these diagnostic groups to classify individuals according to the specific combination of conditions each individual has. The approaches that each model uses, and the way that information is used to generate predictions differ in the three models.

As described in Weiner et al. (1996), the Ambulatory Care Group methodology begins by classifying a subset of all valid ICD9-CM diagnostic codes into 32 diagnostic groups.[54] Depending on the model specified, various combinations of these diagnostic groups are then used to classify each individual into one of up to 83 mutually exclusive Ambulatory Care Groups. In a few cases the ACGs correspond to specific medical conditions (e.g., Asthma); however in most cases the groups are relatively broad (“Acute: Major”, “Chronic Medical, Unstable”, “Chronic Specialty”). Thirteen of the groups are based on counts of how many of the 32 detailed diagnostic groups the patient has, and hence explicitly reward plans for coding more conditions. Payment weights, based on regression analysis, can be used together with ACG assignments to predict individual or group level resource use.

As suggested by their name, ACGs were originally designed to use only ambulatory diagnoses, and hence the ACG algorithms ignore inpatient episodes. Using different classification systems designed to incorporate inpatient diagnoses, the ACG framework has also been expanded to include inpatient diagnoses in two different ways. The first approach summarizes inpatient conditions using simply the 15 Major Diagnostic Categories (MDCs) into which Medicare program’s Diagnosis Related Groups (DRGs) can be grouped. These MDC categories are nonspecific to severity differences within a broad body system (“Infectious and Parasitic Diseases”, “Cancers”, “Diseases of the Circulatory System”, etc.). The second approach uses what Weiner et al. call ”Hospital Dominant” (HOSPDOM) inpatient diagnoses. The model only recognizes inpatient

[54] ACGs have recently been renamed “Adjusted Clinical Groups”. See http://www.hsr.jhsph.edu/acg/acg.html for references and further details about ACGs.

diagnoses for which at least 50 per cent of the patients with that condition were hospitalized in a benchmark dataset. The goal is to avoid rewarding plans for unnecessarily hospitalizing patients in order to increase payments. However this is a very strict criterion for deciding which inpatient diagnoses will affect model predictions.

Although they are not based on the same classification system as the ACG system, the Payment Amounts for Capitated Systems (PACS) developed by Gerry Anderson et al. (1990) is an inpatient diagnosis based system developed using Medicare data. The most distinctive feature of the PACS system is that it counts how many times a person is hospitalized over a two year period within each of fifteen MDCs. It also notes whether the person has any outpatient visits in the year prior to the year being predicted, and classifies hospitalizations into four chronicity levels. The model does not address incentives or other economic, rather than clinical, criteria.

The Diagnostic Cost Group (DCG) risk adjustment models were originally developed by Arlene Ash et al. (1989) using data from the US Medicare population from 1979–80. At the time of its early development, diagnostic information was not yet routinely coded on outpatient claims, and there were also concerns about the completeness of secondary diagnostic codes even on inpatient records. Therefore, the earliest versions of the DCG models used only principal inpatient diagnoses. A series of reports and papers summarized in Ellis and Ash (1995) built upon this early work, used data from the mid 1980's, and explored a variety of extensions that continued to use only principal inpatient diagnoses.

Early DCG models are "single hierarchy" models. Modeling begins by clustering diagnoses into a large number of clinically homogeneous groups. These diagnostic groups were then further aggregated into a small number (between 9 and 20) of Diagnostic Cost Groups (DCGs), according to empirically determined similarities in the future cost of individuals hospitalized with different diagnoses. Some diagnostic groups are ignored in the classification process because they are viewed as being too discretionary or too ambiguously coded. Individuals with multiple hospitalizations in a given year are uniquely assigned to the most expensive DCG in which any of their hospitalizations fell, thus establishing a "single hierarchy". Individual DCG scores are included as categorical variables in linear regression models and used to predict future costs.

A more recent series of studies by the same group [Ellis et al. (1996a, 1996b), Ash et al. (1998), Pope et al. (1998a, 1998b)] has significantly expanded the original DCG framework. One fundamental change is that instead of using only principal inpatient diagnoses, these recent DCG models use all diagnoses from encounters with clinically trained medical professionals, including secondary hospital inpatient diagnoses, hospital outpatient facility diagnoses, and other diagnoses assigned by clinicians. A second fundamental change is that instead of only noting the most serious diagnosis, the models capture multiple conditions. Instead of a single hierarchy used to rank all diagnostic groups, the recent models use information about multiple conditions, and impose hierarchies on diagnostic groups only when they are clinically related to each other. Numerous other important changes were also made. Considerably more clinical input was used to identify selected subsets of diagnoses to use in the first stage of the classification system. The system now includes 543 detailed diagnostic groups, which are further collapsed into 118 groups that are now called HCCs – Hierarchical Condition Categories. The populations studied were expanded from being simply Medicare enrollees to include privately insured and Medicaid eligibles. Instead of clustering diagnoses into cost groups before running a regression, selected diagnostic clusters – the HCCs – are included directly in regressions, so that estimated regression coefficients reflect the incremental cost of specific medical conditions. Concerns about discretionary admission and creating inappropriate incentives were incorporated by excluding selected HCCs from inclusion in regression models that predict subsequent year costs.[55]

In addition to the developmental work in Boston on DCGs, considerable exploration and further developments using the DCG framework have taken place in the Netherlands. Van Vliet and van de Ven (1993) evaluated ten different risk adjustment models, including the original principal inpatient DCG models, DCG models that exclude certain diagnoses due to concerns about discretion noted above, and DCG models customized for the Netherlands. They also draw useful comparisons to models that are based on the PACS system using dummy variables for the MDCs in which each hospitalization falls, and models using prior year expenditures.

Lamers and Van Vliet (1996) expanded the DCG framework by considering multiple years of hospitalizations. The rationale for this is twofold. First, having had a serious hospitalization in a given year might induce predictably above-average expenditures not only in the year directly following but also, to a diminishing degree, in the years thereafter (without necessarily resulting from a new hospitalization). Secondly, by giving higher premium subsidies for people who have been hospitalized for certain diagnoses during one of the previous years (instead of only during the last year), the probability increases that a health plan will receive an appropriate premium for its chronically ill enrollees. The results of Lamers and Van Vliet indicate that the predictive accuracy improves when DCGs over a longer period are incorporated in the subsidy formula. For example, for the five per cent enrollees with the highest costs in year $t - 4$ the predictable losses in year t decreased from 88 per cent (demographic model) to 62 per cent (1-year DCG model) and to 43 per cent (3-year DCG model) of the predicted costs.

Although the DCG-models outperform a model based on age and gender only, there still exist subgroups with substantial predictable losses. Lamers (1999a) showed that when the sponsor uses a (1-year or 3-year) DCG-model, a group of about 30 per cent "bad risks" can be formed by using selected information from a health survey, such as perceived health, having functional disabilities, consultation of the general practitioner, use of home nursing and the number of prescribed drugs. These "bad risks" on average have a predictable loss of more than half the overall mean per capita expenditures. More than 90 per cent of the "bad risks" were not hospitalized in the previous year.

The Disability Payment System (DPS) of Kronick and Dreyfus was developed with the specific aim of risk adjusting payments for persons eligible for Medicaid by reason

[55] See http://www.DxCG.com for references and further discussion about DCG models.

of medical disability. The DPS system is similar to the DCG/HCC system in using all diagnoses from clinical encounters, incorporating hierarchies and concern about incentives, and explaining particularly well the upper tail of the health expenditure distribution. It is somewhat more additive than the DCG/HCC model, taking note of how many conditions a person may have within certain body systems. Because the DPS system has mainly been used for persons with disabilities, it is not clear how well it works for other population subgroups.

One important advantage of diagnosis based risk adjustment is that data often exist for large samples on which models can be developed and calibrated. Although diagnosis based risk adjusters tend to do well in predictive accuracy and feasibility, they do less well on fairness to providers or plans with different levels of completeness in recording diagnoses. Diagnosis based systems almost invariably reward plans that more actively encourage patients to seek treatment. For instance, if a plan screens more aggressively for certain conditions, then they are more likely to detect them, hence increasing payments. Similarly, a plan coding baby deliveries so as to justify performing more Caesarian deliveries will tend to have an enrollee mix that looks sicker than a plan that does not encourage Caesarians. The distortionary effects of using diagnoses for risk adjustment is potentially compounded if the risk adjustment system only notes hospitalizations: if a patient will only be eligible for a higher payment if hospitalized for a condition predictive of higher subsequent year costs (e.g., HIV/AIDS or colon cancer) then a risk adjustment system based on only hospital diagnoses (and not on outpatient care) will encourage unnecessary hospital admissions. This undesired incentive is smaller in a 3-year DCG-model than in a 1-year DCG-model.

3.3.4. Information derived from prescription drugs

Another approach for extracting health status information from prior utilization data is to infer the presence of chronic conditions from the use of prescription drugs. Since pharmacy information is often available with a short time lag, this is another attraction of drug information. Hornbrook et al. (1991) classified drugs into different therapeutic classes. In each class the number of drug orders was counted. Adding 19 drug classes to the adjusters age and gender yielded an increase in R^2 from 0.021 to 0.050. Von Korff et al. (1992) used outpatient pharmacy data to develop the Chronic Disease Score (CDS). The CDS weights were based on physician judgment of disease severity. This CDS was found to predict hospitalization and mortality after controlling for age, gender and health care visits. Clark et al. (1995) revised the CDS by empirically estimating the weights for individual drug classes. They distinguished 28 different conditions. By adding 28 CDS dummy variables as additional risk adjusters to age and gender the predicted variation in total medical expenditures of adults in the next half year increased from 3 per cent to 10 per cent. The results of Clark et al. also suggest that adding information derived from ambulatory diagnoses to the revised CDS adds little additional explanatory power.

Lamers et al. (1999) built on the revised CDS developed by Clark et al. To prevent manipulation Ahey put “alike” conditions (for example hypertension and cardiac disease) in the same group. The chronic conditions could be clustered into six so-called *Pharmacy Cost Groups* (PCGs) on the basis of empirically determined similarities in future costs without affecting the predictive accuracy of the model. Lamers et al. conclude that although PCGs are good predictors of future health care costs, their usefulness as risk adjusters may be restricted because of inappropriate incentives. The additional subsidy for a PCG-classified enrollee (far) exceeds the costs of the prescribed drugs that form the basis for PCG-assignment.[56] A similar result was reported by Ellis (1985), who found that each dollar spent on drugs predicts $3.73 of health care expenditures the following year. Given the high predictive value of CDS and PCGs future research should be directed at minimizing the information surplus of health plans that have access to information about prescription drugs of their enrollees.

3.3.5. Self-reported health information

A fundamentally different approach to risk adjustment is using self-reported measures derived from surveys. Survey-based information has several advantages over diagnosis-based systems [see, e.g., Gruenberg et al. (1996), Hornbrook and Goodman (1996)]: most information is not contingent on having come in contact with a medical provider, no prior history of claims or enrollment is needed to generate predictions, measurement of consumer perceptions of need and anticipated use, uniformity across health plans, and socioeconomic (lifestyle, taste, employment) variables can be measured in addition to health status. There are also some important disadvantages of self-reported measures. Surveys are relatively costly to collect (typical numbers for the USA range upwards from $30 per completed survey). Response rates can be unacceptably low and correlated with medical risk. Large samples generally do not exist on which to develop reliable prediction models. Some survey questions raise confidentiality and accuracy concerns (e.g., questions about HIV/AIDS or mental illness). Although surveys do not require providers or health plans to provide claims information, in many cases health plans are expected to assist with implementation of the surveys, raising concerns about nonrandom sampling or follow up. As shown below, self-reported measures generally do not have as high an explanatory power as diagnosis based systems.

The most common type of information collected through surveys is perceived health status. In its simplest form, it can be a single self-reported health summary of excellent/very good/good/fair/poor. Asking how health status has changed since a year earlier also can be significant. More elaborate surveys such as the Short Form 36 (SF36) [Thomas and Lichtenstein (1986), Ware and Sherbourne (1992)] or the closely related

[56] Given the conclusion by Clark et al. (1995) about the similarity in predictive value of the revised CDS and information derived from ambulatory diagnoses, it is interesting to know whether such perverse incentives also exist in case of risk adjusters based on outpatient diagnostic information.

RAND-36 survey [Hornbrook and Goodman (1995)], measure perceived health status along eight dimensions. A second class of information is functional health status, which assesses how well can the individual perform various activities. The two most common instruments are the Activities of Daily Living (ADLs) and Instrumental Activities of Daily Living (IADLs). A third class of self-reported measures relates to chronic conditions (e.g., diabetes, high blood pressure, asthma, etc.). Such measures, while not coded by a physician, may require that the individual has received a diagnosis from a clinician, and hence to that extent depend on contact with providers. Other self-reported measures include information such as lifestyle (smoking, drinking, food), marital status, employment, education, and whether a person can drive.

Several studies have compared the predictive power of self-reported and diagnosis based risk adjustment models. Table 3 presents R^2 measures from six studies that included various self-reported measures. While all of the models that incorporate self-reported measures are superior to models that include only age and sex, the self-reported models have lower R^2 than the models that include diagnostic information. We note that all of the sample sizes are relatively small, so that overfitting is a concern. Only Lamers and Pope et al. present R^2 which are based on out-of-sample predictions, and hence are robust to overfitting.

3.3.6. Mortality

Mortality has been suggested as an additional risk adjuster because of the high health care expenditures prior to death [see, e.g., Tolley and Manton (1984), Lubitz (1987)]. There are different opinions about its usefulness. Van Vliet and Lamers (1998) conclude that mortality should not be used as a risk adjuster. Their argument is that most of the excess costs associated with the high costs of dying are unpredictable. Even with their most comprehensive regression model ($R^2 = 0.189$; acute care, general population) the actual costs of decedents are still 250 per cent above predicted costs. Furthermore, they found that in the Dutch situation the allocative effect of a mortality-based prospective adjustment based on standardized mortality ratios for the past 5 years, would be very modest: a sickness fund with an extreme excess mortality of 10 per cent could expect an increase of its age-gender based premium subsidies of about 0.25 per cent. One may also wonder whether it is politically and socially acceptable for a health plan to receive a higher subsidy when more of its members die.

In the US, Tolley and Manton (1984) studied the possibility of using cost-weighed cause-specific local mortality for setting Medicare-subsidies to at-risk HMOs. They concluded that it would require a difference in mortality experience of more than 30 per cent to result in a 6 per cent change in the mortality-adjusted AAPCC. Van Vliet and Lamers (1998) note that, although cause-of-death information is theoretically attractive, practical concerns include reliability, validity, availability, manipulation, auditing and privacy of the data.

Another opinion about mortality as an additional risk adjuster is expressed by Beck and Zweifel (1998). They conclude that a dummy variable indicating death during the

Table 3
Comparison of R^2 from various risk adjustment models from six papers

Study	Newhouse et al. (1989)	Van Vliet and van de Ven (1992)	Fowles, Weiner et al. (1996)[b]	Physician Payment Review Commission (1994)	Pope et al. (1998a)	Lamers (1999a)
Sample population	US Privately Insured	Netherlands	US HMO enrollees	US, Medicare	US Medicare	Netherlands sickness fund
Sample period	1974–1979	1981–1982	1991–1993	1991–1992	1991–1993	1991–1994
Sample size	$N = 7{,}690$	$N = 20{,}000$	$N = 5{,}780$		$N = 10{,}893$	$N = 10{,}570$
Age/sex	0.016	0.028	0.058	0.016	0.007	0.038
All socioeconomic[a]		0.037				
Functional status[a]					0.0252	
Self reported chronic conditions[a]		0.071	0.111	0.032	0.0274	
Self reported health[a]	0.028			0.03	0.0311	
Short-form 36 like[a]			0.111	0.033	0.0405	
Prior year spending[a]	0.064				0.0413	
Comprehensive survey[a]		0.114		0.062	0.0418	0.060
Diagnosis based[a]	0.045		0.124[c]		0.0727[d]	0.080[e]
All variables[a]	0.09			0.07	0.0785	0.086

Notes:
[a] All models include age and sex as well as variables shown.
[b] Dependent variable was truncated at $25,000, which inflates R^2.
[c] ACG/ADG model.
[d] DCG/HCC model.
[e] Three-year DCG-model.

observation period should be included in the subsidy formula. They suggest to retrospectively compensate health plans with a prospectively determined payment per death.

The only country, as far as we know, that applies mortality as a risk adjuster is Belgium. The mortality adjuster in Belgium is based on the average number of deaths per 1000 enrollees in prior years at the health plan level [Schokkaert et al. (1996)].

3.3.7. Models using other information

A wide array of alternative risk adjustment information has been examined in the literature. See Epstein and Cumella (1988) for an early review. Van Vliet and Van de Ven (1992) evaluated many demographic variables, including employment, family size, and region, and found that these additional explanatory variables improved the R^2 of 0.028 from an age-sex model to only 0.037 (see Table 3). It is interesting to note that regional dummies increased the R^2 by 0.006 from 0.028 to 0.034. This small improvement may reflect that the Netherlands has less regional variation than many other countries.

Disability and functional health status have been shown to be relatively good predictors of future expenditures [Thomas and Lichtenstein (1986), Hornbrook and Goodman (1996)]. Indicators of functional health status reflect someone's ability to perform various activities of daily living and the degree of infirmity. Disabled and functionally impaired persons appeared to have roughly twice the health care expenditures of those who are unimpaired [Lubitz et al. (1985), Gruenberg et al. (1989)]. Impairment level continued to be a significant contributor to high Medicare expenditures after controlling for demographic factors and prior utilization [Gruenberg et al. (1989)]. Newhouse (1986) considered disability to be an almost ideal adjuster. In 1998 an indicator of disability is used as a risk adjuster in Belgium, Germany and the Netherlands.

As highlighted in Section 2.2, input price variation is a cost factor that is likely to be attractive to include in risk adjustment formulas. In the United States, counties are used as the geographic basis for risk adjustment under the 1998 Medicare HMO payment formulas. Instead of using a county based price index, however, the Medicare program in the US in 1998 used a function of the county level average indemnity payments, by age-sex groups to calculate payments. These county averages reflect not only input price variation, but also practice style variation. Hence they almost certainly overstate the variation that policy makers should use in calculating risk-adjusted subsidies.

Figure 5, based on data in Ellis et al. (1996b), highlights that much of the geographic variation in the US Medicare capitation payment is systematically related to differences in input price variation, although clearly other factors are also varying geographically. The scatter plot highlights that there is a considerable amount of the variation in average costs across metropolitan areas that is explainable by factor price variation. (The simple correlation coefficient between input prices and average costs is 0.61 ($p < 0.001$).) Based on this evidence there appears to be a clear rationale for geographic adjustment for input price differences.

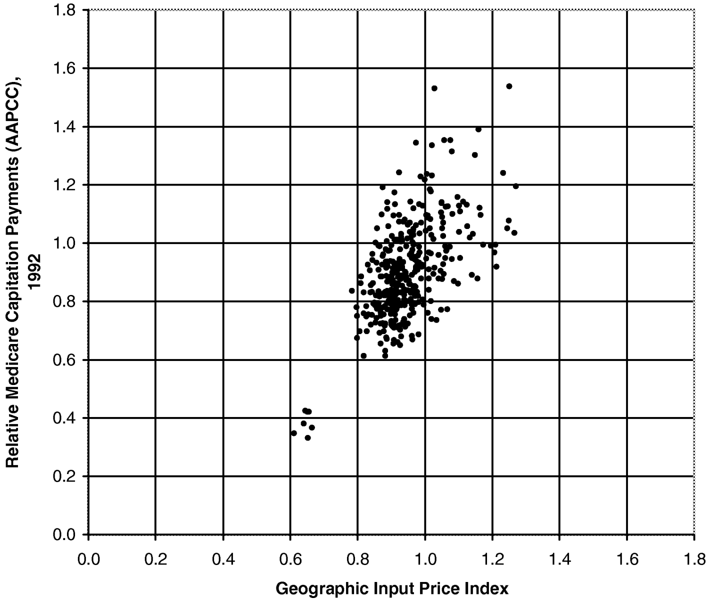


Figure 5. Scatter Plot of Geographic Input Price Index and Relative US Medicare Capitation Payment (AAPCC), 1992, (rho = 0.61). Source: Ellis et al. (1996b, Table G-1).

3.4. Predictive power

3.4.1. Issues with R-square measures

This chapter has so far presented only the conventional R^2 when contrasting the predictive power of different models. In this section, while comparing selected models more fully, we highlight three factors that influence the R^2 that have not always been emphasized in the risk adjustment literature: the impact of truncating the dependent variable, the impact of a logarithmic transformation, and the possibility of overfitting models in small sample sizes. We then present selected additional ways of assessing risk adjustment models.

Figure 6 presents results from five different studies that used different truncation points of the dependent variable. Truncation is also sometimes called top-coding, to distinguish it from censoring, which is the dropping of observations, and is also sometimes done. Whether truncation is appropriate depends on how the risk adjustment system will handle “high outliers”. Truncation is often justified on the basis that health plans may reinsure, in which case very high costs really are not borne by the health plan. While this may be appropriate, it can muddle comparisons across different studies.

Figure 6 makes the important point that truncating health expenditures at a maximum level can have a major impact on predictive power. The predictive power of all of the

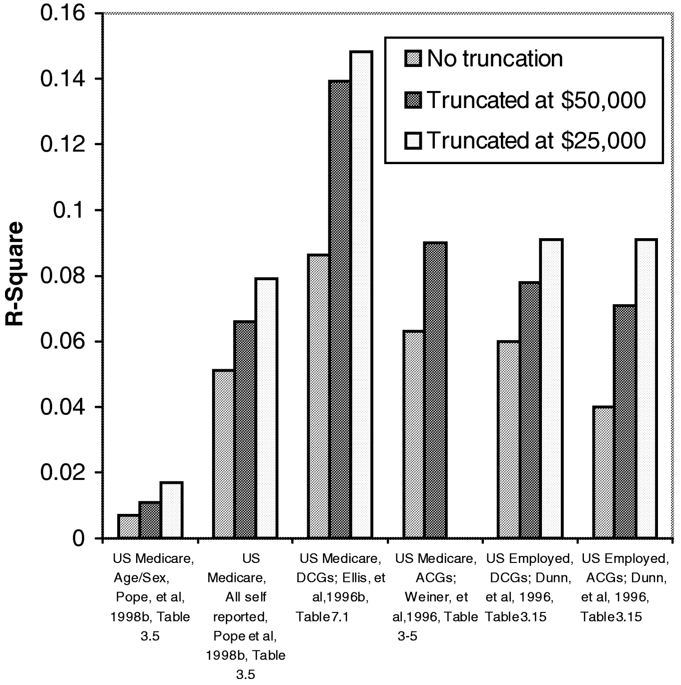


Figure 6. Effects of truncating expenditures on R-square.

models increase on the order of 30–70 per cent once the upper tail of the expenditure distribution is truncated. Note that the difference of truncation is less pronounced with the self-reported measures, and appears to be the greatest for the diagnosis-based models. The comparisons made between DCGs and ACGs by Dunn et al. (1996) suggests that DCGs and ACGs do similarly if health spending is truncated at $25,000, but that DCG models achieve a higher R^2 if one uses untruncated spending.

A second factor complicating comparisons across different studies is that many studies use the natural logarithm of expenditures for some or all of their calculations of R^2. Figure 7 highlights that using a log transformation inflates the conventional R^2 by about 100 per cent, and this holds whether one is using age sex models, self-reported or diagnostic information. As highlighted above, Manning (1998) has demonstrated that the log transformation has undesirable statistical properties for predicting spending in absolute levels, and is difficult to use for generating unbiased predictions even with retransformation.

Figure 8 highlights a problem that has only recently been presented systematically. Ellis and Azzone (1998) used Monte Carlo draws of various sizes from a larger sample to demonstrate that the ordinary R^2 is systematically overstated, even when relatively large samples of 10–50,000 people are used. The problem is more pronounced for the

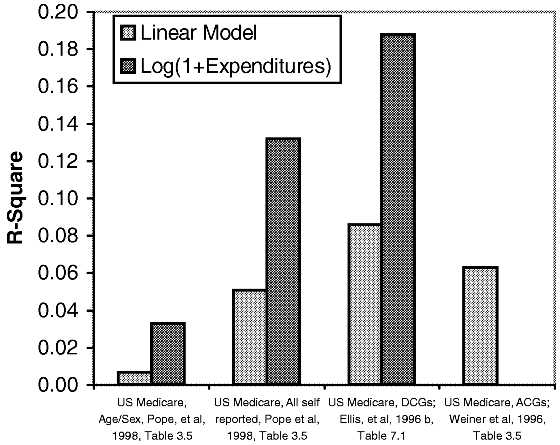


Figure 7. Effects of Log transformation on R-square.

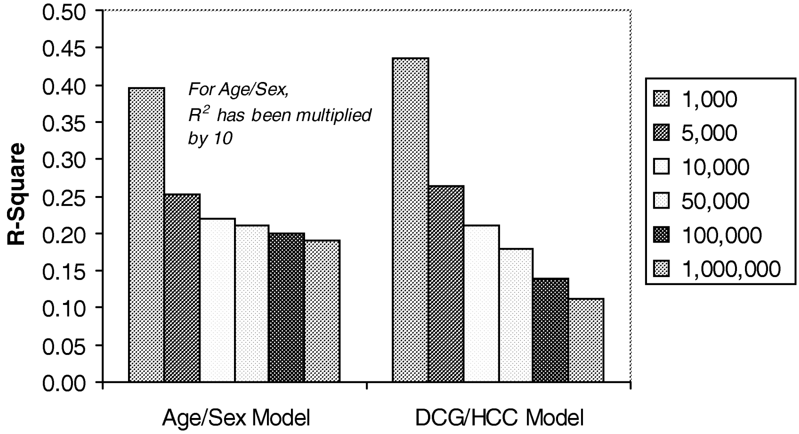


Figure 8. Effects of sample size on R-square. (Mean OLS R^2 from 100 Monte Carlo draws of size shown.) Source: Ellis and Azzone (1998).

more detailed models such as the DCG/HCC model using all diagnostic categories. Yet even simple age-sex models overfit the data for small sample sizes, and overstate the R^2. (Note that the R^2 for the age-sex models has been multiplied by a factor of ten to facilitate being able to see the result on the same scale.) The implication of this finding is that R^2 from diagnosis based models estimated on samples of 50,000 without a (repeated) split sample analysis should be deflated by a factor of 1.3 when comparing to models based on a million observations.

Instead of simply using the individual R^2, Ash et al. (1989) developed the concept of a Grouped R^2. Instead of squaring the difference between actual and predicted spending for each individual, they explored methods of summarizing predictive power using the weighted squared deviations for the averages of exhaustive subgroups. Using this technique, Ash and Byrne-Logan (1998) found that diagnostic-based models such as ACGs and DCGs can explain about 80 per cent of the variation that is explainable by prior year spending once prior year spending is summarized by averaging people into 50 equal-sized samples of 384 people after sorting by year one spending. This measure results in much higher estimates of predictive power because it averages out much of the random error.

3.4.2. Comparisons using other than R^2

The conventional R^2 attaches enormous weight to large outliers: the one person in a sample costing a million dollars more than expected will add as much to the variance as 1,000,000 people with prediction errors of $1,000. To offset this, several researchers have also presented comparisons of different models using the Mean Absolute Error (MAE) [Dunn et al. (1996), Ellis et al. (1996b), Ettner et al. (1998)]. The main difficulty in using this measure is that it is less commonly used, and hence it is more difficult to assess what a good level or improvement of the MAE is.[57]

A further approach that is widely used is to assess the predictive power of various models using selected subsamples of the population being predicted. For example, actual expenditures can be compared to predicted expenditures using each model, for either random or nonrandom subsamples. Comparisons using randomly sampled groups of people provide information about stability of payments and their standard errors. Given that even the best risk adjustment models leave perhaps 90 per cent of the variation unexplained, once samples reach sizes such as 5,000 enrollees, differences between the predictive power of risk-adjusted versus non-risk-adjusted payments for random samples are hardly worth noting. Ingber (1998) provides a good example of the use of this approach.

Given that one of the most important criteria to use in selecting among risk adjustment models is the incentives they create for selecting or dumping certain types of people, one of the most useful assessments that can be done is to compare actual and predicted expenditures for selected nonrandom groups of interest. Figures 9, 10, and 11 present three such comparisons using a privately insured sample from the US from Ash et al. (1998). Figure 9 provides an out-of-sample comparison of actual and predicted expenditures for ten chronic diseases, as identified from insurance claims from inpatient and outpatient bills from hospitals, physicians and other clinically trained professionals. As is readily seen, an age-sex model does not distinguish among the high cost chronic

[57] Also, from the perspective of assuring access for sicker people, larger errors may be disproportionately important.

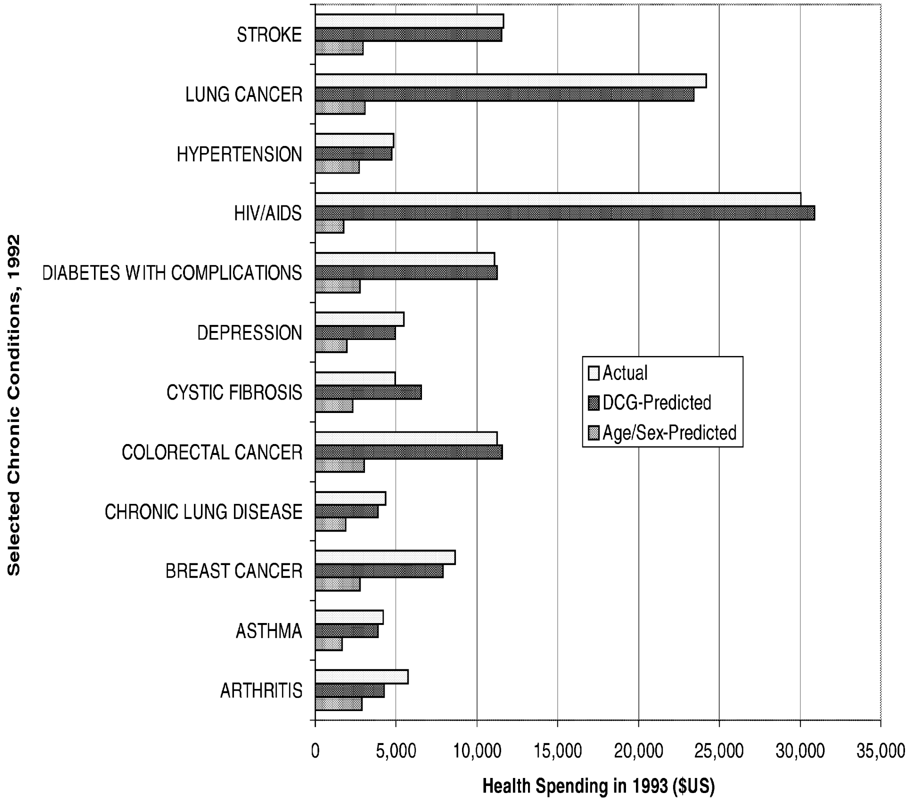


Figure 9. Comparison of actual versus predicted health spending by selected chronic conditions. US privately insured sample ($N = 346{,}466$). Source: Ash et al. (1998).

conditions, and will not reduce incentives to avoid enrolling such individuals. In contrast, a diagnosis based risk adjustment model, such as the DCG/HCC model shown here, can pick up a substantial amount of the variation across different chronic disease groups.[58]

Figure 10 compares predicted and actual spending for nonrandom groups, defined by sorting the sample in ascending order by the predicted level of spending in year two. In Figure 10, this comparison is made using predicted costs from the DCG/HCC model in a privately insured sample. Such a figure is useful for assessing whether the risk adjustment model can identify both very high cost and very low cost individuals, and

[58] The chronic disease groups shown in this figure were selected by HCFA in 1995, and overlap with, but do not coincide with the classification system used by the DCG system. The same chronic groups have been used in Ellis et al. (1996a), Weiner et al. (1996), Ash et al. (1998), and Pope et al. (1998a, 1998b, 1999).

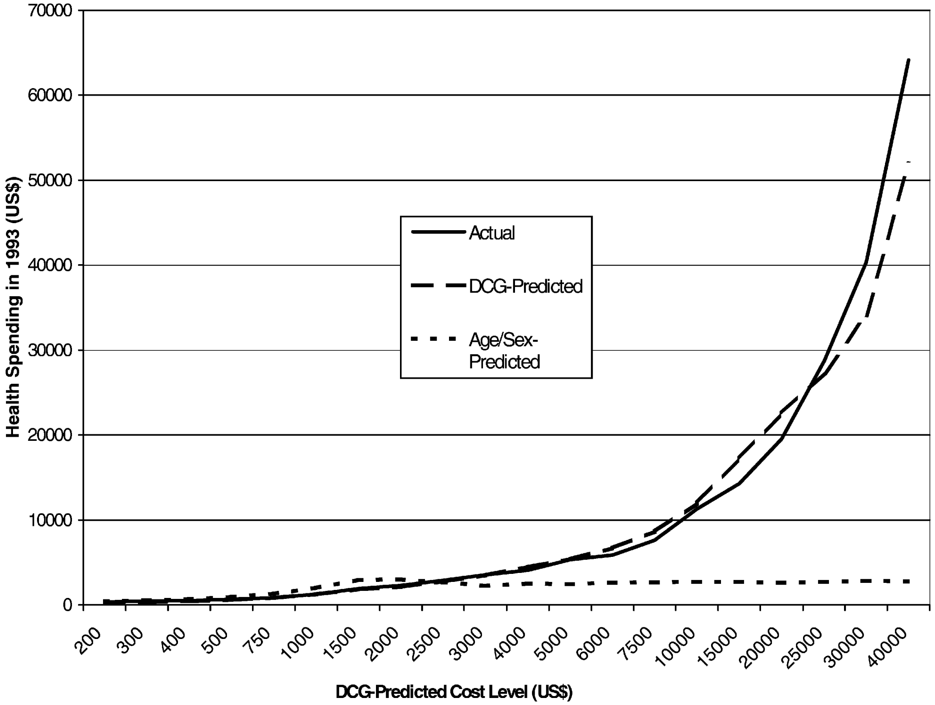


Figure 10. Comparison of actual versus predicted health spending by DCG predicted cost intervals. US private insured sample ($N = 346{,}466$). Source: Ash et al. (1998).

whether the people predicted to be high or low cost do in fact incur these costs. The risk model predicts people with expected costs that range from $270 in the lowest cost interval, to $51,962 in the highest cost group, a factor of nearly 200 to 1. The DCG/HCC predictions track differences in average actual spending quite well for groups defined in this way. In contrast the age-sex model (whose average predicted costs are also plotted) varies by a factor less than ten to one.

While Figure 10 is informative about how well a model can identify high and low cost people, it is likely to be biased in favor of the risk adjustment model that is used to create the intervals shown on the horizontal axis. Since the prediction groups shown on the horizontal axis are defined using groupings from the DCG/HCC model, they are likely to make the DCG model look better than any other model whose predictions are plotted against the DCG defined groups, such as the age-sex model. In a similar way, groups that are defined using self-reported measures are biased in favor of making self-reported measures look good, and groups defined over prior year spending are biased in favor of making prior year spending models look good. [See Pope et al. (1999), and Ash and Byrne-Logan (1998) for a discussion of this issue.]

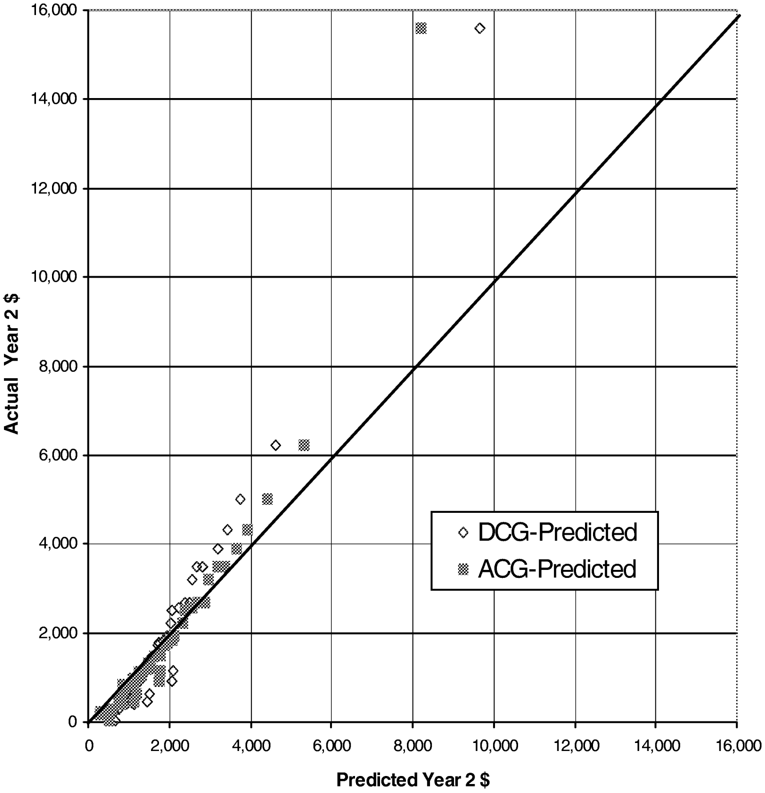


Figure 11. Predicted versus actual year 2 costs with each observation calculated for a 2%-ile group based on year 1 cost. Source: Ash and Byrne-Logan (1998).

Figure 11 presents another way of examining two competing models on neutral territory. Here the population is divided into 50 same-sized groups, based on actual year 1 spending. Each observation is defined by sorting people by year 1 actual spending, and then dividing them into 50 groups. For each group, average year 2 spending is plotted against average predicted spending. The figure compares the predictions of Ambulatory Care Groups (ACGs) with Diagnostic Cost Groups (DCGs). The population is an independent sample of 192,000 privately insured people covered by the Massachusetts Group Insurance Commission [Ash and Byrne-Logan (1998)]. Here ACGs do better than DCGs for many of the low cost groups, but worse for the highest two per cent.

3.5. Directions of ongoing development

Thus far the risk adjustment models discussed have used information to predict individual annual health spending for broad population groups. Several other directions of empirical research have been explored in the literature. We group them into three approaches: those that use information to predict less than annual patterns of spending, those that predict only specific subpopulations, and those that carve out certain services. Finally we discuss ongoing developments around conventional versus optimal risk adjustment.

3.5.1. Timing information

There is potentially a considerable amount of information contained in the date at which new diagnostic or other claims based information is coded. This information is not utilized by any of the risk adjustment models currently in use. For instance, both the ACG and DCG models use diagnostic information without distinguishing whether a person has the diagnosis at the beginning or the end of the base period. Information arriving near the end of the base period will in general be more predictive of spending patterns the following year, and hence it is easy to imagine using this information in making predictions. Ellis and Ash (1989) developed a "Continuous Update DCG model" which used information from a twelve month base period to predict ahead only one month. By making a series of twelve one month predictions, they were able to substantially improve upon the overall predictive power of a prospective model. When the twelve monthly predictions are aggregated to a calendar year, then the resulting DCG models, using only the principal inpatient diagnoses, achieved an R^2 that was more than double that of the simpler model that predicted a full year ahead (0.089 versus 0.039). Ellis et al. (1996a) extended this to use all diagnoses and achieved an R^2 of 0.24. Using a two step estimation algorithm, Ellis (1990) developed a "Time Dependent DCG model" using principal inpatient diagnoses in which nonlinear time profiles for each diagnostic group were estimated. This model also achieved an R^2 of 0.24.

3.5.2. Selected subpopulations

A different dimension of research has been to develop risk adjusters for selected subpopulations. For example, Weiner et al. (1991, 1996) have estimated their ACG models for privately insured, Medicare, and Medicaid populations in the US. Ash et al. (1998) and Pope et al. (1998a) have also done so for DCGs. As previously discussed, the Disability Payment System (DPS) of Kronick et al. (1996) has been developed primarily for Medicaid populations eligible for reasons of disability. Risk adjustment models have also been estimated separately for pediatric populations [Newhouse et al. (1993), Fowler and Anderson (1996)], for persons with HIV/AIDS [Kahn et al. (1995)], and for end stage renal disease (ESRD) patients [Farley et al. (1996)]. The rationale for developing of separate models for each of these groups is that they are vulnerable populations. Incentives under a system that does not specifically distinguish among the characteristics of these subgroups may result in health plans not wanting to enroll them. To the extent that they are unprofitable, or their costs are more uncertain, then there is also a greater risk that payments will be unfair.

3.5.3. Carveouts

Another dimension for ongoing research has been to predict costs for specific sets of medical services. Services such as pharmacy costs, behavioral health care (mental health and substance abuse treatment), dental coverage, and neonatal costs are frequently "carved out" of the expenditures being subsidized. The economic rationale for carveouts can be related to both demand side considerations (they may be harder to manage under a capitated system) or supply side incentive concerns (the carved out services may be more vulnerable to underprovision as health plans attempt to attract profitable enrollees). Frank et al. (1997) and Ettner et al. (1998) develop the rationale for carveouts and separate risk adjustment formulas for behavioral health services in light of the danger that selection incentives will be particularly strong in this group.

3.5.4. Conventional versus optimal risk adjustment

An important dimension of ongoing research is based on the notion that there are two broad approaches to calculating risk adjusters: statistical and economic. More recently, this has been given the names "conventional" versus "optimal" risk adjustment [Glazer and McGuire (forthcoming)]. There is a small but rapidly growing literature that is examining optimal risk adjustment and over the next decade may dominate risk adjustment research.

Conventional risk adjustment modeling has focused primarily on how well various risk factors can predict current health spending. In most cases regression models are estimated so as to generate unbiased predictions conditional on available information. If risk selection is viewed as a problem, as it has been in the US Medicare program, then the sponsor may choose to pay some proportion of the predicted amounts, so as to recover some of the distortion or cost savings from selection efforts. But a key implicit assumption of conventional risk adjustment is that the pattern and level of health spending on a given individual is exogenous, and is not itself affected by the risk adjustment formula.

Another way to think about risk adjusters is to regard them as prices that can be set by regulation, in order to achieve efficiency and equity objectives. In a new literature, researchers are making explicit assumptions about the objective of risk adjustment policy, and the market conditions in which risk adjustment takes place, in order to characterize "optimal" risk adjustment. This optimal risk adjustment literature asks questions such as, "If X is the set of the risk adjusters feasible to use, what are the optimal weights on these adjusters to minimize selection-related inefficiencies". In general, the answer need not be the conditional means that come from statistical research. In an analysis of adverse selection, Glazer and McGuire develop a model in which there is a supply-side moral hazard problem: plans can distort the services offered in order to attract profitable enrollees. They show that if an insurance market contains any element of "separation" of risks in equilibrium, risk adjustment can be improved over statistical average risk adjustment by putting more weight (paying more) for adjusters associated with high costs. In their example, age, an imperfect signal of true severity, is available for risk adjustment. Conventional risk adjustment on age overpays for the healthy and underpays for the sick (generating the selection problem). The plan attracting the sick will be providing too few services in equilibrium. The regulator can do something about this because the sick persons' plan has more old people. By paying more for the elderly, the regulator can increase the spending on the sick. (Corresponding efficiency gains appear in the plan for the healthy too, who may no longer need to separate themselves from the sick.) The new insight of their work is that socially optimal risk-adjusted payments should not simply reflect expected costs, but should also reflect the demand side process through which patients choose health plans and providers, as well as the supply side process through which health plans adjust the service mix to attract specific types of patients.

Notwithstanding the theoretical elegance of their work, its practical relevance depends on the strength of their assumptions. Glazer and McGuire assume that health plans select the profitable enrollees *only* via the distorting services offered and not via other tools for selection (such as those mentioned in Section 2.4.2). If health plans use these other tools, which is quite realistic, paying more for the elderly and less for the young may create a new selection problem: health plans will strive to avoid the young. It also makes it even more rewarding for a health plan to select the healthy elderly (e.g., by selective advertising using the addresses of the fitness club for elderly) than it already was. Further, their assumption of a positive correlation between age and risk factors that are known to the consumers but for which the subsidies are not adjusted, may not always be fulfilled in practice. It may be true for diabetes and cancer, but not for AIDS (young men) and neonatal care (which parents may anticipate because of genetic information or tests). So distorting payments based on an imperfect signal could potentially increase the selection against certain types of patients.

Glazer and McGuire (forthcoming) develop a Bayesian framework in which the sponsor uses an imperfect signal to detect among two patient types, and distorts along a single dimension of services. Frank, Glazer and McGuire (1998) extend this work by examining its empirical implications with multiple services. They show that the sponsor should optimally take into account not only the marginal expected cost of each service, but also the predictability of the service, and how highly correlated the service is with other services. Their empirical specification suggests that the services most vulnerable to being under-supplied are those which predict high future costs, have less uncertainty about this prediction, and are highly correlated with total spending. They use this framework to identify services such as mental health and pharmaceuticals that are most prone to undersupply, but do not try to empirically estimate whether in fact these services are particularly undersupplied. They also show that the method used to risk adjust health plan payments can influence the incentives plans have to distort services.

Shen and Ellis (1998) develop a model in which plans are able to act on private information that cannot be used for risk adjustment by the sponsor. (For example, a plan may know that an individual is hypochondriac.) In their model, plans can perfectly cream skim, but there is no moral hazard problem, i.e. individual expenditures are exogenous characteristics of consumers. They show that conventional risk adjustment can be improved upon in this scenario as well, although interestingly in the opposite direction from Glazer and McGuire. They show that in order to minimize total health payments by the sponsor, payments to plans attracting a favorable selection should be considerably less than that implied by conventional risk adjustment using imperfect risk adjusters.

In one more recent paper from the optimal risk adjustment literature, Encinosa (1998) examines optimal risk adjustment in a world with both moral hazard (HMOs can choose effort) and cream-skimming (HMOs can identify and perfectly select low risk types in the population). He demonstrates not only that risk adjustment may not be able to achieve the social optimum, but that under certain conditions conventional risk adjustment can be worse than no risk adjustment if there is market power by health plans (he examines the duopoly case). It is disturbing, but not surprising that risk adjustment cannot achieve the first best in the presence of market power.

The hallmark of this optimal risk adjustment literature is that conventional risk adjustment can lead to biased predictions of actual spending. This bias can arise either because health plans distort spending patterns, so that spending is not exogenous to payments, or because plans have private information or can otherwise distort enrollment, so that the observed risk factors are biased predictors of actual costs. While a great deal of research remains to be done in this area, we speculate that there may be a reconciliation between conventional and optimal risk adjustment which is that once the behavioral response by the plans has taken place, in general it will be necessary to recalibrate the risk adjustment model to reflect actual spending patterns *within the competing, risk-adjusted health plans*, rather than spending patterns from before risk adjustment occurred, or from some fee-for-service sector.

4. Risk sharing

If the risk-adjusted premium subsidies are not sufficiently refined to reduce selection, a complementary strategy is risk sharing between the sponsor and the health plans [Gruenberg et al. (1986), Newhouse (1986)]. *Risk sharing* implies that the health plans are retrospectively reimbursed by the sponsor for some of the acceptable costs[59] of some of their members. Consequently the risk-adjusted premium subsidies have to be adjusted to the health plans' new financial risk. There is a clear analogy between such risk sharing and the outlier payments in the system of diagnosis-related group (DRG-)payments to hospitals [see, e.g., Keeler et al. (1988)]. Although risk sharing effectively reduces a health plan's incentive for selection, it also reduces its incentive for efficiency. So, given

[59] Because it is hard to define the acceptable cost level of a health plan, in practice some proxy is generally used.

some restrictions related to the premium contribution and given an open enrollment requirement, there is a tradeoff between selection and efficiency [Newhouse (1996)]. The goal of risk sharing as discussed in this section is to reduce the health plans' *predictable* losses and profits, while preserving their incentives for efficiency as much as possible. It is not the goal of risk sharing to reduce a health plan's financial risk by reducing the random variation of its expenditures. This may be achieved by traditional reinsurance.

An essential difference between traditional reinsurance and risk sharing as discussed in this chapter, is that for reinsurance a health plan has to pay a *risk-adjusted* premium to the reinsurer. Consequently, reinsurance does not reduce the health plan's predictable losses on high-risk individuals. It even increases them because of the loading fee included in the reinsurance premium. Therefore, traditional reinsurance cannot be a tool to reduce the health plans' incentives for selection.[60] Risk sharing, as discussed here, could be described as a "mandatory reinsurance program with regulated reinsurance premiums" as distinct from voluntary reinsurance with risk-adjusted reinsurance premiums. The retrospective payments from the sponsor are comparable with a reinsurer's retrospective payments, and the payment that a health plan forgoes because of the financing of the sponsor's retrospective payments (see next paragraph) can be considered a "mandatory reinsurance premium".

There are at least three ways to finance the sponsor's retrospective payments to the health plans: First, the sponsor may reduce the premium subsidy. The reduction of the subsidy per risk-group could be equal to the mean per capita predicted ex post payments that all health plans together receive for consumers in this risk-group. Alternatively, all premium subsidies could be reduced by a certain percentage.[61] Second, the sponsor may ask a non-risk-adjusted payment ("mandatory reinsurance premium") from the health plans.[62] Third, the sponsor may ask higher solidarity contributions from the consumers; at the aggregate level the additional contributions should equal, apart from transaction costs, the reductions in premium contributions that the health plans offer the consumers because of their reduced expenditures due to the risk sharing. The choice of how to finance the retrospective compensation may depend on the institutional context, the regulatory framework, the modality of the subsidy system (see Figure 1) and the precise form of the restrictions on the premium contributions. When introducing risk sharing the sponsor should change the weights in the subsidy formula to adjust the premium subsidies for the new financial risks that health plans bear.

[60] The major functions of traditional reinsurance are to protect a health plan against insolvency and to increase its financial capacity to underwrite coverage [Bovbjerg (1992)].

[61] The sponsor's retrospective payments to the health plans can be considered a second type of subsidy from the sponsor to the health plans (see Section 1.2).

[62] In this case the sponsor's role with respect to risk-sharing could be taken over by an independent insurer entity.

4.1. Forms of risk sharing

4.1.1. Risk sharing for all members

Risk sharing between the sponsor and the health plans can take several forms. The sponsor's retrospective payments may depend on the plan's acceptable costs, which serve as the basis for setting the risk-adjusted premium subsidies. Because it is hard to define a plan's acceptable cost level, in practice the sponsor's retrospective payments often depend on the plan's actual incurred expenses or its imputed spending based on a price or fee schedule applied to observed utilization. In the latter case the incentives to produce the units of service efficiently are preserved. Further, the sponsor's retrospective payments may depend in different ways and to various degrees on the plan's acceptable cost or its proxy. For example, the sponsor may retrospectively reimburse each health plan a fixed percentage, e.g., 50%, of all its acceptable costs. This type of risk sharing has been proposed by Ellis and McGuire (1986, 1993), Gruenberg et al. (1986), and Newhouse (1986, 1994). They variously referred to it as "supply-side cost sharing", "partial capitation" and "a blend of capitation and fee-for-service". They discussed risk sharing in the context of modality A of the subsidy system (see Figure 1) with community-rated premium contributions. For the general case of any form of restrictions on the premium contributions and any modality of risk-adjusted subsidies we will refer to this type of risk sharing as *proportional risk sharing*. In the US, the widespread practice of experience rating health premiums at the employer (i.e. sponsor) level is a form of proportional risk sharing.

Another form of risk sharing is for the sponsor to compensate each health plan only for a certain percentage of the acceptable expenditures above a certain annual threshold, for example $ 20,000, per member.[63] Generally speaking, we will refer to this as *outlier risk sharing*.[64] There are clear analogies between outlier risk sharing (between the sponsor and the health plans) and the outlier pools for hospital and physician reimbursement [Keeler et al. (1988), Ellis and McGuire (1988)]. Outlier risk sharing requires that health plans account for all acceptable expenditures for each of their members. For the time being this requirement may reduce its practical applicability in some countries.

Risk sharing reduces both the incentive to deter nonpreferred risks and the incentive to attract preferred risks. The predicted losses on nonpreferred risk are reduced because

[63] Another variant is that the sponsor compensates each health plan for a certain percentage of all acceptable expenditures above a *cumulative* annual threshold for *all* its members together. We refer to this as *stop-loss* risk-sharing. Although a stop-loss risk-sharing arrangement would provide the health plans with good solvency protection, the effect on the reduction of selection will probably be low. Only if a health plan expects its future annual expenditures to exceed the stop-loss limit is there no incentive for selection. Probably, however, the sponsor would want health plans to have an incentive for efficiency and therefore will set the stop-loss limit at such a level that the majority of the health plans do not exceed it.

[64] Beebe (1992) referred to it as outlier-pooling.

the retrospective payments that a health plan expects to receive for persons who are above-average-risks within their premium-risk-group generally exceed their contribution to finance the sponsor's retrospective payments. Because the opposite holds for the low risks within each premium-risk-group, their predicted profits are reduced by risk sharing.[65]

4.1.2. Risk sharing for high-risks

A common feature of proportional and outlier risk sharing is that the health plan retrospectively receives compensation also for members whom it ex-ante did not consider high-risks, e.g., healthy persons who had a car accident. The retrospective compensation from the sponsor for those members does not contribute to reduce the *predictable* losses and therefore does not reduce the incentives for selection; it only reduces the health plan's incentive for efficiency. To improve the effectiveness of risk sharing Van de Ven and van Vliet (1992, p. 38) proposed that a health plan itself decides for which members it will share the risk with the sponsor. According to their proposal each health plan would be allowed to *ex-ante* designate a specified percentage of its members (for example, 1 or 4 per cent) for whom the sponsor retrospectively would reimburse all or some acceptable expenditures.[66] In advance of the contract period (e.g., a year) each health plan would inform the sponsor which of its members it will share the risk with the sponsor. The group of selected members may change every contract period. A rational health plan will assign those members for risk sharing whom it predicts will have the highest losses. The risk sharing for these high-risk members could apply to a certain percentage of their expenses, or to their expenditures above a threshold, or to a combination. "*Risk sharing for high-risks*" is an effective tool to reduce the incentives for selection if health plans can predict *very high* losses for a *small* group of (potential) members, e.g., because they know the results of lab tests or genetic testing, or they know some specific medical conditions not accounted for by the risk adjusters.

"Risk sharing for high-risks" can be considered a form of pro-competitive arrangement that, from the health plans' point of view, tries to simulate the free competitive market. A free health plan market may lead to discontinuity of coverage, through the

[65] These effects hold on average. The precise effect in an individual case depends on the form of risk sharing, the level of the predictable profit/loss without risk sharing, and the way that the sponsor's retrospective payments are financed.

[66] An alternative is that each health plan is allowed to *ex post* designate a specified percentage of its members, e.g., 1%, for whom the sponsor retrospectively reimburses all or some acceptable expenditures [Van de Ven et al. (1994, p. 130)]. For a statistical analysis of this alternative, including a comparison with 'risk sharing for high-risks', see Van Vliet (1999). Because under this alternative the selected members would be all persons with losses above a certain threshold whose value a health plan may accurately predict, such an alternative is similar in its incentives for efficiency to a system of outlier risk-sharing. A difference is that the threshold amount is not the same for all health plans. Another difference concerns the incentives for selection [Van Barneveld (2000)].

refusal to insure some high-risk applicants or to the exclusion of pre-existing medical conditions [Light (1992)]. Instead of terminating the contract or refusing high-risk applicants, a health plan can now assign high-risk persons for risk sharing with the sponsor. This risk sharing arrangement can be organized such that the high-risk persons themselves are not aware of the risk sharing.

To improve the effectiveness of the risk sharing an information system could be set up to reduce the market imperfection that exists in case a health plan cannot accurately assess the financial risk that a new applicant generates [Newhouse (1994)]. Health plans could receive relevant (standardized) information from the prior plan, or from the sponsor (e.g., whether or not the person was selected for risk sharing in the prior year). Alternatively, a health plan could be allowed to have a health interview with newly enrolled members.

An advantage of "risk sharing for high-risks" is that it may reduce the health plan information surplus vis à vis the sponsor. Because health plans use their information surplus for selecting the high-risk applicants, an actuarial analysis of the risk-profile of the assigned high-risk members, when compared with that of non-assigned members, may provide the sponsor with useful information which can improve the risk adjustment mechanism in successive years. In this way the sponsor can progress in its attempts to incorporate in the risk adjustment mechanism as much information as the health plans have.

A problem with "risk sharing for high-risks", that does not occur with the other forms of risk sharing, is setting the premium contribution for the high-risk members. Without risk sharing a health plan will ask high-risk applicants to pay the *maximum* premium contribution that is allowed under the regulation. On the other hand, if the expenses for a high-risk member are retrospectively reimbursed by the sponsor, the appropriate premium contribution would be the *minimum* premium contribution allowed under the regulation. Because a health plan selects its members to be assigned for risk sharing after it knows all the members in its portfolio for the next contract period, the question is, which premium contribution should ex-ante be offered to a high-risk applicant: the maximum or the minimum premium contribution? The extent of this problem diminishes, of course, as the difference between the maximum and minimum premium contribution narrows. In the extreme case of community rated contributions all members per health plan should be quoted the same premium contribution.

The concept of "risk sharing for high-risks" has been studied by Van Barneveld et al. (1996, 1998) for modality A (see Section 2) of the subsidy system with community-rated premium contributions within each health plan.[67] Van Barneveld et al. (1996) suggest that another variant in the case of poor risk adjusters is to have the percentage

[67] Van Barneveld et al. (1996, 1998) refer to it as mandatory high-risk pooling. Major differences with the high-risk pools in the US [see Bovbjerg and Koller (1986), Zellner, Haugen and Dowd (1993)] are that under mandatory high-risk pooling the high-risk members pay the same (community-rated) premium as others, they have the same benefits package and the same copayment-structure as others, and they are unaware that their health plan shares their risk with the sponsor.

of members to be selected for risk sharing to depend on a health plan's average loss per member – before risk sharing and adjusted for the difference between the health plan's premium contribution and the national average premium contribution – in a preceding year. The rationale is that, especially with crude risk adjusters, these losses are caused mainly by the inability of the crude risk-adjusted capitation to compensate adequately for health status.

When discussing Medicare's method for reimbursing at-risk managed care plans in the US, Newhouse et al. (1997) proposed to implement "risk sharing for high-risks" in addition to proportional risk sharing for all members. They expect "risk sharing for high-risks" to be especially useful for dealing with the terminally ill.

4.1.3. Condition-specific risk sharing

So far we have discussed three forms of cost-based risk sharing. An alternative is to retrospectively reimburse the health plans some *prospectively* determined payments dependent on the occurrence of some medical problems [Luft (1986), Enthoven (1988)]. We refer to these arrangements as "*condition-specific risk sharing*".[68] The payments can be based on diagnoses that are relatively invulnerable to manipulation and for which high cost treatment is relatively non-discretionary. Because the amount of the payment is prospectively determined, and not dependent on a health plan's actual costs, condition-specific risk sharing does not change the plan's incentive to produce the units of service efficiently.[69] Given that the goal of risk sharing is to reduce the incentives for selection, condition-specific risk sharing contributes to this goal only insofar as health plans ex ante know that there are individuals with an above average probability within their premium-risk-group of having or developing the specific condition, or as far as consumers ex ante know they have an above average probability within their premium-risk group to do so. In that case condition-specific risk sharing may be a valuable addition to the DCGs developed by Ash et al. (1989) and the DCG/HCCs developed by Ellis et al. (1996a). Otherwise, condition-specific risk sharing, just like traditional reinsurance, only reduces the health plan's financial risk, without reducing the incentives for selection.

An advantage of "risk sharing for high-risks" over condition-specific risk sharing is that it prevents "diagnosis-inflation" and political battles over which conditions are to be compensated [Swartz (1995)], and it does not retrospectively reimburse expenses for members whom the health plan ex ante did not consider to be high-risk persons. An advantage of condition-specific risk sharing over "risk sharing for high-risks" with a fixed percentage of the selected members is that plans that specialize in certain high-cost

[68] Condition-specific risk sharing differs subtly from retrospective risk adjustment (Section 3.2.2) because with the latter the weights or payments are *retrospectively* determined. In practice this difference will be negligible, so that condition-specific risk-sharing and retrospective risk adjustment come to the same thing.

[69] Related to condition-specific risk sharing is Beck and Zweifel (1998) proposal to give health plans retrospectively a prospectively determined payment for each member who dies.

Table 4
Forms of risk sharing

	Reimbursement rate	Threshold	Ex-ante percentage of members to whom the risk sharing applies
Proportional risk sharing	r	0	100
Outlier risk sharing	r	T	100
Risk sharing for high-risks	r	T	p

treatments and that are therefore flooded by high-risk members (because of adverse selection), receive appropriate compensation for *each* of these high-risk members. Health plans who have no high-risk members at all (possibly because of selection), do not receive any retrospective compensation.

4.2. Empirical results

In principle there exist many forms of risk sharing between the sponsor and the health plans. In this section we discuss some empirical results with respect to forms of risk sharing which have been reported in the literature. As illustrated in Table 4, many of these forms of risk sharing can be described by three parameters [Van Vliet (1997)]:

r, the reimbursement rate;

T, the threshold amount;

p, the percentage of members, to be ex-ante assigned, to whom the risk sharing applies.

There is no common terminology in the literature. Different authors use different terms for the variants of risk sharing. Here we adopt the terminology given in Table 4.[70]

The first empirical study, to our knowledge, on risk sharing in the context of risk adjustment is Beebe's (1992) analysis of outlier risk sharing, which he called an outlier pool. The pool would pay 45 per cent of the expenditures above a threshold amount. Beebe varied the threshold between $100,000 and $10,000 (1992 US-dollars). He analyzed US Medicare data, so most of the sample were above the age of 65. The percentage of persons exceeding the threshold varied from 0.07 to 11.1 per cent. The pool's payments varied from 0.14 to 19.5 per cent of the total expenditures. Beebe concluded that an outlier pool payment method could provide some protection against the risk of an unexpectedly high proportion of high-cost users at a relatively modest cost. He did not examine the outlier pool's ability to reduce the incentives for selection.

Van Barneveld et al. (1996) analyzed the effects of "risk sharing for high-risks" for modality A of the subsidy system with age/gender-adjusted premium subsidies and with

[70] An alternative is to adopt a common term, e.g., Risk-Sharing, and to specify the value of the above three parameters, r, T and p.

community-rated premium contributions. Their findings indicate that under these conditions the mean per capita loss for the 1 per cent of individuals with the highest prior-year expenditures are 8.5 times the overall mean per capita expenditures. This illustrates the great potential of "risk sharing for high-risks" to reduce the health plans' incentives for selection, because health plans could select their members simply on the basis of the prior year's costs. This appears to be an effective selection strategy.[71] To the extent that health plans are able to better predict which of their members belong to the long right tail in the distribution of residual costs not accounted for by the risk adjusters used by the sponsor, the more effective is "risk sharing for high-risks". For the next 1 per cent group with highest prior-year costs, the mean per capita predictable loss falls to 4.5 times the overall mean per capita expenditures. For the next two percentiles this ratio falls to about 2.5, and after that it falls below 1.5. From these figures Van Barneveld et al. (1996) conclude that risk sharing for less than 4 per cent of the members would be most meaningful. If still more members were allowed to be selected, the marginal reduction of predictable losses would be small, while the incentives for efficiency would be lowered further.

Another illustration of the effectiveness of "risk sharing for high-risks" is that with age/gender adjusted capitation the mean per capita predictable loss for the 8 per cent of individuals who were hospitalized *two years ago* is about 1.1 times the overall mean per capita expenditures. If the sponsor would allow risk sharing for 4 per cent of the members, and if health plans would select these 4 per cent members on the basis of *prior-year* expenditures, the predictable loss on those members who have been hospitalized two years ago, would be reduced by about two-thirds [Van Barneveld et al. (1996, Table 3)]. That is, the gross returns on potential selection activities based on hospitalizations two years ago would be reduced by two thirds. Because of the costs of selection strategies, the net returns would go down even more. This reduction of the incentive for selection should not come at the expense of substantially reduced incentives for efficiency because the health plan remains fully at risk for the 75 per cent of the total expenditures caused by the 96 per cent non-selected members. Further, because a health plan remains responsible for the high costs of persons with *unpredictable* high expenditures, which comprise the majority of all high costs, the selected members may be "free riders" as far as the health plan's managed care activities are concerned. If the sponsor turns out to bear the major financial responsibility for certain medical treatment programs, e.g., transplantation, open heart surgery or HIV-treatment, the sponsor should be involved in the process of managing these types of care.

In another study Van Barneveld et al. (1998) compared, under the same conditions as above, the effectiveness of proportional risk sharing, outlier risk sharing with

[71] If health plans would select their members for "risk sharing of high-risks" on the basis of available information on prior hospitalizations and prior costs in the three preceding years, the mean per capita predictable loss for the 1% individuals with the highest losses would increase by less than 10 per cent [Van Barneveld et al. (1998, Table 2)].

100 per cent reimbursement and "risk sharing for high-risks" with 100 per cent reimbursement and no threshold. Because of the tradeoff between efficiency and selection, they chose the values of the parameters r in proportional risk sharing, T in outlier risk sharing and p in "risk sharing for high-risk" such that on average the percentage of the total costs for which the health plans are at risk, was the same for each form of risk sharing. Based on the prior discussion, an approximately optimal choice of p appeared to be 4 per cent. The corresponding values for r and T were 20 per cent and 10 times the mean spending, respectively. For each form of risk sharing the premium subsidies were proportionately reduced to keep the sponsor's total outlay the same. (This reduces the incentive to attract good risks.) As an indicator of the effectiveness of the different forms of risk sharing they used the reduction of the mean per capita predictable loss for the 20 per cent of individuals who had been hospitalized in the four preceding years. Without risk sharing this predictable loss is slightly more than the overall mean per capita expenditures. Proportional risk sharing reduced this predictable loss by 20 per cent, outlier risk sharing reduced it by 41 per cent and "risk sharing for high-risk" by 51 per cent. The predictable profits on the 80 per cent individuals who had not been hospitalized in the four preceding years, are reduced by the same percentages. Therefore, Van Barneveld et al. (1998) conclude that "risk sharing for high-risks" is more effective in reducing the incentive for selection than the two other forms of risk sharing.

Van Vliet (1997) concluded that the effectiveness (in terms of reducing the plans' incentive for selection) of "risk sharing for high-risks", relative to proportional and outlier risk sharing, can be further increased by reducing the reimbursement rate, e.g., from 100 per cent to 80 per cent, while at the same time increasing the percentage of selected members (keeping total retrospective payments constant).

A different type of empirical study has been done by Keeler et al. (1998). They simulated the effect of several forms of risk sharing on the adverse selection that occurs if consumers have an annual choice among three different health plans with varying generosity of coverage. The three simulated plans differed only in the cost-effectiveness ratio that their treatments should surpass. The expenses of the generous plan are nearly double the expenses of the stingy plan for an average case-mix population. It is assumed that health plans cannot use treatment policy to discriminate against the sick. The sponsor requires the plans to ask a community-rated premium contribution from their members (modality A of the subsidy system). Further it is assumed that health plans are not actively selecting healthy enrollees by other forms of selection than the differentiation of the generosity of their coverage. Consumers with different health status, income and tastes for health care are assumed to choose their most preferred health plan during the annual open enrollment period. The acceptable costs, on which the sponsor's subsidy (capitation) is based, are the costs of the middle plan. A first conclusion from this simulation is that flat capitation, i.e. no risk adjustment at all, results in severe adverse selection. The healthy individuals are overrepresented in the stingy plan and the sick in the generous plan. Without the assumption that half the persons will stay in their original plan, there would be no equilibrium. Because the sponsor's subsidy is the same for each consumer, the stingy plan appears to be overcompensated by 30 per cent, relative to its risk-profile, and the generous plan undercompensated by 37 per cent. Keeler et al. (1998) simulated the effects of several forms of risk sharing on the extent of the sponsor's overpaying and underpaying. They found that proportional risk sharing at a 25 per cent reimbursement rate and outlier risk sharing of in total 10 per cent of the sponsor's outlay, each reduced the sponsor's over- and underpaying by 35 to 50 per cent. A form of condition-specific risk sharing which compensates about 25 per cent of the plans' expenses, reduced the sponsor's over- and underpaying by about two-thirds. With all forms of risk sharing the capitation payments are proportionately reduced to keep the sponsor's total outlay the same.

4.3. Discussion

4.3.1. What form of risk sharing is optimal?

Given the effectiveness of different forms of risk sharing to reduce selection Van Barneveld et al. (1998) conclude that "risk sharing for high-risks" should be preferred rather than proportional or outlier risk sharing. This conclusion seemingly conflicts with the view of Newhouse et al. (1997) who argue for proportional risk sharing rather than for other forms of risk sharing. The explanation for these seemingly different conclusions is that Van Barneveld et al. consider risk sharing only as a tool to reduce the incentives for selection in case of imperfect risk adjustment with restrictions on the premium contributions, while Newhouse et al. consider it also as a tool to reduce the incentive for quality skimping, which may occur even in the absence of any incentive for selection.

By *quality skimping* we mean the reduction of the quality of care to a level which is below the minimum level that is acceptable to society. The argument is that if a health plan's marginal revenue is zero for the additional services that its members receive, the plan may have a incentive for quality skimping. Although perfect information and competition in the plan market would prevent underprovision, one should not exclude the possibility that the same information problems that enabled fee-for-service providers to order and profit from excess care, may prevent patients or their agents from punishing underprovision [Keeler et al. (1998)]. According to Newhouse et al. (1997) the ideal form of risk sharing pays a plan a prospectively set marginal cost and a capitation such that the plan breaks even on that case. This would address both concerns of selection and quality skimping.[72] In practice, however, we do not have marginal cost and thus will not have an ideal payment system.[73]

The extent to which risk sharing can be an effective tool to reduce incentives for quality skimping, even in the absence of any incentive for cream skimming, depends (at least) on the type of health plan.[74] We can discern several types of health plans.

[72] Concerns of moral hazard remain [Newhouse (1986)].

[73] For an extensive discussion on the relation between capitation and quality of care, see the chapters on physician payment by McGuire (2000) and Pauly (2000).

[74] It may also depends on the type of benefits included in the health plan coverage [see Van de Ven and Schut (1994)].

Our definition of health plan (see Section 1) is "a risk-bearing entity that performs at least some insurance function, but that may also manage or provide health care". So, on the one hand, a health plan can be a traditional commercial insurer that has no contractual relationship with the providers of care and which (partly) reimburses the fee-for-service bills sent by the providers to the consumers [what Enthoven (1994) calls a "remote third-party payer"]; and on the other hand a health plan can be a managed care organization, e.g., a capitated provider group, that itself delivers the care to its members. Only for the latter type of health plan is the literature on optimal provider reimbursement relevant. In this literature it is argued that, rather than full (risk-adjusted) capitation, some form of reduced supply-side cost-sharing is optimal [Ellis and McGuire (1986, 1993); see also Pauly (1980)]. In our terminology this could be a form of risk sharing between the sponsor and the health plans.[75] However, if the health plan is a traditional "remote third-party payer", the arguments about optimal provider reimbursement do not influence the optimal structure of the insurance premium [Ellis and McGuire (1990), Selden (1990)] and risk sharing between the sponsor and the health plans cannot be considered a tool for reducing the incentives for quality skimping.

4.3.2. Proportional risk sharing or prior costs as a risk adjuster?

So far we have considered risk sharing as complementary to insufficient risk adjustment. So, we discussed risk adjustment and risk sharing as separated issues. However, there is a similarity between the two, as Newhouse (1986) argued. He compared the situation where the sponsor subsidy to the health plans depends on prior cost (or prior utilization) with the situation that it depends on current cost (or current utilization). Prior costs is used as a risk adjuster in the premium subsidy formula. Current cost is applied in the form of a blend of capitation (not dependent on prior costs) and current costs. We refer to the latter as proportional risk sharing. Newhouse (1986) argued that prior cost and current cost are similar in their incentive effects, or can be made so (except for those who die or switch plans).[76] Given this similarity, Newhouse favors proportional risk sharing rather than prior-costs-adjusted capitation because current utilization shows recognition of changes in health status as they occur, rather than with a delay. Newhouse considers current utilization a more sensitive measure of predictable variation in expected cost then prior utilization.[77]

Some other points can be noted if the sponsor subsidy depends on either prior costs or current costs (or utilization). First, the way that the retrospective payments from

[75] An alternative is direct consumer cost-sharing with a coinsurance rate of, e.g., 30% [Manning and Marquis (1996)] or prior-cost as a rating factor in the premium contribution model with a weight of e.g., 0.30.

[76] The same similarity exists between outlier-risk-sharing and a risk-adjuster "high prior cost", i.e. prior costs only as far they exceed a certain threshold, e.g., the 99 percentile of the empirical distribution of costs [as applied by Lamers and Van Vliet (1996)].

[77] Another argument is that in the Medicare system in the US no prior cost (or utilization) data are available for a new cohort of enrollees.

the sponsor in the case of proportional risk sharing are financed (see above) may have distributional effects that differ from those of prior-cost-adjusted subsidies. Second, the weights given to the other adjusters may change after inclusion of prior costs in the subsidy formula. Third, the premium subsidy need not be a linear function of prior costs, and it may depend on several years claims records rather than one year's claims records. Fourth, prior costs or prior utilization as a risk adjuster may need some adjustment in case of the opening/closing of hospitals and other health care facilities in the region.

4.3.3. Ongoing research

The research on risk sharing as a tool to reduce the incentives for selection is in an early stage. Because risk adjustment in principle is the preferred option to prevent selection, it is not surprising that the research on risk sharing started some ten years later than the risk adjustment research. Given the growing consensus in the literature about the need for some form of risk sharing to complement imperfect risk adjustment, and given the primitive forms of risk adjustment currently used in most venues (see Section 5), and given the growing awareness that "it is now clear that risk adjustment is a very complex technical issue, and that it will be extremely expensive to try to build the capability to create close to perfect risk adjusters" [Swartz (1995)], there is a growing need for further research on risk sharing.[78]

Future research should focus on getting to know the terms of the tradeoff between efficiency and selection. A conceptual framework could be built to weigh efficiency against selection, taking into account the different types of managed care strategies, e.g., case by case management versus the management of special treatment programs, as well as the different types of selection strategies as mentioned in Section 2.4. Prediction models should be developed that health plans could use in practice, given the information in their administration. Future research should then try to answer questions like: What is the distribution of predictable profits and losses if health plans use their best prediction model, given a certain subsidy formula and given certain restrictions on the premium contribution? How do we value an overall reduction of the predictable losses versus a selective reduction of the highest predictable losses only? [See Van Barneveld et al. (1999).] What are the optimal values of the parameters of risk sharing for several subsidy-formulae and several forms of restrictions on the premium contribution? Future research could also focus on the consequences of risk sharing on the subsidy formula, i.e. the recalculation of the premium subsidies, and on the health plan's premium setting. Finally, attention should be paid to the similarities and differences between forms of risk sharing (proportional and outlier risk sharing) and prior costs as a risk adjuster.

[78] For some work in progress concerning risk-sharing, see Van Barneveld (2000).

5. The practice of risk adjustment and risk sharing

5.1. International comparison

In the late 1990s risk adjustment and risk sharing are being applied in competitive health plan markets in at least eleven countries (see Tables 5 and 6).[79] All countries, except the US, implemented these financing mechanisms in the 1990s. In all countries, expect Ireland and Switzerland, the solidarity contributions are income-related. Most countries use age and gender as risk adjusters. In addition, some countries adjust the sponsor premium subsidy also for region and disability. The most predictive risk adjusters mentioned in Section 3, are not yet in common use. The major exception is the US, where some programs have implemented diagnosis-based risk adjustment [Dunn (1998)] and where the Medicare program has announced that it will use diagnosis-based risk adjustment in the year 2000 to pay HMOs for their enrollees [Greenwald et al. (1998)]. We speculate that three major barriers have contributed to the delayed implementation of risk adjustment in many countries: the recency of the most promising research, the non-availability of relevant data, and inertia. Because in most countries risk adjustment is a dynamic process,[80] over time we may expect to see research results to be implemented in practice.

All countries listed in Tables 5 and 6 have stringent restrictions on the variation of premium contributions. In four, the sponsor requires the premium contribution to be zero, that is, the premium must equal the premium subsidy. In the other seven countries the health plans are allowed to ask from their members a community-rated premium contribution. As discussed in Sections 2 and 3 the combination of poor risk adjusters and stringent restrictions on the premium contributions results in large predictable losses on high-risk individuals. Given this conclusion, it is surprising to see that one half of the countries mentioned in Tables 5 and 6 have no form of risk sharing to reduce the predictable losses and profits.

Despite the strong financial incentives created by capitation payments, selection and its adverse effects have only infrequently been reported as a major problem in most countries.[81] The primary example where selection is pervasive is the US, where there is considerable evidence that health maintenance organizations (HMOs) enjoy favorable selection[82] and where concern is growing about the adverse effects of selection on the quality of care, especially for high-risk patients. In its recent Report to Congress, on

[79] In other countries proposals for a competitive health plan market are under discussion, e.g., Poland (to be implemented in 1999), Argentina, Chile, Portugal and Taiwan.

[80] For example, in the Netherlands and the United Kingdom in 1991 and 1992 the subsidy was based on (estimated) prior costs at the plan level. Subsequently more risk-adjusters were implemented.

[81] For an academic discussion on the potential threat of cream skimming, see, e.g., Newhouse (1982), Luft (1986), Van de Ven and Van Vliet (1992) and Matsaganis and Glennerster (1994). This discussion is summarized in Section 2.

[82] See, e.g., Hellinger (1995), Lichtenstein et al. (1991), Luft and Miller (1988) and Robinson et al. (1993).

Table 5
The practice of risk adjustment and risk sharing in 10 countries*

	Belgium	Colombia	Czech Republic	Germany	Ireland	Israel	Netherlands	Russia	Switzerland	United Kingdom
Risk-adjusters	age/gender, region, disability, unemployment, mortality	age/gender, region	age	age/gender, disability	age/gender, hospitalization, both weighed with current expenses	age	age/gender, region, disability	many different regional experiments	age/gender, region	age/gender, prior utilization, local factors
Restrictions on premium contribution	community rating	zero premium contribution	community rating	community rating	community rating	zero premium contribution	community rating	zero premium contribution	community rating per region	zero premium contribution
Risk sharing	proportional risk sharing, at least 85%	no	no	no	see risk adjusters (above)	severe diseases (6 per cent of expenses)	outlier risk sharing & proportional risk sharing	many different regional experiments	no	outlier risk sharing (£6000) (mid 1990s)
Number of health plans	6	24	26	1200	2 (until 1997:1)	4	25	100s	166	2500 (early 1996)
Modality A or B	A	B	B	B	B	A	A	A	B	A
Open entry for new health plans? (subject to certain conditions)	no	yes	yes	yes	yes	yes	yes	yes	yes	yes
Open enrollment every month/. . ./year	quarter	year	year	year	year	half-year	year	year	half-year	no open enrollment guarantee
Is long-term care included in benefits package?	yes	no	no	no	no	no	no	no	no	no
Mandatory or voluntary membership?	M	V	M	M	V	M	M	M	M	V
Year of implementation	1995	1994	1993	1994	1996	1995	1991	1993	1993	1991

* Source: Chernichovsky and Chinitz (1995), Files and Murray (1995), Ham (1996), Kennedy (1996), Kesenne (1996), Londono (1996), Matsaganis and Glennester (1994), McCarthy et al. (1995), Schneider (1996), Schokkaert et al. (1996), Sheiman (1994, 1995), Sheldon et al. (1994) and Shirom (1995).

Table 6
The practice of risk adjustment and risk sharing in the US*

	Medicare program, HMOs in 1997	Medicare, proposed for HMOs in year 2000	Federal Employees Health Benefits' Program (FEHBP)	New York State	Health Insurance Plan of California (HIPC)	Minnesota Buyers Health Care Action Group	Washington Health Care Authority
Risk-adjusters	age/gender, region (county), institutional status, welfare status	age/gender, region (county), welfare status, Principal Inpatient Diagnostic Cost Groups (PIP-DCGs)	no risk-adjusters. Each consumer's subsidy is based on 60% of the average premium of the six largest plans	age/gender, region	gender, number of children, 120 marker diagnoses, risk adjustment only applied if plan scores deviate from 1 by ±5 per cent	ACGs	age, sex, employee status since 1989, DCGs announced for 2000
Restriction on premium contribution	community rating	community rating	community rating	community rating	premium contribution depends on age, region and family/single within a rate band (±10%)	premium contributions set by competitive bidding	premium contributions based on competitive bids
Risk sharing	no	no	no	condition-specific risk sharing	no	stoploss for catastrophic individuals	yes
Number of health plans	100s	100s?	100s	?	28	15	3
Modality A or B	A	A	A	B	B	A	A
Open entry for new health plans?	yes	yes	yes	yes	yes	yes	no
Open enrollment every month/ . . . /year	month	month, with proposed transition to year	year	?	year	year	year
Is long term care included in benefits package	no	no	no	no	no	no	no
Mandatory or Voluntary membership?	V	V	V	V	V	V	V
Year of implementation	1972	2000	1960	1993	1992	1997	1989

* Source: Buchmueller (1997), Butler and Moffit (1995), Dunn (1998), Lee and Rogal (1997), McCarthy et al. (1995) and Shewry et al. (1996).

Medicare Payment policy, the Medicare Payment Advisory Commission (1998) highlights that new enrollees in Medicare managed care plans cost about 35 per cent less than the Medicare fee-for-service average in the six months prior to enrollment. In contrast, Medicare expenditures on persons disenrolling from HMOs averaged 60 per cent above average in the six months following disenrollment. This finding is also supported by comparisons using self-reported health status measures from the 1994 Medicare Current Beneficiary Survey [Riley et al. (1996)].

Several arguments explain why selection may not be a major issue in the early stage of the implementation of a risk adjustment mechanism in a (potentially) competitive health plan market, and why over time selection may increasingly become a problem. First, in the early stage many players may be unfamiliar with the rules of the game. For example, in the Netherlands, even five years after the implementation of the consumer's right to change health plans annually, many consumers were unaware of this right. In addition, people often associate changing health plans with a potential non-acceptance, the exclusion of pre-existing medical conditions, or higher premiums. Also in the early stage not all managers working within in a health plan fully understand the financial incentives of the financing mechanism. So, in the beginning this lack of knowledge, which is enlarged by the complexity of the risk adjustment system and by the annual changes in the system, may restrict the selection problems. However, over time consumers and managers will be better informed and can be expected to react to large incentives for selection that occur in a system without adequate risk adjustment. Secondly, in the early stage in most countries the differences among health plans with respect to benefits package, premium contribution and contracted providers are relatively small. Over time, especially when less stringent government regulation with respect to planning facilities and medical pricing permits health plans to diversify the conditions of the contracts with their members, we may see more market segmentation. Thirdly, in most countries the risk adjustment mechanism has been implemented in the (mandatory) social health insurance sector. Traditionally most of the health plans working in that sector are highly driven by social motives rather than by financial incentives. However, with open entry for new health plans (subject to certain conditions), as is the case in all countries except Belgium, new health plans may make the behavior of the traditional health plans more incentive driven. As the chief-executive-officer of a large Dutch sickness fund said: We are administering the social health insurance now with one additional limiting condition, i.e. "our expenses should not exceed our revenues".[83] Finally, one may argue that selection is not so much of a problem because doctors may be reluctant to discriminate among risks because of medical ethics. However, present ethics may change if the entire delivery system becomes more competitive. We share Newhouse's (1982) skepticism that medical ethics are sufficient to make selection effects unimportant. In our opinion, appropriate financial incentives and appropriate rules of the game should provide the ultimate safeguard against the adverse effects of selection.

[83] Until 1991 the Dutch sickness funds received from the sponsor a full reimbursement of all their expenses.

From Tables 5 and 6 we see that a variety of forms of risk sharing are applied in practice. In Belgium (1998) the health plans are at risk for about 3 per cent of their expenses.[84] In the Netherlands (1998) proportional risk sharing (with $r = 0.95$) for the fixed (i.e. production independent) hospital costs[85] (about one third of the total health care expenses) is combined with proportional risk sharing ($r = 0.40$) and outlier risk sharing (with $r = 0.90$ and $T = 4500$ guilders) for all other health care expenses.[86] New York State applies a form of condition-specific risk sharing in its Medicaid program. In the United Kingdom (UK) the general practitioners (GP) fundholders, who in our terminology can be considered health plans, were in 1998 fully compensated for all expenses above £6000 per person per year. The fundholders' budgets are determined by negotiations around an "activity-based capitation bench-mark" based on the age/gender characteristics of the practice, and also the practice's historic activity [McCarthy et al. (1995), Sheldon et al. (1994)]. Because negotiations allow the influence of other local factors or expenditure to be included in the budgets, some implicit forms of risk sharing are already incorporated into the budgets.

Most countries have an annual open enrollment period. Two major exceptions are the US and the UK.[87] Until 2002 Medicare members in the US may change health plans every month, which gives more opportunities for selection than does annual choice. In the UK the GP-fundholders could refuse to accept new patients and they had the right to request that any patient should be removed from their list without explanation [McCarthy et al. (1995)]. In 1992–93 78,000 patients, about one in 600 of the population of England, were removed from a GP's list specifically at the request of the GP [Bevan and Sheldon (1996)].

In seven countries membership in a capitated health plan is a mandatory aspect of the social health insurance system. In three countries consumers have alternative options within the system: they may choose the traditional public system (Colombia and UK) or the traditional fee-for-service system (US Medicare). In Ireland the risk adjustment mechanism applies to voluntary private health insurance, which is complementary to the National Health Service. Although the whole Irish population is entitled to receive public health care, about one-third of the population has a voluntary supplementary insurance, primarily to receive private care and thus bypass the queue in the public system. The various selection effects (see Table 1) may depend on *voluntary* or *mandatory* membership in a capitated health plan. On the one hand the voluntary character of supplementary insurance with community-rated consumer contributions in Ireland speeds

[84] Another unique aspect of the Belgium system is that the benefits package for which the health plans are at risk, also covers long term care. This may be an (additional) argument for not giving the health plans too much financial risk [see Van de Ven and Schut (1994)].

[85] The subsidy for the fixed hospital costs is based on a plan's prior year costs.

[86] The threshold of 4500 guilders is slightly more than double the average per capita total annual expenditures. In 2000 the threshold went up to 10,000 guilders.

[87] Another exception is Germany, where the firm-based sickness funds are exempted from the open enrollment requirement.

up adverse selection. The low risks simply do not buy the insurance (and thereby do not pay a solidarity contribution), resulting in a continuously upward premium spiral. On the other hand, the voluntary character of membership in a capitated health plan in the US Medicare system may dampen some of the adverse effects of selection. With mandatory membership high-risk consumers would be forced to choose one of the competing capitated health plans, each of which has a disincentive to be responsive to their preferences. This could potentially result in poor service and poor quality of care for them. Due to the voluntary character of membership in a capitated plan, the high-risk persons are able to choose, as an alternative, the traditional fee-for-service sector, where physician fees likely exceed their marginal costs.

In the US risk adjustment is applied by different types of sponsors (federal government, states, employer groups). In addition to the projects mentioned in Table 6 the Ohio Medicaid program applies risk adjustment and risk sharing with contracted health plans [Kronick et al. (1995, p. 20)], and risk adjustment programs using diagnosis-based information have been implemented in the late 1990s in Washington, Minnesota, and Colorado [(Lee and Rogal (1997), Dunn (1998)].

In addition to the above mentioned differences in risk adjusters and forms of risk sharing, there also appear to be a great variety in the number of competing health plans, the modality of the subsidy system (A or B), and the institutional context and regulatory regime. For example, there are large differences in the extent to which health plans are allowed to manage the care, e.g., by selective contracting, by negotiating prices, by building new facilities, or by buying new equipment. There are also differences in the conditions to be fulfilled by new health plans entering the market and differences in the benefits to be covered. Because of all these differences it is hard to compare and evaluate the effects of different forms of risk adjustment and risk sharing in practice.

5.2. *Problems in practice*

A major practical problem is the availability of data with which to risk adjust. In some countries (e.g., Belgium, Colombia, Germany, Israel and Russia) there are no data available that link individual consumer characteristics with individual health care expenditures. So the first generation risk adjusters in these countries are based on available aggregated data. Some subsidy formulae are based on utilization data per age/gender group for major types of care, weighted by their relative proportion in total health care spending.[88] Over time, as individual expenditure data become available, better subsidy formulae can be applied.[89] Another problem is that for some risk adjusters the average per capita expenditures are known only per subgroup of this risk adjuster, but not

[88] For the details of the subsidy formula and the risk sharing arrangement in eight countries, see McCarthy et al. (1995), Kennedy (1996) and Schokkaert et al. (1996).

[89] For the development of the subsidy formula the sponsor should ideally have at its disposal a large data base with individual consumer information about expenditures and as many risk factors as possible. If the sponsor had the same information as the health plans routinely administer, the sponsor can simulate the predictable losses and profits for subgroups known to the health plans. For the day-to-day operational administration

for the sub-subgroups in interaction with the other risk adjusters. Consequently some sub-subgroups are over- or undercompensated because of correlations between risk adjusters.

The (non-)availability of data largely influences the type of risk adjusters or risk sharing to be used. For example, the application of DCGs or HCCs as risk adjusters requires that the sponsor has access to the relevant diagnostic information of the members of each health plan. In the US Medicare system this is less of a problem than elsewhere [Ellis et al. (1996b)]. For example, in the Netherlands the specialists working in the hospital refuse, for privacy reasons, to provide the sickness funds with diagnostic information about individual hospital admissions. On the other hand, in the Netherlands detailed information on health care expenditures (also per subsidy-risk-group) per sickness fund is routinely available to the sponsor. As a result the marginal administrative costs of risk sharing between the sponsor and the sickness funds are relatively low. Because these cost data are not routinely available in HMOs in the US, Beebe (1992) proposed to use hospital stays as the basis for outlier risk sharing. So, the availability of either diagnostic information or prior costs may influence the sponsor's preference for either DCGs/HCCs or prior costs as a risk adjuster, as well as for either condition-specific risk sharing or some other form of risk sharing.

A practical problem with risk sharing is assessing the acceptable costs that form the basis for risk sharing and for setting the risk-adjusted premium subsidies. In the Netherlands there is a very detailed specification of the basic benefits package in combination with a standardized fee schedule and the sponsor closely monitors all the sickness funds' expenditures and decides which expenses are acceptable and which are not. This procedure will become more complicated the more degrees of freedom the plans have for managing the care and negotiating different price- and quality-levels, and the more the health plans integrate the specified basic benefits coverage with supplementary health insurance (as HMO's do in the US Medicare system). The problem of the acceptable cost-level is related to the lack of clinical consensus on the treatment of certain conditions. It is also related to the distinction between the S-type and N-type risk factors (see Section 2.2.), i.e. the risk factors for which solidarity is desired or not desired. This distinction appears to be an issue especially in Belgium [Schokkaert et al. (1998)].

The problem of the non-availability of data for risk sharing may be reduced by using data that the health plans routinely collect for their own reinsurance. With respect to risk adjustment the sponsor may tackle the data problem by announcing that, after a reasonable period of time, higher subsidies for certain subgroups will be given only if the consumer or the health plan provides the sponsor with certain information. This provides the health plans with an incentive to routinely collect the required data.

A conclusion we can draw from the experience in practice is that even the simplest risk adjustment mechanisms are complex [Gauthier et al. (1995)] and that start up "surprise problems" can be expected [Bowen (1995)]. However, several sites in the US in

of the risk-adjusted subsidies it is sufficient if the sponsor knows the per capita normative expenditures per subsidy-risk-group and each health plan's number of members per subsidy-risk-group.

the 1990s have made progress with the implementation of health based risk adjustment, suggesting that it is indeed possible to overcome both technical and political hurdles [Rogal and Gauthier (1998)].

6. Directions for future research

We have already covered so much that it would be very difficult to try to summarize it. Instead, we would like to end by identifying a few topics that have not yet received significant attention, but seem likely to be the focus of significant research in the future.

Risk adjustment has already come a long way over the past two decades, increasing both in its predictive power and in its sensitivity to creating appropriate incentives. It appears likely that the next decade will also see large improvements in predictive power, with the improvements coming in many areas. Those that seem most promising to us include using more refined clinical information (such as the results of lab tests or clinical chart information); pharmaceutical data; combining claims with self-reported information; or building better models of patterns of service use over time. Episode-based models, and models that make better use of the timing of new information generated during the year also hold out promise of improving the predictive power of risk adjustment models. Models that predict individual level expenditures on specific services instead of in aggregate terms may also hold out promise. We may also expect to see more realistic simulations of health plans' incentives for selection. This might entail weighting profits and losses unequally, for example, by giving greater weight to larger profits and losses than to small ones, or by giving greater weight to profits and losses that persist over time than those that occur only in the short term. In addition, models may be developed that compensate high-risk individuals for their above average fluctuations around the expected spending.

We have spent a considerable amount of effort in this chapter documenting the many diverse ways in which health plans may behave strategically in order to attract or retain profitable enrollees. Clearly as risk adjustment is implemented in more and more countries in more and more settings, it will be important to generate both theoretical models and empirical measures of the magnitudes of this behavioral response.

One reason for understanding health plan behavioral response to the premium subsidy calculations is to better inform the theory and empirical implementation of optimal risk adjustment research. Analogously to the extensive research of the last decade that has attempted to model and understand provider response to the method used to reimburse them, we anticipate that the next decade will see a proliferation of research that examines how premium payments, premium subsidies, and ex post risk sharing to health plans influences plan level behavior.

In order to understand how health plan behaviors will influence the enrollment patterns of consumers that choose among competing health plans, it would be useful to understand how well consumers can anticipate their own health care needs (e.g., using information from genetic testing), and how willing they are to change health plan enrollment in order to act on these expectations. Indeed there is a significant literature on individual choice of health plans, but greater attention could be paid to how these choice variables and consumer information are related to expected spending. A great deal of attention has focused on determining how much of the variation in health care spending is potentially explainable using individual level information. Yet it may turn out that consumers are either more naive than the researcher's predictive models, or consumer inertia or noneconomic factors result in selection problems that are less serious than predicted by the models.

We have highlighted that regulation of plan level competition and standardization of many plan features is an important mechanism for constraining cream skimming and other forms of selection activities. In addition to studying selection behaviors, studying the effectiveness of different regulations would be very helpful.

Risk adjustment and risk sharing are two different strategies for reducing risk selection incentives. Although there is a considerable literature on each, we are not aware of a literature that has examined the tradeoff between the two approaches or carefully examined how risk-sharing arrangements alter the desired risk adjustment formulas. This line of research seems particularly relevant given that in practice risk sharing arrangements are very common at the time that risk adjustment is introduced. In addition, future research could develop criteria for comparing different forms of risk sharing that aims at reducing both the incentives for selection and the incentives for quality skimping.

This paper has made a first step at assembling a few tables that compare risk adjustment and risk sharing internationally. Perhaps as interesting as studying settings where risk adjustment is being introduced is to understand why it has to date been so rarely used. As further experimentation goes on, it will also be helpful if countries could learn from the mistakes and successes of other countries. This would require that comparisons take place on a regular and systematic basis.

References

Altman, D., D.M. Cutler and R.J. Zeckhauser (1998), "Adverse selection and adverse retention", American Economic Review 88(2):122–126.

Anderson, G.F., E.P. Steinberg, N.R. Powe, S. Antebi, J. Whittle, S. Horn and R. Herbert (1990), "Setting payment rates for capitated systems: a comparison of various alternatives", Inquiry 27:225–233.

Arrow, K.J. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53(5):940–973.

Ash, A.S., and S. Byrne-Logan (1998), "How well do models work? Predicting health care costs", Proceedings of the Section on Statistics in Epidemiology (American Statistical Association) 42–49.

Ash, A.S., R.P. Ellis, W. Yu et al. (1998), "Risk adjusted payment models for the non-elderly", Final report for Health Care Financing Administration, June.

Ash, A.S., F. Porell, L. Gruenberg et al. (1989), "Adjusting Medicare capitation payments using prior hospitalization data", Health Care Financing Review 10(4):17–29.

Ashley, J., and W. McLachlan, eds. (1985), Mortal or morbid? – A diagnosis of the Morbidity Factor (Nuffield Provincial Hospitals Trust, London).

Beck, K., and P. Zweifel (1998), "Cream-skimming in deregulated social health insurance: evidence from Switzerland", in: P. Zweifel, ed., Health, the Medical Profession and Regulation (Kluwer, Dordrecht, The Netherlands) 211–227.

Beebe, J.C. (1992), "An outlier pool for Medicare HMO payments", Health Care Financing Review 14:59–63.

Beebe, J.C., J. Lubitz and P. Eggers (1985), "Using prior utilization to determine payments of Medicare enrollees in health maintenance organizations", Health Care Financing Review 3(6):27–38.

Bevan, G., and T. Sheldon (1996), "Review of methods of risk rating in the UK National Health Service", in: J. Hermesse, ed., Risk Structure Compensation, Proceedings of an AIM Workshop (Maastricht).

Bovbjerg, R.R. (1992), "Reform of financing for health coverage: what can reinsurance accomplish?", Inquiry 29:158–175.

Bovbjerg, R.R., and C.F. Koller (1986), "State health insurance pools: current performance, future prospects", Inquiry 23:111–121.

Bowen, B. (1995), "The practice of risk adjustment", Inquiry 32:33–40.

Bradshaw, J. (1972), "A taxonomy of social need", in: G. McLachlan, ed., Problems and Progress in Medical Care (Oxford University Press, London) 71–82.

Buchmueller, T.C. (1997), "Managed competition in California's small-group insurance market", Health Affairs 16(2):218–228.

Buchmueller, T.C., and P.J. Feldstein (1997), "The effect of price on switching among health plans", Journal of Health Economics 16:231–247.

Butler, S.M., and R.E. Moffit (1995), "The FEHBP as a model for a new Medicare program", Health Affairs 14(Winter):47–61.

Carter, G., J. Newhouse and D. Relles (1990), "How much change in the case mix index is DRG creep?", Journal of Health Economics 9(4):411–428.

Carter, G.M., R.M. Bell, R.W. Dubois et al. (1997), "A clinically detailed risk information system for cost", Report for HCFA, RAND report number DRU-1731-1-HCFA, November.

Chapman, J.D. (1997), "Biased enrollment and risk adjustment for health plans", unpublished doctoral dissertation (Harvard University).

Chernichovsky, D., and D. Chinitz (1995), "The political economy of health system reform in Israel", Health Economics 4:127–141.

Chinitz, D., A. Preker and J. Wasem (1998), "Balancing competition and solidarity in health care financing", in: R.B. Saltman, J. Figueras and C. Sakellarides, eds., Critical Challenges for Health Care Reform in Europe (Open University Press, Buckingham, UK) 55–77.

Clark, D.O., M. von Korff, K. Saunders, W.M. Baluch and G.E. Simon (1995), "A chronic disease score with empirically derived weights", Medical Care 33(8):783–795.

Cochrane, J.H. (1995), "Time-consistent health insurance", Journal of Political Economy 103(3):445–473.

Cutler, D.M., and S.J. Reber (1998), "Paying for health insurance: the tradeoff between competition and adverse selection", Quarterly Journal of Economics 113(2):433–466.

Cutler, D.M., and R.J. Zeckhauser (2000), "The anatomy of health insurance", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 11.

Diamond, P. (1992), "Organizing the health insurance market", Econometrica 60(6):1233–1254.

Duan, N., et al. (1983), "Smearing estimate: a nonparametric retransformation method", Journal of the American Statistical Association 78:605–690.

Dunn, D.L. (1998), "Applications of health risk adjustment: What can be learned from experience to date?", Inquiry 35:132–147.

Dunn, D.L., A. Rosenblatt, D.A. Tiaira et al. (1996), "A comparative analysis of methods of health risk assessment", Final report to the Society of Actuaries.

Efron, B. (1978), "Regression and ANOVA with zero-one data: measures of residual variation", Journal American Statistical Association 73:113–121.

Ellis, R.P. (1985), "The effect of prior-year health expenditures on health coverage plan choice", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services, Vol. 6 (JAI Press, Greenwich, CT) 247–272.

Ellis, R.P. (1990), "Time dependent DCG model", Report prepared for Health Care Financing Administration, June (Boston University).

Ellis, R.P. (1998), "Creaming, skimping, and dumping: provider competition on the intensive and extensive margins", Journal of Health Economics 17:537–555.

Ellis, R.P., and A.S. Ash (1989), "The continuous update Diagnostic Cost Group model", Report prepared for the Health Care Financing Administration, June (Boston University).

Ellis, R.P., and A.S. Ash (1995), "Refinements to the Diagnostic Cost Group model", Inquiry 32:418–429.

Ellis, R.P., and V. Azzone (1998), "OLS, loglinear and two part models of health expenditure: what do the data tell us?", unpublished working paper (Boston University).

Ellis, R.P., and T.G. McGuire (1986), "Provider behavior under prospective reimbursement: cost sharing and supply", Journal of Health Economics 5:129–151.

Ellis, R.P., and T.G. McGuire (1988), "Insurance principles and the design of prospective payment systems", Journal of Health Economics 7:215–237.

Ellis, R.P., and T.G. McGuire (1990), "Optimal payment systems for health services", Journal of Health Economics 9:375–396.

Ellis, R.P., and T.G. McGuire (1993), "Supply-side and demand-side cost sharing in health care", Journal of Economic Perspective 7:135–151.

Ellis, R.P., G.C. Pope, L.I. Iezzoni, J.Z. Ayanian, D.W. Bates, H. Burstin and A.S. Ash (1996a), "Diagnosis-based risk adjustment for Medicare capitation payments", Health Care Financing Review 12(3):101–128.

Ellis, R.P., G.C. Pope, L. Iezzoni et al. (1996b), "Diagnostic Cost Group (DCG) and Hierarchical Coexisting Conditions and Procedures (HCCP) Models for Medicare risk adjustment", Final Report, Health Economics Research for Health Care Financing Administration, April.

Encinosa, W. (1998), "Risk adjusting in imperfectly competitive markets", Agency for Health Care Policy Research, unpublished paper.

Enthoven, A.C. (1978), "Consumer-choice health plan", New England Journal of Medicine 650–658 and 709–720.

Enthoven, A.C. (1986), "Managed competition in health care and the unfinished agenda", Health Care Financing Review (Annual Supplement) 7:105–120.

Enthoven, A.C. (1988), Theory and Practice of Managed Competition in Health Care Finance (North Holland, Amsterdam).

Enthoven, A.C. (1994), "On the ideal market structure for third-party purchasing of health care", Social Science and Medicine 39:1413–1424.

Epstein, A.M., and E.J. Cumella (1988), "Capitation payment: using predictors of medical utilization to adjust rates", Health Care Financing Review 10(1):51–69.

Ettner, S.L., R.G. Frank, T.G. McGuire, J.P. Newhouse and E.H. Notman (1998), "Risk adjustment of mental health and substance abuse payments", Inquiry 35:223–239.

Farley, D.O., G.M. Carter, J.D. Kallich et al. (1996), "Modified capitation and treatment incentives for end stage renal disease", Health Care Financing Review 17(3):129–142.

Files, A., and M. Murray (1995), "German risk structure compensation: enhancing equity and effectiveness", Inquiry 32:300–309.

Fowler, E.J., and G.F. Anderson (1996), "Capitation adjustment for pediatric populations", Pediatrics 98(1):10–17.

Fowles, J.B., J.P. Weiner, D. Knutson, E. Fowler, A.M. Tucker and M. Ireland (1996), "Taking health status into account when setting capitation rates", Journal of the American Medical Association 276(16):1316–1321.

Frank, R.G., J. Glazer and T.G. McGuire (1998), "Measuring adverse selection in managed health care", NBER Working Paper No. 6825, December (Cambridge).

Frank, R.G., T.G. McGuire, J.P. Bae and A. Rupp (1997), "Solutions for adverse selection in behavioral health care", Health Care Financing Review 18(3):1–14.

Garber, A.M., T.E. MaCurdy and M.B. McClellan (1998), "Persistence of Medicare expenditures among elderly beneficiaries", in: A.M. Garber, ed., Frontiers in Health Policy Research I (MIT Press, Cambridge, MA) 153–180.

Gauthier, A.K., J.A. Lamphere and N.L. Barrand (1995), "Risk selection in the health care market: a workshop overview", Inquiry 32:14–22.

Glazer, J., and T.G. McGuire (forthcoming), "Optimal risk adjustment in markets with adverse selection: an application to managed care", American Economic Review.

Greenwald, L.M., A. Esposito, M.J. Ingber and J.M. Levy (1998), "Risk adjustment for the Medicare program: lessons learned from research and demonstrations", Inquiry 35:193–209.

Gruber, J. (2000), "Health insurance and the labor market", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 12.

Gruenberg, L., S.S. Wallack and C.P. Tompkins (1986), "Pricing strategies for capitated delivery system", Health Care Financing Review (Annual Supplement) 7:35–44.

Gruenberg, L., C. Tompkins and F. Porell (1989), "The health status and utilization patterns of the elderly: implications for setting Medicare payments to HMO's", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 10 (JAI Press, Greenwich) 41–73.

Gruenberg, L., M.H. Kaganova and M.C. Hornbrook (1996), "Improving the AAPCC with health status measures from the MCBS", Health Care Financing Review 17:59–76.

Ham, C. (1996), "Managed markets in health care: the U.K. experiment", Health Policy 35:279–292.

Hamilton, G.J., ed. (1997), Proceedings of the AIM-symposium "Competition and solidarity" (Alliance Nationale des Mutualités Chrétiennes (ANMC) of Belgium, Brussels) March.

Hay, J.W., and R.J. Olsen (1984), "Let them eat cake: a note on comparing alternative models of the demand for medical care", Journal of Business and Economic Statistics 2:279–282.

Hellinger, F.J. (1995), "Selection bias in HMOs and PPOs: a review of the evidence", Inquiry 32:135–142.

Hornbrook, M.C., and M.J. Goodman (1995), "Assessing relative health plan risk with the Rand-36 health survey", Inquiry 32:56–74.

Hornbrook, M.C., and M.J. Goodman (1996), "Chronic disease, functional health status, and demographics: a multi-dimensional approach to risk adjustment", Health Services Research 31(3):283–307.

Hornbrook, M.C., M.J. Goodman and M.D. Bennett (1991), "Assessing health plan case mix in employed populations: ambulatory morbidity and prescribed drug models", in: R.M. Scheffer and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 12 (JAI Press, Greenwich) 197–232.

Hornbrook, M.C., R. Meenan, D. Bachman et al. (1998), "Forecasting health plan liability for defined populations: A Global Risk-Assessment Model", Working Paper, Kaiser Permanente, June.

Iezzoni, L., ed. (1994), Risk Adjustment for Measuring Health Care Outcome (Health Administration Press, Ann Arbor, MI).

Ingber, M.J. (1998), "The current state of risk adjustment technology for capitation", Journal of Ambulatory Care Management 21(4):1–28.

Jensen, G., and M. Morrisey (1990), "Group health insurance: a hedonic approach", Review of Economics and Statistics 72(1):38–44.

Jones, A.M. (2000), "Health econometrics", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 6.

Kahn, J.G., H. Luft and M.D. Smith (1995), "HIV risk adjustment: issues and proposed approaches", Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 8(Suppl.1):S53–S66.

Keeler, E.B., G.M. Carter and S. Trude (1988), "Insurance aspects of DRG outlier payments", Journal of Health Economics 7:193–214.

Keeler, E.B., G.M. Carter and J.P. Newhouse (1998), "A model of the impact of reimbursement schemes on health plan choice", Journal of Health Economics 17:297–320.

Kennedy, A. (1996), "Private health insurance in Ireland; the advent of a competitive market and a risk equalisation scheme", in: J. Hermess, ed., Risk Structure Compensation, Proceedings of an AIM Workshop, Maastricht.

Kesenne, J. (1996), "Health care reform and risk structure compensation: the case of Belgium", in: J. Hermesse, ed., Risk Structure Compensation, Proceedings of an AIM Workshop, Maastricht.

Kronick, R., Z. Zhou and T. Dreyfus (1995), "Making risk adjustment work for everyone", Inquiry 32:41–55.

Kronick, R., T. Dreyfus, L. Lee and Z. Zhou (1996), "Diagnostic risk adjustment for Medicaid: the disability payment system", Health Care Financing Review 17:7–34.

Lamers, L.M. (1998), "Risk adjusted capitation payments: developing a diagnostic cost groups classification for the Dutch situation", Health Policy 45:15–32.

Lamers, L.M. (1999a), "Risk adjusted capitation based on the diagnostic cost group model: an empirical evaluation with health survey in formation", Health Services Research 33(6):1727–1744.

Lamers, L.M., R.C.J.A. van Vliet and W.P.M.M. van de Ven (1999), "Farmacie Kosten Groepen: een verdeelkenmerk voor normuitkeringen gebaseerd op medicÿngebruik in het verleden" (Pharmacy costs groups: a risk adjuster for capitation payments based on the use of prescribed drugs), Report (Institute Health Policy and Management, Erasmus University, Rotterdam).

Lamers, L.M. (1999b), "The simultaneous predictive accuracy for future health care expenditures of DCGs, PCGs and prior costs", Paper presented at the iHEA-Conference, 6–9 June 1999, Rotterdam.

Lamers, L.M., and R.C.J.A. van Vliet (1996), "Multiyear diagnostic information from prior hospitalizations as a risk adjuster for capitation payments", Medical Care 34:549–561.

Lamphere, J.A., P. Neuman, K. Langwell and D. Sherman (1997), "The surge in Medicare managed care: an update", Health Affairs 16:127–133.

Lee, C., and D. Rogal (1997), "Risk adjustment: a key to changing incentives in the health insurance market", Special Report (Robert Wood Johnson Foundation, Alpha Center, Washington) March.

Lichtenstein, R., J.W. Thomas, J. Adams-Watson, J. Lepkowski and B. Simone (1991), "Selection bias in TEFRA at-risk HMOs", Medical Care 29:318–331.

Light, D.W. (1992), "The practice and ethics of risk-rated health insurance", Journal of the American Medical Association 267:2503–2508.

Londono, J.L. (1996), "Managed competition in the tropics?", Paper presented at the International Health Economics Association Inaugural Conference, Vancouver, May.

Lubitz, J. (1987), "Health status adjustments for Medicare capitation", Inquiry 24:362–375.

Lubitz, J., J. Beebe and G. Riley (1985), "Improving the Medicare HMO payment formula to deal with biased selection", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich) 101–122.

Luft, H.S. (1982), "Health maintenance organizations and the rationing of medical care", Milbank Memorial Fund Quarterly/Health and Society 60:268–306.

Luft, H.S. (1986), "Compensating for biased selection in health insurance", Milbank Quarterly 64:566–591.

Luft, H.S. (1996), "Modifying managed competition to address cost and quality", Health Affairs 15(1):23–38.

Luft, H.S., and R.H. Miller (1988), "Patient selection in a competitive health care system", Health Affairs 7(Summer):97–119.

Manning, W.G. (1998), "The logged dependent variable, heteroskedasticity, and the retransformation problem", Journal of Health Economics 17(3):283–296.

Manning, W.G., and M.S. Marquis (1996), "Health insurance: the tradeoff between risk pooling and moral hazard", Journal of Health Economics 15:609–639.

Manning, W.G., et al. (1987), "Health insurance and the demand for medical care: evidence from a randomized experiment", American Economic Review 77:251–277.

Marquis, M.S. (1992), "Adverse selection with multiple choice among health insurance plans: a simulation analysis", Journal of Health Economics 11:129–151.

Matsaganis, M., and H. Glennerster (1994), "The threat of cream skimming in the post-reform NHS", Journal of Health Economics 13:31–60.

McCall, N., and H.S. Wai (1983), "An analysis of the use of Medicare services by the continuously enrolled aged", Medical Care 21:567–585.

McCarthy, T., K. Davies, J. Gaisford and U. Hoffmeyer (1995), Risk Adjustment and Its Implications for Efficiency and Equity in Health Care Systems (National Economic Research Associates (NERA), Los Angeles/London) December.

McClure, W. (1984), "On the research status of risk adjusted capitation rates", Inquiry 21:205–213.

McGuire, T.G. (2000), "Physician agency", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 9.

Medicare Payment Advisory Commission (1998), "Report to Congress: Medicare payment policy", March.

Morrisey, M. (1992), Price Sensitivity in Health Care: Implications for Health Care Policy, Monograph (National Federation of Independent Businesses).

Mullahy, J. (1998), "Much ado about two: reconsidering retransformation and the two-part model in health econometrics", Journal of Health Economics 17(3):247–282.

Newhouse, J.P. (1982), "Is competition the answer?", Journal of Health Economics 1:109–115.

Newhouse, J.P. (1984), "Cream skimming asymmetric information, and a competitive insurance market", Journal of Health Economics 3:97–100.

Newhouse, J.P. (1986), "Rate adjusters for Medicare capitation", Health Care Financing Review (Annual Supplement):45–55.

Newhouse, J.P. (1994), "Patients at risk: health reform and risk adjustment", Health Affairs 13:132–146.

Newhouse, J.P. (1996), "Reimbursing health plans and health providers: efficiency in production versus selection", Journal of Economic Literature 34:1236–1263.

Newhouse, J.P. (1998), "Risk adjustment: where are we now?", Inquiry 35:122–131.

Newhouse, J.P., W.G. Manning, E.B. Keeler and E.M. Sloss (1989), "Adjusting capitation rates using objective health measures and prior utilization", Health Care Financing Review 10(3):41–54.

Newhouse, J.P., E.M. Sloss, W.G. Manning and E.B. Keeler (1993), "Risk adjustment for a children's capitation rate", Health Care Financing Review 15(1):39–54.

Newhouse, J.P., M.B. Buntin and J.D. Chapman (1997), "Risk adjustment and Medicare: taking a closer look", Health Affairs 16:26–43.

Pauly, M.V. (1980), Doctors and Their Workshops: Economic Models of Physician Behavior (University of Chicago Press, Chicago, IL).

Pauly, M.V. (1984), "Is cream skimming a problem for the competitive medical market?", Journal of Health Economics 3:87–95.

Pauly, M.V. (1986), "Taxation, health insurance, and market failure", Journal of Economic Literature 24(3):629–675.

Pauly, M.V. (1988), "Efficiency, equity and costs in the US health care system", in: C.C. Havighurst et al., eds., American Health Care: What Are the Lessons for Britain? (IEA Health Unit, London) 23–45.

Pauly, M.V. (1992), "Risk variation and fallback insurers in universal coverage insurance plans", Inquiry 29:137–147.

Pauly, M.V. (2000), "Insurance reimbursement", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 10.

Physician Payment Review Commission (1994), "A comparison of alternative approaches to risk measurement", Selected External Research Series, No. 1 (Washington DC).

Pope, G.C., K.W. Adamache, R.K. Khandker and E.G. Walsh (1998a), "Evaluating Alternative Risk Adjusters for Medicare", Health Care Financing Review 20(2):109–129.

Pope, G.C., R. Ellis, C.F. Liu, A. Ash, L. Iezzoni, J. Ayanian, D. Bates and H. Burstin (1998b), Revised Diagnostic Cost Group (DCG)/Hierarchical Coexisting Conditions (HCC) Models for Medicare Risk Adjustment, Final Report to HCFA, Health Economics Research, Waltham, MA, February.

Pope, G.C., C.F. Liu, R.P. Ellis, A.S. Ash, L. Iezzoni, J.Z. Ayanian, D.W. Bates, H. Burstin, K. Adamache and B. Gilman (1999), Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment, Final Report to HCFA, Health Economics Research, Waltham MA, February.

Porell, F.W., and W.M. Turner (1990), "Biased selection under the senior health plan prior use capitation formula", Inquiry 27(1):39–50.

Price, J.R., J.W. Mays and G.R. Trapnell (1983), "Stability in the Federal Employees Health Benefits Program", Journal of Health Economics 2:207–223.

Price, J.R., and J.W. Mays (1985), "Biased selection in the Federal Employees Health Benefits Program", Inquiry 22:67–77.

Riley, G., C. Tudor, Y. Chiang et al. (1996), "Health status of Medicare enrollees in HMOs and fee-for-service in 1994", Health Care Financing Review 16(4):65–76.

Robinson, J.C., L.B. Gardner and H.S. Luft (1993), "Health plan switching in anticipation of increased medical care utilization", Medical Care 31:43–51.

Rogal, D.L., and A.K. Gauthier (1998), "Are health-based payments a feasible tool for addressing risk segmentation?", Inquiry 35:115–121.

Rothschild, M., and J. Stiglitz (1976), "Equilibrium in competitive insurance markets: an essay on the economics of imperfect information", Quarterly Journal of Economics 90:629–649.

Schneider, B. (1996), "Risk structure compensation in Switzerland", in: J. Hermesse, ed., Risk Structure Compensation (Proceedings of an AIM Workshop, Maastricht).

Schokkaert, E., P. Kestens, G. Dhaene, H. Lustig, C. van de Voorde, J.M. Laasman, B. de Lange and M. Lona (1996), "Normuitkeringen voor de Belgische Ziekenfondsen: de eerste fase", Openbare Uitgaven 28(4):195–204.

Schokkaert, E., G. Dhaene and C. van de Voorde (1998), "Risk adjustment and the trade-off between efficiency and risk selection; an application of the theory of fair compensation", Health Economics 7:465–480.

Schokkaert, E., and C. van de Voorde (1998), "Risk adjustment and the fear of markets: the case of Belgium", unpublished working paper (University of Leuven).

Schut, F.T. (1995), "Competition in the Dutch health care sector", dissertation (Erasmus University, Rotterdam).

Selden, T.M. (1990), "A model of capitation", Journal of Health Economics 9:397–409.

Sheiman, I. (1994), "Forming the system of health insurance in the Russian Federation", Social Science and Medicine 39:1425–1432.

Sheiman, I. (1995), "New methods of financing and managing health care in the Russian Federation", Health Policy 32:167–180.

Shen, Y., and R.P. Ellis (1998), "Cost-minimizing risk adjustment and selection", Working Paper (Boston University) October.

Sheldon, T.A., P. Smith, M. Borrowitz, S. Martin and R. Carr Hill (1994), "Attempt at deriving a formula for setting general practitioner fundholding budgets", British Medical Journal 309:1059–1064.

Shewry, S., S. Hunt, J. Ramey and J. Bertko (1996), "Risk adjustment: the missing piece of market competition", Health Affairs 15(Spring):171–181.

Shirom, A. (1995), "The Israeli health care reform: a study of an evolutionary major change", International Journal of Health Planning and Management 10:5–22.

Swartz, K. (1995), "Reducing risk selection requires more than risk adjustments", Inquiry 32:6–10.

Thomas, J.W., R. Lichtenstein, L. Wyszewianski and S. Berki (1983), "Increasing Medicare enrollment in HMO's: The need for capitation rates adjusted for health status", Inquiry 20:227–239.

Thomas, J.W., and R. Lichtenstein (1986), "Including health status in Medicare's adjusted average per capita cost capitation formula", Medical Care 24:259–275.

Tolley, H.D., and K.G. Manton (1984), "Assessing health care costs in the elderly", Transactions of the Society of Actuaries 36:579–603.

US General Accounting Office (1994), "Medicare: changes to HMO rate setting method are needed to reduce program costs", GAO/HEHS-94-119, September (Washington, DC).

US General Accounting Office (1996), "Private health insurance; millions relying on individual market face cost and coverage trade-offs", GAO/HEHS-97-8, November (Washington, DC).

US General Accounting Office (1998), "Health insurance standards; new federal law creates challenges for consumers, insurers, regulators", GAO/HEHS-98-67, February (Washington, DC).

van Barneveld, E.M. (2000), "Risk sharing as a supplement to imperfect capitation in health insurance", Working Paper (Ph.D. Dissertation), in progress.

van Barneveld, E.M., R.C.J.A. van Vliet and W.P.M.M. van de Ven (1996), "Mandatory high-risk pooling: an approach to reducing incentives for cream skimming", Inquiry 33:133–143.

van Barneveld, E.M., R.C.J.A. van Vliet and W.P.M.M. van de Ven (1997), "Risk adjusted capitation payments for catastrophic risks based on multi-year prior costs", Health Policy 39:123–135.

van Barneveld, E.M., L.M. Lamers, R.C.J.A. van Vliet and W.P.M.M. van de Ven (1998), "Mandatory pooling as a supplement to risk adjusted capitation payments in a competitive health insurance market", Social Science and Medicine 47:223–232.

van Barneveld, E.M., L.M. Lamers, R.C.J.A. van Vliet and W.P.M.M. van de Ven (1999), "Ignoring small predictable profits and losses: a new approach for measuring incentives for cream skimming", Health Care Management Science (forthcoming).

van de Ven, W.P.M.M., and F.T. Schut (1994), "Should catastrophic risks be included in a regulated competitive health insurance market?", Social Science and Medicine 39:1459–1472.

van de Ven, W.P.M.M., and R.C.J.A. van Vliet (1992), "How can we prevent cream skimming in a competitive health insurance market? The great challenge for the '90's", in: P. Zweifel and H.E. French, eds., Health Economics Worldwide (Kluwer Academic Publishers, the Netherlands) 23–46.

van de Ven, W.P.M.M., and R.C.J.A. van Vliet (1995), "Consumer information surplus and adverse selection in competitive health insurance markets: an empirical study", Journal of Health Economics 14:149–169.

van de Ven, W.P.M.M., R.C.J.A. van Vliet, E.M. van Barneveld and L.M. Lamers (1994), "Risk adjusted capitation: recent experiences in The Netherlands", Health Affairs 13:118–136.

van de Ven, W.P.M.M., R.C.J.A. van Vliet, F.T. Schut and E.M. van Barneveld (1997), "Access to coverage in a competitive individual health insurance market: via premium regulation or via risk adjusted vouchers?", Report (Department of Health Policy and Management, Erasmus University, Rotterdam) September, to be published in Journal of Health Economics (forthcoming).

van Vliet, R.C.J.A. (1992), "Predictability of individual health care expenditures", The Journal of Risk and Insurance 59(3):443–460.

van Vliet, R.C.J.A. (1997), "Beperkte vormen van pooling als aanvulling op demografisch bepaalde normuitkeringen (Mandatory pooling as remedy for crude risk adjusted capitation payments to health plans)", Tijdschrift Sociale Gezondheidszorg 75:244–251.

van Vliet, R.C.J.A. (1999), "A statistical analysis of mandatory pooling across health insurers", Journal of Risk and Insurance (forthcoming).

van Vliet, R.C.J.A., and L.M. Lamers (1998), "The high costs of death: should health plans get higher payments when members die?", Medical Care 36:1451–1460.

van Vliet, R.C.J.A. and W.P.M.M. van de Ven (1992), "Towards a capitation formula for competing health insurers: an empirical analysis", Social Science and Medicine 34:1035–1048.

van Vliet, R.C.J.A. and W.P.M.M. van de Ven (1993), "Capitation payments based on prior hospitalization", Health Economics 2:177–188.

von Korff, M., E.H. Wagner and K. Saunders (1992), "A chronic disease score from automated pharmacy data", Journal of Clinical Epidemiology 45(2):197–203.

Ware, J.E., and C.D. Sherbourne (1992), "The MOS 36 item short form health survey (SF-36)", Medical Care 30:473–483A.

Weiner, J.P., A. Dobson, S. Maxwell et al. (1996), "Risk adjusted Medicare capitation rates using ambulatory and inpatient diagnoses", Health Care Financing Review 17:77–100.

Weiner, J.P., B. Starfield, D. Steinwachs and L. Mumford (1991), "Development and application of a population oriented measure of ambulatory care case mix", Medical Care 29:452–472.

Welch, W.P. (1985), "Regression towards the mean in medical care costs: Implications for biased selection in HMOs", Medical Care 23:1234–1241.

White House (1993), Task Force on Health Risk Pooling for Small-Group Health Insurance, Washington, DC.

Wilson, C. (1977), "A model of insurance markets with incomplete information", Journal of Economic Theory 12:167–207.

Wouters, A.V. (1991), "Disaggregated annual health services expenditures: their predictability and role as predictors", Health Services Research 26:247–272.

Zeckhauser, R. (1970), Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives", Journal of Economic Theory 2(1):10–26.

Zellner, B.B., D.K. Haugen and B. Dowd (1993), "A study of Minnesota's high-risk health insurance pool", Inquiry 30:170–179.

Zweifel, P., and W.G. Manning (2000), "Moral hazard and consumer incentives in health care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 8.

This Page Intentionally Left Blank

Chapter 15

GOVERNMENT PURCHASING OF HEALTH SERVICES*

MARTIN CHALKLEY

University of Dundee, UK

JAMES M. MALCOMSON

University of Oxford, UK

Contents

*This chapter has benefited greatly from very helpful comments by the editors, Joe Newhouse and Tony Culyer, by the discussants, Marty Gaynor and Mark Satterthwaite, and by other participants in the Handbook Workshop. The authors gratefully acknowledge the support of the UK Economic and Social Research Council (award number R000236723) for their research on health care issues.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This chapter reviews the literature on payment schemes for government purchases of health services. It focuses on four themes: (1) the tension between obtaining appropriate quality of services and keeping the cost of those services at an acceptable level; (2) the role of cost sharing by the payer when there is asymmetric information between purchaser and supplier about costs or case-mix; (3) the importance of commitment in purchasing; and (4) the role of reputation in maintaining quality in long term relationships between purchasers and suppliers.

Keywords

prospective payment, quality of service, cost sharing, commitment, reputation

JEL classification: I11, I18

1. Introduction

In many parts of the world, a substantial proportion of expenditure on health services is paid for by the public sector.[1] Government expenditure on health services has also been increasing as a percentage of GDP. Both the extent and changes in these expenditures can be gauged from Figure 1. Some of these expenditures are on services provided by publicly-owned hospitals and clinics. Others take the form of payments by government agencies to privately-run hospitals and clinics (*for-profit* or *not-for-profit*) for supplying services. In either case, the basis on which payments are made to the suppliers of services has an impact on the services delivered. This chapter is concerned with the nature of that impact.

In the literature on government purchasing of health services, there are two approaches to analyzing government agencies. The first is a normative perspective. It is concerned with what actions and decisions are appropriate for a government agency that is concerned with social welfare, defined in an appropriate way. The second is a political economy perspective. It is concerned with analyzing the actions and decisions of government agencies given that those who are in charge of the agencies have their own interests and are subject to political influences of various sorts. This chapter is concerned with the first approach. The political economy approach is discussed in Besley and Gouveia (1994).

A government agency purchasing health services with the objective of improving social welfare will be concerned about who receives those services and whether what they receive is appropriate to their condition. "Appropriateness" may have many aspects, for example, that they receive the right medical treatment, that they are treated in a sympathetic and understanding way, and so on. A number of terms have been used for this in the literature, including intensity and quality. In this chapter we simply refer to all these as aspects of the *quality* of service, in which we include any aspect of service that benefits the recipients, whether during the process of treatment or in the health outcome after treatment. We are not concerned here with how to measure the benefits from provision of different services – that involves issues of medical effectiveness, ethics and equity, issues that are discussed in other chapters. The concern in this chapter is with the impact of purchasing arrangements on the benefits, however assessed. Because purchases by government agencies are paid for by tax revenues that involve deadweight losses, such agencies will also be concerned to keep down the *cost* of providing services.

The quality issues of concern here arise because it is not usually feasible in practice for a purchaser to specify in advance precise levels for all the many aspects of quality for every condition and severity of condition in such a way that it is clear after the delivery of services whether the specification has been met. Similarly, the lengths to

[1] The chapter by Gerdtham and Jönsson (2000) provides a detailed analysis of international public and private health expenditures.

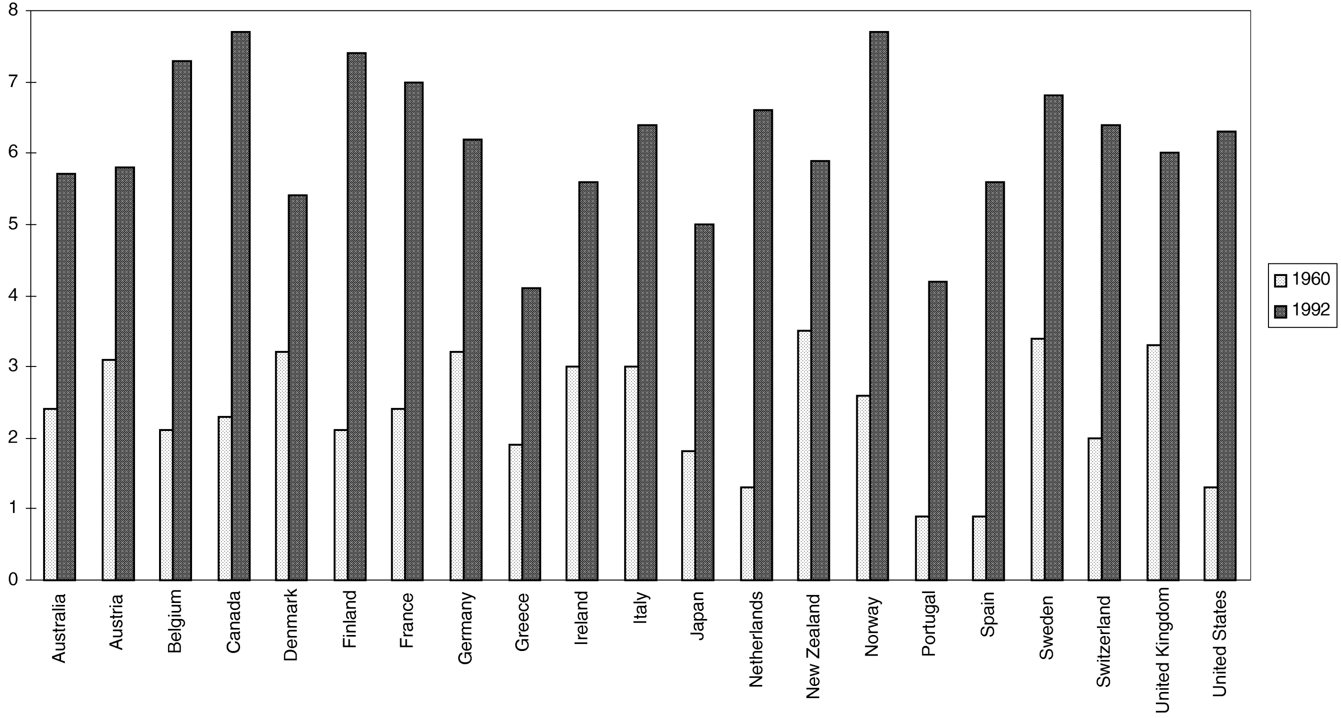


Figure 1. Health expenditure by public sector, 1960 and 1992 (% of GDP at market prices). Source: Office of Health Economics, OECD.

which a supplier goes to keep down cost, which we refer to as *cost-reducing effort*, are difficult to specify in advance. The terminology used in the literature is that quality and cost-reducing effort are *unverifiable* to a third party acting as an arbitrator or, in the case of a formal contract, a court. If quality and effort were verifiable, the purchasing agency could ensure appropriate provision by specifying the quantity and quality of services it wants and rewarding the supplier sufficiently to make supply worthwhile. It would not need to *induce* the supplier to provide the appropriate services by choice of the arrangements for payment. When some dimensions of quality and effort are not verifiable, the issue of how to induce the supplier to act appropriately is an example of what Holmstrom and Milgrom (1991) term a *multi-task agency* problem. In common with the literature, we use the term *contract* loosely to refer to arrangements for payment even where (as with National Health Service (NHS) hospitals in Britain) the supplier is a publicly-owned body with whom any such arrangement is not a formal legal contract. A traditional concern of agency theory [see, for example, Hart and Holmström (1987)] is with risk sharing in the face of uncertainty. Whilst uncertainty pervades many aspects of the provision of health services, other issues have dominated the literature on health contracts and we reflect that emphasis in this chapter.

Issues of unverifiable quality are not unique to health services. They arise in many areas of government procurement. What is distinctive about much of the literature on purchase of health services by government agencies is its concern with issues that arise when at least some dimensions of quality are more readily observed by agents who are independent of the purchaser (patients, relatives, referring doctors, etc.) than by the purchaser itself. That a dimension of quality is observed does not imply that it is verifiable. Verifiability requires that it can be specified enforceably in a contract, observability only that subjective assessments can be made about it.

The chapter is organized as follows. Section 2 provides a framework within which to analyze the conflict between incentives to induce appropriate quality and incentives to keep costs down. It then considers payment arrangements discussed in the literature for the case in which the supplier cannot influence the number of patients wanting treatment from it. Section 3 considers the case much studied in the literature in which quality of service is observed by patients, or those whom they consult in making health care choices, and thus influences the demand for treatments by patients. Section 4 then assesses the extent to which evidence from changes in payment systems in the US, particularly the change from cost reimbursement to prospective payment for Medicare, is consistent with the conclusions of this literature.

An important practical issue ignored in the framework of Section 2 is that, even within a category of service such as a *diagnosis related group* (DRG) for Medicare, there is typically substantial variation in the appropriate provision of services. Patients with the same basic diagnosis, for example, differ in the severity of their conditions and thus may require different treatments. No practical classification system will be sufficiently fine to differentiate between all the possible variations in advance and specify the appropriate treatment and quality for each. Even once treatment has taken place, it may well not be straightforward to identify whether the treatment was appropriate given the information available at the time it was carried out. This heterogeneity complicates the purchasing process in two ways. First, different treatments, and different costs of treatment, are appropriate for different people in the same identified category. Second, each supplier may have a different *case-mix*, that is, a different proportion of more and less expensive cases. The issues to which this gives rise are taken up in Section 5.

Section 6 considers a number of issues to do with advance commitment. One such issue is that of *supply assurance*, an agreement for the supplier to reserve a certain capacity for a specific purchaser's patients. The other commitment issues concern the terms on which services are to be supplied in the future. With short term contracts that may (as in the British NHS) be renewed, the purchaser may capture part of the return on a supplier's specific investments, the so-called *hold-up* problem. Also with short term contracts, the purchaser may negotiate different terms for future contracts in the light of information acquired about a supplier's costs, the *ratchet effect*. These two issues have featured extensively in the contracting literature but have been little discussed in the context of health service provision. Another issue about which there has been little discussion in the context of health service provision is the role of a supplier's concern for its future reputation as a mechanism for maintaining quality of provision. We discuss some implications for an on-going relationship between a purchaser and a supplier in Section 7.

Health services differ in the extent to which government purchases are from publicly owned and from privately owned suppliers. There has been considerable debate about the relative merits of these. There is a growing literature on why ownership matters [see, for example, Hart (1995)] and applications of these ideas to regulated firms [see, for example, Laffont and Tirole (1993, Chapter 17)] and to certain specific public services [see, for example, Hart, Shleifer and Vishny (1997)] but no application that we know of specifically to the supply of health services. Nor is the best way to apply these ideas to health provision clear to us. We do not, therefore, discuss this issue.

Many of the issues discussed in this chapter also apply to other types of purchasers, such as health insurance companies, that pay for services they do not directly receive. There are, however, a number of respects in which the discussion here is oriented towards government agencies. The normative analysis is concerned with purchasers whose objective is to maximize a social welfare function, rather than profit or individual utility. It is, moreover, not concerned with the impact of competition between purchasers to attract customers, an important element with competing insurers. In addition, the discussion here is concerned with the case in which customers make no payments for services (apart, of course, from their contributions to taxation). Finally, the analysis for the most part assumes that the purchaser can make credible "take it or leave it" offers to suppliers. When Congress specifies the Medicare payment system, it specifies the system unilaterally for all suppliers – despite the lobbying system, it does not actually negotiate with individual suppliers. That is less likely to be the case with private sector purchasers.

2. Quality and costs

When quality and cost-reducing effort are unverifiable, there is a potential conflict for purchasers between high quality and low costs that goes deeper than the obvious one that higher quality requires more inputs that must be paid for. This conflict is illustrated by two forms of payment to suppliers that have been widely used for health services. One is *cost reimbursement* in which the purchaser reimburses the supplier for the actual cost of supplying the service. This corresponds to standard insurance based payments and to Medicare payments for elderly patients in the US before the Medicare reforms in 1983. The other is a *fixed price* for each person treated that is independent of the actual costs of treatment. This corresponds to the *prospective payment system* (PPS) introduced for Medicare in the US in 1983 and the *cost per case* contracts used by some health authority purchasers in Britain since the reform of the National Health Service (NHS) that started in 1990. Intuitively, cost reimbursement provides no incentive to keep costs down and, when it also allows a mark-up over costs to cover overheads, provides a strong incentive to increase quality in order to increase costs and so have a higher mark-up revenue. In contrast, prospective payment provides a strong incentive to keep costs down because the supplier keeps all the cost savings, but it also provides an incentive to cut costs by reducing quality unless there is some mechanism to prevent that.

In this section, we provide a framework to formalize this intuition and then discuss payment mechanisms to mitigate the conflict.

2.1. The framework

Let x denote the number of patients with a specific diagnosis treated by a supplier, q the quality of treatment, and e the effort of the supplier to control costs. In the case of hospital services, a common measure of numbers treated is the *finished consultant episode* (FCE). In the case of primary services paid for with a capitation fee as in the British NHS, the measure of numbers treated is patients registered with the physician. Indicators such as *length of stay* in hospital (LOS) are potentially verifiable measures of inputs that may affect quality of service but, we assume, do not capture everything about quality. For our initial analysis, we assume that there is only a single unverifiable dimension to quality reflected in q and a single dimension to effort. We return to the case of multi-dimensional quality and effort later.

The total monetary cost of the treatments consists of fixed cost F and variable cost $c(x, q, e)$. Variable cost increases with the number treated and the quality of treatment and decreases with cost-reducing effort, so $c_x(.), c_q(.) > 0$ and $c_e(.) < 0$ for all (x, q, e). Denote by P the payment from the purchaser. In principle, this payment can depend on anything that is verifiable – the focus of this chapter is with how payment should be determined. The supplier is concerned about its financial surplus $P - F - c(x, q, e)$, either because it is a for-profit institution or because the surplus can be spent on perks for staff or on improving facilities. But it may also be concerned about numbers, quality and effort directly, either because the administrators care about patients, or because treating more patients and offering higher quality involves more work, or because effort to reduce cost has disutility. We capture this by writing the supplier's objective as

$$u = P - F - c(x, q, e) - v(x, q, e), \tag{1}$$

where $v(.)$ corresponds to non-monetary cost. We denote by $\bar{u}$ the lowest payoff for which the supplier is prepared to provide the service. To simplify the exposition, we assume $c(.) + v(.)$ is strictly convex for all (x, q, e).

The function $v(.)$ may be different for different types of suppliers. Suppliers vary a lot. For a physician practicing alone, $v(.)$ is a reflection of that physician's utility measured in money terms. Increasing quality of services may require spending more time with patients, time that must be taken away from other things. Seeing more patients also involves more time. Costs can be reduced by effort to check out the prices of different medications. For a complex institution like a major hospital, $v(.)$ reflects the outcome of interaction between employees, contracted physicians and owners. Quality of service may be increased by senior managers working harder to keep the institution running efficiently in a way that may not be reflected in monetary costs. Costs may be similarly reduced by hard work. Medical staff may care about the quality of service provided to patients quite apart from its financial consequences. Then $v(.)$ may be a decreasing function of quality, possibly even negative overall. In the case of not-for-profit institutions, it has been argued that the quantity and quality of services are of intrinsic concern – see, for example, the classic statement in Newhouse (1970). Empirical evidence from Frank and Lave (1989) and Dranove and White (1994) indicates that certainly not all hospitals are profit maximizers. The objectives of not-for-profit institutions are discussed further in the chapter by Sloan (2000).

Because of this diversity, we want to avoid unnecessary assumptions about properties of the function $v(.)$. For the present discussion we simply require that there is a genuine issue of getting the supplier to provide the appropriate level of quality at the appropriate cost. We can capture this formally in the following way. Let $b(x, q)$ denote the benefit the purchasing agency attaches to having x patients treated with quality q. We take this to be concave and increasing in both its arguments. The purchaser would like to maximize social welfare consisting of the benefit of treatment $b(x, q)$, plus the payoff to the supplier u, less the cost to taxpayers of the total payment P made to the supplier. It is conventional to assume that the purchasing agency attaches a premium of $\alpha > 0$ to these payments to account for distortions from raising revenue from taxation. Thus social welfare is

$$b(x, q) + u - (1 + \alpha)P. \tag{2}$$

Substitution for P from (1) enables the purchaser's objective to be written

$$\max_{x,q,e,u} b(x, q) - (1 + \alpha)\big[F + c(x, q, e) + v(x, q, e)\big] - \alpha u, \tag{3}$$

subject to feasibility constraints on the number of patients wanting treatment and $u \geqslant \bar{u}$. We refer to the solution to the problem in (3) as the *efficient* or *first best* outcome. It always involves $u = \bar{u}$ because the maximand in (3) is decreasing in u. For the purposes of discussion, we assume that it involves strictly positive quantity, quality and effort, denoted by x^*, q^* and e^*, respectively. Then the service is actually worth providing and x^*, q^* and e^*, which are independent of $\bar{u}$, satisfy the first order conditions

$$b_x(x^*, q^*) - (1+\alpha)\big[c_x(x^*, q^*, e^*) + v_x(x^*, q^*, e^*)\big] = 0, \tag{4}$$

$$b_q(x^*, q^*) - (1+\alpha)\big[c_q(x^*, q^*, e^*) + v_q(x^*, q^*, e^*)\big] = 0, \tag{5}$$

$$c_e(x^*, q^*, e^*) + v_e(x^*, q^*, e^*) = 0. \tag{6}$$

We capture the issue of concern about quality with the assumption

$$c_q(x^*, q^*, e^*) + v_q(x^*, q^*, e^*) > 0. \tag{7}$$

This assumption ensures that there is always positive marginal cost (monetary plus non-monetary) to quality at an efficient outcome. For a supplier maximizing the objective in (1) who receives no financial recompense for higher quality, there is no corresponding marginal revenue. Thus quality is always below the efficient level when efficient quantity and effort are supplied. We emphasize that we make this assumption because without it there would be no concern about underprovision of quality.

Even where governments are the major purchasers of a health service, there is often a significant amount of private purchasing alongside. We discuss in an appendix how that can be incorporated into the framework.

2.2. *Quality and effort*

Suppose the supplier cannot influence the number of patients *wanting* to be treated but can decide how many actually to treat. It then follows directly from (7) that a fixed price payment for each patient treated cannot achieve the efficient outcome (x^*, q^*, e^*). If the payments are designed to induce the supplier to treat x^* patients and exert effort e^*, quality is less than q^*. In the extreme case of $c_q(.) + v_q(.) > 0$ for all (x, q, e), which corresponds to higher quality always involving higher monetary plus non-monetary costs, the supplier sets quality at the lowest level it can without, for example, being sued for malpractice.

Consider the alternative of cost reimbursement that corresponds to $P = F + c(x, q, e)$. It follows directly from (1) that a supplier concerned only with monetary profits (and for whom, therefore, $v(.) \equiv 0$ for all (x, q, e)) is completely indifferent about the number of patients treated, quality and effort. Such a supplier has no incentive to deviate from efficient levels (although no positive incentive not to deviate either). That no longer applies if the supplier is concerned with non-monetary costs. Newhouse (1970), for the case of non-profit hospitals that care about patient welfare, and Ellis and

McGuire (1986), for suppliers with quality decisions made by physicians rather than administrators, argue that $v_q(.) < 0$. Such suppliers provide too high quality under cost reimbursement because they are indifferent to the monetary costs and thus supply the higher quality that patients like. This conclusion is consistent with the view that the high costs of care under the US Medicare system before the reforms of 1983 were the result of cost reimbursement, see Weisbrod (1991), though Newhouse (1992) argues that such considerations are unlikely to account for the sharp *increase* in costs over time.[2]

When $v_q(.) < 0$ but decisions about effort e are either unimportant or incur no non-monetary cost, it is straightforward to design a payment system that achieves the efficient outcome, as Ellis and McGuire (1986) show. All that is required is *cost sharing* by the purchaser, a payment scheme with less than 100% of costs reimbursed plus a fixed payment per patient treated that ensures the supplier is willing to treat the appropriate number of patients. Such payment schemes have also been termed *supply side cost sharing* to distinguish them from *demand side cost sharing* in which patients pay part of the cost of treatment. By appropriate choice of the share of costs reimbursed, the purchaser can reduce the supplier's incentive to provide excessive quality. This involves setting the share of costs reimbursed, denoted γ, so that $(1-\gamma)c_q(x^*, q^*, e^*) = -v_q(x^*, q^*, e^*)$.

One problem with cost sharing is that actual costs may not be easily measurable or, even if measurable, may not be directly attributable to patients with a particular diagnosis or paid for by a particular purchaser. An obvious example studied by Glazer and McGuire (1994) is the allocation of costs not attributable to particular patients (for example, fixed costs) when a supplier treats patients of more than one purchaser. Purchasers may then choose contracts in order to shift part of these costs to other purchasers, which may result in inefficient provision.

A second problem with cost sharing arises when cost-reducing effort is important and involves non-monetary costs so that $v_e(x^*, q^*, e^*) > 0$. Then full cost reimbursement and cost sharing have adverse effects on effort decisions. For any share $\gamma > 0$ of cost reimbursed, the supplier does not choose the efficient effort e^* when the number and quality of treatments are x^* and q^*, respectively, because the supplier's first order condition with respect to effort differs from (6) in having $c_e(.)$ multiplied by $(1-\gamma)$. This is another reason why cost reimbursement (which is equivalent to $\gamma = 1$) may result in overly costly medical care. In the absence of any other way to maintain quality, it is still typically worthwhile having some element of cost sharing by the payer, as explained in Chalkley and Malcomson (1998a). The intuitive reason is as follows. Without cost sharing, quality is inefficiently low but effort is efficient conditional on the level of quality achieved. Introducing a small amount of cost sharing improves quality. It will distort

[2] It is important for this conclusion that the supplier's concern is with patient welfare, not social welfare which also takes account of the costs to taxpayers. In the latter case, the provider would recognize that quality above the efficient level does not increase social welfare and thus not supply it. Ellis and McGuire (1990) consider the case in which patients pay a share, but not the whole, of the cost of treatment and bargain with the provider over the level of treatment. Such bargaining provides another channel by which quality can be influenced.

effort. However, because effort is conditionally efficient, the change in effort results in only a second order reduction in social welfare, whereas the improvement in quality results in a first order improvement. Hence, a small amount of cost sharing improves on a fixed payment per patient treated.

Such cost sharing does not, however, fully overcome the basic conflict between keeping costs down to the efficient level and keeping quality up to the efficient level. The next section considers other ways of overcoming that conflict.

3. Quality, demand and fixed price contracts

A mechanism for overcoming this potential conflict has been identified in the case in which the quality of service offered by a supplier influences the demand for treatment that it faces. See, for example, Pope (1989), Allen and Gertler (1991), Hodgkin and McGuire (1994), Ma (1994), Rogerson (1994), Ma (1997, 1998), Chalkley and Malcomson (1998b) and Ellis (1998). Demand may not respond to quality for all services. In the case of emergency treatments, for example, speed of attention is all important and so the supplier's location may be more important than its quality of service. Moreover, for demand to respond to quality, some measure of quality has to be available to patients or to those from whom they seek advice, which may be problematic in the case of aspects of quality that can be assessed only during the process of treatment (*experience* quality as it is termed). Provided some such measure is available, demand may increase with quality either because quality influences where patients go for treatment or, in non-emergency cases, because it affects whether they go for treatment at all. The latter may be significant – in general, quality reflects the probability of complications, cross infection, and so on. Thus, even where there is no choice about where a patient goes for treatment, quality may still affect demand. As long as it does, a supplier who offers higher quality attracts more patients and, even with a contract that has a constant price per patient, increases its revenue by treating them. The higher is the price, the more worthwhile it is for the supplier to offer higher quality in order to attract more patients. Thus, by appropriate choice of price, the purchaser can influence the quality supplied. Moreover, although the price is set to induce quality, the payment to any individual supplier depends only on the number of patients treated, so there is no blunting of incentives to reduce costs.

The impact of price on quality is illustrated in Figure 2. The supplier's marginal cost of quality *MC* is increasing in quality. Because increasing quality increases demand, the supplier's marginal revenue $MR(p)$ from quality is positive and proportional to the price p per treatment. The quality chosen is that at which marginal revenue equals marginal cost. For price $p' > p$, the marginal revenue curve shifts upwards, so the quality provided is higher. By manipulating the price, the purchaser can induce any quality level at which demand is increasing in quality.

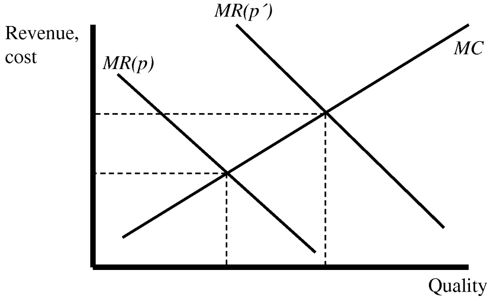


Figure 2. Marginal cost and revenue from increasing quality.

3.1. *Efficiency with fixed price contracts*

The arguments above can be formalized in the following way. Suppose the demand for treatment by the supplier $n(q)$ is increasing in q, so $n'(q) > 0$ for all q. If there is only one supplier, the increased demand reflects more people deciding it is worth having treatment if quality is higher. If there is competition between suppliers, it also reflects the effect of increased quality by the supplier in attracting patients from other suppliers. In the latter case, there is competition in quality of the type discussed by Spence (1975) and applied to health services by Pope (1989) that we do not formalize here. See the chapter by Dranove and Satterthwaite (2000) for a discussion of market interactions between suppliers. What is important for the present discussion is that each supplier's own demand increases with its own quality for given quality provided by other suppliers, even if overall market demand as a whole is insensitive to quality. To simplify notation, define

$$B(q) \equiv b\big[n(q), q\big], \quad \text{for all } q, \tag{8}$$

$$C(q, e) \equiv F + c\big[n(q), q, e\big] + v\big[n(q), q, e\big], \quad \text{for all } q, e. \tag{9}$$

We assume that $n(q)$ is such that $B(q)$ is concave and $C(q, e)$ is strictly convex. (The precise conditions for this are tedious to derive.) In this case, the purchaser's objective from (3) becomes

$$\max_{q,e} B(q) - (1+\alpha)C(q, e) - \alpha\bar{u}, \tag{10}$$

with first order conditions that determine q^* and e^*

$$B'(q^*) - (1+\alpha)C_q\big(q^*, e^*\big) = 0, \tag{11}$$

$$-(1+\alpha)C_e\big(q^*, e^*\big) = 0. \tag{12}$$

Suppose the purchaser agrees to pay the supplier a lump sum a (which may be positive or negative) plus a fixed payment p per patient treated. Then the supplier chooses q and e to

$$\max_{q,e} a + pn(q) - C(q,e), \tag{13}$$

with first order conditions

$$pn'(q) - C_q(q,e) = 0, \tag{14}$$

$$-C_e(q,e) = 0. \tag{15}$$

It follows directly from (15) that, as argued above, the supplier chooses efficient effort for any given quality under fixed price payment. Moreover, it is clear from comparison of (14) with (11) that, with the price set so that

$$p = \frac{B'(q^*)}{(1+\alpha)n'(q^*)}, \tag{16}$$

the supplier chooses both efficient quality q^* and efficient effort e^*. The lump sum a can then be set to ensure $u = \bar{u}$, so the purchaser pays no more in total than the minimum necessary for the efficient level of service. Standard comparative static analysis establishes that quality is an increasing function of price so the purchaser can, by manipulation of the price, achieve any quality level at which demand is increasing in quality. To relate this to Figure 2, note that the first term on the left hand side of (14) is the marginal revenue from increasing quality. The marginal cost of quality is given by the second term with effort at the level satisfying (15). A special case of this result was derived in Ma (1994), the more general version presented here in Chalkley and Malcomson (1998b). Rogerson (1994) considers pricing rules analogous to this when a supplier provides more than one type of service.

Since the payment schedule has a lump sum component a as well as a fixed payment p per patient treated, it is a two-part tariff rather than a pure fixed-price tariff, though we refer to it loosely as a fixed price contract in what follows. The value of a is zero if suppliers receive no rents at the fixed price p, though this may not result in the efficient number of suppliers if the absence of rents is the result of free entry and p is not equal to minimum average cost. See Edlin (1997) on this issue.

The efficiency of fixed price contracts when quality affects demand is a striking result. Before making use of it in practice, however, it is important to be aware of its limitations. The rest of this section discusses extensions and limitations of the basic result.

3.2. Quality and effort: perceptions and dimensions

One issue is whether patients, or even those from whom they seek advice, correctly perceive the quality of treatment on offer. The efficiency of fixed price contracts does not, however, depend on correct perception. If quality is perceived perfectly and if the purchaser is concerned with the welfare of patients as assessed by patients themselves, then there is a close relationship between the demand function $n(q)$ and the benefit function $b(x, q)$. But that relationship was not used in deriving the result. The only role of patients' perceptions is to make demand increase with quality. Thus, as long as those perceptions are positively correlated with actual quality, the purchaser can choose the price to ensure efficiency.

This conclusion is, however, somewhat less reassuring than it might at first seem. In any practical situation, there are likely to be many dimensions to quality – health care combines clinical, nursing, hotel, and many other services. These are easily incorporated into the framework because q can be a vector $(q_1, \ldots, q_n)$ of n different dimensions to quality. Moreover, a two-part tariff can still induce efficient choices provided the relative valuations of the different quality dimensions by patients (with the guidance of those from whom they seek advice) are the same as the relative valuations of the purchaser, see Chalkley and Malcomson (1998b). The essential reason is the following. The supplier's revenue depends on quality only as expressed through demand. How demand responds to the different dimensions of quality depends on patients' perceptions, so the supplier is concerned with the relative importance of quality dimensions as perceived by patients. The purchaser can influence the overall level of quality by changing p but not the relative provision of the different dimensions. This is a serious limitation if patients are more aware of some dimensions of quality (for example, hotel services) than of others. There is, moreover, considerable evidence that patients do not perceive or respond to important dimensions of quality even when measures of these are available. See, for example, Chernew and Scanlon (1998), Haas-Wilson (1994), Hibbard and Jewett (1997) and Mennemeyer, Morrisey and Howard (1997). The same applies to dimensions of quality that provide externalities to other patients (such as those that prevent transmission of a patient's infectious disease) that the patient may care about less than the purchaser. It is not clear what the purchaser can do about this problem other than either try to ensure that patients are well-advised and well-informed, or focus monitoring of standards on those dimensions that patients perceive or value least relative to the purchaser. However, the result on fixed price contracts implies that even then it is unnecessary to monitor *all* dimensions of quality, only those to which problematic circumstances apply.

Multiple dimensions to cost-reducing effort do not, in contrast, raise complications for fixed price contracts because the first order conditions corresponding to (15) for each dimension ensure efficient effort for given quality.

3.3. *Efficient treatment numbers*

A second issue is that it may not in fact be efficient to treat all those patients who want treatment at the quality offered. An assumption implicit in the analysis above is that all those demanding treatment at efficient quality are actually treated, that is $x^* = n(q^*)$. But this may well not be efficient when patients do not themselves pay for treatment. Patients demand treatment as long as the benefits are positive but it is efficient to treat only those for whom the benefits are greater than the costs. Whenever, at the quality offered, there are some patients on the margin between choosing to be treated and choosing not to be treated at all, it will not be efficient to treat them. That may well be the case for various types of elective treatment such as hip replacements, treatment of varicose veins, and so on. Moreover, if the supplier's services are used to capacity, the number of treatments may necessarily be less than the demand for them and the supplier may thus not incur the cost of increasing quality in order to increase demand. These issues are taken up in Chalkley and Malcomson (1998b) who show that fixed price contracts then necessarily result either in too low quality or in too many patients being treated. If, however, there are measures of demand separately identifiable from numbers actually treated (for example, patients referred for treatment who are not actually treated), these can be used to enhance the efficiency of provision.

3.4. Uncertainty

A third issue is uncertainty about costs and about the relationship between quality and demand. The efficiency of a two-part tariff does not in fact depend on these being known with certainty, provided the supplier has no better information at the time of deciding on quality and effort than the purchaser has at the time the payment terms are set and provided the parties are risk neutral. To see this, suppose the functions $n(.)$, $c(.)$ and $v(.)$ contain stochastic components (which need not be additive). Then, if the functions $B(.)$ and $C(.)$ in (8) and (9) are defined in terms of the expected values, everything goes through as before provided $n(q)$ and $n'(q)$ in (13)–(16) are replaced by their expected values. This applies equally to uncertain case-mix if the supplier does not learn the case-mix before deciding quality and effort, and cannot *dump* (that is, decline to treat) patients once it learns the cost of treating them, [see Ma (1994, 1997)].

It does not, however, apply if either or both of the parties are risk averse. Government purchasing agencies may be risk averse if, as is not uncommon in practice, they must stick within fixed budgets. Suppliers may be risk averse if they are unable to diversify risk via, for example, the stock market. A risk averse supplier who faces uncertainty about the cost function will want insurance against uncertainty in net revenues which can be provided by the purchaser agreeing to pay some share of the actual costs incurred in addition to the fixed payment per patient treated. Such cost sharing results in a trade off between insurance and incentives that is familiar from the principal-agent literature. It is clear from the analysis above that any sharing of costs with the payer reduces the supplier's incentive to incur effort to reduce costs and thus induces inefficient effort when quality and numbers treated are at the efficient level. As noted in the Introduction, however, issues other than risk have dominated the literature on health contracts.

It may also be optimal for the payer to share part of the actual costs if the supplier has information at the time it makes treatment decisions that the purchaser does not have at the time the payment terms are set. Such asymmetric information may occur for a number of reasons. First, the supplier may have better information about monetary and non-monetary costs than the purchaser. If this applies only to fixed costs, a two-part tariff can still achieve efficient quality and effort but the purchaser may incur unnecessarily high expenditure if fixed costs are in fact low. The reason is the need to set a sufficiently high to ensure that the supplier would provide the service even if fixed costs had been high. It may then be worth basing payment on actual costs in order to reduce expected expenditure, even though that may not induce efficient quality and effort. Second, as noted in the Introduction, within any set of service categories such as DRGs that are not too fine for practical use, there are typically substantial variations in the appropriate provision. If different suppliers have different case-mixes, this will affect their variable costs in a way that may be unknown to the purchaser. If, in addition, there are dimensions to quality that can be varied between patients in the same category, a supplier may offer inefficiently low quality to high cost patients or, as pointed out by Newhouse (1983), decline to treat them at all if it has discretion to do that. This issue has been of popular concern with the fund-holding arrangements for some general medical practitioners (GPs) in the British NHS. What payment arrangements it makes sense to adopt under these conditions has been a major concern in the literature. It is taken up in Section 5.

Before exploring that issue further, however, we discuss some of the empirical evidence to assess the practical relevance of the framework set out here.

4. Empirical findings

The previous two sections provide a theoretical framework for analyzing the impact of different purchasing mechanisms on the decisions of health service providers and, hence, on the implications of these decisions. That framework suggests a number of testable hypotheses. There has been a substantial body of research concerning the effect of purchasing mechanisms on the cost and quality of health services, particularly the replacement of cost reimbursement with a prospective payment system (PPS) for Medicare and some Medicaid services in the US. In this section we first review that evidence on cost and quality to assess whether the predictions of the framework are borne out and, if so, what are the quantitative magnitudes involved. After that, we briefly review preliminary research that looks more directly at whether the forms of payment mechanism that have evolved in both the US and the British NHS are consistent with the theoretical framework.

4.1. Prospective payment, cost and quality

The most widely researched evidence on the effect of different payment mechanisms on quality and cost of health services concerns the switch from cost reimbursement to PPS under Medicare in the US. For most patients, PPS rewards suppliers with a fixed payment per patient treated, with the amount of the payment depending on the diagnosis related group (DRG) to which a patient is assigned. There is provision for additional payment for those patients who are unusually expensive within the DRG but these outlier payments apply to only a small proportion of patients and account for only 5% of Medicare payments to hospitals [see McClellan (1997)]. This switch provides experiences that can be used to assess the relevance of the framework discussed in Sections 2 and 3 above. We discuss two issues. First, what happened to costs with the switch to PPS and, second, whether any change in costs can be attributed to changes in quality or whether it results from what, in our framework, would be changes in effort. There is also evidence on the effects of per diem payments adopted by some Medicaid programs for nursing home placements. Nursing homes, however, have special characteristics (capacity regulation, licensing, etc.) that may affect their responses to payment systems, so we leave discussion of that evidence to the chapter by Norton (2000).

4.1.1. Evidence on costs

In the case discussed in Section 2 in which quality does not affect demand for services, the framework set out above predicts that a shift from cost reimbursement to PPS will reduce costs. The same applies even if quality affects demand for services as long as prices under PPS are not set so high as to induce an even higher quality of service than prevailed under cost reimbursement, which seems unlikely given the concern of policy makers to contain costs at the time of the shift.

Because hospitals produce a multiplicity of services the composition and quantity of which vary over time, direct observation of hospital costs for any one service is notoriously difficult. Furthermore, since hospitals typically supply a mixture of Medicare and other patients, isolating the effect of the payment system for Medicare patients on costs is compounded by the difficulties involved in assigning costs to individual patients. In the empirical literature, this has almost invariably led to the implicit assumption that costs are linear in the number of patients treated, that is, $c(x, q, e)$ is linear in x, and are well correlated with indicators of real resource use such as length of stay in hospital (LOS). The problem with this approach is that there may be unobserved resources used in the production of health services so that observed measures of resource use do not fully reflect changes in underlying costs of treatment. Nevertheless, there are some health services, for example, psychiatric services, for which LOS is often thought to be a sufficient statistic of resource use, so these services have featured prominently in many studies of costs, including Freiman, Ellis and McGuire (1989), DesHarnais, Kobrinski, Chesney, Long, Ament and Fleming (1987) and Ellis and McGuire (1996).[3] Other measures of resource use that have been studied include measures of the use of intensive care facilities (either frequency or duration of use), the number of separate consultations per patient, and measures of staff-patient ratios, often adjusted by skill mix.

[3] LOS should, in principle, be verifiable and it might thus be optimal to contract on it explicitly. However, the studies cited consider cases in which that was not in fact done, so the change in LOS is an appropriate measure of the response of suppliers to the change in payment mechanism.

The evidence on the effect of a switch to prospective payment on LOS is substantial. Freiman et al. (1989), Frank and Lave (1989), DesHarnais et al. (1987), Newhouse and Byrne (1988), DesHarnais, Wroblewski and Schumacher (1990), Manton, Woodbury, Vertrees and Stallard (1993) and Ellis and McGuire (1996) all report studies in which LOS is shown to fall significantly following a move from cost reimbursement to prospective payment. In many cases the decline in LOS is considerable. For example, Ellis and McGuire (1996) consider data which displays a decline in LOS for psychiatric patients in New Hampshire, following the introduction of prospective payment, of 30%.

The view that a movement to prospective payment reduces resource usage and hence costs receives further support from those studies that consider a broader range of measures of resource use. DesHarnais et al. (1987), for example, examine a sample of 729 short term general hospitals and consider measures of resource use including LOS, frequency of intensive care use per patient, and the number of consultations per patient. To address the obvious problems of changing technology and health care trends between pre and post prospective payment regimes, these researchers use a forecasting model to predict resource usage on the basis of historical data relating to a period of cost reimbursement. They find that, following the switch to prospective payment, resource usage is significantly less than forecast. Part of the apparent cost savings may, however, be illusory for a number of reasons. First, as Newhouse and Byrne (1988) note, there is evidence of increases in LOS at hospitals within Medicare that are not subject to PPS, suggesting that costly patients are being re-directed towards suppliers who benefit from cost reimbursement. Second, within hospitals that are subject to PPS there is evidence of *cost shifting* [Eldenburg and Kallapur (1997)] where resources actually used on PPS patients are recorded as being used on non-PPS patients (for example, outpatients) so that they can be charged for under cost reimbursement. Glass and Sappington (1998) discuss the incentives for cost shifting provided by Medicare financial regulations.

A recurring problem in interpreting reductions in resource use, which is illustrated by DesHarnais et al. (1987), is disentangling the effect of the payment mechanism from the many other influences that impact upon costs. In DesHarnais et al. (1987), for example, a linear forecasting scheme is used to predict resource use for 1984 based on trends from 1980 to 1983, with the discrepancy between predicted use and actual use then attributed to prospective payment. Whilst a method such as this is useful for measuring the impact of prospective payment on costs *following* a reform in the payment mechanism, it cannot inform policy makers as to the potential cost savings from adopting prospective-like mechanisms in areas of health service provision that have not been subject to relevant experience. This suggests that caution is needed before concluding that prospective payment can provide large cost savings wherever it is applied. A consideration of this issue is provided by Miller and Sulvetta (1995) who decompose costs for outpatient services into exogenous (beyond the control of hospitals) and endogenous categories. This study attributes 69% of costs to the exogenous category which suggests a rather limited scope for cost savings in this particular area.

Notwithstanding real questions of exactly how much of the observed reductions in resource usage can properly be attributed to changes in the payment mechanism and concerns as to how useful existing experience is in predicting cost savings from, for example, the expansion of prospective payment in the future, there is a substantial body of evidence indicating that real cost savings resulted from the adoption of prospective payment in place of cost reimbursement. That in itself does not, however, establish whether the cost savings are the result of a move towards more efficient cost-reducing effort (e), as predicted in the framework set out above, or simply the result of reductions in the quality of services provided (q). The next section addresses that question.

4.1.2. Evidence on quality

Direct measures of cost-reducing effort are typically unavailable. Thus to assess the extent to which cost reductions can be attributed to that and attributed to reduced quality, we focus on measures of quality. Even with quality, there seems little prospect of obtaining direct measures that capture all that purchasers are concerned about. If such measures existed, they could be used in payment systems to avoid the problem of potential underprovision of quality. Thus, the measures available are at best imperfect indicators of quality. There has recently been increasing emphasis on the measurement of health outcomes, see Whynes (1996). However, the most frequently used outcome measures, such as mortality rates, tend to be rather crude measures of the benefits of treatment. Using readmission rates and mortality rates, DesHarnais et al. (1987) find no deterioration in outcomes following the introduction of prospective payment. This constancy of outcome following the introduction of PPS is confirmed in the study by DesHarnais et al. (1990), whilst Cutler (1995) finds that the introduction of PPS changed the timing of mortality but not the overall rate.

One interpretation of this evidence is that the previously discussed cost savings are the result of greater cost-reducing effort rather than reductions in quality. This is however too simplistic because there is no great support for the view that outcome measures of the kind discussed here are measuring the benefits of treatments that are of concern to purchasers. Hence, whilst $b(x, q)$ may plausibly partly depend on such outcome measures, our notion of quality is such that it is possible for $b(x, q)$ to decline whilst outcome measures are constant.

The logic of the framework set out in Section 2 suggests that there may be ways of assessing quality, at least as it is perceived by patients, through a consideration of the demand for treatments. As noted there, it is not necessary for patients to be fully able to assess the quality of treatment they are offered for demand to provide a useful indicator of quality, only for demand and quality to be positively correlated. At the time when PPS was to be introduced for Medicare, many commentators predicted an increase in the volume of treatments that would be carried out as hospitals responded to restrictions of their revenues. But the evidence reviewed by Hodgkin and McGuire (1994) is to the contrary. Treatment numbers declined significantly following the introduction of PPS. One explanation for this is that hospitals chose to dump (not treat) costly patients. Another possibility is that the decline in treatment numbers reflects a transfer of patients to non-PPS institutions. A third possibility is that the decline in treatment numbers reflects a decline in demand that is a consequence of reduced quality of treatment.

Further indications of quality effects come from studies that attempt to break down the reduction in resource use experienced under PPS into constituent parts. Notable in this regard is the previously referenced work of Ellis and McGuire (1996) on reduction in LOS for psychiatric patients in New Hampshire. Econometric methods are used to identify the proportion of this LOS reduction that is due to reduction in quality of service as opposed to a change in the style of treatment, as measured by the number of patients whose treatment is diverted to other types of hospital (with already shorter LOSs) rather than being curtailed within a given hospital. Clearly, there are problems in defining quality in this context and plausibly 'style of treatment' effects in Ellis and McGuire (1996) might be associated with either cost-reducing effort or quality. Ellis and McGuire (1996) find that approximately 40% of the reduction in LOS that followed the introduction of PPS for this group of patients can be attributed to reductions in treatment intensity (their term for one aspect of what we call quality) leaving 60% of the reduction in LOS to be attributed to either other aspects of quality or effort.

4.2. Payment mechanisms in practice

The prospective payment system for Medicare allows outlier payments for patients who are unusually expensive for the DRG to which they are assigned. In that respect, it contains an element of cost reimbursement. Research by McClellan (1997) shows that the effective degree of cost reimbursement is in fact even greater. The reason is that some DRGs are defined not only in terms of the patient's identified condition but also by the treatment received, with more expensive treatments being in DRGs with higher payments. Where treatment in such cases is at the discretion of the supplier, the supplier in effect receives some cost reimbursement for choosing a more expensive treatment. Malcomson (1999) discusses the theory of when it makes sense to define DRGs by the treatment given. McClellan (1997) provides a detailed empirical analysis of the extent to which PPS allows, in practice, for cost sharing and shows that, for some patients with some conditions, effective cost sharing is in fact quite substantial. McClellan (1994, 1995) consider how this finding can help account for the continuing increases in the costs of Medicare which, given the evidence reported above on the effect of PPS on resource use, seems paradoxical.

In the British NHS, the NHS Executive has in most circumstances prohibited contracts that incorporate explicit cost sharing. In the early stages of the post-1990 reforms, there was widespread use of *block* contracts in which payment to the supplier is simply a lump sum for provision of services. More recently, purchasers have been encouraged to use fixed price contracts (known as *cost per case* contracts), contracts with payment a non-linear function of the number of patients treated (*cost and volume* contracts), or *sophisticated* block contracts with specified upper and lower bounds on patient numbers and a mechanism for renegotiating the payment if numbers treated are not within those bounds. National Audit Office (1995) summarizes the advice given to NHS institutions on this. Sophisticated block contracts are by far the most common, see Chalkley and McVicar (1998). Moreover, as Chalkley and McVicar (1998) argue, many of those contracts are written in such a way as to allow *de facto* cost sharing because one of the things specified as a trigger for renegotiation is that costs have changed substantially from those anticipated. Thus, as with the PPS in Medicare, the degree of cost sharing in practice is greater than initial appearances might suggest.

The use of cost sharing suggests that the contracting parties regard it as valuable. As already noted, it typically is valuable when there is asymmetric information about either the supplier's cost function itself or the cost composition (case-mix) of the patients to be treated. We therefore turn next to the implications for payment systems of asymmetric information about costs.

5. Asymmetric information on costs and case-mix

In this section, we consider in more detail the implications for payment systems that arise from the purchaser having less good information about costs or case-mix than the supplier. The literature has adopted two approaches to this issue. The first is the mechanism design approach which considers the best that can achieved given the asymmetric information and what forms of payment system achieve it. This approach has been used only for simpler scenarios. The second approach analyses what can be achieved with particular types of payment systems. Table 1 lists studies in the order discussed here, starting with the mechanism design approach, and summarizes the cases they consider.

Table 1
Theoretical models of asymmetric information about case-mix

Study	1	2	3	4	5	6	7	8
Laffont and Tirole (1993, Ch. 1)				x		c		RT
Lewis and Sappington (1998)				x		2		RT
Laffont and Tirole (1993, Ch. 4)	x		x	x		c		RT
Newhouse (1983)					x	2		FP
Dranove (1987): Scenario I						c		FP
Dranove (1987): Scenario II					x	c		FP
Ma (1997, 1998): dumping	x	c	x	x	x	c		FP, CR
Allen and Gertler (1991)	x	x	x		x	2		FP
Ma (1994): cream skimming	x	x	x	x	x	2		FP, CR
Ellis (1998)	x	x			x	c	x	LCS

x – model has this feature.
1. Supplier chooses unverifiable quality.
2. Quality can be different for different patients.
3. Quality affects demand.
4. Supplier chooses unverifiable effort to reduce cost.
5. Supplier chooses which patients to treat after knows cost.
6. Number of cost types (c – continuous interval).
7. Supplier some benevolence [$v_q(.) < 0$ for some (x, q, e)].
8. Payment: RT – based on revealed type; CR – cost reimbursement; FP – fixed price; LCS – linear cost sharing (FP and CR special cases).

(Payment based on revealed type corresponds to the mechanism design approach.) All the studies listed there consider either a single supplier or competition between suppliers with the choice between them made by patients. In provision of health care, there has also been much interest in competitive bidding, where the choice between suppliers is made by the purchaser. We discuss that at the end of this section.

5.1. The mechanism design approach

As noted above, when the purchaser has less good information about costs than the supplier it is generally beneficial for the purchaser to make payment depend on the actual costs incurred in treatment, not just on the number of patients treated, provided monitoring of actual costs is not too expensive. The basic intuitive reason is as follows. Suppose the only asymmetric information concerns the fixed cost F, the purchaser knows this to lie in the interval $[\underline{F}, \bar{F}]$, and even at $\bar{F}$ treatment is worthwhile. Then the efficient numbers, quality and effort involved in treatment can still all be calculated by the purchaser because the solution given by (11) and (12) depends only on marginal costs and is thus independent of F. Moreover, these efficient levels can still be achieved by a two-part tariff with payment per patient treated given by (16). However, to ensure treatment is actually carried out, the lump sum payment a has to be high enough to enable the supplier to break even at the highest possible level of fixed costs. If the supplier's fixed cost is actually below this highest level, the supplier receives payoff $u > \bar{u}$. Thus, the supplier receives an *informational rent* because the information about fixed costs is private, the purchaser pays out more than the minimum necessary to have the service delivered and, because of the deadweight loss from raising funds via taxation, there is a welfare loss. Provided actual costs can be monitored, total payment is on average lower if there is some element of cost sharing because that lowers the total payment if the supplier's fixed cost is below the highest anticipated level. It also reduces incentives for cost-reducing effort because the supplier shares the cost savings with the purchaser. But, starting from the efficient level, a small reduction in cost-reducing effort has only a second-order welfare effect, whereas the reduction in expenditure that results from some cost sharing has a first-order effect. Thus, provided actual costs can be monitored sufficiently cheaply, it is always worthwhile to have some element of cost sharing.

The nature of optimal contracts in such situations is studied in Laffont and Tirole (1993). Their Chapter 1 analyses a model that is, in essence, similar to the case with quantity x and quality q given exogenously, though they use additional assumptions not detailed here. Their main results are as follows. Because of the revelation principle, we know that anything that can be achieved by a contract can be achieved by one that induces the supplier to correctly reveal its fixed cost. The supplier can be induced to do this because the lower is the fixed cost, the less is the effort (and, hence, the less both the disutility and the marginal disutility of effort) required to keep costs down to a given level. Thus, faced with a payment schedule that gives greater monetary reward (net of monetary cost) for lower observed cost, the supplier has greater incentive to keep cost down when fixed cost is lower. By facing the supplier with an appropriate trade-off between higher net reward (conditional on a low cost) and lower net reward (conditional on a high cost), the purchaser can ensure that only when the fixed cost is sufficiently low will the supplier expend the effort to reduce cost. Thus, one characteristic of an optimal contract is that effort increases, and total cost decreases, as fixed cost decreases. It turns out that effort is at the efficient level e^* for fixed cost at its lowest possible level $\underline{F}$. In all other cases, effort is below the efficient level. Other characteristics of the optimal contract follow from the concern to keep total payments as low as possible while still ensuring both that the service is provided and that truthful revelation is preserved. Thus, the optimal contract gives payoff $\bar{u}$ to the supplier when fixed cost is at its highest level $\bar{F}$ – anything higher would use more tax revenue than necessary. Moreover, for fixed cost below the highest level, the payoff is kept just high enough for the supplier to be indifferent between revealing the true fixed cost and claiming the next higher level of fixed cost. This enables the purchaser to keep the rent to the supplier below $\bar{F} - F$, its lowest level without cost sharing, but not eliminate it entirely. The rent necessarily increases as the fixed cost decreases because the supplier can achieve the same cost level as if fixed cost were higher with less effort, and thus with a higher payoff, than if the fixed cost actually were higher. For a precise statement of these characteristics and a formal derivation, see Laffont and Tirole (1993, Proposition 1.3).

This analysis applies to risk neutral purchasers and suppliers. With a risk averse supplier, of course, nothing changes unless there is some randomness in cost or its measurement. Baron and Besanko (1988) point out that it makes a difference whether such randomness is a forecast error about actual cost or an accounting error in the measurement of actual cost. Forecast errors introduce variance into the supplier's costs. The greater the share of costs reimbursed by the payer, the less the variance in the supplier's payoff. When the supplier has "mean-variance" preferences, therefore, greater risk aversion results in a higher optimal share of costs reimbursed. Accounting errors introduce variance into the supplier's revenue whenever there is some reimbursement of measured costs but not into actual costs. Thus the less the share of costs reimbursed by the payer, the less the variance in the supplier's payoff and higher risk aversion results in a smaller optimal share of costs reimbursed. For further discussion and references, see Laffont and Tirole (1993, Chapter 1).

Lewis and Sappington (1998) consider the case in which the supplier must incur a screening cost to ascertain whether the cost of treating the patient population is high or low but can also control costs by increased effort. There are two possible cost levels. If the screening cost is sufficiently high that it is not worth the supplier incurring it when the purchaser pays a fixed amount equal to the average cost of treatment with efficient effort, then a fixed payment of that amount is efficient. This is because, for reasons discussed previously, the fixed payment results in efficient cost-reducing effort and all patients are then treated at an average cost that reflects efficient effort. If, however, the fixed payment is at a level that results in screening, treatment will be carried out in the high cost case only if the payment is high enough to cover its cost. Then the supplier covers cost on high cost patients and makes a profit on low cost ones, thus receiving a rent overall. Cost sharing can improve on this because, if sufficient cost is shared with the purchaser, the supplier can be induced not to incur the screening cost and this enables the purchaser to lower the expected payment even though cost sharing induces too little cost-reducing effort. The purchaser then has the choice between this cost sharing arrangement and a fixed payment high enough to deter screening. However, for sufficiently low screening cost, the purchaser can do even better with a payment scheme that allows the supplier to opt for either a fixed payment or a smaller lump sum together with partial cost reimbursement. The precise amounts are selected to ensure that the supplier chooses the fixed price when costs turn out low and partial cost reimbursement when high. Then, as in the model of Laffont and Tirole (1993) discussed above, effort is efficient for low costs but kept too low for high costs in order to ensure truthful revelation.

These results do not consider the implications for unverifiable quality of service. Laffont and Tirole (1993, Ch. 4) consider the case in which the supplier chooses quality which, together with a random variable that is unobserved by the purchaser, affects demand. (The model also has demand affected by a regulated price that must be set to zero when patients do not pay for treatment.) They show that, for particular forms of the demand and cost functions in which effort affects marginal (and not fixed) costs, the optimal quantity and quality for *given* effort are the same as if the purchaser could condition the contract on all the relevant information. However, as in the case with no quality dimension, effort is distorted downwards except at one end of the distribution of types. Whether quality is distorted upwards or downwards depends on whether the difference between the marginal benefit and the marginal cost of quantity decreases with quality (net substitutes) or increases with quality (net complements). The intuition is that lower effort increases marginal cost and reduces output which makes higher quality services desirable in the net substitutes case. When effort affects fixed (but not marginal) costs, the degree of cost sharing is independent of the demand for quality. The conditions used to make this model tractable are, however, quite restrictive.

5.2. Particular payment mechanisms

An alternative to the mechanism design approach in the literature is that of analyzing the properties of particular payment mechanisms. A concern in the literature that has not been treated in the mechanism design approach is the effect of payment mechanisms on a supplier's decision of whether to treat only some patients. In the studies discussed so far, the supplier treats either all the patients or none of them. When, however, some patients are more costly to treat than others, a supplier that can identify costly patients may choose not to treat them. In particular, as Newhouse (1983) points out, a fixed payment per patient makes it unprofitable for suppliers to treat patients whose cost of treatment is higher than the payment. A profit-maximizing supplier will refuse to treat such patients if it can, and this has been a concern with payments (which are for the most part not based on actual costs of treatment) to GP fundholders in the British NHS. Even where such dumping is formally illegal (where, for example, hospitals cannot simply refuse to treat patients because of cost), ways may be found around that by, for example, claiming that the hospital does not have the facilities necessary to treat very severe cases. Dranove (1987) analyses the optimal fixed payment per patient when there is no cost sharing and suppliers can neither reduce costs by increased effort nor change quality. He considers the effect of increased case-mix variation in two scenarios, one in which hospitals must treat either all or none of the patients in their catchment area, the other in which they can choose to treat only the less costly patients. Patients who are refused treatment at their favored hospital incur access costs to an alternative hospital and, possibly, higher treatment costs at a hospital of last resort. Lowering the fixed payment shifts some patients from high cost to low cost hospitals but increases access costs and shifts some to high cost last-resort hospitals. In the first scenario, increased case-mix variation reduces social welfare for certain plausible distributions of costs, though it may increase or decrease the optimal payment per patient. In the second scenario, similar welfare results hold, though the relevant distributions are perhaps less plausible in this case.

Of course, when suppliers have no choice of quality or cost-reducing effort and the number of patients wanting treatment is inelastic, cost reimbursement can avoid dumping without having an adverse effect on costs. Ma (1997, 1998) add quality and effort decisions, both unverifiable but the same for all types of patients. An example of quality of this form is investment in better equipment that, if available for one patient, is available for all. An example of effort of this type is that involved in devising a better organizational structure. These papers give conditions under which dumping can be avoided without adverse consequences for quality and effort.[4] The contract to which these conditions apply has a fixed payment per patient with costs below a specified level and reimbursement of costs, plus a fixed amount per patient, for patients with cost above that level. The conditions, however, require that cost-reducing efforts affect costs only of less severe cases.

With quality and effort different for each type of patient, it is obviously much harder in general for the purchaser to induce the supplier to take the desired decisions because there are so many different dimensions to those decisions. If, however, suppliers can reduce the cost of treating individual patients by reducing quality for only those patients, they have less incentive to dump patients simply because they are too costly. An additional issue that arises in this case is that suppliers may bias quality towards those types of patients that are most profitable to treat in order to attract them, a phenomenon known as *cream-skimming*. Unless it turns out that the optimal price for all types if they were separately distinguished (given by (16)) happens to be the same, then a fixed price payment system cannot ensure efficient quality for all types. Allen and Gertler (1991) show that quality is inefficient with fixed prices when there are just two types of patients. Ma (1994), also for the case with two types of patients, compares a fixed price payment system with cost reimbursement and shows that one cannot, in general, conclude that one is preferable to the other.

Ellis (1998) analyses the implications of competition in quality for a duopoly with patient types along both a continuum in severity and a continuum in location, with travel

[4] There are problems with the discussion of this case in Ma (1994).

costs that depend on distance from the supplier, but with no possibility of the supplier reducing costs by increasing effort. If there are no non-monetary costs and benefits (formally, the function $v(.) \equiv 0$ everywhere), full cost reimbursement has no adverse effects on cost-reducing effort and leaves the supplier indifferent between quality levels, so the supplier has no positive incentive to deviate from efficient quality for each type of patient (although no positive incentive not to deviate either). Ellis (1998), however, considers the case in which the supplier cares directly about quality received by patients (corresponding to $v_q(.) < 0$), so full cost reimbursement results in quality for all types above the efficient level. Fixed price payment, while reducing incentives to oversupply quality, nevertheless still results in cream-skimming because suppliers provide too high quality to less severely ill patients (defined as those who benefit less from a given quality of treatment) in order to attract them. Quality for more severely ill patients, however, is typically less than efficient. Ellis (1998) argues that *linear cost sharing* (payment schemes with a combination of a fixed payment and a constant share of the actual cost of treatment for each patient) can improve on fixed price payment because, for a given total expenditure, cost sharing reduces the profitability of treating low severity patients (with lower costs) relative to high severity patients (with higher costs). It can also improve on cost reimbursement because it reduces the incentives to oversupply quality.

5.3. Competitive bidding

One obvious way to reduce costs is to use competitive bidding to select a low cost supplier. Laffont and Tirole (1993, Ch. 7) consider auctioning of incentive contracts for the model of their Chapter 1, similar in essence (as noted above) to the case in which quantity x and quality q are given exogenously, the differences between suppliers correspond only to differences in the fixed cost F, which the purchaser knows to lie in the interval $[\underline{F}, \bar{F}]$, and even at $\bar{F}$ treatment is worthwhile. Not surprisingly, the optimal auction awards the contract to the supplier that reports the lowest expected cost. Less obviously, the contract between the purchaser and the selected supplier is the same as with a single supplier except for a reduced fixed payment. The intuition for this is the following. For a single supplier, as explained above, the optimal degree of cost sharing trades off less efficient effort if fixed costs are high against reduced rent if fixed costs are low. At each level of fixed cost F, that trade-off depends on the distribution of fixed cost only for values below F. Competitive bidding eliminates some suppliers from this distribution, but only those with fixed costs higher than the selected supplier. Thus, it does not affect the trade-off (and, hence, the optimal degree of cost sharing) at the fixed cost level of the selected supplier. Moreover, for a supplier to win, its fixed cost must be below that of the second lowest cost bidder, which is in general strictly below $\bar{F}$. That enables the purchaser to lower the fixed payment required to ensure that the winner is willing to provide the service. For other models of competitive bidding, see the surveys by Milgrom (1987) and McAfee and McMillan (1987).

Competitive bidding has been used for some years by the US Centers for Disease Control and Prevention for the purchase of vaccines. Salkever and Frank (1996) do not find a significant effect of the number of actual or potential bidders (licensees) in reducing the purchase price relative to the catalog price, though they note that this may be because catalog prices themselves reflect the number of producers or because of collusion by suppliers. Recently, some Medicaid programs have adopted elements of competitive bidding for the provision of psychiatric treatments. A discussion of the methods of bidding used can be found in Schlesinger, Dorwart and Pulice (1986), Robinson and Phibbs (1989) and Fisher, Lindrooth, Norton and Dickey (1998). In contrast with most models of bidding, Medicaid programs delegate the bidding process to managed care organizations which then in turn attract bids for the provision of services and select suppliers on the basis of their bids. The purpose of these reforms is to lower the cost of service provision and there is some evidence that they are successful in this regard, see Schlesinger et al. (1986).

There is, however, concern that the quality of services supplied may suffer if bidders are selected only on the basis of price. In the context of psychiatric services, Fisher et al. (1998) find that there is no tendency for suppliers with the lowest historic costs to be more likely to win in the bidding process. They further find that a bid is more likely to be accepted when a supplier has relevant experience in dealing with complex cases. These findings suggest that price alone is not the basis for selecting a supplier in practice. This accords with theoretical analysis. Manelli and Vincent (1995) consider purchasing when different suppliers offer different qualities of service that are not observed by the purchaser. In their model, quality is an inherent characteristic of the supplier, not something the supplier can choose, and there is no choice of cost-reducing effort. Although the technical details are complicated, it is clear in this case that the purchaser may not wish to select the lowest cost bidder if the quality-cost combination of low cost bidders is unattractive. Indeed, Manelli and Vincent (1995) show that it may be better to allocate the contract to a randomly chosen supplier. However, we know of no theoretical results for the case in which unverifiable quality is chosen endogenously by the supplier.

5.4. Asymmetric information: conclusion

Optimal contracts are very sensitive to the nature of information asymmetry and to precisely what is under the control of the supplier. Even the performance of a *given* type of contract depends crucially on such things as whether the supplier observes the costs of treating individual patients before deciding whether to treat them. These are the kinds of things that may well vary for different types of health services. Thus, we have no general conclusion to this section beyond the rather obvious one that, provided costs can be monitored without too much expense, it is typically optimal for the payer to share some of the costs actually incurred in treatment.

Monitoring costs accurately is, however, not trivial. It requires setting up a mechanism to do so. There is also a concern that suppliers may indulge in *cost padding*, accounting manipulations that inflate the costs incurred, so that cost sharing is less powerful as a way to reduce total payments to suppliers. Cost padding, and ways to handle it, are of serious concern in the literature on government procurement, see for example, Laffont and Tirole (1993, Chapter 12). In severe cases, it may simply not be worthwhile to monitor costs. An important issue for empirical analysis is to assess how much is really gained in practical contexts by having the payer share part of the costs so that the gain can be weighed against the undoubted difficulty and expense of monitoring what costs actually are. Chalkley and Malcomson (1999) provide a preliminary assessment of this gain using distributions of costs taken from Medicare DRGs that indicates substantial cost savings of between 15% and 30%.

6. Commitment, hold-up, and the ratchet effect

Two issues that have been the focus of much attention in the economics literature on contracts are *hold-up* and the *ratchet effect*. These are as relevant to the purchase of health services as to other areas of contracting. Both are related to the extent to which the purchasing agency commits itself to the terms on which it will purchase services in the future. In this section, we explain the basic issues in the context of the framework set out above. Because, however, they have not been widely discussed in the context of health services, references are to the contracting literature in general rather than to that specifically on purchasing health services. First, however, we discuss the health purchasing literature on commitment to purchasing treatments in advance of knowing demand, termed *supply assurance*.

6.1. Supply assurance

The arrival of individuals requiring treatment in practice occurs continuously over time. Commitment can be important if there is stochastic demand and the purchaser wishes to ensure that capacity is available to treat patients whenever they arrive as, for example, with emergency care and life-saving medical interventions. Unless the purchaser has ensured adequate availability of capacity from a supplier, it may be required to purchase treatments at short notice from unfamiliar suppliers on the equivalent of ‘spot’ markets. To avoid this, the purchaser may commit to pay for a given number of treatments, some of which will go unused if few patients require treatment. The importance of supply constraints for hospital beds is considered by Joskow (1980). Supply assurance has been analyzed in the context of health services by Fenn et al. (1994).

Suppose that, in contrast to the models discussed in Sections 2.2 and 3, both quality and cost-reducing effort are verifiable. The purchaser can then make payment conditional on efficient quality and effort being supplied. Supply assurance in the sense used by Fenn et al. (1994) involves a commitment to pay $F + c(N, q^*, e^*) + v(N, q^*, e^*)$ independently of the number of patients who actually require treatment, up to a maximum of N, the assured supply. The purchaser anticipates cost $\bar{c}$ for each patient in excess of N who requires treatment, where $\bar{c}$ exceeds the average cost per patient that results from assured supply. Formally, $\bar{c} > [F + c(N, q^*, e^*) + v(N, q^*, e^*)]/N$ for some $N > 0$. This may be for a number of reasons. First, transacting with an alternative supplier may itself be costly. Second, the purchaser may be less able to exert bargaining power over outside suppliers and may therefore have to pay in excess of cost. Propper (1996), for example, finds evidence that the prices paid by health authorities in the British NHS are higher when they are negotiated with outside suppliers. Third, if spot market treatments are with more distant suppliers, the purchaser may have to pay for additional transport costs. The purchaser then has a trade-off to consider. By increasing its commitment to the primary supplier, the purchaser ensures that one extra treatment if needed is provided at a lower cost than would be incurred in purchasing that treatment elsewhere, but increases the risk that it pays for a treatment that is not required. With demand for treatments x a random variable that has density function $g(x)$ and distribution function $G(x)$, the purchaser chooses N to

$$\min_N c(N, q^*, e^*) + v(N, q^*, e^*) + \frac{\int_N^\infty xg(x)\,dx}{1 - G(N)}\bar{c}. \tag{17}$$

Given the assumptions, this has an interior solution with the purchaser committing to some N^* greater than the lowest level of demand.

Fenn et al. (1994) analyze a number of variations of this model in which the supplier has fixed capacity that may be bid away by rival purchasers and there is regulation of external prices. The determinants of N^*, the optimal level of supply assurance, are thereby derived. In Csaba and Fenn (1997) this model is applied to data relating to *cost per case* contracts in the British NHS, which exhibit an element of supply assurance.

6.2. Investment and hold-up

Suppliers make investments that reduce the cost, or improve the quality, of the services supplied. In many cases those investments are to some extent specific to the relationship with a particular purchaser. A British NHS hospital investing in specialist facilities, for example, is unlikely to be able to make full use of those facilities for other purchasers if the district health authority decides not to contract with it.

Suppose the fixed cost F, instead of being given exogenously, results from investment expenditure that reduces the cost of supplying a given quantity and quality of services with given cost-reducing effort. To capture this, we rewrite variable cost as $c(x, q, e, F)$ with $c_F(.) < 0$ and $c(.)$ a strictly convex function. However, as is the nature of investment, this expenditure has to be made before a decision is taken on (x, q, e). For any given choice of (x, q, e) anticipated at the time the investment decision is made, the efficient investment can be derived from the appropriately amended version of the maximand in (3) and is given uniquely by the first order condition

$$-c_F(x, q, e, F) = 1. \tag{18}$$

This equates the marginal cost reduction with the marginal cost of the investment which, since the investment is measured in money terms, is just one.

Suppose the investment is entirely specific (that is, it provides no return to the supplier other than when supplying the particular purchaser) and the supplier decides the amount of investment before negotiating terms with the purchaser. At the time terms are negotiated, the investment has already been made and its cost is a bygone. Thus the supplier's decision whether to accept the terms offered depends only on whether they enable a payoff of at least $\bar{u}$ from the future. The amount that has been invested in the past is irrelevant to this. The purchaser still wishes to minimize total expenditure for given (x, q, e) in order to minimize the deadweight loss from taxes, as explained in Section 2.1. With no asymmetric information, the purchaser therefore sets the terms so as to give the supplier a payoff of exactly $\bar{u}$. That payoff is independent of the amount of investment. Thus the supplier has no incentive at all to invest. In effect, the purchaser captures the return on the supplier's investment, which is what is meant by *hold-up*. Even if the supplier can get a return on the investment from supplying to another purchaser, as a result of which $\bar{u}$ increases with the amount of the investment, the level of investment will be efficient only if the return, at least at the margin, equals the cost savings that accrue with the particular purchaser (that is, $d\bar{u}/dF = -c_F(x, q, e, F)$). That corresponds to the investment being general in the sense of Becker (1975).

An obvious way to respond to the hold-up problem is for the purchaser and supplier to agree terms before the investment is made. If the purchaser can commit to not capturing any of the return on the supplier's investment, then the supplier will invest efficiently. Moreover, it is in the purchaser's interest to make such a commitment because making the investment at this stage enhances social welfare. The contracting literature shows that it is not always easy to set the terms of a contract in such a way that the purchaser never captures any of the return on the supplier's investment, especially if the precise details of the investment cannot be specified verifiably in advance and if, between making the investment and provision of the service, unverifiable random events occur that affect efficient provision or costs. See Laffont and Tirole (1993, Sections 1.8 and 1.9) for a discussion of these issues. Moreover, any such arrangements can work for only as long as the purchasing agency is prepared to commit itself in advance. Some specific investments have returns extending far into the future. If new terms are negotiated before the returns have all been received, the purchaser will typically capture some of those returns which, if anticipated, results in inefficient investment.

How important is this issue for government purchases of health services? Two conditions are crucial for hold-up to arise. The first is that the return on the supplier's investments is lower if the purchasing agency does not actually purchase the anticipated services from it. That is likely to be particularly important where one government agency is a major purchaser of services within the geographical area of the supplier. The second is that there is individual negotiation between purchaser and supplier about the terms of provision, as a result of which the terms are influenced by the individual supplier's investment. This will not be the case where an individual supplier provides only a small proportion of the total supply of that service and the terms set are the same for all suppliers. In that case, any one supplier's investment has an insignificant effect on the terms set and each individual supplier reaps all the returns on its own investment. Then, for the same reason as in the model of yardstick competition in Shleifer (1985), the supplier's decision is efficient. On both these counts, one might plausibly expect hold-up to be a more important issue in the British NHS, with dominant area-based purchasers negotiating terms with individual suppliers, than in the US Medicare system.

A related issue arises with investment in research and development (R&D) for new pharmaceuticals. Danzon (1997, 1998) argue that pricing of pharmaceuticals at marginal cost will typically not cover investments in R&D (which accounts for roughly 30% of total costs) and that, therefore, differential pricing in different national markets along Ramsey pricing principles is appropriate. If government purchasers in each national market negotiate contracts with price reduced to marginal cost, the incentive to invest in R&D is correspondingly reduced. As in the hold-up model, the buyer uses the fact that investment is a bygone to capture part of the return on the investment. Of course, pharmaceutical R&D is not typically a specific investment and it generates new products, rather than cost reductions for existing products. However, the purchaser's monopsony power gives it the ability to capture part of the return on the investment in a way similar to the monopsony power a purchaser acquires when the supplier makes a specific investment. Moreover, when there are a number of purchasers with monopsony power, not the single monopsony purchaser in the case of a specific investment, commitment by individual purchasers may not overcome the inefficiency of investment because each may individually gain by refusing to commit and then free riding on the investment that results from commitment by other purchasers. There are, of course, other major issues in inducing efficient investment in R&D for pharmaceuticals, issues wider than the purely contracting ones raised here – for example, patent protection and its implications for patent races.

6.3. The ratchet effect

Commitment can also be important when there is asymmetric information about levels of cost or case-mix and these levels persist over time. The essential reason can be illustrated with the case in which the fixed cost F is known to the supplier but unknown to the purchaser and both x and q are given exogenously. With only a single interaction as in Section 5, the supplier's cost observed *ex post* by the purchaser reveals the supplier's fixed cost. Thus, if the same purchaser and supplier were subsequently to negotiate a new contract for a second period, the purchaser would know the supplier's fixed cost and set the terms such that the supplier receives a payoff of only $\bar{u}$ in that period. So, if the purchaser offers the optimal contract for a one period relationship in the first period and the supplier responds in the way that is optimal for a one period relationship, the supplier receives the same rent in the first period as from a one period relationship but no rent in the second period. This is the *ratchet effect*. As a result of observing a low cost in the first period, the purchaser ratchets up the performance required in the second period to reduce the rent it pays to the supplier.

But the supplier can always do better than truthfully revealing costs in these circumstances. To see why, consider the case in which the fixed cost takes on one of two values $\underline{F}$ and $\bar{F}$. In a single period relationship as discussed in Section 5, the optimal contract leaves the supplier indifferent between revealing the true fixed cost and claiming the next highest level of fixed cost. That is, if fixed cost is $\underline{F}$, the supplier receives the same payoff from revealing that fixed cost is $\underline{F}$ as from claiming that fixed cost is $\bar{F}$ and lowering effort so that total cost is consistent with its claim. Thus if, in the first period of a two period relationship, a supplier with fixed cost $\underline{F}$ is faced with the optimal contract for a one period relationship, that supplier receives the same rent (call it R) in the first period from misrepresenting fixed cost as from truthfully revealing it. Moreover, if the supplier adopts the strategy of *always* claiming that fixed cost is $\bar{F}$, the purchaser has learned nothing about the true fixed cost at the start of the second period. Thus, the optimal contract for the second period is then simply the optimal contract for a one period relationship, from which the supplier receives rent R in the second period also. The conclusion is that, from truthfully revealing fixed cost in the first period, a supplier with fixed cost $\underline{F}$ receives rent R in the first period but no rent in the second, whereas from not truthfully revealing fixed cost in the first period, the supplier receives rent R in both periods. Thus it is never optimal for the supplier to reveal the true fixed cost. That though is not the end of the problem. As explained in Section 5, the optimal contract in the one period relationship ensures that the supplier exerts efficient cost-reducing effort when the fixed cost is $\underline{F}$ but less than efficient effort when fixed cost is $\bar{F}$. Thus, to disguise the true fixed cost, the supplier reduces effort below the efficient level even when fixed cost equals $\underline{F}$. In this way, the existence of the second period actually makes things worse for the purchaser in the first.

It is clear from this discussion that the purchaser would do better if it could commit itself to offering the optimal contract for a one-period relationship in both periods. The overall rent for a low fixed cost supplier would be the same but that supplier would provide efficient effort in the first period because, with advance commitment from the purchaser to the second period contract, there is no reason not to reveal the true fixed cost in the first period. In fact, this is the best the purchaser can do, see Laffont and Tirole (1993, Chapter 9).

Both hold-up and the ratchet effect illustrate the types of problems that can arise from using short-term contracts to govern long-term relationships. Thus, where specific investments and/or asymmetric information about costs are important, inefficiencies may arise if, as with most of the contracting currently done in the British NHS, purchasers use only short-term contracts. Of course, purchasers may perceive problems in committing themselves to long-term contracts, particularly when there is substantial uncertainty. Political pressures may, for example, make it hard to stick to a long-term contract that is optimal under asymmetric information about costs if costs actually turn out to be extremely low and suppliers make very large surpluses as a result. But there are also economic benefits to having short-term contracts because they increase the importance to suppliers of maintaining a good reputation. We take up that issue in the next section.

7. Reputations

An important issue neglected in the framework of Section 2.1 is that the relationship between purchaser and supplier is, in many cases, an on-going one with the purchaser making not just a single arrangement for services to be provided but expected to continue buying services from the same supplier year after year. The on-going nature of the relationship provides another mechanism for maintaining unverifiable quality of services, namely the supplier's concern for its future business via the effect on its reputation.

The literature has discussed two forms of reputation. The first is reputation for a characteristic that is either inherent to the supplier or that, once acquired, is long lasting. An example is when the supplier makes an investment that, instead of simply reducing the cost of providing services as in Section 6.2, enhances the quality of services now and in the future. The second is reputation for past behavior that does not necessarily imply anything about future quality but that nevertheless affects future equilibrium outcomes. An example is when an increase in quality in one period does not necessarily result in higher quality in subsequent periods but plays a role in sustaining beliefs that high quality will be provided in the future. We discuss each of these briefly.

7.1. Reputation for characteristics

For a profit-maximizing supplier to be concerned about future reputation, there must be some rent in the future from having a reputation. Such a rent can arise when information about quality is acquired during the *experience* of being treated and passed on to others seeking treatment, who then make decisions on the basis of it. When investments have a long-term effect on quality, future patients know they will receive benefits from those investments experienced by current patients. Higher quality now may thus result in higher patient demand in the future. That may make investment in quality characteristics now worthwhile.

It does not, however, guarantee it. Exactly as in the discussion of hold-up in Section 6.2, once the investment has been made, the purchaser still wishes to minimize total expenditure for given levels of service. If it can, it will therefore set terms to give the supplier a payoff of exactly $\bar{u}$ in the future and the supplier will not receive any return on the reputation. That can be avoided if the purchaser commits in advance to allowing a rent for high demand in the future or if there is something the purchaser does not know about the supplier that allows the supplier to obtain an informational rent in the future. The former case is similar to that discussed in Section 3 in which current demand is influenced by quality except that the effect on demand is in the future. Thus the rate at which the supplier discounts the future (or, alternatively, how long it takes for information about quality to get around) becomes important for the incentive effects of demand. Laffont and Tirole (1993, Section 4.6) give an example of the latter with asymmetric information about costs. Again, for obvious reasons, the discount rate plays an important role in determining incentives, in this case the extent to which the payer shares the actual costs of treatment. Whether that share is larger or smaller than when quality is verifiable is, however, ambiguous.

7.2. Reputation for past behavior

We turn next to reputation for past behavior, rather than for long-lasting characteristics. To enforce specific quality standards by a formal agreement or contract requires that the quality standards are verifiable, that is, they are specified unambiguously in advance in such a way that it is clear after the event whether they have been met. But in an on-going relationship with contracts that are renewed, a purchaser can make use of information about quality of provision in the past, even if not verifiable, to affect whether the contract is renewed and, if so, on what terms. By offering less good terms in the future, or threatening to switch to another supplier, the purchaser can make it costly for the supplier to skimp on quality now.

To see how powerful this mechanism can be, consider the following scenario. Suppose patient demand is completely unaffected by quality, so there is no scope for maintaining quality via the demand mechanism analyzed in Section 3. Suppose, moreover, that quality is unverifiable but, to take the extreme case, is observed correctly by the purchaser after each round of treatments. Suppose, in addition, both purchaser and supplier envisage that the relationship could, in principle, continue for ever. Finally suppose that the number of patients to be treated and all the functions are constant over time and known to both parties.

We use the same notation as before but with a subscript t attached to each variable to denote the time period. For the present discussion, it is convenient to define $e(x_t, q_t)$ as the cost-reducing effort that minimizes $c(x_t, q_t, e_t) + v(x_t, q_t, e_t)$ for given x_t and q_t, and $q(x_t)$ as the quality that minimizes the same expression given x_t and effort $e(x_t, q_t)$. We know from Section 2.1 that, whatever x_t and q_t actually are, $e(x_t, q_t)$ is the effort the purchaser would wish the supplier to provide and that the supplier will in fact provide this level if there is no element of cost reimbursement. With $e(x_t, q_t)$ and the expression for u in (1) substituted into the social welfare function in (2), the purchaser's payoff can be written

$$W(x_t, q_t, P_t) \equiv b(x_t, q_t) - F - c\big[x_t, q_t, e(x_t, q_t)\big] - v\big[x_t, q_t, e(x_t, q_t)\big] - \alpha P_t. \tag{19}$$

The supplier's payoff can be written

$$U(x_t, q_t, P_t) \equiv P_t - F - c\big[x_t, q_t, e(x_t, q_t)\big] - v\big[x_t, q_t, e(x_t, q_t)\big]. \tag{20}$$

The present discounted value of the supplier's payoff from any date τ on is given by

$$\sum_{t=\tau}^{\infty} \delta^{t-\tau} U(u_t, q_t, P_t), \tag{21}$$

where δ $(0 \leqslant \delta < 1)$ is the supplier's discount factor.

Suppose the service is provided by this supplier for only one period. Then, since patient demand is unaffected by quality, with any reward scheme that pays only on the basis of numbers treated the supplier sets quality at $q(x_t)$.[5] Suppose the purchaser specifies total payment of P if x patients are treated, but nothing otherwise. Then, provided it results in a payoff of at least $\bar{u}$, the supplier treats x patients at quality $q(x)$ and with cost-reducing effort $e[x, q(x)]$. Anticipating this behavior, the purchaser chooses x to maximize social welfare in (19) given $q(x)$ and $e[x, q(x)]$, and sets P to yield the supplier the lowest payoff $\bar{u}$ at which the service will be provided. Denote these levels by $\bar{x}$ and $\bar{P}$, respectively, and let $\bar{q} = q(\bar{x})$. The strategies in which the purchaser sets the contract $(\bar{P}, \bar{x})$, and the supplier responds with $\bar{q}$, form an equilibrium for a single provision of the service. Applied to every period, they also form an equilibrium for repeated provision because at each decision node the purchaser is making the best choice given the supplier's behavior and *vice versa*.

With repeated provision, however, the parties can typically do better than this. Suppose they were to make the following informal agreement about P, x and q. The purchaser is to set the formal contract specifying total payment of P conditional on x patients being treated, and nothing otherwise. The supplier is to respond by choosing quality $q > q(x)$. As long as each has always stuck to this agreement in the past, it is to be continued into the future. However, if either ever departs from the agreement, the purchaser is to set the contract $(\bar{P}, \bar{x})$, and the supplier quality $\bar{q}$, in every subsequent period. If either departs from the agreement, this continuation behavior is equilibrium behavior because, as already explained, it corresponds to an equilibrium of the repeated game.

We are interested in the agreements about P, x and q that both parties will actually stick to. Such agreements are called *self-enforcing*. Clearly, necessary conditions for the agreement to be self-enforcing are that neither party receives a payoff lower than with $(\bar{x}, \bar{q}, \bar{P})$, that is

$$W(x, q, P) \geqslant W(\bar{x}, \bar{q}, \bar{P}) \tag{22}$$

$$U(x, q, P) \geqslant \bar{u}. \tag{23}$$

Given the response of the supplier, the best deviation from the agreement for the purchaser is to set the contract $(\bar{P}, \bar{x})$. Thus the purchaser will stick to the agreement as long as (22) is satisfied. Moreover, given the response of the purchaser, the best deviation for the supplier is to set quality at $q(x)$ because the payoff in each future period

[5] It is possible to achieve quality above $q(x_t)$ by reimbursing costs sufficiently generously to compensate the supplier for the additional monetary and non-monetary costs of the higher quality [see Chalkley and Malcomson (1998a)]. However, the supplier then sets cost-reducing effort different from $e(x_t, q_t)$. The discussion in the text assumes that it is better for the purchaser to put up with quality $q(x_t)$ than with inefficient cost-reducing effort. The argument can be reformulated if the opposite is the case.

following a deviation is $\bar{u}$ no matter what the supplier does now. Thus the supplier will stick to the agreement at every date τ if and only if

$$\sum_{t=\tau}^{\infty} \delta^{t-\tau} U(x, q, P) \geqslant U\big[x, q(x), P\big] + \sum_{t=\tau+1}^{\infty} \delta^{t-\tau} \bar{u}, \quad \text{for all } \tau. \tag{24}$$

This condition can be rewritten

$$U(x, q, P) + \sum_{t=\tau+1}^{\infty} \delta^{t-\tau} U(x, q, P) \geqslant U\big[x, q(x), P\big] + \sum_{t=\tau+1}^{\infty} \delta^{t-\tau} \bar{u}, \quad \text{for all } \tau, \tag{25}$$

or, with the summation terms moved to the left hand side and evaluated using the formula for the sum of a geometric progression, and the other terms moved to the right hand side,

$$\frac{\delta}{1-\delta}\big[U(x, q, P) - \bar{u}\big] \geqslant U\big[x, q(x), P\big] - U(x, q, P), \quad \text{for all } \tau. \tag{26}$$

Equation (26) has a direct interpretation. The left hand side is the present discounted value from $\tau + 1$ on of the gains to the supplier from sticking to the agreement. The right hand side is the largest short term gain in period τ that the supplier can make by defaulting on the agreement. Note that it is always non-negative because $q(x)$ maximizes $U(x, q, P)$ for given x and P. Thus (26) corresponds to the requirement that the long term gains to the supplier from sticking to the agreement exceed the short term gains from cheating on it.

For $\delta = 0$ (that is, the supplier does not care about future payoffs at all), the left hand side of (26) is zero and thus the only agreement that satisfies (26) is $(\bar{x}, \bar{q}, \bar{P})$. Then an informal agreement cannot deliver anything better than treating each period independently as it comes. As long as $\delta > 0$, however, the left hand side is positive for any agreement that gives the supplier a payoff greater than $\bar{u}$ and an informal agreement can then do better than treating each period independently. (The supplier does not actually have to receive a total payoff from the agreement greater in present discounted value than $\bar{u}$ in each period because the purchaser may be able to impose an upfront payment from the supplier to make the agreement in the first place. This has not been included in the formal model.) In the limit as $\delta \to 1$, the left hand side of (26) tends to infinity for any (x, q, P) for which (23) is satisfied with strict inequality. The right hand side is independent of δ. Thus, for δ sufficiently close to 1, any (x, q, P) that satisfies (22) and (23) with strict inequality is self enforcing. This is an example of the well known "Folk Theorem" result for repeated games. The sustainable levels include, in particular, the efficient levels x^* and q^* (which imply efficient effort e^*) with the supplier's payoff arbitrarily close to $\bar{u}$.

That is, of course, an extreme case. What can be achieved in practice is limited by a number of things. First, to the extent that the supplier discounts the future it may not be possible to get close to the efficient outcome. Moreover, the requirement that $\delta > 0$ corresponds to there not being a last period in which the service is supplied. If there were, the supplier would no longer care about the future. That limitation may, however, be less important in practice than in theory. It is a robust empirical finding with experimental games of this type that players behave in the early stages as if the game will continue for ever even though it is in fact finite [see Roth (1995, Section III.A.1)].

A second important practical limitation concerns measurement error in quality observed by the purchaser and uncertainty about the efficient number of patients to treat at the time the contract for each period is agreed. Quality indicators such as mortality statistics are affected by things other than the actions taken by the supplier. The problem this creates is that the purchaser will not then be sure whether the supplier provided the quality level that was agreed. Uncertainty about efficient treatment numbers does not create a problem if the realization of the random variable that determines it (for example, demand for treatment) is verifiable *ex post* because the payment can then be made conditional on treatment of the number that is efficient given the realization. Even if it is not verifiable but is observed by the purchaser, continued cooperation in the future can be made conditional on treatment of that number. Where that is not the case, however, the contract would have to specify a schedule of payments conditional on the number treated and that may give the supplier more scope for profitable deviations from what is agreed. One way for the purchaser to respond to both measurement error in quality and unobservable demand is to set up a statistical test and withdraw cooperation only if the probability the supplier has not supplied the agreed quality is above a specified level. For δ sufficiently close to 1, there are then conditions under which the Folk Theorem result holds, see Fudenberg and Levine (1991) and Fudenberg, Levine and Maskin (1994), because a deviating supplier will eventually be discovered. For lower values of δ, however, measurement error in quality and unobservable demand may limit further what can be achieved. Observation of costs may help in reducing measurement error because it provides an alternative way to estimate the quality level given that effort is always chosen in a myopically optimal fashion. If, of course, there is asymmetric information about costs and case-mix, issues of the type discussed in Section 5 arise, though repetition of the relationship over time can still be expected to enable the parties to get closer to an efficient outcome.

It is, however, important to remember that *any* outcome that satisfies (22), (23) and (26) is an equilibrium, even those that are Pareto dominated. Thus, although repetition of the relationship over time *may* enable the parties to do better than if they are concerned with only one period at a time, there is no guarantee that it *will* do so. A sour relationship in which neither party trusts the other to continue to cooperate in the future may be no better than if the parties treat each period independently.

8. Conclusion

In many countries, governments are substantial purchasers of health services and in this role need to pay attention to the implications of different forms of payment mechanism for the nature and cost of services that are delivered. In this chapter we have focused on a number of different themes concerning the relationship between government purchasers and suppliers, either that have featured prominently in the literature on contracts for health services or that we think have not received the attention in that literature that they deserve.

The first, and most pervasive, theme concerns the tension between obtaining appropriate quality of services and keeping costs at an acceptable level. This tension goes deeper than the obvious one that higher quality requires more inputs that must be paid for. When payment fully reimburses suppliers' costs, as with traditional medical insurance, suppliers have no incentive to go to any trouble to keep costs down for any *given* level of quality (and may also indulge patients' wishes for quality beyond the level at which marginal benefit equals marginal cost). It is therefore natural that much of the literature has been taken up with examining mechanisms for containing cost whilst maintaining quality. In this capacity, fixed price contracts have been the subject of particular attention, both practically in the form of prospective payment systems, and theoretically. One conclusion is that, if conditions are right – which has a number of specific connotations here – fixed price contracts have much to commend them. Moreover, the evidence from the now widespread use of fixed prices by government purchasers, particularly in the US, supports the view that both quality and cost dimensions of health services respond in a way consistent with the framework that underlies this conclusion. There are, however, good reasons to think that the very specific conditions under which fixed price contracts are the best payment mechanism are not universally applicable. Consistent with this is the evidence that, even where explicit cost sharing is not used, the contracting parties introduce some degree of cost sharing in other ways.

Under some conditions, cost sharing by the purchaser can help improve the cost and quality trade-off. It is also valuable in reducing the expected cost of service provision when there is asymmetric information between purchaser and supplier about costs or case-mix, the second theme in this chapter. Asymmetric information about case-mix arises naturally when there are problems in specifying a treatment precisely enough given the variation in the medical conditions that can be encountered and the superior information that suppliers have in reaching a precise diagnosis. The literature on asymmetric information has analyzed a variety of different cases, with little in the way of general results beyond the usefulness of *some* role for cost sharing.

The third theme – the importance of commitment in contracts – has not attracted as much attention in the health economics literature on government purchasing. Other than for analysis of the desire to commit to a given volume of purchasing in order to ensure supply of vital services, comparatively little attention has been directed towards the need of purchasers to commit in advance. We have considered additional commitment issues that would appear to be of practical relevance to government purchasers of health services. Investment by suppliers of health services may be specific to the government purchaser, if only because the government agency purchases a significant proportion of services in the locality and thus alternative purchasers are hard to find if the agency withdraws its custom. In the presence of specific investments, hold-up is a real concern. Likewise, circumstances in which suppliers would wish to inflate costs in order to avoid subsequent downward pressure on prices – the ratchet effect – are a concern where, as in the British NHS, there is widespread use of short term contracts. Short term contracts and the problems they entail have not as yet received much attention in the literature on health contracting but we conjecture that they can be expected to feature in the future. The same applies to analysis of the role of reputation in maintaining quality in long term relationships, the final theme we have identified. We are not aware of any applications of the literature on this to public purchasing of health services and we have, therefore, provided some preliminary analysis that we hope will stimulate health economists to pursue this issue in the future.

The approach we have followed in this chapter is that of normative economics – it has been concerned with what it is appropriate for a government agency to do given that it is concerned with social welfare. We, in common with much of the literature, have not considered the practical issues that result from those agencies having goals that may differ from social welfare or the problems that arise when those responsible for running government agencies are subject to pressure or influence. Some approaches to this issue are discussed in Besley and Gouveia (1994) but it, together with the themes we have highlighted, remain important areas for future research.

Appendix: The framework with a private market

When suppliers of health services provide for private as well as publicly funded patients, a social welfare maximizing purchaser needs to be concerned with the benefits of all recipients of services. The framework set out in Section 2.1 can be extended to allow for this concern. Let x_p denote the number of privately paying patients, q_p the quality with which they are treated, p_p the price each private patient pays, and e_p the cost-reducing effort the supplier puts into providing private services. The government purchasing agency is concerned with social welfare, so its benefit function should include those patients paying privately and we write that function as $\hat{b}(x, q, x_p, q_p)$. The variable monetary and non-monetary cost functions must be similarly extended to patients paying privately, so we write these as $\hat{c}(x, q, e, x_p, q_p, e_p)$ and $\hat{v}(x, q, e, x_p, q_p, e_p)$, respectively. The objective of the supplier in (1) is then replaced by

$$u = P + p_p x_p - F - \hat{c}(x, q, e, x_p, q_p, e_p) - \hat{v}(x, q, e, x_p, q_p, e_p). \tag{27}$$

In general, the demand for private services will depend on the quantity x and quality q provided publicly as well as the price p_p and quality q_p offered in the private market. Contracts with a government purchasing agency are revised only periodically, hence their terms are typically known in advance of patients deciding whether to be treated privately and thus before private market price, quality and effort are determined. For any terms set for public provision, the supplier selects p_p, q_p and e_p to maximize the objective function in (27) given the resulting private demand and the market structure for private supply. We suppose that two properties are satisfied. First, the private market outcome is unique given any publicly provided outcome x, q, and e. Second, cost-reducing effort for public provision e does not affect the marginal monetary and non-monetary costs of supplying quantity or quality to the private market. We can then represent the outcome of the private market as

$$\begin{aligned} x_p &= x^p(x,q), \\ q_p &= q^p(x,q), \\ e_p &= e^p(x,q), \\ p_p &= p^p(x,q). \end{aligned} \tag{28}$$

With a private market, the expression for social welfare differs from (2) in that $b(.)$ is replaced by $\hat{b}(.)$ and the payment by private patients, $p_p x_p$, must be subtracted because the supplier's benefit from these payments is included in u, see (27), and their cost to private patients must therefore be deducted. Substitution from (27) for P in the expression for social welfare gives the purchaser's objective function

$$\begin{aligned} &\hat{b}(x,q,x_p,q_p) + \alpha p_p x_p \\ &\quad - (1+\alpha)\big[F + \hat{c}(x,q,e,x_p,q_p,e_p) + \hat{v}(x,q,e,x_p,q_p,e_p)\big] - \alpha u. \end{aligned} \tag{29}$$

Then, if we define the benefit and cost functions $b(.)$, $c(.)$ and $v(.)$ by

$$\begin{aligned} b(x,q) &\equiv \hat{b}\big[x,q,x^p(x,q),q^p(x,q)\big] + \alpha p^p(x,q)x^p(x,q), \\ c(x,q,e) &\equiv \hat{c}\big[x,q,e,x^p(x,q),q^p(x,q),e^p(x,q)\big], \\ v(x,q,e) &\equiv \hat{v}\big[x,q,e,x^p(x,q),q^p(x,q),e^p(x,q)\big], \end{aligned} \tag{30}$$

we get exactly the same purchaser's objective as in (3). Thus, provided the compound functions in (30) have the appropriate concavity and convexity properties, the analysis can proceed as in the text. Note that the payment by private patients multiplied by α in (29) is properly represented as a social benefit because it corresponds to the reduction in deadweight loss from not having to raise these funds via taxation.

The assumption that cost-reducing effort for public provision e does not affect the marginal monetary and non-monetary costs of supplying quantity or quality to the private market may not, however, be appropriate in many cases. Then the private market outcome in (28), and hence the compound benefit function $b(.)$, in general depend on e through the cost side. The intuition here is that effort in public provision affects marginal monetary and non-monetary costs in private provision, and thus the quantity and quality of private provision. Those, in turn, affect social benefit even though the benefit function does not itself depend directly on effort. In that case, the analysis in the text needs to be amended in a number of places. We do not pursue that issue here.

The private market may, of course, suffer from market failure for any of the reasons traditionally associated with markets for health services. On this, see the chapter by Dranove and Satterthwaite (2000). Then the government may wish to intervene in the market to increase social welfare. That, however, is an issue of market regulation that falls outside the scope of the present chapter.

References

Allen, R., and P. Gertler (1991), "Regulation and the provision of quality to heterogeneous consumers: the case of prospective pricing of medical services", Journal of Regulatory Economics 3:361–375.

Baron, D.P., and D. Besanko (1988), "Monitoring of performance in organizational contracting: the case of defense procurement", Scandinavian Journal of Economics 90(3):329–356.

Becker, G.S. (1975), Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, 2nd edn. (Columbia University Press, New York).

Besley, T., and M. Gouveia (1994), "Alternative systems of health care provision", Economic Policy 19:199–258.

Chalkley, M., and J.M. Malcomson (1998a), "Contracting for health services when patient demand does not reflect quality", Journal of Health Economics 17(1):1–19.

Chalkley, M., and J.M. Malcomson (1998b), "Contracting for health services with unmonitored quality", Economic Journal 108(449):1093–1110.

Chalkley, M., and J.M. Malcomson (1999), Cost Sharing in Health Service Provision: An Empirical Assessment of Cost Savings (Department of Economics, University of Oxford).

Chalkley, M., and D. McVicar (1998), Contracts in the National Health Service: An empirical study (University of Southampton, Department of Economics).

Chernew, M., and D.P. Scanlon (1998), "Health plan report cards and insurance choice", Inquiry 35(1):9–22.

Csaba, I., and P. Fenn (1997), "Contractual choice in the managed health care market", Journal of Health Economics 16(5):579–588.

Cutler, D.M. (1995), "The incidence of adverse medical outcomes under prospective payment", Econometrica 63(1):29–50.

Danzon, P.M. (1997), "Price discrimination for pharmaceuticals: welfare effects in the US and the EU", International Journal of the Economics of Business 4(3):301–321.

Danzon, P.M. (1998), "The economics of parallel trade", Pharmacoeconomics 13(3):293–304.

DesHarnais, S.I., R. Wroblewski and D. Schumacher (1990), "How the Medicare prospective payment system affects psychiatric patients treated in short-term general hospitals", Inquiry 27:382–388.

DesHarnais, S.I., E. Kobrinski, J. Chesney, M. Long, R. Ament and S. Fleming (1987), "The early effects of the prospective payment system on inpatient utilization and the quality of care", Inquiry 24:7–16.

Dranove, D. (1987), "Rate-setting by diagnosis related groups and hospital specialization", RAND Journal of Economics 18(3):417–427.

Dranove, D., and M. Satterthwaite (2000), "The industrial organization of health care markets", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 20.

Dranove, D., and W.D. White (1994), "Recent theory and evidence on competition in hospital markets", Journal of Economics and Management Strategy 3(1):169–209.

Edlin, A.S. (1997), "Do guaranteed-low-price policies guarantee high prices, and can antitrust rise to the challenge?", Harvard Law Review 111(2):528–575.

Eldenburg, L., and S. Kallapur (1997), "Changes in hospital service mix and cost allocations in response to changes in Medicare reimbursement schemes", Journal of Accounting and Economics 23(1):31–51.

Ellis, R.P. (1998), "Creaming, skimping and dumping: provider competition on the intensive and extensive margins", Journal of Health Economics 17(5):537–555.

Ellis, R.P., and T.G. McGuire (1986), "Provider behavior under prospective reimbursement: cost sharing and supply", Journal of Health Economics 5:129–151.

Ellis, R.P., and T.G. McGuire (1990), "Optimal payment systems for health services", Journal of Health Economics 9(4):375–396.

Ellis, R.P., and T.G. McGuire (1996), "Hospital response to prospective payment: moral hazard, selection, and practice-style effects", Journal of Health Economics 15(3):257–277.

Fenn, P., N. Rickman and A. McGuire (1994), "Contracts and supply assurance in the UK health care market", Journal of Health Economics 13:125–144.

Fisher, W.H., R.C. Lindrooth, E.C. Norton and B. Dickey (1998), How Managed Care Organizations Develop: Selective Contracting Networks for Psychiatric Inpatient Care: A Case Study (University of Massachusetts Medical School, Department of Psychiatry).

Frank, R.G., and J.R. Lave (1989), "A comparison of hospital responses to reimbursement policies for Medicaid psychiatric patients", Rand Journal of Economics 20:88–102.

Freiman, M.P., R.P. Ellis and T.G. McGuire (1989), "Provider response to Medicare's PPS: reductions in length of stay for psychiatric patients treated in scatter beds", Inquiry 26:192–201.

Fudenberg, D., and D. Levine (1991), "An approximate folk theorem for repeated games with private information", Journal of Economic Theory 54:26–47.

Fudenberg, D., D. Levine and E. Maskin (1994), "The folk theorem with imperfect public information", Econometrica 62(5):997–1039.

Gerdtham, U.-G., and B. Jönsson (2000), "International comparisons of health expenditure: theory, data and econometric analysis", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 1.

Glass, D., and D.E.M. Sappington (1998), Cost shifting under Medicare's financial regulation of managed care operators (University of Florida).

Glazer, J., and T.G. McGuire (1994), "Payer competition and cost shifting in health care", Journal of Economics and Management Strategy 3(1):71–92.

Haas-Wilson, D. (1994), "The relationship between the dimensions of health care quality and price: the case of eye care", Medical Care 32(2):175–182.

Hart, O. (1995), Firms, Contracts, and Financial Structure (Clarendon Press, Oxford, UK).

Hart, O., and B. Holmström (1987), "The theory of contracts", in: T.F. Bewley, ed., Advances in Economic Theory: Fifth World Congress (Cambridge University Press, Cambridge, UK and New York, NY) Chapter 3, 71–155.

Hart, O., A. Shleifer and R.W. Vishny (1997), "The proper scope of government: theory and an application to prisons", Quarterly Journal of Economics 112(4):1127–1161.

Hibbard, J.H., and J.J. Jewett (1997), "Will quality report cards help consumers?", Health Affairs 16(3):218–228.

Hodgkin, D., and T.G. McGuire (1994), "Payment levels and hospital response to prospective payment", Journal of Health Economics 13:1–29.

Holmstrom, B., and P. Milgrom (1991), "Multitask principal-agent analyses: incentive contracts, asset ownership, and job design", Journal of Law, Economics, and Organization 7:24–52.

Joskow, P. (1980), "The effects of competition and regulation on hospital bed supply and the reservation quality of the hospital", Bell Journal of Economics 11(2):421–447.

Laffont, J.-J., and J. Tirole (1993), A Theory of Incentives in Procurement and Regulation (MIT Press, Cambridge, MA).

Lewis, T.R., and D.E.M. Sappington (1998), Using Subjective Risk Adjusting to Prevent Patient Dumping in the Health Care Industry (University of Florida).

Ma, C.-t.A. (1994), “Health care payment systems: cost and quality incentives”, Journal of Economics and Management Strategy 3(1):93–112.

Ma, C.-t.A. (1997), Cost and Quality Incentives in Health Care: Altruistic Providers (Boston University, Department of Economics).

Ma, C.-t.A. (1998), “Cost and quality incentives in health care: a reply”, Journal of Economics and Management Strategy 7:139–142.

Malcomson, J.M. (1999), The Specification of Diagnosis-Related Groups (Department of Economics, University of Oxford).

Manelli, A.M., and D.R. Vincent (1995), “Optimal procurement mechanisms”, Econometrica 63(3):591–620.

Manton, K.G., M.A. Woodbury, J.C. Vertrees and E. Stallard (1993), “Use of Medicare services before and after introduction of the prospective payment system”, Health Services Research 28:269–292.

McAfee, R.P., and J. McMillan (1987), “Auctions and bidding”, Journal of Economic Literature 25(2):699–738.

McClellan, M. (1994), “Why do hospital costs keep rising? A model of hospital production and optimal payment regulation”, Working Paper (NBER).

McClellan, M. (1995), “Uncertainty, health-care technologies, and health-care choices”, American Economic Review 85(2):38–44.

McClellan, M. (1997), “Hospital reimbursement incentives: an empirical analysis”, Journal of Economics and Management Strategy 6(1):91–128.

Mennemeyer, S.T., M.A. Morrisey and L.Z. Howard (1997), “Death and reputation: how consumers acted upon HCFA mortality information”, Inquiry 34(2):117–128.

Milgrom, P.R. (1987), “Auction theory”, in: T.F. Bewley, ed., Advances in Economic Theory – Fifth World Congress (Cambridge University Press, Cambridge, UK).

Miller, M.E., and M.B. Sulvetta (1995), “Growth in Medicare’s hospital outpatient care: implications for prospective payment”, Inquiry 32:155–163.

National Audit Office (1995), “Contracting for acute health care in England”, HC 261 Session 1994–95 (National Audit Office).

Newhouse, J.P. (1970), “Toward a theory of nonprofit institutions: an economic model of a hospital”, American Economic Review 60(1):64–74.

Newhouse, J.P. (1983), “Two prospective difficulties with prospective payment of hospitals, or, it’s better to be a resident than a patient with a complex problem”, Journal of Health Economics 2:269–274.

Newhouse, J.P. (1992), “Medical care costs: how much welfare loss?”, Journal of Economic Perspectives 6(3):3–21.

Newhouse, J.P., and D.J. Byrne (1988), “Did Medicare’s prospective payment cause lengths of stay to fall?”, Journal of Health Economics 7:413–426.

Norton, E.C. (2000), “Long term care”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 17.

Pope, G.C. (1989), “Hospital nonprice competition and Medicare reimbursement policy”, Journal of Health Economics 8(2):147–172.

Propper, C. (1996), “Market structure and prices: the responses of hospitals in the UK National Health Service to competition”, Journal of Public Economics 61(3):307–335.

Robinson, J.C., and C.S. Phibbs (1989), “An evaluation of Medicaid selective contracting in California”, Journal of Health Economics 8:437–455.

Rogerson, W.P. (1994), “Choice of treatment intensities by a nonprofit hospital under prospective pricing”, Journal of Economics and Management Strategy 3(1):7–51.

Roth, A.E. (1995), “Introduction to experimental economics”, in: J.H. Kagel and A.E. Roth, eds., Handbook of Experimental Economics (Princeton University Press, Princeton, NJ) Chapter 1, 3–109.

Salkever, D., and R. Frank (1996), “Economic issues in vaccine purchase agreements”, in: M.V. Pauly, C.A. Robinson, S.J. Sepe, M. Sing and M.K. Willian, eds., Supplying Vaccines: An Economic Analysis of Critical Issues (IOS Press, Amsterdam) Chapter 5, 133–150.

Schlesinger, M., R. Dorwart and R. Pulice (1986), “Competitive bidding and states’ purchase of services: the case of mental health care in Massachusetts”, Journal of Policy Analysis and Management 5:245–263.

Shleifer, A. (1985), “A theory of yardstick competition”, RAND Journal of Economics 16(3):319–327.

Sloan, F.A. (2000), “Not-for-profit ownership and hospital behavior”, in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 21.

Spence, A.M. (1975), “Monopoly, quality, and regulation”, Bell Journal of Economics 6:417–429.

Weisbrod, B.A. (1991), “The health care quadrilemma: an essay on technological change, insurance, quality of care, and cost containment”, Journal of Economic Literature 29(2):523–552.

Whynes, D.K. (1996), “Towards an evidence-based National Health Service”, Economic Journal 106(439):1702–1712.

AUTHOR INDEX

n indicates citation in footnote.

SUBJECT INDEX