VATT-JULKAISUJA
45
VATT PUBLICATIONS

FINANCE AND INCENTIVES OF THE HEALTH CARE SYSTEM

Proceedings of the 50th Anniversary Symposium of the Yrjö Jahnsson Foundation

Edited by

Antti Suvanto

Hannu Vartiainen

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Helsinki 2007

ISBN 978-951-561-677-7 (nid.)
ISBN 978-951-561-678-4 (PDF)

ISSN 0788-4990 (nid.)
ISSN 1795-3332 (PDF)

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Arkadiankatu 7, 00100 Helsinki, Finland

Gummerus Kirjapaino Oy
2007

Kansi: Markku Böök
Taitto: Nina Intonen

Foreword

The Yrjö Jahnsson Foundation was established in 1954. Mrs. Hilma Jahnsson, the spouse of the deceased professor of economics Yrjö Jahnsson, donated a generous amount of their joint estate as initial capital for the Foundation.

Initially the purpose of the Foundation was to promote Finnish research in economics, and medicine, and to maintain and support Finnish educational and research facilities in these fields. Research in these two disciplines remains the main target of the Foundation's funding. In 1974 the Board of the Foundation decided to extend the funding to health economics.

Health economics was a natural extension of the Foundation's activities given its established position in funding research in economics and medicine. In the middle of 1970s health economics was a new discipline in Finland. As a matter of fact the Yrjö Jahnsson Foundation imported the whole discipline to the country. The first symposium in health economics was organised in 1975. In 1977 the Foundation started to finance post graduate studies in health economics. In the succeeding ten years the Foundation sent one to three graduate students to the University of York to participate in the newly launched graduate programme in health economics. This York-trained group later formed the core of the health economics profession in Finland. Today the discipline is well established in the country. This volume comes from an international symposium on health economics that was arranged in Helsinki in August 9 to 8, 2004 to celebrate the 50th anniversary of the Foundation. As the title indicates, the aim of the conference was to discuss on the incentives and finance of health care system.

I want to express my gratitude to all the lecturers and commentators for interesting presentations, and also for dutifully contributing to this volume.

I also want to thank the chairman of the meeting, Dr. Antti Suvanto, as well as the staff of the Foundation, Dr. Hannu Vartiainen and Mrs Aila Palmu, for organizing the successful occasion, and for compiling this volume. Finally, I want to thank Prof. Aki Kangasharju, Prof. Jaakko Kiander, and Mr. Iikko B. Voipio of the Government Institute for Economic Research (VATT) for making it possible to publish this volume in VATT Publications Series

Helsinki November 14, 2006

Timo Laatunen

Chairman of the Board

Contents

Introduction: Is Health Different?

1

Antti Suvanto

Bank of Finland and Yrjö Jahnsson Foundation

Hannu Vartiainen

Yrjö Jahnsson Foundation

Health economics is a young sub-discipline of economics but, as Alan Maynard discusses in this volume, it has proven remarkably successful. Health is now one of the big issues in economics (for example, in the 2005 American Economic Association meeting, health or health care is the subject of four out of 24 invited sessions).

There are practical reasons for why health has become such a hot topic. Improved computational resources and statistical techniques together with remarkably rich data sets (especially in Nordic countries) have opened an exciting domain of health related questions that can be analyzed. This progress alone will occupy health economists for a long time.

Another reason for the growing interest in health is the increase – and foreseeable increase – in the demand for health services. Concrete new problems concerning the functioning of the health care system abound. Designing the health care system is, in principle, a standard microeconomic resource allocation problem, and many questions can be fruitfully approached from this angle. For example, research on hospital payment methods, on hospital competition and hospital mergers, or on physicians' compensation schemes could not have

done appropriately without the microeconomics machinery (see e.g. Newhouse, 1996; Krishnan 2001). An interesting recent example of direct application of economic theory to health service production is reported in Roth et al. (2006).

But, from this viewpoint, is health economics "just" an important subfield of microeconomics, one of many others (e.g. labour economics, banking and finance, law and economics,…)? As Finnish economist Jouko Paunio argued in 1975 in one of the first health economics symposia in Finland, health *is* different. The distinctive features of health sometimes call for non-standard thinking. In this introduction, we discuss some special characteristics of health and health economics.[1]

Measuring well-being

The key problems in economics are often conceptual, i.e. how to describe economic variables, how to model their relations, and how to compare outcomes. Economists often seek generality at the expense of concreteness. The aim is to "understand". One consequence of this tendency is the reliance on the *homo economicus* assumption: that economic agents are rational maximizers. While not many believe that the assumption is descriptively true, it is useful for modeling purposes. More importantly, it integrates different models together, and the models to data. Given the rationality assumption, preferences of individuals can be traced from their behavior. In fact, a hard-boiled economist believes that preferences are the only empirically tractable characteristic of a decision maker's mental activity.[2]

This is where health economics differs. A central task of a health economist is to serve as a social engineer: to back public decision makers with the relevant information and recommendations. Since some decisions have to be made and some recommendations have to be given, a health economist cannot afford the luxury of relying only on preference-related information. To evaluate the consequences of

1 We are looking at the issue from the angle of methodology of economics.
2 See e.g. Gul and Pesendorfer (2005).

different interventions, a health economist has to form a view of the individual *value* of health.

The central feature of the value concept of health is that it has to be *comparable* across individuals. Otherwise it would not give a rationale for allocating resources to individuals with infinite needs. Economists feel nervous about the idea of comparing utilities.[3] The economists' way to compare policies is to rely on the opportunity costs; how much of a common denominator – money – an individual is willing to sacrifice for a good or service. Since the willingness-to-pay is evaluated against the market data, this method would not work well in health markets. Nothing guarantees that the market data reveals relevant information concerning needs for health services as the needs need not match with financial resources. Thus the value of health has to be measured *directly*.

Measuring the value of health is not only a problem for deciding how to allocate resources efficiently through public policies but, as discussed by Robinson and Megerlin in this volume, it also contributes to the functioning of markets. In particular, it affects how private contractual relationships can be designed. Without a general and transparent measure it is impossible to verify the extent to which health services are provided. And without verifiable criteria, contracts cannot be drawn.

Measuring the value of health requires one to impose a concrete structure on the sources of well-being. Various health-related-quality-of-life (HRQoL) measures have been proposed that are composed of physical, mental and social aspects of well-being (see Brommels and Sintonen, 2001). While the HRQoL criteria are not derived from preferences, as a hard-boiled economist would demand, they are for the most part consistent with psychological evidence. Under the conjecture that this evidence is telling, the measures do provide valuable information of aspects of well-being.

3 In the tradition of logical positivism, they feel that only ordinal, not cardinal, measures of payoffs can be meaningfully elicited via choices. However, this view has recently been challenged by the happiness literature (see e.g. Frey, 2002).

Perhaps the more important feature of a HRQoL measure than its descriptive accuracy is its usability. Through a HRQoL measure health related well-being becomes an operationalizable quantity, against which one can base policy evaluations. Cost-benefit analysis is the cornerstone of the contemporary health economics research.

Thus one difference between mainstream economics and health economics is that where the former is skeptical towards any specific notion of well-being and wants to avoid analyzing it in concrete terms,[4] the latter addresses precisely the question of *what* the benefits could be. Thus health economists make questions that are *qualitatively* different from the mainstream economics.

Paternalism

Health economics does not typically see markets as a reliable steering mechanism of health services. Functioning of the market has to be improved by public interventions. But public interventions are not only targeted against the traditional market failures. Their aim is also to steer individual behavior to the direction that is good for *themselves*. This often means limiting or affecting individual's choice set. In this sense a clear *paternalistic* tone is characteristic to health economics in many of its policy recommendations.[5]

The issue of paternalism arises when the public authority enhances individuals to maximize, say, their life expectancy. While it is easy to argue why banning or taxing addictive substances, reducing adult obesity, or providing citizens with a proper public health insurance is a natural thing to do by a responsible public authority, these measures are effective precisely because they reduce the choices that are available to the individuals (at least through their budget set). Any such measures can therefore be interpreted, in varying degree, paternalistic.[6]

4 If he does assume a specific utility function etc., he implicitly means any neighbouring function.
5 Paternalism = Interfering with a person's freedom for his or her own good.
6 There may be additional reasons for intervention, e.g. externalities.

From the viewpoint of mainstream economics, purely paternalistic interventions could never be welfare improving. Given the available choices, a rational actor always chooses the optimal decision, and what an outsider can do is at most as good for him but never better. In the rational actor framework, only coordination failures of some sort can justify a policy intervention. Conversely, if an intervention is justified in without a coordination failure, then there has to be problems with the individuals' decision-making ability. This seems be the underlying rationale in many of health related policy interventions.

One reason for why the rational choice model may have only limited use in health economics could concern the nature of uncertainty in health related decision-making. Many people have very limited capacity to deal with information concerning their health status, to the contrary of, say, their financial status. New evidence, technologies, and views about the effectiveness of health procedures keep arising with a speed that make it difficult for even professionals to keep up with. To make a good decision, an individual should nevertheless evaluate correctly what the health signals mean but also what it would mean *not* to observe them. It is not surprising that individuals in health market often follow decision procedures that look suboptimal to an outside observer (see McCall, 1996). When one cannot count on individuals making wise decisions concerning themselves, the standard economics approach is no longer valid

Public intervention such as education may improve the individuals' decision-making ability. But it may not help always. If the underlying problem is in the individuals' capacity to process information, then education may not help and a need for a paternalistic intervention would be permanent. Hence, if one believes paternalistic policies to improve well-being, there is no reason to expect the justification for them to go away in the long run.

Another motivation for a paternalistic intervention in health markets stems from the individuals' inability to commit to long-term goals.

Self-control problems such as addiction or obesity are difficult to capture in the rational-choice framework.[7] The problem with an addict is that a monetary transfer to the addict, which would enlarge his choice set, may not improve his well-being in the long run whereas a rehabilitation program of the same value might. Thus restricting the choice set may be beneficial in the presence of self-control problems.

Many health policies seem to have a paternalistic component. This is the case whenever the effectiveness of an intervention is evaluated on other grounds than what is communicated by individuals through their actions, i.e. on *ad hoc* grounds. However, given the potential problems with individuals' decision-making ability, such approach can be justified. But this requires one to believe in a model that is distinct from the standard homo economicus model of economics.

Human interaction

Problems in understanding the value of treatments emphasize the role of health professionals. One thing that makes the health production process special is the physician-patient relationship. This was observed already by Arrow (1963), who motivates health institutions through the principal-agent problem between a physician, a patient, and an insurer.

However, if patients or physicians(!) fail to fully understand the situation, it is questionable whether the principal-agent framework does full justice to the physician-patient relationship.[8] While human interaction seems important in health related decision-making, it is difficult to tackle in the standard economics framework. Let us consider some examples.

7 For a classical attempt to understand addiction or self-harmful behaviour in a rational choice framework, see Becker and Murphy (1988). See also the recent literature on time inconsistent preferences.

8 For a recent attempt to capture essential features of physician -patient relationship, see Koszegi (2001).

There are typically many recommendations a physician may justifiably give to a patient from the same verifiable data. Because of this, the physician cannot avoid deciding of how much to reveal and which way to manipulate the patient. To cooperate with the physician, the patient must trust that the physician uses his authority in favor of him. Therefore, trust plays a central role in a well functioning physician-patient relationship. Since altruistic people are presumably more trustworthy than others, altruism has been thought as an important character of physicians (see Ma in this volume). But trust without substance – e.g. possibility to punish or reward – has no meaning in economics. Thus it seems that to analyze physician-patient relationship, one must go beyond the standard economics framework.

While it is easy to see why a patient departures from the rational man paradigm, physicians are not perfect either. They take actions on the basis of rules of thumb, stereotypes (e.g., believe that certain kind of patients do not comply with recommendations), common practice, the need to "do something" (the physician feels he cannot send the patient home without even a worthless recommendation), the patient's expectations (e.g., prescribe antibiotics just because it is what the patients asked). From the perspective of economics, it is far from clear how one should organize the incentives of a boundedly rational physician?

A good example of why the standard economics approach is insufficient is the "treatment effect". It is easy to figure out reasons why it not only matters *what* the treatment is but also *how* it is provided. Anxiety, procrastination, suspiciousness, and denial are common features of patient's decision making.[9] How and when such states of mind show up may be sensitive to the treatment strategy. More generally, the success of human interaction can be sensitive to the motives, or suspected motives, of individuals.

Human interaction is hard to capture in a standard economics framework. One of the key principles in economic modeling is

9 See Koszegi (2003).

that only the outcomes matter. This rules out all kinds of treatment effects. Of course one could try to enlarge the model to capture also the unmodeled payoff consequences. However, it is safe to argue that there is not rich enough model that could ever approximate all potentially relevant features of a physician-patient relationship. It is possible the correctness of an elaborate physician-patient model may be as sensitive to unmodeled features as a sketchy model. Hence keeping the model simple when analyzing health care markets, as is customary in health economics, may actually be without much loss of generality.

Concluding remarks

To summarize, health economics is an important and growing field, getting its inspiration from real world health problems that also contain an economic dimension. Much of health economics can hence be categorized as "applied economics". Nevertheless, since economics should be interpreted as a language rather than a domain, and the language is based on the paradigm of rationality, the questions these two disciplines make are often qualitatively different. For example, measuring well-being, which much of the normative analysis in health economics relies on, is quite unique to this discipline.

All in all, health economics has value precisely because it gives practical solutions on concrete problems. Keeping the bridge between data and recommendations narrow and transparent has its virtues.

This volume

This volume, coming from a symposium organized by the Yrjö Jahnsson Foudation in August 2004, consists of discussion of leading health economists. The theme of the symposium was Incentives and Finance of the Health Care System.

Alan Maynard gives an overview of the health economics. Health economics has developed rapidly and impressively over the last 50 years. He argues that part of its success has been the development of

techniques of particular interest to physicians. The theory and policy challenges now are to ensure that evidence is translated into practice, and that further evidence of the cost effectiveness of competing therapies is produced efficiently. This requires investment in the systematic and careful testing of performance information systems (particularly the measurement of outcomes) and innovative use of mixed systems of incentives. The failure to meet this challenge in recent decades has condemned some citizens to avoidable mortality and morbidity.

James Robinson and *Francis Megerlin* examine innovations in physician payment systems in the United States and France. They point out that the primary social goal with respect to payment for physician services has been to promote patient access to care and the improvement of quality through new clinical interventions, and that fee-for-service compensation mechanism directly supported the achievement of these goals. However, as social concerns shifted from to controlling costs, fee-for-service lost its luster and found itself blamed for many of the system's ills. Experiments with capitation, which rewards the physician based on the outputs rather than on the inputs, transferred excessive risk to physicians and thereby created incentives for under-treatment and the avoidance of particularly sick patients. The contemporary hybrid method of physician compensation seeks to incorporate elements of both prospective and retrospective payment. Nevertheless the authors point out that while the new pay-for-performance initiatives together with technological innovations offer insights into the manner by which incentives influence behavior, they are limited by the fundamental difficulties in measuring performance. The quantity of inputs is poorly linked to the quality of outputs, and patient-derived measures of satisfaction and functional ability are only weakly related to specific interventions by specific physicians. Also, the organizational challenges facing physician payment reforms stem from the multi-agent, multi-task nature of medicine, especially for patients with chronic conditions, are deep and difficult to overcome. Groups must develop internal mechanisms for linking group-level payments to individual physicians or forgo the incentive benefits of pay-for-performance. Finallym the authors

speculate on the predictions that can be made with respect to the trajectory of physician payment methods in the future.

Ching-To Albert Ma studies the effect of interaction between the public and private sectors on physician incentives. Providers of health care in public and private sectors are often subject to different incentives. While the market mechanism is expected to work in the private market, the public system is typically characterized by simple and low-powered incentives. Ma notes that a conventional model that relies on profit maximizing preferences for economic agents is inappropriate for the public sector. Such a model would predict uniformly poor service quality and work effort there, which is inconsistent with empirical observations. Ma then assumes that physicians are not just simple maximizers of their own material payoffs but possess a degree of altruism. The key hypothesis is that some healthcare providers are either sincere or have altruistic preferences. Ma asks who among the heterogenous physicians will work in the public system, who in the private market? His basic conclusion is that public policy should steer sincere and altruistic physicians to work in the public system, while the marketplace disciplines those who seek personal gains. It is a straightforward conclusion due to the use of low-powered incentives in the public system. Allowing the self-interest seeking providers the opportunity of higher profits in the private sector actually alleviates the inefficiency in the public sector.

Robert Evans focuses on the conflict of interest stemming from the incidence of health costs. Conflict of interest, in turn, is a cause for providing ideologically motivated information concerning the functioning of the health sector. In this conflict, economists have no legitimate role in pretending to offer objective, professionally-based determinations of "optimal" arrangements. Economists should play a more useful role by providing assessments of the likely consequences of different policies and their distributional impacts – and unmasking the distributional agenda lying behind the proposals for reform. Careful economic analysis can be very powerful for this purpose, so long as it is grounded in the institutional and behavioural realities of working health care systems. As an example of crippled analyses

that serves merely as a part of the propaganda of those promoting a regressivity agenda, Evans sees the "welfare burden" literature advanced e.g. by Arrow (1973). In Evans' opinion, this literature is incapable of formulating the relevant questions that drive the real-world debates over health care policy. He argues that any analysis of health care financing that ignore distributional considerations does not contribute the debate.

William Hsiao analyzes health care financing problems in the western countries. There clearly is no common agreement between different countries on the best finance method. Health care financing involves trade-offs between equity and efficiency. Often cited inefficiencies include the tax payment distortions and moral hazard produced from health insurance. But health care financing influences both equity and cost containment, and the influence on the latter is often overlooked. Cost containment effects inherent in any health financing method influence efficiency of a national health system because they impact allocative and productive efficiency and thus the total amount of resources one has to spend for health. Consequently, certain financing methods may have superior features that enhance both equity and efficiency that can minimize the trade-off between them. The focus of this paper is to analyze the different degrees of trade-off of several major financing methods.

References:

Arrow, K. (1963), Uncertainty and the Welfare Economics of Medical Care, American Economic Review, 53, 941-73.

Becker,G. and K. Murphy (1988), A Theory of Rational Addiction, Journal of Political Economy, 96 no. 4.

Brommels, M. and H. Sintonen (2001), Health related Quality of Life in Chlinical Strudies, Annals of Medicine, 33.

Frey, B. and A. Stutzer, (2002). What Can Economists Learn from Happiness Research?, Journal of Economic Literature, American Economic Association, 40, 402-435.

Paunio, J. (1975), Terveyden taloustiedettä – alustavia näkökohtia, Kansantaloudellinen aikakauskirja 4, 344-50.

McCall T.B. (1996) Examining Your Doctor: A Patient's Guide to Avoiding Harmful Medical Care, Carol Publishing Corporation.

Newhouse, J., (1996), Reimbursing health plans and health providers: Efficiency in production versus selection, Journal of Economic Literature, 34, 1236-63.

Koszegi, B. (2005), Emotional Agency: The Case of the Doctor-Patient Relationship, manuscript, UC Berkeley.

Koszegi, B. (2003) Health Anxiety and Patient Behavior, Journal of Health Economics, 22, 1073-84.

Krishnan, R. (2001), Market restructuring and pricing in the hospital industry, Journal of Health Economics, 20, 213-237.

Roth, A., Sönmetz, T. and U. Unver (2006), Efficient Kidney Exchange: Coincidence of Wants in a Structured Market, Journal of Political Economy.

Health Economics in the Past, the Present and the Future

2

Alan Maynard

Department of Health Sciences, University of York*

2.1 Introduction

"In this world nothing can be said to be certain, except death and taxes"

Benjamin Franklin (1789)

Every citizen has a terminal, sexually transmitted disease called life, which they generally value highly and seek to protect. The purpose of health economics is to protect individual citizens by reducing death and taxes. The former can be achieved if scarce health resources are targeted efficiently to produce health i.e. increases in the length and quality of life. The latter can be achieved if such efficiency results in the removal of inefficient old procedures and the prevention of the adoption of inefficient new interventions, even if these are championed by powerful commercial and professional interest groups who are not focused on the evidence base about the clinical and cost effectiveness of therapeutic and diagnostic interventions that are competing for funding. Health economics is the optimistic component of the dismal science of economics!

Health care is a large multi-faceted industry, which has grown rapidly in the last half century, often consuming 10 percent of the Gross Domestic Product of many West European and North American

*Area 2, Seebohm Rowntree Building, York, YO10 5DD, England.

countries. The original architects of public health care systems, such as Bismarck in Germany in the nineteenth century and Lloyd George in the early twentieth century in Britain, had three clear goals in their reforms:

I. to improve the supply of fit young men for the armed forces

II. to improve the supply of fit young workers for industry

III. to protect the middle classes from infectious diseases carried by the working poor.

These equity and efficiency goals continue to dominate health care policy in middle and low income countries. For instance, after 15 years of rapid (on average 9 percent per annum) real economic growth in China poverty has increased and the income differential between the urban and rural workers is 3.1. The collapse of the Maoist system of public health and rural health care (e.g. barefoot doctors) has led to a situation where price barriers reduce utilisation, leaving clinics and hospitals underused, has led to growth in infectious diseases (TB, SARS and HIV-AIDS) that now threaten the economic and political stability of the country.

Despite the dominance of basic Bismarckian objectives in such countries today, equity issues in richer countries have tended to be submerged by policy concerns about microeconomic efficiency and macro-economic expenditure control. The work of health economists reflects this policy interest, although a strong strand of research activity continues in equity (e.g. see Jones and van Doorslaer (2004)), even if its prominence in policy is more often rhetorical than real in some countries.

After a discussion of the nature of health economics (Section 1), a brief review of the sub-discipline in the past and present will be presented. There is then a discussion of the major policy problems

facing health care systems at present and how these might be resolved by the application of economics in the future.

2.2 The Nature of Health Economics

Williams illustrates the nature of health economics nicely in his "plumbing" diagram (figure 1) (Williams (1987)). This depicts the subject matter in a series of boxes, each linked in different ways by a variety of 'pipes' that show how parts of the sub-discipline feed into and affect each other. Each of the boxes represents one of the principal subject areas of the sub-discipline. Boxes A, B, C and D are the core elements of health economics, with the other 4 boxes (E, F, G and H) being areas in which the core elements are applied.

Box A is concerned with the health production function i.e. what influences the creation of health for the individual and populations, and what are the relative contributions of income, wealth, education, genetic endowment, leisure and work clinics and other behaviours? This sphere of activity is closely associated with the human capital school of Gary Becker as applied to health particularly by researchers such as Michael Grossman (Grossman (1972)). It analyses how initial endowments of health capital can be augmented by investment and how it depreciates over the life cycle, and produces a stream of discounted benefits or 'healthy years'.

Box B is a core issue: health. What is health? And what is its value? This involves study of the perceived attributes of health, health status indexes, the valuation of life and utility scaling of health. It is a subject area populated by psychologists, sociologists, epidemiologists, operational researchers and economists, who sometimes work collaboratively! This is an area where value judgements need to be explicit and dealt with carefully and where, although there have been substantive methodological advances (e.g. health related quality of life measures such as Short Form 36 (www.sf36.org) and EQ5D (www.euroqol.org) and they have been translated into dozens of languages, are quite well validated and have

used extensively in thousands of clinical trials, routine measurement and management of health related quality of life remains absent in all health care systems. This failure to measure "success" in medicine makes evaluation of health care reform universally incomplete (Kind and Williams (2000), Appleby and Devlin (2004)).

Box C is focused on the demand for health care, which is a derived demand exercised in the expectation by patients that health care will increase their health and produce more healthy years. What are the determinants of patients' health seeking behaviour and how are these modulated by the agency relationships and providers' (especially doctors') capacity to both reinterpret and increase demand? How do barriers to the utilisation of health care such as time (travel distance) prices and financial prices affect demand e.g. what are the price, income and cross elasticity's of patient demand and how do these health care behaviours affect health? This is the area where externalities (e.g. the infectious disease threat from the poor to the middle classes) are evident and where the arguments between those concerned with 'need' debate their values with those concerned with treating health care as a normal economic good and analysing it simply in the content of neoclassical economics.

Box D covers a wide range of supply side issues ranging from the behaviour of provider institutions, the needs of client groups and the nature of public, private and informal suppliers. This includes the nature of hospital and other provider groups (e.g. primary care practices) production functions, "skill mix" and factor substitution, incentive systems for labour groups and organisations and the impact of trust and regulation on the performance of institutions (e.g. the pharmaceutical industry) and labour groups of providers. Perhaps the biggest inherent policy challenge in this box is the coordination of fragmented systems of funding and provision to ensure the supply of integrated care of major client groups such as the elderly and the mentally ill who may experience chronic and multiple diseases.

Box E is concerned with the micro evaluation of health care in terms of screening diagnostics, therapeutics, rehabilitation and

alternative modes of delivering care. An essential ingredient into evaluative studies and modelling is the evidence base about clinical effectiveness as revealed by systematic reviews and the web pages of the Cochrane Collaboration (www.cochrane.org). The techniques of cost effectiveness and cost utility analysis have colonised clinical trials practice, as regulators increasingly require economic data to inform reimbursement choices about new and old technologies. This practice of adding economic components to stage two clinical trails may yield useful economic and clinical information that can be used in Bayesian modelling in stage three trials. This shift from established techniques of economic evaluation to Bayesian modelling is a source of continuing debate amongst practitioners.

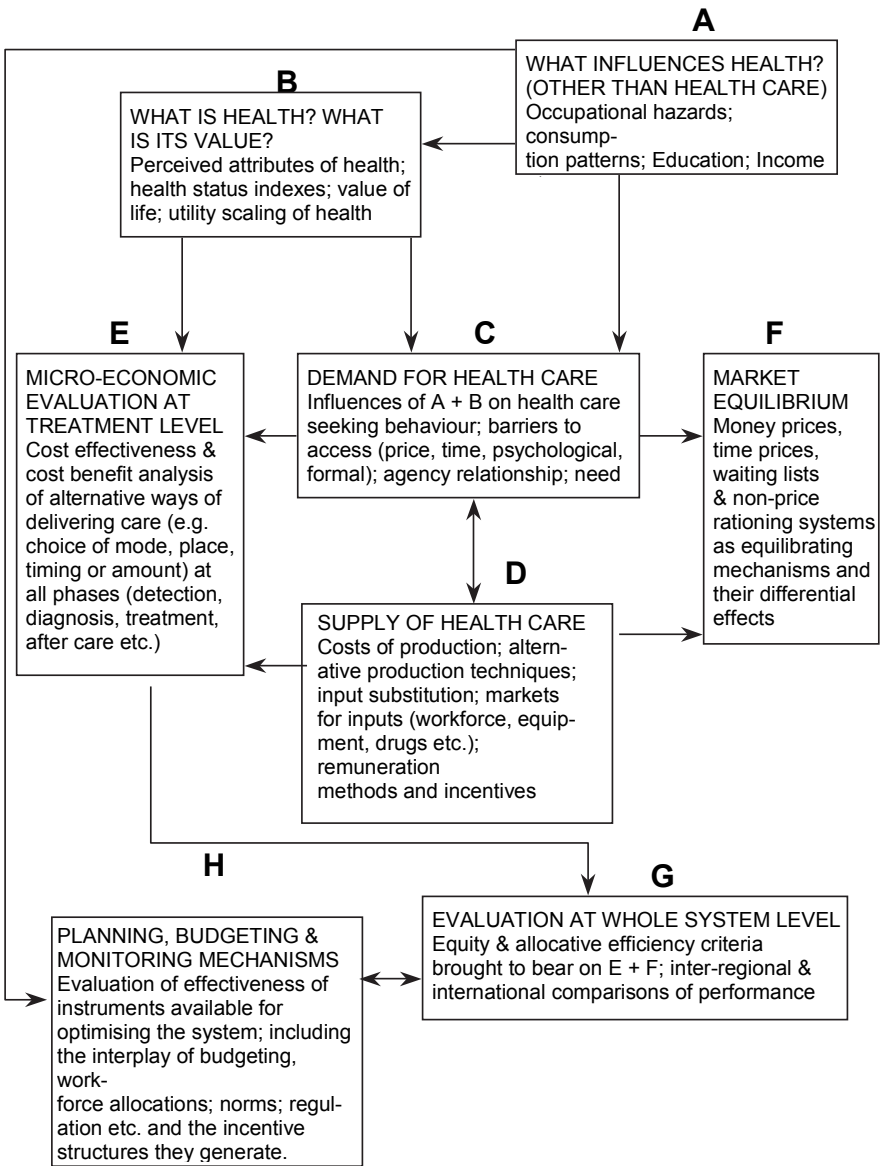
The subject matter of **Box F** addresses the issue of market equilibrium i.e. how markets "clear" and equilibrate the demand for health care with its supply. Public and private institutions "clear" markets in different ways e.g. the former may use time rationing and waiting lists, and the latter is more dependent on money prices and restrictions on benefit packages. Policy analysis of such mechanisms combines positive (what is?) issues with normative (what ought to be) statements e.g. time rationing equilibrates supply and demand, and waiters could be prioritised on the basis of normative assessments of relative need as practiced in New Zealand in the recent past.

The subject matter of **Box G** is the high level evaluation of the systems that create the input health care and the outcome, improved individual health. A significant characteristic at the micro level and the macro level is the variation in the distribution and use of resources. Inter-regional and international variations in expenditure, activity and outcomes challenge practitioners in terms of methodology (are you comparing like with like?) and also in terms of the policy lessons, if any, that can be learnt and to inform reform processes.

The final **Box H**, is concerned with the study of planning, budgeting and monitoring mechanisms and their impact on both equity and efficiency. How can systems be better optimised by the coordination of market incentives, planning and regulatory mechanisms, and

workforce planning? The erosion of trust in the medical profession, not so much amongst patients but amongst purchasers, creates a nice dilemma. If professionals in the health care market cannot be trusted to pursue efficiency on the basis of duty and moral obligation, what other incentives, financial and non-financial, can be used to create economising behaviour? (Maynard and Bloor 2003 (a)). Would giving doctors budgets, with some property rights in surpluses generated create greater efficiency in health care? The lessons from US managed care and UK experiments with general practice fund-holding and innovations in payment systems appear to support the use of such incentives (Dushieko et al (2003), Maynard and Bloor (2003 (b)).

Figure 1. The Willams' 'Plumbing' Diagram of Health Economics



The interactions between the eight boxes are shown in figure 1 by the arrowed relationships. For instance, the core issue of what is health and what is its value (Box B) is a central input to boxes C (the demand for health care), E (microeconomic evaluation), and H (planning and budgeting). Each of the boxes involves consideration of positive and normative issues, with ideology, and preferences for public or private action, permeating debate often in subtle ways.

2.3 From the Early Days to the Present

It is now over 50 years since the publication of a section of the papers and proceedings of the annual American Economic Association meeting on the "Economics of Medical Care" (American Economic Association (1951)). Some of the authors of papers in this collection, for instance, Eli Ginzberg, Herb Klarman, Seymour Harris and Rothenberg, went on to publish further substantial contributions to the development of the sub-discipline.

Since its inception in the 1950s, health economics has been a remarkably successful part of the economics discipline. In the 1960s the growing interests in public economics and the economics of human resources led to the development of sub-disciplines such as the economics of health and the economics of education. Yet while the former flourished to become an "industry" in its own right, the latter has developed more slowly. The causes of this differential success, for instance, individual inspirational intellectual leadership by people such as Joe Newhouse and the group around him who conducted the largest randomised controlled trial in history, the Rand Insurance Experiment (Newhouse (1993)) and Alan Williams (e.g. see Culyer and Maynard (1997)) together with the sub-discipline's capture of influential interest groups such as Government policy makers and civil servants, the pharmaceutical industry and the medical profession, makes a nice subject for research in itself! The characteristics of health economists have included not merely a willingness to apply economic theory innovatively to the subjects of health and health care, but also to engage directly with decision makers in advisory

roles (e.g. Abel-Smith in the UK in the late 1970s was adviser to Ministers of Health), consultancies (the source of some affluence to those working with the pharmaceutical industry!) and public debates (e.g. in the media).

The 'birth' of health economics is often defined as 1963 when Arrow's article on the "Welfare economics of medical care" was published (Arrow (1963)). This remarkable paper always rewards re-reading for its insightfulness and comprehensiveness. However others were contributing before this. For instance, Mushkin (1958) defined the task of health economists as "appraising the efficiency of the organisation of health services, and to suggest ways of improving this organisation". Mushkin (1964) also edited the proceedings of a conference on the economics of medical care and noted that the central issue was the allocation of resources at a time when health care costs were rising, the Government's role increasing and medical developments were increasing. Themes that are very familiar today!

In Britain around the same time, the initial innovators were more multi-disciplinary with generalist funders supporting the work of economists who made pioneering initial contributions to the sub-discipline (e.g. Dennis Lees (e.g. Lees (1962)) and Mike Cooper (undated)). Their involvement with Gordon McLachlan at the Nuffield Provincial Hospital Trust (see e.g. McLachlan (1964), the first of a series of nine volumes up to 1974 reporting Nuffield research) and George Teeling-Smith at the Office of Health Economics (e.g. OHE (1962 -) helped lay the foundations of health economics in the UK. This work was also complemented by the right wing policy 'think-tank', the Institute of Economic Affairs (IEA), whose authors supported competitive markets as an alternative to the NHS (e.g. Lees 1966).

The York economics group developed this initial policy orientated work with vigour. Alan Peacock and Jack Wiseman, who both researched and taught in public finance and the emerging sub-

discipline of public choice, created the Department of Economics at York in 1962. Their York colleague, Alan Williams was seconded from York to the UK Treasury working with, as a consequence, Archie Cochrane and other radical and innovative physicians in the Ministry of Health. This fired his interest in health economics, supported by coincidence, by the recruitment to York of Culyer and Maynard (both whom Williams taught) in the early 1970s.

The seminal contributions of Williams both to the health economics literature and to the development of the sub-discipline are remarkable. His contacts from his Government days helped him compete for funding not only for am ambitious research programme that grew into the Centre for Health Economics in 1983 but also for training, with the creation of the Graduate Programme in Health Economics at York in 1977.

Even before the creation of this programme, Williams links with Scandinavia where he worked on a PhD about the work of a Swedish public finance economist) had led him into contact with Finnish colleagues and the arrival of graduate students supported by the Yrjő Jahnsson Foundation. This exchange continues and has created a remarkable cadre of York trained health economists in Finland.

Indicators of the increasing 'maturity' of the sub-discipline were the emergence of bibliographies and text books in the late 1970s. The bibliographies of innovators such as Culyer et al (1977), Griffith et al (1980) and Blades et al (1986) have evolved into a variety of electronic databases e.g. the NHS economic evaluation database at www.york.ac.uk/inst/crd and the Health Economics Evaluations Database (HEED) available on CD-Rom from the Office of Health Economics, London. The first three textbooks in health economics (Newhouse (1978), Cullis and West (1979) and Evans (1984)) have been followed by steady flow of new authors over the last 25 years.

The two principle journals in the sub-discipline are the Journal of Health Economics (JHE was established 1981) and Health Economics (HEC was established in 1992). An analysis of the articles published

in these two journals to mid-1999 as classified in the William's diagram (figure 1) is reproduced in table 1. There is no reason why there should be a balance of activity across these groups and as can be seen the largest category is D, the supply of health care. As shown in the original paper (Maynard and Kanavos (2000)), this is a product of a higher publication rate on the supply of health care in the JHE, reflecting US interests. There was a higher level of publication in microeconomic evaluation in HEC, reflecting European interests.

Table 1. The Activities of Health Economists

	Category	Number of Articles	Articles as percentage of total
A	What influences health?	84	11.41
B	What is health? What is its value?	57	7.74
C	Demand for health care	96	13.04
D	Supply of health care	152	20.65
E	Microeconomic evaluation	108	14.67
F	Market equilibrium	57	7.74
G	Evaluation of the whole system	76	10.33
H	Planning, budgeting and monitoring mechanisms	90	12.33
	Overview	16	2.17
		736	100.00

In citation indices JHE and HEC both score highly, usually occupying positions in the top dozen of both the economics and health care science listings. The former achievement is increasingly attracting "mainstream" economists to publish applied and methodological work in these journals.

The volume of studies reviewed in table 1, 736, is only part of a literature explosion in health economics. New journals have entered the market (e.g. European Journal of Health Economics and Applied Health Economics and Policy). Practitioners also publish in the core journals of economics and extensively in medical journals, both generalist (e.g. British Medical Journal and the Journal of the American Medical Association) and specialist.

Thus the achievements of health economists during the last half century have been quite spectacular in terms of creating a demand for their services in the public and private sectors and meeting that demand by creating a professional group significant in its contributions to both methodological development and the application of economic methods to clinical problems and policy choices. However the remaining challenges are significant. How can health economists as practitioners of the 'optimistic science' contribute more to the reduction death and taxes?

2.4 Health Economics in the Future

As well as the spectacular achievements in the development of health economics in the last 50 years, there have been curious failures. These problems will challenge the profession in the future.

2.4.1 Health care challenges

Despite increased levels of investment in health services research and health economics, health care delivery is characterised by well evidence problems that have been known and unresolved for decades.

To vary is normal but how much variation is efficient?

In every health care system there is evidence of variation in medical activity and practice. Perhaps the longest and best known tradition of research in this area is associated with Jack Wennberg and his US colleagues at the Dartmouth College. Wennberg's analysis of small area variation in New Haven and Boston concluded nearly 20 years ago that Boston did more to its patients, spent twice as much per capita but had similar mortality outcomes. Wennberg's US work has been replicated in Europe, showing similar variations, and he has produced an atlas, mapping variations across the USA (Wennberg et al (1987), (1989)).

More recently Wennberg's colleagues have charted the national variations in medical practice, finding that US Medicare per capita spending on 2000 was $10,550 per enrolee in Manhatten and $4,823 in Portland, Oregon. The east is expensive – west is cheap divide in this Federal programme of health care is due to differences in the volume of care delivered and not due to differences in illness socio-economic status or the price of services (Fisher et al (2002)).

Fisher concluded, "Residents in high spending regions receive 60 percent more care but do not have lower mortality rates, better functional status or better patient satisfaction". Separately he noted that there were potential savings of 30 percent of Medicare expenditure if high spenders reduce expenditure and provide the safe practices of conservative treatment regions (Fisher 2003).

Whilst economists know of such variations from the health services research literature (recently reviewed nicely in a web exclusive of Health Affairs (www.healthaffairs.org, October 7th, 2004), their engagement with their mitigation has been marginal. Charles Phelps, in an editorial responding to the work of Fisher and his colleagues emphasised the significance of this work and the need for greater understanding of it.

i. To err is human

It is normal for human beings to make mistakes, but in health care errors can create avoidable morbidity and mortality. The measurement of error rates is poor in most health care systems. Small local studies in the United States show errors to be 3–5 percent of hospital admissions. Extrapolating these limited studies to the national level creates estimates that errors kill 44,000 to 98,000 Americans each year.

Estimation of error rates in England and Australia show even higher rates (e.g. Vincent et al 2001, (Wilson et al (1995)). As in the USA, these errors are related to mistakes in medication (wrong drug administered and/or wrong dose), in surgery and in staff hygiene: hospital acquired infections from failure to wash hands creates substantial financial burdens on health care systems as well as creating avoidable morbidity and mortality for patients (Pittet et al (2000)).

This literature raises a range of nice economic issues. For instance, what is the optimal or efficient level of errors? With the major users of hospital and primary care being elderly, error prevention may only have a marginal benefit in times of delaying time to death. For instance an American study estimated that many of those upon whom medical errors were inflated would have been dead within 3 months (Hayward et al (2001)). The related issue is the evidence base about the cost effectiveness of competing methods of reducing errors. This is extremely limited. Despite this poor evidence base, the "quality industry" is growing and investing heavily often in interventions where there is no evidence of cost effectiveness.

Remedies for medical errors may involve improved staff practices and alternatives in staff ratios. For instance there is a literature that concludes that patient outcomes (measured in terms of mortality) is determined not only by the volume of nurse staffing but also by its grade (e.g. Aiken (2003) and Needleman (2002)). These authors have concluded from their cross section analysis of administrative data

about nurse staffing levels that increased levels of nurse staffing and graduate nurses give better patient outcomes. More recent time series analysis of such data shows an effect, but one that is weaker than the one asserted by Aiken and Needleman. (Mark et al (2004)) and in a systematic review of the literature, Lancashear et al (forthcoming 2005) concluded that the relationship may be curvilinear, exhibiting diminishing returns.

Such work raises some nice methodological issues in particular the appropriateness of multivariate and multi-level estimation techniques and the strengths and weaknesses of such methods when using cross section and time series administrative data in estimation linear and non-linear relationships.

The policy issues are also of great interest. The Aiken literature has led to California legislating minimum nurse-bed staffing ratios in an effort to reduce inpatient mortality. Since January 2004 that ratio has been 1 to 6, and from January 2008 it is to be 1 to 5 (after being deferred from implementation in 2005). This compares with, for instance, a ratio of 1.11 in England and widely varying ratios in the rest of the European Union.

Another policy derivative of the error problem has been the adoption of techniques of engineering safety, in particular "six sigma" management. This involves focusing managerial effort on the tails of distribution (i.e. 3 standard deviations above and below the mean) to correct inadequate performance. Such techniques can be very helpful, but only if the evidence base on inefficient practice is good and it is cost effective to remedy such deviance.

The nice issue arising from the errors literature and its relation to staffing is the issue of the relative cost effectiveness of investments in labour and capital to reduce such failures. Would investment in computerising and "robotising" pharmacy reduce medication errors more than investing in the number and quality of pharmacy staff? Does investing in nursing increase health more than investing in hospital specialists or generalist physicians (GPs)? Some argue

that investment in physician specialists has now reached "flat of the cure" or diminishing returns, but there is some uneven evidence that investments in GPs may enhance population health effectively.

ii. Time to measure success?

The literature on medical practice variations and errors focuses on the failures of physicians and health care systems i.e. it measures mortality, hospital readmission and error levels. Cochrane decades ago was rightly sceptical about the impact of medical care on population health (Cochrane (1972)) and likened health care systems to crematoriums i.e. these were high levels of "input" but relatively little "output"! However now there is better evidence that some interventions are both clinically and cost effective.

Often such interventions are cheap, well known but not delivered efficiently to potential beneficiaries e.g. a recent estimate by the Rand Corporation showed that only 55 percent of Americans had their routine health needs met by the US public and private health care systems (The American health care systems (e.g. private (managed care) insurance and Federal and State programmes such as Medicare) like their counterparts in Europe and the rest of the world fail to register, monitor and treat chronic illness such as hypertension, asthma, diabetes, cholesterol and heart disease, thereby condemning their citizens to high levels and avoidable morbidity and mortality (Nolte and McKee (2004), Maynard, McKee and Nolte (2004) and Kerr et al (2004)).

The emphasis on the failures of the health care systems is curious, in particular as for over 30 years health economists have advocated the application of measures of the success of health care. President Nixon funded the Rand Health Insurance Experiment. This was one of the (if not the) largest randomised controlled trials ever undertaken, costing $75 million in 1970s currency. Part of this investment produced the Rand 36 item health survey (Hayes et al (1993)), later known as short form 36 (SF36). Subsequent work in Europe has produced another generic health related quality of life measure, the Euroqol or

EQ5D. These two measures (www.sf36.org and www.euroqol.org) have been used in thousands of clinical trials to measure how patients physical, social and psychological functioning changes with medical and surgical treatments. The concept of the quality adjusted life year (QALY) emerged from the US Office of Technology Assessment in the late 1970s, and was later developed in the UK context by Williams (1984). The concept of the QALY remains contentious but its use by regulatory agencies such as the English National Institute of Clinical Excellence is now extensive and having significant effects on policy (see e.g. Maynard, Bloor and Freemantle (2004)).

Both the QALY, and the two generic health related quality of life measures (SF36 and the Euroqol) have been used extensively but no health care system has used such measures to quantify the success of its investments. For instance, if ambulatory (GPs) doctors used such a measure routinely for each patient-visit, variations in the HRQoL could be measured over time. Use of electronic methods could present data onto physicians' screens before a consultation and with fluctuation in HRQoL identified in such a way consultations might be more efficient. Patient "journeys" outside the ambulatory sector could be assessed using HRQoL and this would not only identify the 'success' of interventions in terms of improving patients' quality of life, but would also identify the relative success of physicians and surgeons in improving or stabilising patients' health status.

Like variations and errors, advocacy of outcome measurement has a long history of being ignored. The English nurse Florence Nightingale advocated such measurement nearly 150 years ago, arguing for the measurement of success in terms of whether patients were "dead, relieved or unrelieved" (Nightingale (1862)). More recently Kind and Williams (2004) and Appleby and Devlin (2004) have re-iterated the call to measure success in health care. To date only one private UK insurer (the British United Provident Association) has adopted such measures (Vallance-Owen and Cubbin (2002)).

2.4.2 Overview

In their drive to reduce death and taxes, health economists (and their colleagues in related disciplines) have had little success in reducing a series of well evidence problems i.e. medical practice variations, including observable and significant differences in health between different groups, medical errors including the failure to deliver evidence based, cost effective care, and the measurement of success (i.e. improvements in patient related quality of life or outcomes). Why have they failed in these areas when their 'empire' has flourished in recent decades? How will the sub-discipline develop in the future?

The future of health economics

Given the failures of health economics and health services research chronicled in the preceding section, the natural response of economists is to emphasis the roles of quantitative analysis and modelling for analysing activity at the micro level (i.e. the level of the physician and the patient) and the creation of appropriate incentives to alter behaviour in ways that ensures physicians and other decision makers deliver what is needed to maximise population health at least cost.

Wennberg and his colleague Gittelshohn argued 30 years ago as follows:

"The amount and cost of hospital treatment in the community has more to do with the number of physicians there, their medical specialties and the procedures they prefer than the health of patients"

(Wennberg and Gittelsohn (1973))

Thus then, as now, the physician determines the pattern of care that is delivered and in no country is this systematically measured and managed. Typically physicians do not have good administrative information systems by which they can routinely monitor:

- What they produce in terms of activity, case mix and patient outcomes

- How much they produce relative to their peers

- How they produce i.e. what criteria they use to abandon old and adopt new technologies

- To whom do they deliver care, by income group

In the absence of such information, physicians practice in the fragmented, isolated Hippocratic tradition, trusted by patients but increasingly interrogated and regulated by public and private purchasers aware of the evidence base and sceptical of the efficiency or the "value for money" in terms of health creation delivered from using 10 percent of the GDP.

The scepticism of purchasers is leading to their investment in IT systems and the increasing use of financial incentives to reinforce traditional regulatory methods based on trust in the medical profession and mutual ignorance of whether that trust is merited and resources are used efficiently. British and German changes in physician remuneration may have significant effects on behaviour and require careful evaluation (e.g. Maynard and Bloor (2003)).

The choice of the efficient incentive package depends not only on a clear specification and ranking of policy goals, processes always elusive in the politically contentious area of public and private health care, but also on evidence of the cost effectiveness of alternative mixes of remuneration for practitioners and institutions, such as hospitals. Robinson remarked that " there are many ways of paying physicians; some are good and some are bad. The three worst are fee for service, capitation and salary" (Robinson (2001)). While this may be true, devising, testing and creating an evidence base to identify the appropriate "blend" of payment methods requires innovatory and systematic experimentation, which is rarely popular with politicians or the commercial organisations concerned with health care delivery.

If health economists are to inform the reformation of health care systems and improve both efficiency and equity, basic questions

have to be addressed by creating better information and improved incentives for decision makers to act upon that information:

(i) What are the objectives of the health care system? What ordering or weighting do these objectives get and how are these changing over time? Efficiency and equity in terms of distribution and access are clearly important elements in this definition of objectives.

(ii) Who is really responsible for control of the system(s), and who controls resource use at the margins or boundaries of care? I.e. who rations what and how in each part of the system.

(iii) What information about activity, case mix and patient outcomes do those at these margins need in order to achieve their goals?

(iv) How can performance be measured and managed appropriately?

(v) What incentives (monetary and non-monetary) are these (or lack of them) organisations for decision makers in public and private to achieve efficiency and equity?

These questions are very familiar (e.g. Maynard (1982)) and Williams over his career (e.g. Williams (1993)). Their resolution requires the further exploitation of mainstream economics theory (e.g. labour economics) and the application of quantitative techniques using appropriate methods of analysis (e.g. multi-level modelling) and very large data sets. Recent examples of this are Dusheiko et al (2003), which use differences in differences before and after methods of analysis, and Bloor et al (2004), which uses multi-level modelling. In both cases the techniques were applied to large data sets with millions of observations (Hospital Episode Statistics). As ever the challenge for health economists when deploying these methods of analysis is to

ensure that the right question or hypothesis is set out at the beginning, i.e. the results of analysis inform responses to (i) to (v) above.

2.5 Conclusions

Health economics has developed rapidly and impressively over the last 50 years. Part of its success has been that some of its practitioners possessed techniques (economic evaluation: box E in figure 1) of particular interest to powerful groups in the health care market (especially physicians and the pharmaceutical industry). The investments in this sphere of activity have enhanced the evidence base about what works (although with some corruption of that evidence base (e.g. Maynard (2002)).

The theory and policy challenges now are to ensure that further evidence of the cost effectiveness of competing therapies is produced efficiently and that the evidence is translated into practice i.e. the delivery of cost effective care in an efficient and equitable way. This requires investment in the systematic and careful testing of performance information systems (particularly the measurement of outcomes) and innovative use of mixed systems of incentives.

In 1753 the Scottish doctor James Lind published his "Treatise on Scurvy". This reported the findings of a trial of alternative ways of treating scurvy during a voyage to India. He showed that oranges and limes prevented scurvy. In 1795 the British Admiralty tardily began to issue limes and oranges to its sailors.

The efficient and timely translation of evidence into practice then as now is a nice challenge! The failure to meet this challenge in recent decades has condemned citizens, particularly the poor, to avoidable mortality and morbidity. Such outcomes for health economists driven to reduce health and taxes are a nice challenge.

Bibliography

Aiken LH, (2003), Achieving an Interdisciplinary Workforce in Health Care, The New England Journal of Medicine, 348: 164-166.

Aiken LSM, Clarke SP, Cheung RG, Sloane DM, Silber JH (2003), Education levels of hospital nurses and surgical patient mortality, Journal of the American Medical Association, 290, 1617-1623.

American Economic Association (1951), Economic of medical care, American Economic Review, 41, 2, 617-96 (papers and proceedings).

Appleby and Devlin, N, Measuring Success in the NHS: using patient-assessed health outcomes to manage the performance of health care providers, Dr Fosters' Ethics Committee and the Kings Fund, London.

Arrow K (1963), Uncertainty and the welfare economics of medical care, American Economic Review, 53, 5, 941-73.

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

Bloor K, Maynard A and Freemantle N. Variation in activity rates of consultant surgeons and the influence of reward structures in the English NHS. Journal of Health Services Research and Policy, 2004, 9, 2, 79-84.

Cochrane AL (1972), Effectiveness and Efficiency: random reflections on health services research, Nuffield Provincial Hospitals Trust, London.

Cullis J and West PA, (1979) The Economics of Health: an introduction, Martin Robertson, Oxford.

Culyer AJ, Wiseman J and Walker A (1977), An Annotated Bibliography of Health Economics: English language sources, Martin Robertson, London.

Culyer AJ and Maynard A (eds) (1997), Being Reasonable about the Economics of Health: selected essays by Alan Williams, Edward Elgar, Cheltenham.

Dusheiko M, Gravelle H, Jacobs R, Smith P (2003), The effect of budgets on doctor behaviour: evidence from a natural experiment, Centre for Health Economics, University of York, CHE Technical Paper Series 26.

Evans RG (1984), Strained Mercy: the Economics of Canadian Health Care, Butterworth, Toronto.

Fisher ES, Wennberg DE et al. The implications of regional variations in Medicare spending, two parts, Annals in Internal Medicine, 138, 273-87 and 288-98, February 2003.

Fisher ES. Is more always better? New England Journal of Medicine, 349, 166507, 2003.

Griffith DAT, Rigou R, Tacier P and Prescott NM (eds), (1980), An Annotated Bibliography of Health Economics: Western European Sources, Martin Robertson, Oxford.

Grossman M (1972), The Demand for Health: a theoretical and empirical investigation, National Bureau for Economic Research, Occasional Paper 119, New York and London.

Hays RD, Sherbourne CD and Mazel RM (1993), The Rand 36-Item Health survey 1.0, Health Economics, 2, 3, 217-28.

Hayward RA, Hofer TP (2001), Estimating Hospital Deaths Due to Medical Errors, Preventability is in the Eye of the Reviewer, Journal of the American Medical Association, 286, 4, 415-20.

Jones AM and van Doorslaer E, (date), Income related inequalities in health and health care in the EU, Health Economics, 13, 7, 605-8.

Kerr EA, McGlynn EA, Adams J, Keesey J and Asch SM (2004), Profiling the quality of care in twelve communities: results from the CQI survey, Health Affairs, 23, 3, 247-256.

Kind P and Williams A (2004), Measuring success in health care: the time has come to do it properly, Health Policy Matters, 9, Department of Health Sciences, University of York, (http://www.york.ac.uk/healthsciences/pubs/hpmindex.htm).

Lancashear, A, Sheldon, T and Maynard, A, (2005), Nurse staffing and healthcare outcomes: a systematic review of the international research literature, Advances in Nursing Science, 28,2, forthcoming.

Lees DS (1966), Economic Consequences of the Professors, Institute of Economic Affairs, London.

Mark BA, Harless DW, McCue M, Xu Y (2004), A Longitudinal Examination of Hospital Registered Nurse Staffing and Quality of Care, Health Services Research, 39, 2, 279-300.

Maynard A and Bloor K (2003 (a)), Trust and performance management in medical marketplace. Journal of the Royal Society of Medicine, 96, 532-539.

Maynard A and Bloor K (2003 (b)), Do Those Who Pay the Piper Call the Tune? Health Policy Matters, Issue 8, Department of Health Sciences, University of York, (http://www.york.ac.uk/healthsciences/pubs/hpmindex.htm#Iss8).

Maynard A and Kanavos P (2000), Health economics: an evolving paradigm, Health Economics, 9, 3, 183-190.

Maynard A, McKee M and Nolte E (2004), Models of Care: some European perspectives, mimeograph.

Maynard A, Bloor K and Freemantle N (2004), Nice challenges, British Medical Journal, 7459, 227-29.

Maynard A (2002), Evidence-based medicine e conflitto di interessi. Politiche Sanitaire, vol 3, N.1, Gennario-Marzo 2002, page 30-33.

McLachlan G and Maynard A (1982), The public-private mix for health: the emerging lessons, in G McLachlan and A Maynard (eds), The Public-Private Mix for Health, Nuffield Provisional Hospitals Trust, London.

Mushkin SJ (1958), Towards a definition of health economics, Public Health Report, 73, 9, 785-93.

Needleman J., Buerhaus P., Mattke S., Stewart M., Zelevinsky K (2002), Nurse-Staffing.

Levels and the Quality of Care in Hospitals, The New England Journal of Medicine, 346: 1715-1722.

Newhouse JP (1978), The economics of medical care: a policy perspective, Addison Wesley.

Newhouse, JP and the Insurance Experiment Group (1993), Free for all: lessons from the Rand health insurance experiment, Harvard University Press.

Nightingale F, (1863), Some Notes on Hospitals, Longman, Third Edition.

Nolte E and McKee M (2004), Does health care save lives? Avoidable mortality revisited, Nuffield Trust, London.

Office of Health Economics (1962 -), Studies in Current Health Problems, London.

Pittet D, Hugonnet S, Harbarth S, Mourouga P, Sauvan V, Touveneau S (2000), Effectiveness of a hospital-wide programme to improve compliance with hand hygiene, The Lancet, 356: 9238.

Vallance-Owen, A and Cubbin, S. Monitoring national clinical outcomes: a challenging programme, British Journal of Health Care Management, 8, 11, 412-16, 2002.

Vincent C, Neale G, Woloshynowych M (2001), Adverse events in British hospitals: preliminary retrospective record review, British Medical Journal, 322: 517-19.

Wennberg JE, Freeman, JL and Culp WJ (1987), Are hospital services rationed in New Haven or over-utilised in Boston, The Lancet, I, 1185-8.

Wennberg JE, Freeman JL, Shelton RM., Bubolz TA (1989) Hospital use and mortality among Medicare beneficiaries in Boston and New Haven, The New England Journal of Medicine, 321:1168-1173.

Wennberg JE and Gittelsohn A (1973), Small area variations in health care delivery: a population information system can guide planning and regulatory decision making, Science, 182, 1102-8.

Williams A (1988), Health Economics: The cheerful face of a dismal science? (in A J Culyer and A Maynard (eds) Being Reasonable about the Economics of Health, Edward Elgar, Cheltenham, 1997).

Williams A (1993), Priorities and research strategies in health economics in the 1990s, Health Economics, 2, 295-302.

Wilson RM, Runuman WB, Gibberd RW, Harrison BT, Newby L, Hamilton JD (1995), The quality of Australian healthcare study, Medical Journal of Australia, 163: 458-71.

Discussion I

Hannu Valtonen

Department of Health Policy and Management,
University of Kuopio

As Voltaire said, *"The art of medicine consists in amusing the patient while nature cures the disease."*

Or, in modern parlance: most drugs work in only 30% or 50% of people. (S. Connor, Glaxo chief: our drugs do not work on most patients. Independent 2003 Dec 8:1, cited by Doust and Del Mar 2004)

The history of health economics as a discipline is not very long, although the basic problems that attribute to health economics are old. As we can see, Voltaire posed the question of the cost-effectiveness of health care. There are two questions that have been asked about health economics during this history, and both of these are questioning its legitimacy as a science. The older one, and now almost forgotten, of these questions is: "Why is economics penetrating into the health field?" (One example of this discussion is: (2)). There was a suspicion that economists were importing tools or values that were not compatible with medical ethics or the objectives of health care. Alan Williams has several times defended health economics against accusations of being unethical (3–7). In his presentation prof. Maynard is showing how health economics has a reasonably well defined map of the object of economic study, and it also has a morally sound basis for its existence, *i.e.,* the achievement of both efficiency and equity through proper economic organisation of health care. This map has been very useful in providing a structure to our thinking

about our own discipline, and it gives an answer to what economists think they are doing in health care.

The second question that has been asked throughout history and even today is: 'Is health economics useful?' Health economics has proved to be successful as a research program both in the UK and Finland, but is it also useful for the whole society? One of the most interesting and provocative formulations of this question was done by a group of sociologists at the university of York – health economics 'colonising the minds of policy makers' (8,9).

One additional question that might increase our understanding about the usefulness of health economics has not been asked that often: Why did health economics start to develop in the UK from 1970s onwards and in Finland a little later? In principle, the birthplace of health economics in Europe could have been any other country. In Maynard's paper the information supply side is well covered. The answer to this might also provide an answer to the question whether or not health economics is useful. From the history of science we know that science and scientific inventions need suitable conditions in the society.

A tentative answer to the question could be similar to the answer to the question about the birth of medical sociology in the UK. After the creation of the National Health Service, the information thought necessary was sociological information, and this need for information was one of the reasons for the birth of medical sociology in the UK. A growing health care system is also a growing consumer of control, planning, management, and policy making and information, including economic information. At the same time the general position of 'the economy' in western societies strengthened, and a natural consequence of all this was an increased demand for health economics. Hurst (10) is also showing that health economics has had an impact not only on policy but there is an impact in the direction of health economics, too (5). In Finland, we were living in the same kind of situation from 1970s onwards with a growing public health care system. Therefore, after the first generation of Finnish health economists returned from

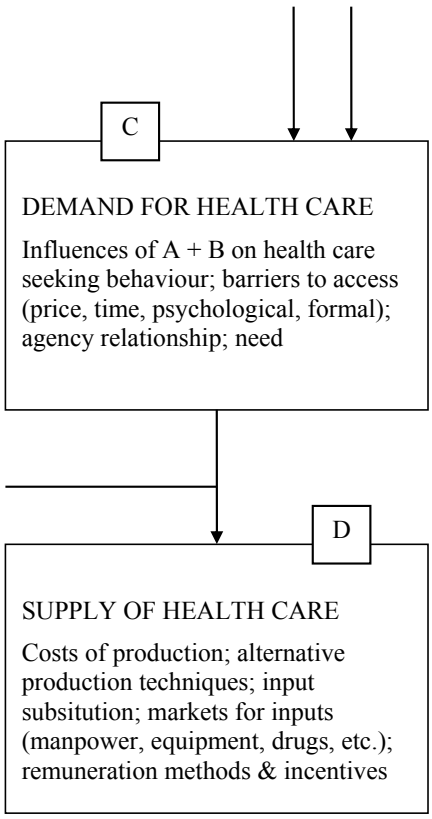
York, there was also demand for their services created by the control, planning, management, and policy making needs of the public health care system. The supply of economic information in health services was created by Yrjö Jahnsson Foundation and the University of York. The Finnish response to this demand for information can be seen, *e.g.*, in the Finnish bibliography of health economics. (11)

Some comments of future challenges:

I return to the map, to one part of it – to the meeting of supply and demand. Up to the beginning of the 90s before the decentralisation of health care system management, we in Finland could think supply as a policy variable – the supply of services was planned. Thus, there was no need to question the determination of supply. Our knowledge of the functioning of health care leads easily to the conclusion that it is a 'supply-led' system. *I.e.*, when the relative powers of demand (or even need) and supply are compared, there is much more societal power on the supply side (the economic, political, and professional power, *e.g.*, in hospitals or the trade union force of medical associations). In the map the boxes meet nicely, but when we start the orientation with this map, we easily see that there is much more market or societal power behind the supply than behind the demand. Therefore, in addition to the themes in the supply box, we as economists should in future try to better understand the supply and demand not only technical dependencies but as social forces, as carriers of power, as human activity.

What hope is there for not using treatments and tests that don't work? Medicine is not just a science – it is a human activity. It entails ritual, custom, and the expectations of doctors, patients, and society. (1)

Figure 1. Demand meets Supply



References:

(1) Doust J, Del Mar C. Why do doctors use treatments that do not work? BMJ 2004 2/28;328(7438):474-475.

(2) Klein R. The role of health economics, Rosencrantz to medicine's Hamlet. British medical journal 1989;299:275-276.

(3) Williams A. Economics, QALYS and medical ethics - A health economists perspective. Health Care Analysis Aug 1995;3(3):221-226.

(4) Williams A. Cost-effectiveness analysis: is it ethical? Journal of Medical Ethics 1992;Mar;18(1):7-11.

(5) Williams A. The Role of Health Economics in Clinical Decision-Making - is it Ethical. Respiratory medicine 1991 SEP;85:3-5.

(6) Williams A. One economist's view of social medicine. Epidemiology and community health 1979;33:3-7.

(7) Williams A Being reasonable about the economics of health: selected essays by Alan Williams. : Elgar; 1997.

(8) Ashmore M, Mulkay MJ, Pinch TJ. Health and efficiency : a sociology of health economics. Milton Keynes [England]: Open University Press; 1989.

(9) Mulkay M, Ashmore M, Pinch T, Hutton S, Pinch T, Shiell A. Health, Dependency and Despair, Economics and the Health Care System. In: Hutton J, editor. Dependency to Enterprise: a challenge to policy, London: Routledge; 1991.

(10) Hurst J. The impact of health economics on health policy in England, and the impact of health policy on health economics, 1972-1997. Health economics 1998 AUG;7:S47-S61.

(11) Mikkola H. (ed.) Suomalaisen terveystaloustieteen bibliografia. Helsinki: Terveystaloustieteen seura; 2001.

Discussion II

Harri Sintonen

Department of Public Health, University of Helsinki

First of all, on behalf of the Finnish Society for Health Economics I would like to congratulate the Yrjö Jahnsson Foundation for its 50th Anniversary and to thank the Foundation for the grants and support, from which the Society and most of its members personally have benefited over the years. I do hope that this fruitful collaboration will continue also in the future.

The presentation by Alan was as insightful and at the same time entertaining as ever. I have come across many definitions of the purpose of health economics, but now Alan came up with yet another saying that the purpose is "to protect individual citizens by reducing death and taxes". This sounds very ambitious, almost a mission impossible, since at least until today the death rate of the population has been 100 % and only market forces connected with the enlargement of the European Union and globalisation have forced Finnish government to reduce taxes with the consequence that local communities have increased their taxes. I might have chosen a slightly more modest mission for health economics than Alan.

The already familiar plumbing diagram shows the areas and issues that health economists are addressing in their work. As Alan indicated, there are differences in the main areas of interest and research in different countries, most notably between the US and Europe, reflecting different value climates and public-private mix in these areas. For instance equity has been a concern in Europe, not that much in the US. Also otherwise Alan reviews the past, present and future of health economics largely from the British and American perspective. This is understandable, since the roots are there and

much of health economics research takes place in those countries. However, it may be useful to say a few words about the past and present of health economics in Finland as well. Antti Suvanto in his opening address already highlighted some milestones from the viewpoint of Yrjö Jahnsson Foundation, but I can amend it a little from inside the profession.

The first Finnish language textbook on health economics was published in 1987. The first professor in health economics was nominated in 1988, now there are three full or part-time professorships in three universities. The Finnish Society for Health Economics was established 1992, now it has about 60 selected members, one third of them have studied at York. Since its inception the Society has organised, with financial support for example from the Yrjö Jahnsson Foundation, an annual open Day for Health Economics, which has been a great success with an annual audience of 250-300. In 2001 the Society published a bibliography of its members. The bibliography includes about 800 entries. Following the classification of the plumbing diagram, most publications are in the categories of microeconomic evaluation, planning, budgeting and monitoring mechanisms as well as demand and utilisation. Health economics plays an important role in the Finnish Office for Health Care Technology Assessment (FinOHTA) established in 1995 at Stakes, since it has an explicit mission to collect, generate and deliver information not only on the effectiveness, but on cost-effectiveness of health care technologies. In 2002 Centre for Health Economics (CHESS) was established at Stakes with a staff of about 20. In CHESS the research focus is on health care organisation, economic incentives, health care financing and productivity. Finland was one of the first countries to introduce mandatory economic evaluation of new pharmaceuticals in 1998.

Returning to Alan's review I felt somewhat uneasy, when Alan at least indirectly blames health economists that despite increased levels of investment in health services research and health economics, they have not been able to solve all problems in health care. As to the level of investment in health economics in Finland, it is next to nothing, at maximum 0.04 % of the annual health care expenditure. A lot has been

achieved that cheap, but it is too low a price to make a full revolution. In spite of hundreds of years of effort worldwide to develop a good health care system, I am not aware that anywhere they managed to get it right, with or usually without health economists.

I think all the other parties in health care but health economists are to be blamed. In Finland health care is solidly based on the holy trinity of three principles: local self-administration, clinical freedom and multiple sources of financing. The well over 400 local communities running the health service can basically do whatever they want under local self-administration. The doctors working in health service can do whatever they want under clinical freedom and multiple sources of financing create a lot of perverse incentives that work against achieving an efficient and equitable system. It is hard to imagine a more fragmented system with more health politicians that the Finnish one. The dilemma is how to manage a basically unmanageable health care system towards efficiency and equity. The avoidable mortality and morbidity, that is, avoidable loss of length and quality of life may be unacceptable to health economists, but they seems to be acceptable to health profession and politicians inferring from the reluctance or at least slowness of introducing systems by which health care could be provided more efficiently and equitably. Obviously creating such information and management systems is not in the interest of all parties connected in one way or another with health care. At least it should be of a high interest to patients and taxpayers.

As to the future, I strongly agree that introduction of routine measurement of and management by health-related quality of life and QALYs is a key element in an effort to create a more efficient and also more equitable health care system. However, I was disappointed that Alan only advocates EQ-5D and SF-36 for the measurement of health-related quality of life, although there are other and better generic instruments available. Finland is surely among the leading countries in the world, if not in the pole position in introducing such a measurement and management system. Integrating a better generic health-related quality of life instrument, namely the 15D, in the electronic medical records of Turku University Hospital District is

underway, a large experiment of measuring health-related quality of life routinely before and after treatment is ongoing in the Helsinki and Uusimaa Hospital District and in some other smaller districts and there is hope that the 15D would be included in the uniform electronic medical records to be created in the near future for Finnish health care.

In addition to creating a uniform system for measuring outcomes in generic terms, uniform systems for measuring the resource consumption (costs) are also needed and a sizeable group of health economists and statisticians need to be harnessed to analyse such data for instance in FinOHTA and universities. Increasing amounts of money and resources are poured into health care without knowing what they produce in terms of health. It is hard be imagine any other industry of that scale, where the costs and output of production are known so poorly. But I also agree fully with Alan that management by information is not adequate alone, but improved incentives to act upon that information need to be created. Considering the cemented peculiar features of Finnish health care already mentioned it is easier said than done.

In spite of difficulties of various kind, we keep on working for more efficient and therefore also more ethical and equitable health care systems until we retire and some, like Alan Williams, has even retired from retirement to promote that goal. I see light in the end of the tunnel and look forward to a brighter future as Alan outlined in his presentation.

Physician Payment Innovations In The United States And France

3

James C. Robinson*

School of Public Health, University of California

Francis Megerlin**

Maitre de Conferences, Droit et Economie de la Sante,
Universite de Paris V

3.1 Introduction

Fee-for-service, long the dominant method of physician payment in the United States and France, has come under criticism as incompatible with the incentives needed to shift towards a health care system that emphasizes cost-effective care coordination for an older and chronically ill population. During the 1990s many critics saw an alternative in prospective methods of payment, especially capitation, according to which physicians or physician organizations are paid on a per-member-per-month basis for primary care and, in some cases, specialty care services. The subsequent experimentation with capitation, however, highlighted the incentive problems latent in prospective payment and the difficulties faced by physicians in building organizational structures with the scale and sophistication to manage complex financial responsibilities. Since the turn of the

*jamie@berkeley.edu
**francis.megerlin@univ-paris5.fr

millennium, capitation has been in retreat throughout the United States, re-igniting the fires of cost inflation and lending greater urgency to the development of payment methods that encourage cooperation among physicians, compliance with clinical protocols, and the diffusion of information technologies. In France, capitation in its pure form was never attempted, and the contemporary experimentation is not a response to disillusion with prospective payment. The French health insurance system is seeking payment methods that promote performance measurement and reward in the face of continuing cost inflation and the perceived need for new forms of health care organization.

In light of the limitations inherent in both capitation and fee-for-service, it is not surprising that the United States and France are engaged in discussions concerning new methods of physician payment. In the U.S., emerging compensation structures combine prospective and retrospective approaches, paying physicians partly on the basis of the number of procedures performed and partly on the basis of capitation and other measures independent of clinical inputs. Some of the experiments seek to reward specific dimensions of physician performance, particularly dimensions that reflect the quality of the services provided. The rubric of "pay for performance" includes financial incentives directed at individual physicians or physician groups, incentives that cover "process" and "outcome" measures of clinical quality, and incentives for use of information technology, cooperation with disease management programs, and responsiveness to patient concerns. In France, discussions of payment reform have bypassed capitation and proceeded to flat rate (*forfait*) payment for episodes of care, usually blended with elements of fee-for-service.

Whereas the decentralized health insurance system in the U.S. facilitates experimentation, the diverse but high regulated French insurance system channels initiatives into the policy and political arena. The modification of physician payment systems must be agreed to at the national level through the structured relationships between the public insurer (Assurance Maladie, a branch of Social Security, the dominant purchaser for health services) and various associations

representing the interests of the physicians. This centralized bargaining structure traditionally fostered continuity and inertia, but recently has permitted an accelerating diversity in payment methods at the national, regional, and local levels. Budgetary pressures on the governmental side and restiveness on the physician side have been building for several years, and in 2004 the legislature enacted major reforms in response to an extensive study of the demographic and financial challenges facing the French health care system.

This paper examines contemporary innovations in physician payment in the United States and France, highlighting the variety in approaches as well as the underlying similarities. We begin with a sketch of the economic literature on contractual relationships between principals and agents, which lays the foundation for understanding efforts to base payment on performance in a world of uncertainty and risk, multiple tasks and multiple agents, and imperfect measurement of effort and efficacy. We then describe recent examples of payment methods that blend prospective and retrospective incentives and that seek to reward input or output measures of quality, relying on journalistic accounts, case studies, and survey data.

3.2 Physician Payment In Light Of Economic Theory

The economic literature on principal-agent relationships interprets observed payment methods as efforts by principals (insurers, purchasers, consumers) to balance the advantages of productivity incentives for agents (physicians) with the disadvantages of exposing these agents to risk. Where the effort or the efficacy of the agent cannot be evaluated directly, as is often the case for technical services such as medicine, the principal may seek to pay agents on the basis of outcomes rather than of inputs. Outcomes are influenced by numerous factors, however, and outcome-based performance payment exposes agents to the risk of uncontrollable or unpredictable payment fluctuations. Risk-averse agents demand compensation for bearing risk, and so the cost to the principal of outcomes-based payment

includes a premium that grows with the extent of risk faced by agents. As the required risk premium grows, principals are motivated to shift from outcomes-based payment to input-based methods or to salaried compensation divorced from any immediate measure of performance (Prendergast 1999).

The central empirical implication of the agency framework, that outcomes-based pay should be less prevalent in settings where outcomes are subject to high risk than in more stable and predictable settings, has not been supported by the extant empirical studies. Formal analyses of compensation methods for various occupations and professions, plus journalistic accounts of stock options in technology start-ups, indicate that performance-based pay is often found in highly uncertain environments and is often absent from stable and predictable environments. The variation in payment mechanisms across occupations and industries cannot be explained in a convincing manner by variation in risk and uncertainty. In light of these discrepancies, the conceptual framework has been extended to emphasize the influence of worker selection effects (such as hiring and promotion) in the design of payment methods (Lazear 2000); the avoidance of simple productivity incentives in contexts where agents must allocate effort among multiple tasks (Holmstrom and Milgrom 1991; Baker 1992); the need to support cooperation and joint productivity in contexts where team effort is important (Kandel and Lazear 1992); the salience of task delegation and outcomes-based performance in risky environments where principals cannot evaluate inputs (Prendergast 2000; 2002); the use of subjective evaluations and informal "relational contracts" (Gibbons 1998); the avoidance of performance-based pay within organizations where internal politics adversely affects productivity (Milgrom and Roberts 1988); and the importance of measurement difficulties in all contexts, including allocation of efforts across different tasks, allocation of effort by different team members, responsiveness of agents to particular incentives, and ability of agents to influence measurement methods (Baker 2000).

3.2.1 The Limitations of Fee-for-Service

The importance of protecting physicians from financial risk is evident in the traditional absence of payments linked to health status improvement (clinical outcomes), as these are influenced by patient behaviors, inputs from multiple caregivers, and the natural course of the underlying disease, as well as by the effort and expertise of the physician. The traditional form of physician payment in the United States and France has been fee-for-service, which measures performance based on the number and complexity of inputs rather than the health output. The importance of motivating individual productivity, and the robustness of clinical relative-value scales, is evident in the continued reliance on fee-for-service long after piece-rate payment has disappeared from most occupations (Gibbons 1987; Whyte 1955).

Fee-for-service suffers from serious limitations as a compensation mechanism in the contemporary economic and technological environment (Robinson 2001A). Fee-for-service creates incentives for excessive treatment, in that procedure prices must be set at levels more than twice the marginal cost to the physician of providing the service, given that administrative and clinical overhead consumes more than half of practice revenues. Linked to particular acts by particular agents, fee-for-service interferes with changes in practice methods that rely on non-physician caregivers, supervised patient self-management, patient group visits, and telephonic and electronic consultations between physicians and patients. Episodes of care for patients with chronic conditions typically extend over time and across multiple settings and are poorly supported by compensation methods that target discrete physician interventions without regard to what has gone before or what is likely to come after. Fee-for-service rewards the quantity rather than the quality of the services delivered, rewards unnecessary as well as necessary care, and imposes a 100% tax on practice innovations that reduce the need for subsequent (remunerable) services. At the level of the health care system as a whole, fee-for-service supports a hamster-wheel of payment-induced high volumes, accelerating cost inflation, cuts in payment rates for

each unit of service, and a consequent acceleration by physicians in the number of services for which they submit payment claims.

3.2.2 The Limitations of Capitation

Capitation is an output-based measure of physician performance, in the sense that the physician is paid a predefined sum for all the care needed to maintain the health of those patients who choose him or her as their personal physician. The physician can refer the patient to other providers for care that the capitated physician cannot provide, and these referred services typically are paid by the insurance plan directly to the referral specialist rather than by the capitated physician. By divorcing payment from particular acts and particular actors, capitation facilitates experimentation with non-physician caregivers, electronic "visits", and other alternatives to one-on-one in-person encounters in the multi-task, multi-agent clinical context. However, capitation suffers from its own set of important and often fatal limitations. Most obviously, capitation shifts to the physician the risk of attracting an exceptionally sick patient population (as capitation rates are only partially adjusted for health status) and the risk that the patient's health status will decline over time (thereby necessitating more physician intervention) for reasons independent of the physician's efforts. Capitation payment does not reward high-quality care and rewards the under-provision of services in cases where so doing encourages particularly sick patients to leave the practice. While global capitation covering all professional services creates a collective efficiency incentive for all covered physicians, capitation limited to one physician's services (e.g., primary care capitation) encourages referrals and a narrow scope of practice.

Some of the challenges posed by capitation can be attenuated when prospective payment is applied at the level of the physician organization rather than that of the individual physician (Berwick 1996). Large, multi-specialty medical groups are better able than are individual physicians to spread the financial risk inherent in prospective payment and can develop administrative mechanisms and cultural norms to limit inappropriate referrals and risk selection

by individual clinicians. Prepaid physician organizations can serve as a buffer between the high-powered prospective payment incentives received from health plans and the decision-making by individual clinicians. However, group capitation encounters its own set of challenges, especially the unwillingness of physician-owned medical groups to invest in the administrative and information technology infrastructure to manage complex accounting and financial flows and to retain adequate reserves for medical costs incurred by prepaid patients (Robinson 2001B). Physician organizations often find it difficult to translate the collective incentive for efficiency generated by capitation into a workable incentive for efficiency for the individual physician.

Blends of Fee-for-Service and Capitation

The limitations of fee-for-service and capitation as incentive mechanisms in the multi-task, multi-agent, referral-based, poorly measured, and highly risky business that is medicine have led to efforts to blend retrospective and prospective components into hybrid payment incentives. Indeed, the cost structure of medicine would appear to call for a two-part payment structure, in which the fixed administrative and support expenses that account for half of total practice costs could be paid on a capitated basis while the marginal costs of directly treating individual patients (that account for the other half of total costs) could be paid on a fee-for-service basis. The need to encourage cooperation among physicians in referrals and consultations would seem to favor low-powered salaried or equal-share payments as a supplement to, if not replacement for, fee-for-service incentives for individual productivity.

Evidence of physician payment methods that blend retrospective and prospective elements is found in several U.S. contexts. Journalistic accounts of compensation methods used by particular physician organizations and insurance plans highlight the variation across clinical specialties, geographic regions, and the size and structure of practice organizations in the manner by which physicians are paid. A professional information distribution firm recently published

a summary of journalistic accounts of payment methods used in multi-specialty medical groups, Independent Practice Associations (IPAs), academic medical centers, health insurance plans, hospital-owned practices, and single-specialty groups covering surgery, cardiology, gastroenterology, orthopedics, psychiatry, geriatrics, and other specialties (Atlantic Information Services, 2003). The most striking feature of the journalistic accounts is the heterogeneity of the payment methods, which defy easy categorization, but also what the authors identify as the underlying commonality: efforts to balance incentives for individual physician productivity, on the one hand, and for cooperation by individual physicians with other group members, on the other. Most organizations place the greatest weight on individual productivity, measured in terms of patient care revenues or relative value units, but none rely solely on fee-for-service payment. Rather, the organizations have developed methods to link the financial success of the individual physician to the financial success of the larger organization, including profit sharing, straight salary for clinical work, and supplemental salary for administrative responsibilities.

Case study evidence of hybrid payment methods within the U.S. context is provided by Robinson (1999), who studied the blends of capitation and fee-for-service for individual physicians developed by physician-owned IPAs in the late 1990s. These IPAs tended to prepay physicians for primary care services but to supplement the capitation base payment with fee-for-service supplements for preventive services, for services involving high supply costs (e.g., in-office vaccines), for discretionary services provided outside the office (e.g., visits to patients in nursing homes), and, most importantly, for services that lay on the frontier between primary and specialty care (e.g., flexible sigmoidoscopy) and hence which are subject to the incentive latent in capitation for physicians to narrow their scope of practice and refer complex procedures to others.

Econometric evidence on the blending of prospective capitation and retrospective fee-for-service comes from a study of physician organizations in California by Rosenthal and colleagues (2002) and a

national study of physician organizations by Robinson and colleagues (2004). Rosenthal et al. highlight the "vertical" blending of payment incentives, in which the health plan pays the physician organization (multi-specialty medical group or IPA) a prospective capitated rate for each patient and the physician organization subsequently pays its individual physician members on a salaried basis. Robinson et al. measure "horizontal" payment blends for primary care and specialty physicians across medical groups and IPAs nationally, in terms of the percentage of annual compensation paid to individual physicians through fee-for-service (with the remainder being paid through capitation in the IPA context and salary in the medical group context). Approximately one quarter of medical groups pay their primary care physicians on a pure fee-for-service basis, one quarter of groups base no part of primary care payment on individual productivity, and half pay various blends of salary and fee-for-service. A similar mix of capitation and fee-for-service for primary care physicians is reported in the IPA context. The percentage of total compensation linked to individual productivity is higher for specialists than for primary care physicians in both multi-specialty medical groups and IPAs.

Capitation as a method of physician payment never emerged in France due to its hospital-centered clinical culture and the absence of large medical groups. An initiative in 1997 sought to identify physicians willing to take on the role of coordinating specialty referrals, and to reimburse "referring physicians" on a capitated basis for this activity. This approach to prospective payment made little headway against the patients' desire to self-refer to specialty care and to switch doctors at will, and did not reduce the rate of cost growth faced by the Assurance Maladie. In 2004 the effort to promote payment reform was included in a broader initiative that seeks to develop a physician gatekeeping system and develop physician teams organized around chronic conditions and types of care. Reimbursement is to combine a flat payment for care coordination activities for a defined period of time, on the one hand, with fee-for-service payment for individual visits and procedures, on the other. As the flat rate payment is available only for patients with particularly complex needs and only for a specified period of time, they approximate "episode-of-illness"

payments (Emery 1999) or payments for disease management activities, rather than capitation.

An example of the emerging payment blend in France is to be found in palliative care for home-bound, seriously ill patients. A lead physician is paid a monthly sum to coordinate the care from specialist physicians, nurses, physical therapists, and others. Individual visits and interventions by each provider can be paid on a fee-for-service basis or on an episode-of-illness flat rate. The Assurance Maladie encourages, but does not require, the various providers to agree amongst themselves on a particular structure of payment. In any case, payment is made by the insurer directly to each individual clinician, not to the care coordinator or any other intermediary on behalf of all members of the team. The blending of episode payment with fee-for-service corresponds well to the contemporary organization of medicine in France, where professionals remain largely in solo practice but where efforts are being made to structure clinical networks around particular conditions or forms of care. It is impossible to predict the ultimate effects of the payment reform, which presumably will vary across diagnoses and geographic regions. In any case, innovations in forms of payment are responding to, rather than causing, changes in the form of organization (multi-disciplinary networks).

3.3 Pay-For-Performance

Conventional discussions of health care quality focus on structure (e.g., professional credentials, organizational capabilities), process (services performed), and outcome (change in health status), with the belief that outcome measures are better than process measures, which in turn are better than structural measures. The contemporary movement for quality improvement in health care highlights the desirability of moving towards outcomes-based measures, and pay-for-performance experiments have strived to measure health status. However, as outcomes are influenced by many factors aside from the physician's intervention, outcomes-based measures shift excessive financial risk to physicians unless several modifications are made.

Outcomes-based performance measurement and payment become more viable to the extent a narrow definition of outcome is substituted for broad definitions, measures are based on changes in health status rather than levels of health status, and multiple measures of outcomes are combined into an index.

The measurement and causality problems inherent in outcomes-based measures drive payment methods back towards process-based and, in some cases, structure-based measures of performance. The most common process measures of performance, for purposes of payment reward, include whether appropriate tests, prescriptions, and procedures are ordered for the populations of patients for whom they are indicated (based on evidence-based standards of appropriate care). Process measures of performance can include whether the appropriate intervention is made or, on the contrary, whether inappropriate interventions are made (e.g., prescription of antibiotics for viral infections). They can focus purely on quality or include consideration of cost (e.g., generic versus branded drug prescription). Structural measures of performance seek to reward organizational and technological prerequisites for high-quality care, such as accessibility (e.g., time to obtain an appointment) and use of information technology.

The other dimension of pay-for-performance programs concerns the organizational level at which performance is measured and rewarded. Measuring and rewarding the performance of the individual physician has the advantage of focusing attention on the person most immediately responsible for the care of the patient and avoids "common pool" problems inherent in assigning responsibility to groups comprised of semi-autonomous individuals. However, to the extent that relationships and cooperation among multiple physicians (e.g., appropriate referrals, consultation) are important for quality and efficiency, the appropriate level of focus shifts to the group. Payment at the group level highlights the importance of administrative and information infrastructure (disease management programs, patient education programs, point-of-care electronic prescribing or medical chart access) and channels funds to finance

organizational capabilities rather than to increase the personal income of individual clinicians. On purely statistical grounds it is often more valid to measure clinical performance at the group level, as any one physician may not see sufficient numbers of patients with particular conditions to ensure precise confidence intervals (e.g., percentage of diabetics receiving appropriate vision and hemoglobin tests). On the other hand, most physicians are not members of strongly integrated clinical organizations and it is difficult to apply measurement and reward to loose associations where little effort is made to enforce consistent standards of practice.

3.4 Examples

3.4.1 Blue Cross of California

Blue Cross of California (BCC) is the California subsidiary of WellPoint, one of the largest health insurance plan in the United States,with approximately 16 million enrollees nationally. Of the seven million BCC members in California, approximately 4 million are enrolled in the firm's commercial PPO product (with two million in the commercial HMO and 1 million in the Medicaid HMO). BCC has developed a pay-for-performance program for PPO enrollees, named Physician Quality and Incentive Program (PQIP), with a focus on increasing payment levels to physicians who achieve defined levels of performance on 27 measures, which are weighted and combined into an overall performance index for each eligible physician. In order to be eligible for extra payments, physicians must see a non-trivial number of BCC patients each year (e.g., more than $12,000 in annual claims) and be in primary care or a high-volume medical specialty (allergy, cardiology, endocrinology, gastroenterology, obstetrics/gynecology, psychiatry, pulmonology). Physicians are eligible for annual bonus payments up to 10% above their claims, with a maximum of $5000 per year. Performance is measured in terms of three principal categories: clinical quality (mostly process measures), pharmaceutical cost effectiveness (generic prescribing), and structural measures (accessibility to patients and investment in

information technology). The quality and pharmaceutical indicators each are accorded 40% weight in the final index, with accessibility and information technology together accounting for the remaining 20%.

Process measures that indicate whether patients have received appropriate tests, prescriptions, or procedures face the problem that each patient may see multiple physicians over the course of the year and the payment system must decide whom to reward. In the HMO setting, each patient selects a primary care physician to coordinate all care, and this individual can be held responsible for much of what happens or fails to happen throughout the year (including intervention by referral specialists). In the PPO product, however, patients need not select an individual physician to coordinate care. Blue Cross of California has decided to consider all the physicians that the patient has seen over the course of the year as collectively responsible for ensuring that appropriate care is delivered (without regard to which physician actually delivered the care). The enrollment and claims data systems operated by BCC permit it to create a "virtual medical group" for each PPO enrollee. Each physician's performance then is based on the performance of all of the virtual groups of which that physician is a "member" (in light of the patients' choices of physician). For each physician, the measures of clinical quality are defined in terms of the percentage of their patients (e.g., the patients who have come to them for any reason over the course of the year) who have received the appropriate care from any physician during that year.

The most straightforward process measures of quality concern the administration of vaccinations, tests, and prescriptions for specified patient populations. One measure, for example, focuses on whether children receive vaccination for measles before their second birthday, while another focuses on whether adult patients suffering from congestive heart failure receive ACE inhibitor medication. For the measles measure, criteria for inclusion (denominator of the performance measure) are based on the child's age and continuous enrollment in the health insurance plan for 12 months. For heart

failure patients, inclusion criteria require two outpatient visits to a physician or one inpatient hospital visit. Performance (numerator) data are derived from pharmaceutical prescriptions.

For patients with chronic conditions such as high blood pressure or elevated blood lipids, process measures can be extended from whether the appropriate medication was prescribed to what percentage of the year the patient was actually using the medication (given the tendency for patients to stop using medications for "silent" conditions such as hyperlipidemia). A third Blue Cross performance measure thus focuses on whether patients with high cholesterol are taking lipid-lowering medications for at least 80% of the year. Inclusion criteria (denominator) in this measure cover adults whose physician claims included diagnosis of hyperlipidemia or who were prescribed lipid-lowering drugs at some point in the prior year. Performance (numerator) criteria sum the number of non-overlapping days' supply of lipid-lowering medications prescribed and filled during the year.

A fourth measure of clinical performance focuses on whether patients with asthma receive appropriate physician visits after an acute episode in the emergency room or hospital. The inclusion criteria for this measure cover patients with emergency or hospital visit with primary diagnosis of asthma, and the performance criteria is whether the patient had a follow-up outpatient physician visit for asthma care within 14 days of the acute episode. A final example highlights clinical quality measures that focus on inappropriate rather than inadequate care: the prescription of antibiotic medications for patients suffering from viral upper respiratory tract infections. (Antibiotics are ineffective against viral infections, but 35% of visits for these conditions among Blue Cross enrollees generate antibiotic prescriptions). Here the denominator includes each episode of an office visit for viral upper respiratory infection and the numerator includes patients who filled an antibiotic prescription within 7 days of such visit.

Blue Cross does not include outcomes-based performance measures in its physician performance index currently, although it would consider so doing if cost and data limitations could be overcome.

For example, a shift from measuring whether hemoglobin tests were performed for diabetics to the hemoglobin levels found on those tests would require the insurer to have access not only to claims data (charge for administering the test) but to the electronic results files from the laboratories performing the tests. While laboratory values are available from the largest laboratory companies, many tests are performed in small local labs or in the physician's office and do not report results in comparable electronic formats. Patient-assessed measures of outcome (satisfaction with care, functional ability) would require surveying multiple patients per physician per year, which would be excessively costly for any one insurer to conduct alone. Cooperative approaches towards patient surveying have been developed in the HMO context but have not been extended to the PPO product, which is subject to much less regulation than the HMO counterpart in the U.S.

The pharmaceutical performance measures used by Blue Cross of California includes six classes of medications for which generic substitutes are commonly available (antibiotics, anti-depressants, hyperlipidemia medications, ulcer medications, hypertension medications, anti-inflammatory agents). Each physician's rate of generic prescribing (generic prescriptions divided by all prescriptions) within each therapeutic class is compared against rates for all physicians within the same specialty to obtain a percentile ranking, which then are averaged across the six therapeutic classes. This overall percentile ranking is multiplied against the total "points" (weighting) allocated to pharmaceutical performance in the performance index (40 points) to calculate each physician's score on this class of measures.

Structural measures used in the Blue Cross of California pay-for-performance system include specialty board certification; practice open to new patients; number of years physician has contracted with Blue Cross; whether physician participates in other Blue Cross products (e.g., commercial or Medicaid HMO); the proportion of claims submitted electronically rather than in paper form; participation in an internet-based communication interface between Blue Cross

and physician offices concerning eligibility, benefits, and claims status; and prompt response to the insurer's credentialing process. Many of these structural measures reward physician cooperation with the business objectives of the insurer (broad choice of physician is an important component of product value in the eyes of purchasers; electronic data interchange reduces administrative costs) rather than measuring quality or efficiency in any direct sense.

3.4.2 Integrated HealthCare Association

The Integrated HealthCare Association (IHA), a non-profit organization dedicated to fostering cooperation among health insurers, physician organizations, and hospital systems in California, has developed a pay-for-performance initiative at the level of the physician organization (multi-specialty medical groups and IPAs) rather than individual physicians. The IHA initiative grew out of frustration among physician entities concerning the administrative burdens of complying with different performance measurement and incentive programs for each of the HMO plans with which they contract. The IHA has developed a consistent set of performance measures, covering clinical quality, patient satisfaction, and information technology capabilities, that have been adapted by all the major HMO plans in the state (WellPoint/BCC, PacifiCare, HealthNet, Blue Shield of California, Aetna, CIGNA) with the exception of the Kaiser-Permanente system (which has its own exclusive plan-provider relationships). Each of the health insurers decides autonomously on the level of payment it will attach to each of the commonly agreed metrics (this avoids problems with violation of anti-trust law, which prohibits cooperation among purchasers on pricing). The overall performance index weights clinical measures at 50%, patient satisfaction measures at 40%, and information technology capabilities at 10%. The performance-based payment is structured as a percentage or dollar supplement to the per-member-per-month capitation by which HMOs pay physician organizations in California. Groups are eligible for up to a 5% supplement, with variation by health plan, and with the pool of funds budgeted for the program growing as the health insurance plans grow more confident

of the validity of the administrative data. In 2003, approximately $50 million was disbursed under the IHA program; this is expected to rise to $150 million in 2004.

Process measures of quality included in the IHA pay-for-performance initiative center around the percentage of patients who receive appropriate tests and drugs within the appropriate time interval. Examples include the percentage of children (by age category) who receive various vaccines (e.g., MMR, influenza, hepatitis B, chicken pox); percentage of women aged 50-69 receiving mammogram within two-year period; percentage of women aged 18-64 who receive pap test at least once in three years; percentage of patients with persistent asthma (by age category) who receive prescription for inhaled corticosteroids; percentage of patients with acute myocardial infarction or undergoing cardiac surgery or angioplasty who are screened for LDL levels; and percentage of diabetics screened for hemoglobin levels. In the second year of the program (2004-05), performance on LDL and hemoglobin monitoring is being extended from whether the tests are performed to the levels reported on the tests. Rewarding the level of test results moves money to the best performing groups (who may not be able to substantially improve performance due to ceiling effects), while rewarding changes in levels moves money to the most-improved groups (who might have had poor performance at baseline).

Patient satisfaction is measured for HMO enrollees in California through a standardized survey instrument used for all health plans, and hence the IHA initiative can include patient self-assessed measures of satisfaction with accessibility and quality of care. The survey components adopted by the IHA include satisfaction with access to specialty care (referrals), general access to care, communication with physicians, and global satisfaction with care received.

The measures of information technology among medical groups and IPAs include administrative and clinical capabilities. The administrative measures focus on the ability of the physician organization to integrate electronic data sets, including physician

visits and claims, laboratory claims or results, prescriptions, hospital admissions or emergency department visits, radiology claims or findings, clinical findings such as blood pressure or substance use (tobacco, alcohol), and the ability to report at the patient level to individual physicians. The clinical support measures focus on the use of electronic clinical information in the physician's office at the point of care (e.g., not for billing). Examples of clinical information use at the point of care include electronic checking of drug-drug interactions at time of prescribing; electronic access to clinical notes from other physicians; electronic access to clinical findings such as blood pressure or substance abuse; physician receipt of electronic reminders before patient visit (e.g., for needed vaccinations, cancer screenings, aspirin and ACE inhibitor use for diabetics); patient ability to email with physicians; and primary care physician ability to email with referral specialists.

3.4.3 Assurance Maladie

Assurance Maladie, the health insurance component of the French system of social security and, as such, the dominant purchaser of medical services in the nation, historically maintained a fee-for-service method of physician payment negotiated with the various generalist and specialist physician associations (*Syndicats*). In recent year Assurance Maladie has experimented with performance-based contractual relationships with both physician associations and individual physicians in the hopes of stimulating practice efficiency, cost-effective use of services, and a more appropriate geographic distribution of the physician workforce. In 2002 the health insurance system received legislative authorization to reward collective and individual performance through agreements at the national, regional, and local levels. Legislation in 2004 mandated that the content of performance-based contracts be reviewed by the ANAES, the national agency with responsibility for professional accreditation and evaluation, in order to forestall contracts that focus solely on cost control without a component focused on quality improvement.

Contracts between Assurance Maladie and the physician associations, defined in terms of specialty under the rubric of Effective Care Agreements (AcBUS: *Accords de Bon Usage des Soins*), are binding on all physician members of the associations. Effective Care Agreements target clinical services where there is evidence of resource misallocation and hence the potential for economic savings if average performance can be brought up to the best prevailing standards. The financial structure of these contracts is gainsharing, with individual physicians eligible for monetary (typically flat rate payments, sometimes fee-for-service augmentations) or non-monetary (free equipment or advanced training for their personal practices) if the collective pattern of care changes and generates savings. For example, a contract between Assurance Maladie and a generalist physician association targeted the excessive prescription of broad-spectrum antibiotics and generated a 17% reduction in penicillin volume in the first year. Individual physicians did not receive financial reward proportional to their individual change in prescription patterns but were eligible to receive diagnostic testing equipment for their practices. In another example, Assurance Maladie contracted with the generalist association to reduce utilization of home care visits not necessary given patients' degree of autonomy (based on age, disability, mental health). The 18% savings to Assurance Maladie in the first year were used to raise the case rate paid for home visits that were in accord with guidelines issued by ANAES and the Ministry of Health. Other contracts are focused on appropriate use of dental surgical materials, prenatal care by nurse midwives, diagnosis of renal insufficiency, coordination of medications for elderly patients with multiple conditions, choice of medical transport services, appropriate prescription of psychotropic medications, appropriate prescription of non-steroidal anti-inflammatory medications, appropriate prescription of fertility medications, and appropriate processes of care for morbid obesity.

In addition to collective agreements that mandate individual participation, Assurance Maladie is signing Professional Practices Agreements (*Contrats de Bonne Pratique; Contrats de Pratique Professionelle*) where participation by physician members of

the signing associations in on an individual and voluntary basis. Typically these contracts permit the individual to commit to a higher level of performance than that mandated for all professionals in the AcBUS agreements. For example, one Professional Practices Agreement commits signing specialists to maintain an appropriate volume of procedures, coordinate post-surgical follow-up with other clinicians, and participate in patient education programs. Physicians participating in these Professional Practices Agreements are eligible for payments supplemental to the basic fee-for-service reimbursement from Assurance Maladie. Another set of contracts permits surgeon members of the major specialist associations to commit to minimize pre- and post-surgical complications, reduce hospital-acquired infections, enhance reporting of care process and outcome, and coordinated post-treatment clinical follow-up. Surgeons are paid on a flat rate basis according to each practitioner's overall level of activity (not linked directly to number of procedures); in some instances Assurance Maladie with pay part of the physician's medical malpractice insurance costs.

3.5 Conclusion

For much of the 20th century the primary social goal with respect to payment for physician services was to promote patient access to care and the improvement of quality through rapid diffusion of new clinical interventions. Fee-for-service directly supported the achievement of these goals and offered the supplementary benefit of compatibility with an organizational framework of small, autonomous physician practices. As social concerns shifted from promoting more services to controlling costs and from supporting physician autonomy to encouraging adherence to evidence-based protocols, however, fee-for-service lost its luster and found itself blamed for many of the system's ills. Experiments with capitation, which in principle rewards the outputs of rather than the inputs to physician services, transferred excessive risk to physicians and thereby created incentives for under-treatment and the avoidance of particularly sick patients. The contemporary experimentation in methods of

physician compensation stems from disillusion with both fee-for-service and capitation, in their pure forms, and seeks to fashion hybrid mechanisms that incorporate elements of both prospective and retrospective payment and of both input and output measures of performance. These pay-for-performance initiatives offer beneficial insights into the manner by which incentives influence behavior but are limited by two continuing challenges to the efficiency and quality of physician services: the difficulties in measuring performance and the weaknesses in physician practice organization.

Measurement difficulties are pervasive within medicine. The quantity of inputs is poorly linked to the quality of outputs, evidence-based measures of efficacy are available for only a few discrete interventions, and patient-derived measures of satisfaction and functional ability are only weakly related to specific interventions by specific physicians. The contemporary explosion of interest in quality measurement, coupled with the rapid diffusion of information technologies, offers new hope that the basis of physician payment can be extended from the quantity of services rendered to include the quality of service, measured in terms of inputs (process measures of appropriate care) and outputs (outcome measures of health status improvement).

The organizational challenges facing physician payment reforms stem from the multi-agent, multi-task nature of medicine, especially for patients with chronic conditions. To the extent performance is measured and rewarded at the level of the individual physician, no incentives are provided for the coordination of care across caregivers and institutional settings. To the extent performance is measured and rewarded at the level of the physician group, incentives are provided for coordination, but only at the price of weakening productivity and performance incentives for individual physicians. Each physician in the group receives only a small portion of the reward for his or her good behavior, and must struggle against the temptation to coast on the efforts of colleagues and dissemble responsibility and blame to the collectivity. Groups must develop internal mechanisms for linking group-level payments to individual physicians or forgo the incentive benefits of pay-for-performance.

There are two easy and one hard prediction that can be made with respect to the trajectory of physician payment methods. First, given the widespread discontent with both fee-for-service and capitation, experimentation with performance-based hybrid methods will continue and probably accelerate. Second, given the important linkage between methods of payment, on the one hand, and methods of performance measurement and physician organization, on the other, the experiments in physician payment will vary substantially across the institutional features that differentiate national health care systems. The third, and much more difficult, prediction concerns whether or not the contemporary diversity in payment experiments will converge to a standard approach that achieves the quasi-dominant status long held by fee-for-service. The highly localized nature of innovation and the continued salience of physician autonomy and patient choice tip the scale towards continued diversity, while the consolidation of insurers under governmental and market pressures for cost containment would appear to favor a uniform approach. The United States and France, nations with historically quite different health insurance systems, are engaged in similar processes of experimentation because they are faced with similar challenges in epidemiology, medical technology, and health care finance. For the moment, it appears that convergence dominates divergence, as both focus on payment methods that blend prospective and retrospective elements while seeking to improve each system's ability to measure performance. In the future, however, each nation's method of paying its physicians must remain compatible with its method of organizing physician practice and its structures of regulatory law. In these larger domains one cannot but be impressed by the persistence of difference and by the continuity of each system with its own institutional legacy.

References

Atlantic Information Services. Case Studies of Physician Compensation in Medical Groups. Washington D.C.: Atlantic Information Services, 2003.

Baker G. The Use of Performance Measures in Incentive Contracting. American Economic Review 2000; 90(2):415-420.

Berwick DM. Payment by Capitation and the Quality of Care. New England Journal of Medicine 1996; 335(16):1227-31.

Emery D. (Editor). Global Fees for Episodes of Care. New York: McGraw Hill (HealthCare Financial Management Association), 1999.

Gibbons R. Incentives in Organizations. Journal of Economic Perspectives 1998; 12(4):115-132.

Gibbons R. Piece Rate Incentive Schemes. Journal of Labor Economics 1987; 5(4):413-429.

Hansen DG. Worker Performance and Group Incentives: A Case Study. Industrial and Labor Relations Review 1997; 51(1):37-49.

Holmstrom B and P Milgrom. Multi-Task Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. Journal of Law, Economics, and Organiztion 1991; 7:24-52.

Kandel E and EP Lazear. Peer Pressure and Partnerships. Journal of Political Economy 1992; 100(4):801-817.

Lazear EP. The Power of Incentives. American Economic Review 2000; 90(2):410-414.

Milgrom P and J Roberts. An Economic Approach to Influence Activities in Organizations. American Journal of Sociology 1988; 94 (supplement):154-179.

Prendergast C. The Provision of Incentives in Firms. Journal of Economic Literature 1999; 37(1):7-63.

Prendergast C. The Tenuous Trade-Off between Risk and Incentives. Journal of Political Economy 2002; 110(5):1071-1102.

Robinson JC. Blended Payment Methods in Physician Organizations under Managed Care. JAMA 1999; 282(13):1258-1263.

Robinson JC. Physician Organization in California: Crisis and Opportunity. Health Affairs 2001; 20(4):81-96.

Robinson JC. Theory and Practice in the Design of Physician Payment Incentives. Milbank Quarterly 2001; 79(2):149-177.

Robinson JC, SM Shortell, R Li, LP Casalino, T Rundall. The Alignment and Blending of Payment Incentives within Physician Organizations. Health Services Research 2004;39(5):1589-1606.

Rosenthal MB, RG Frank, JL Buchanan, and AM Epstein. Transmission of Financial Incentives to Physicians by Intermediary Organizations in California. Health Affairs 2002; 21(4):197-205.

Whyte WF. Money and Motivation: An Analysis of Incentives in Industry. New York: Harper and Brothers, 1955.

Discussion

Ismo Linnosmaa*

Department of Health Policy and Management,
University of Kuopio

Physicians have several roles in health care. According to Kuhn (2003), primary care physicians produce care by combining services they provide themselves with secondary care services and pharmaceuticals. Gaynor (1994) puts more emphasis on the role of physicians as specialists. According to Gaynor, physicians diagnose and advise their patients to choose most fit treatment alternatives available. In all these roles, physicians' decisions have an influence on health outcomes and costs of health care which typically have been considered as main performance measures of the health care sector. For this reason it is important to understand physician behaviour and decision-making more generally and the implications of financial payment schemes on physician behaviour in particular.

Robinson and Megerlin provide comprehensive discussion on theoretical implications of fee-for-service and capitation payments. Fee-for-service payment is a retrospective compensation scheme which links the physician remuneration to the quantity of health care services provided. Given that the piece-rate is set sufficiently high and exceeds marginal cost of producing services, fee-for-service scheme induces physicians to increase the quantity of services provided and tends to lead to excessive consumption of health care services. As a prospective compensation scheme, capitation payment creates

* email: ismo.linnosmaa@uku.fi

incentives to cut costs because this results in a higher monetary payoff. Lower costs may be obtained through more efficient organization of service production or lower quantity of services. If the latter case capitation payment may cause quantity of health care to be too low from the social point of view.

Empirical evidence on physician payment schemes supports these theoretical predictions. Empirical evidence indicates that, in comparison to capitation payment, fee-for-service payment leads to a more extensive use of diagnostic and curative services and lower levels of prescribing and referrals (see Scott, 2000). Capitation payment creates incentives to minimize effort at own practice while fee-for-service payment induces physicians to increase the use of services at own practice. Empirical evidence (Gosden et al., 1999) also suggests that salaried (also a prospective payment) physicians provide lower quantities of care, when measured using quantity of tests and the number of patient consultations, than physicians working under fee-for-service payment arrangements.

Although prospective capitation payment may induce efficient production of physician services, it may also have adverse consequences. One such consequence, also discussed in the paper, is patient selection. Capitation payment reimburses physicians same amount for each visiting patient and creates incentives to avoid high cost patients. Dumping refers to situation in which the physician refuses to treat costliest patients (Ma, 1994). Ellis (1998) shows that prospective payment may indeed lead to patient dumping and, furthermore, providers may have an incentive to provide less than the efficient amount of services to severely ill patients (called skimping). Magnitudes of dumping and skimping are reduced as there is a shift from prospective payment to a mixed system combining prospective payment and retrospective cost-reimbursement. Third possible implication of capitation payment discussed in the literature is cream-skimming (see Ellis, 1998). In this case the payment scheme induces the provider to supply excessive amount of services to patients with low severity of illness.

A shift from the retrospective payment mechanism to the capitation payment mechanism can improve total welfare if it induces physicians to produce the same amount of services at lower cost. Patient dumping, on the other hand, might also reduce total welfare. If the marginal health benefit of physician services is sufficiently high and exceeds the marginal cost of producing physician services also for high cost patients and dumping occurs, total welfare is lower than it would be if dumped patients were treated. Capitation payment might then increase or decrease total welfare depending on the magnitudes of its efficiency and selection effects.

Issues of selection have been studied empirically in the context of insurance markets, where adverse selection of consumers has been shown to exist (see Van de Wen and Ellis, 2000). Ellis and McGuire (1996) provide evidence that a shift from the cost-based reimbursement mechanism to the prospective payment system have led to patient selection also in the US hospitals. It seems, however, that there is less empirical research and evidence on patient selection in the provision of physician services (Newhouse, 1996).

The paper also discusses shortly the implications of capitation payment on quality of physician services, which are less clear-cut though. If the production of quality is costly[1], a shift from the fee-for-service remuneration system to capitation payment may induce the physician to reduce the quality of physician services. If the demand for physician services depends on quality levels of physician services and the market for physician services is competitive, individual physicians might increase quality to attract new patients. Hence, the overall impact of capitation payment on the quality of physician services is ambiguous and depends on the degree of competition in the market for physician services (Kuhn, 2003).

It is also important to ask how an optimal payment scheme, which would account for efficiency and selection, is characterized. Several

1 Quality can be costly for the physician firm because it increases the average cost of treating patients or the production of quality takes more of the physician's time in which case opportunity costs of physician time may be marginally higher.

papers in the literature (see e.g. Newhouse, 1996, Ma, 1994, Ellis, 1998) suggest that the optimal scheme is a mixed scheme containing elements of prospective and retrospective payment schemes. As the authors demonstrate, physician remuneration in the US is based on a mixture of payment mechanisms. Total compensation of the health centre physicians in Finland consists of salary (60%), fee-for-service (15%), capitation (20%) and local allowances (5%) (Järvelä, 2002), which also demonstrates the use of mixed payment systems in the practice of health care. Question that remains is what fractions of prospective and retrospective payments are optimal. Answering this question would require empirical research. One should also mention that another approach to deal with the trade-off between efficiency and patient selection is to adjust prospective capitation payments according to risk types of patients (Van de Ven and Ellis, 2000).

The paper describes in detail how US physicians are remunerated on the basis of quality performance. Interesting policy issue here is, of course, what fractions of the total compensation are prospective (capitation payment or salary) and linked to the quality performance. Principal-agent theory provides some answers to this question. Holmstöm and Milgrom (1987) show that under the exponential utility function and a linear and additively separable production function, the optimal payment scheme is linear and takes the form $w = \alpha + \beta q$, where w measures physician compensation, q is physician's quality performance, α stands for prospective payment and β is the incentive parameter. Payment mechanism provides high-powered (low-powered) incentives for the production of quality when the incentive parameter is high (low) and the prospective payment is low (high). Theory suggests that it is optimal to provide low-powered incentives in circumstances in which the variance of quality performance is high and/or the physician's risk-aversion is high and/or the marginal cost of effort for the physician is high.

This can be applied to a situation in which quality is contractible. Although it would be desirable to tie performance payment to health outcome measures, in practice physician remuneration is linked to the quantity of health care provided. A standard explanation

appearing in the theoretical literature for this is that health is not contractible because the *ex post* costs of verifying health states are too high. Ma and McGuire (1997) go even further and argue that neither the quantity of health services is contractible if it is based on *reported* quantity of services. This is not without consequences. The fact that the quantity is reported complicates the derivation of the optimal contract because the regulator has to design the optimal contract under additional constraint which requires the physician to report quantities truthfully.

Robinson and Megerlin cover main theoretical issues related to physician incentive payments in health care and provide several descriptive and illuminating examples on the use of performance payment schemes in the USA and France. On the basis of the article and the material it provides it is looks like there is still plenty of room for research on physician payment schemes in health care.

References

Ellis, R.P., 1998, Creaming, skimping and dumping: provider competition and the intensive and extensive margins, Journal of Health Economics, 17, 537-555.

Ellis, R.P. and McGuire. T.G., 1996, Hospital response to prospective payment: Moral hazard, selection and practice style effects, Journal of Health Economics, 15, 257-277.

Gaynor, M., 1994, Issues in the Industrial Organization of the market for physician services, Journal of Economics & Management Strategy, 3, 211-255.

Gosden T, Pedersen, L and Torgerson, D, 1999, How should we pay doctors? A systematic review of salary payments and their effect on doctor behaviour, Quarterly Journal of Medicine, 92, 47-55.

Holmström, B. and Milgrom, P., 1987, Aggregation and linearity in the provision of intertemporal incentives, Econometrica, 55, 303-328.

Järvelä, J., 2002, Health care systems in transition: Finland, European Observatory on Health Care Systems, WHO.

Kuhn, M., 2003, Quality in primary care: Economic approaches to analysing quality-related physician behaviour, Office of Health Economics, London.

Newhouse, J.P., 1996, Reimbursing health plans and health providers: Efficiency in production versus selection, Journal of Economic Literature, 34, 1236-1263.

Ma, C.A., 1994, Health care payment systems: cost and quality incentives, Journal of Economics & Management Strategy, 3, 93-112.

Ma, C.A. and McGuire, T.G., 1997, Optimal health insurance and provider payment, The American Economic Review, 87, 685-704.

Scott, A, 2000, Economics of general practice, in Culyer A.J. and Newhouse, J.P., Handbook of Health Economics, Elsevier Science B.V., Amsterdam.

Van de Ven, W and Ellis, R.P., 2000, Risk adjustment in competitive health plans in Culyer A.J. and Newhouse, J. (eds.) Handbook of Health Economics, Elsevier Science B.V., Amsterdam.

Altruism and Incentives in Public and Private Health Care*

4

Ching-to Albert Ma

Department of Economics, Boston University

4.1 Introduction

In many countries health care is by provided by firms in the public and private sectors. The effect of interaction between the public and private sectors on physician incentives is the subject of this chapter. Providers of health care in public and private sectors are often subject to different incentives. While the market mechanism is expected to work in the private market, the public system is typically characterized by simple and low-powered incentives. For the public sector, a conventional model, relying on straightforward utility or profit-maximizing preferences for economic agents, is inappropriate. Such a model would predict uniformly poor service quality and work effort there. This prediction is certainly inconsistent with casual observations; despite the lack of incentives, tasks and services are still being performed in the public sector.

The key hypothesis of this paper is that some healthcare providers are either sincere or have altruistic preferences. This deviation from the conventional wisdom may be less readily appreciated by academic

*My thanks to Hans Keiding and Klaus Kultti for their comments. Many ideas in this paper have been developed jointly with Gary Biglaiser of the University of North Carolina, and Philippe Choné of CREST-LEI, Paris.

economists than practitioners. Academic economists tend to believe that simple, consistent maximizing behavior is enough to study most social institutions. I believe that it is a misguided hypothesis for the health sector. It is indeed true that many models predict that maximizing behaviors may actually lead to sincere and altruistic actions. They typically push the "rational" preferences one step higher into the hierarchy of analysis; agents behave sincerely because of their fear for punishment or their expectation of future rewards. I tend to believe that this method obscures rather than enlightens. For policy matters, it is more important to study the practical implications of sincere and altruistic behaviors rather than simply to push the question one step back.

In this paper, I consider models in which healthcare providers may be sincere and altruistic. I assume that physicians consist of heterogenous groups, of which some are more sincere and altruistic than others. I study how a public system can exploit this heterogeneity. Who among the heterogenous physicians will work in the public system, who in the private market? Healthcare reforms seek to enhance efficiency, reduce cost, and maintain access and quality. Healthcare providers play the crucial role in this discussion. How they behave and react against changes in the public and private systems drive many of the conclusions in an analysis of any reform. In this paper, some physicians behave like the conventional "economic man," maximizing their utilities or profits. Others either do not exploit their private information or supply services with good qualities even when monitoring and incentives are absent.

My basic conclusion is that public policy should steer sincere and altruistic physicians to work in the public system, while the marketplace disciplines those who seek personal gains. It is a straightforward conclusion due to the use of low-powered incentives in the public system. Allowing the self-interest seeking providers the opportunity of higher profits in the private sector actually alleviates the inefficiency in the public sector.

I present two models. In the first, the quality of health services in the public system is unmonitored. There are sincere physicians who will supply services at good qualities even when there is no explicit incentive. The profit-maximizing physicians will supply services at a minimal quality. The public system may introduce a referral system, allowing physicians to transfer patients in the public system to their own private practice. The profit-maximizing physicians will exploit this market opportunity when they and their patients can mutually gain. The referral system may save costs in the public sector because the government does not pay for the costs of care of patients referred to the private market.

In the second model, physicians are altruistic, their preferences being combinations of their profits and patient benefits. There is a continuum of altruistic characteristics. Only the physicians know their preferences; the government or firms in the private sector do not. I first present the optimal mechanism that firms in the private sector will offer to these altruistic physicians, who value the services provided to their patients. The optimal mechanism respects the physicians' private information, distorting quantities as well as giving information rent to physicians.

I then let the public sector offer low-powered incentives to attract some of the physicians into public service. In effect, the public system selects the more altruistic physicians, who value the services provided to their patients. This has also the effect of changing the profile of characteristics of physicians remaining in the private sector. The overall effect may tend to reduce information cost in the private sector.

Some papers have studied dual job incentives in the health sector, see Rickman and McGuire (1999), and Gonzalez (2004). The paper by Gonzalez studies incentives of excessive treatment in the public sector when physicians there may signal their abilities to the private market. The distortion for excessive treatment because a physician performing well in the public system may generate a higher private demand. Other papers have looked at that same issue generally, see Che

(1995) and Lewis and Sappington (1989) for example. Che considers the incentives to invest in human capital when a regulator can seek post-tenure appointment at a regulated firm. Lewis and Sappington studied incentive designs when the agent's interests are in many environments. Papers on the multi-task principal-agent problem deal with similar issues, see Holmstrom and Milgrom (1994) for a general model and Ma (1994) for a study on the health sector. The general approach assesses the effectiveness of incentive mechanisms when agents have a lot of instruments available, but they adopt the usual assumption that agents maximize their own personal gains.

Many papers have examined the effect of the private sector on waiting lists in the public sector. Iversen (1997) considers a dynamic model of rationing by waiting lists; see also Barros and Olivella (2002). Besley, Hall, and Preston (1998) show that waiting lists in the UK National Health Services are positively associated with private insurance. Ma (2003) considers cost incentives when the public system rations supplies.

Recent advances in economic theory have incorporated ideas from the psychology literature. My hypothesis of heterogenous behaviors regarding profit and altruistic acts among physicians builds on this literature. Rabin (1993) studies a formal game-theoretic model on fairness. In Alger and Ma (2003) some physicians reveal information truthfully despite the lack of incentives to do so. Jack (2004) uses a model of altruistic providers to study cost and quality incentives. Frank (1996) argues that some workers are willing to give up high wages to serve in companies that provide social services. Finally, Besley and Ghathak (2003) let firms adopt social missions to attract workers who prefer to work in these firms. Brekke, Kverndokk and Nyborg (2003) propose a model of moral motivation, which is supported by agents' preferences for following an endogenous social norm. The next section presents the two models. The following two then contain the analysis. Concluding remarks are in the last section.

4.2 Models of Dual Job Incentives and Physician Agency

I present in this section two models of physician behaviors. I call the first one a dual job model; the second, a physician agency model. Common to both models is a set of consumers; they are indexed by a parameter α, which is positive and follows a distribution F with density f on the positive support $\left[\underline{\alpha}, \overline{\alpha}\right]$. The parameter α measures a patient's valuation of medical services. The benefit that the consumer obtains when service q is provided by a physician is $\alpha B(q)$, where B is a strictly increasing and concave function. The distribution of α captures diverse preferences as well as illness severities. It is assumed that the value of α is only known to the patient and the physician, not to the insurer, regulator or government.

The cost of service q, borne by a physician, is given by the strictly increasing and convex function C [q]. The physician may receive reimbursement to cover this cost. The measure of medical service can be interpreted either as quantity or quality. I adopt an interpretation of quality for q in the dual job model, and an interpretation of quantity in the physician agency model. I use different assumptions on whether the cost of medical service is verifiable; details follow soon.

I write down a benchmark for efficiency. For a given value of α, the service that maximizes the net benefit is given by

$$q(\alpha) \equiv \arg\max_q \alpha B(q) - C(q),$$

or

$$\alpha B'(q(\alpha)) = C'(q(\alpha)).$$

I call this the first-best level of service. The first best refers to the situation where the value of α is common knowledge. The first-best service $q(\alpha)$ is increasing in α. The rest of the paper is concerned with asymmetric information, when α is only known to the patient and the physician.

There is a set of physiciains, and I normalize the total number of physicians so that each doctor treats one and only one patient. There are always more patients than physicians, so that no physician has to be idle. In the dual job model, there are only two types of physicians. The first type I call the dedicated doctors; these are physicians who follow recommendations and work in a sincere fashion. I will let the regulator or government issue an instruction on the provision of medical service in the public sector. So the dedicated doctors will simply follow this instruction; of course, the regulator will have to reimburse them for the costs. The second type of physicians are common economic agents who maximize their utility or profits. Let the fraction of dedicated doctors among all physicians be μ.

In the second model, there is a continuum of physician types. Each physician's preferences depend on a combination of his own profit and his patient's benefit. The precise weight in the combination of physician profit and patient benefit varies across the population of physicians. So some physicians have preferences that put more weights on patient benefits than others. I let the parameter β represent this weight; this parameter follows a distribution G with density g on the positive support $\left[\underline{\beta},\overline{\beta}\right]$. If a physician of type β is paid R after having supplied services q to a patient, then his preferences are given by $R - C(q) + \beta B(q)$. The physician gets to observe the patient's preference parameter α. Both α and β are the physician's private information.

There are two sectors: the public and the private. All consumers are insured and may get services from the public sector. I assume that consumers pay a constant user fee for services in the public sector, and this is normalized to zero. A private sector for medical services may also exist. Also, if the government permits it, a physician working in the public sector may participate in the private sector too.

In the public sector, high-powered incentive mechanisms are not used. As I demonstrate later, this does not imply that the public sector becomes nonviable. My hypothesis is that the public sector uses low-powered incentives to attract some types of providers.

The market mechanism in the private sector is assumed to facilitate the exchange between consumers and physicians, or to make available incentive mechanisms to solve problems due to asymmetric information. For the dual job model, in the private sector, consumers and physicians can freely contract on quality and payment. Most important is the physician-consumer coalition's ability to tailor quality q to the specific consumer preference parameter α. For the physician agency model, I assume that in the private sector an incentive mechanism can be used to extract the physician's private information of α and β.

In each of the two models, I hypothesize unconventional behaviors for healthcare providers. In the dual job model, the dedicated doctors behave in a nonstrategic way. In the physician agency model, doctors have altruistic preferences (although a doctor's degree of altruism is his private information). Both assumptions deviate from the usual and conventional one of pure profit or utility maximization. However, they are important for a study of public services. Strong incentive systems in the public sector are seldom found. Yet, in most countries projects and tasks in the public sector are being carried out. A conventional assumption of complete self-interest would imply that nothing can be achieved in the public sector. This is incompatible with casual observations.

I now describe the various specific assumptions on information, variable verification, and the extensive form for each model. The public system integrates insurance and provision. For insurance purposes, consumers do not bear the full marginal cost of service.

4.2.1 Dual Job Incentives

For the dual job model, I first consider a benchmark in which all physicians work in the public sector. Then the government has to decide if the physicians are allowed to work in the private sector. I assume that quality q is nonverifiable. So the government can only pay a fixed renumeration to a physician who works in the public system. Treating a patient in the public system results in the physician receiving a fixed wage, R. The government will also issue an instruction on the quality that should be supplied to patients in the public sector. Let this be q^r. The instruction is supposed to inform the physicians the level of quality that should be supplied, and the level of compensation R is set at $C(q^r)$, which covers the associated cost of the recommended quality.

The dedicated doctors and profit-maximizing physicians react differently against the fixed wage and quality recommendation. The dedicated doctors follow the instruction, supply services at quality $C(q^r)$, and accept the compensatory wages. The profit-maximizing physician supply services at a minimum quality, $\underline{q}$, because no explicit incentives are present in the public sector; these agents then net a profit of $C(q^r) - C(q)$.

If participation in the private market is allowed, a profit-maximizing physician can refer a patient from the public sector to the private sector. After a referral, the government will not have to pay the physician. In the private sector, the physician will bargain and contract with the patient. I assume that the bargaining is efficient. This, for example, is a property of the Nash bargaining solution. Efficiency implies that the patient will obtain quality $q(\alpha)$. I assume that the physician and the patient will then split the surplus by an appropriate transfer. For simplicity, I assume that the dedicated doctors will not participate in the private market. The purpose of this model is to demonstrate how opening up the private market will alleviate the inefficiency in the public sector, so I will not analyze explicitly change of behavior by the dedicated doctors.

4.2.2 Physician Altruism

For the physician agency model, I assume that q measures healthcare quantity, and is verifiable, and the payment R can be contingent on how much the physician has chosen to supply. I first consider a benchmark in which all physicians work in the private sector. The government then considers hiring some of these physicians to work in the public system.

The physician agency model is one of multi-dimensional uncertainty. The physician gets to observe the patient's valuation parameter α and his own degree of altruism β. I assume that a firm in the private sector will use an optimal mechanism. This takes the form of a direct revelation mechanism with truthful revelation being an equilibrium. A mechanism takes the form of a pair of functions: $(R(\alpha,\beta),q(\alpha,\beta))$. The physician (who knows the true values of α and β) is asked to make a report. If he reports values of α' and β' he is paid $R(\alpha',\beta')$ and required to supply quantity $q(\alpha',\beta')$. Truthful revelation as an equilibrium means that the functions R and q guarantee that misreporting values of α and β is suboptimal.

I require that the physician makes nonnegative profits. Although the physician's utility depends on both profits and consumer benefit, it is unlikely that he can sustain a monetary loss. This assumption is much more natural than a reservation utility constraint. I will present the optimal mechanism in the private sector. When the government attempts to hire physicians, it maintains its low-powered incentive system, simply offering a wage in exchange for the physician providing some quantities. The purpose of this model is to study whether more or less altruistic physicians find working in the public system more attractive.

4.2.3 Policies

I will regard permitting public physicians to work in the private sector, and expanding the public sector as policy choices. The question remains what guides the policy decisions. I use a utilitarian approach: the government is supposed to adopt a policy that maximizes the sum of consumer benefits less any payments made by consumers and the government. In practice a government's agenda may not be so entirely reflected by a utilitarian approach. Nevertheless, it is a useful welfare index to start with. For a more realistic discussion, I will mention distribution issues in the analysis.

4.2.4 Dedication and Dual Job Incentives

In this section, I study the dual job model. I begin with the benchmark when physicians in the public sector are unable to take up jobs in the private sector. A patient is randomly matched to a physician in the public sector. Let the government recommend a quality q^r and offers a payment of $R_. = C(q^r)$ to a physician for providing a treatment to a patient. The dedicated doctors will supply services at quality q^r; they earn zero profits. The profit-maximizing doctors will merely provide a service at minimum quality $\underline{q}$; they receive payment $C(q^r)$, hence each netting a profit $C(q^r) - C(q)$. The welfare index consists of the expected consumer (net) benefit less the payment made by the government

$$\mu E(\alpha)B(q^r) + (1-\mu)E(\alpha)B(\underline{q}) - C(q^r). \qquad (1)$$

where $E(\alpha)$ is the expected value of α. In my setup, the government uses a low-powered incentive system, paying all physicians a fixed wage. It relies on the goodwill of the dedicated doctors to maintain the recommended quality to some consumers. The payments to the profit-maximizing doctors are just transfers. From (1), one can characterize the optimal recommended quality by the first-order condition with respect to q^r: $\mu E(\alpha)B'(q^r) = C'(q^r)$. Because the payments to the profit-maximizing doctors do not yield any quality, the marginal benefit is only proportional to the fraction of dedicated doctors (μ).

Next, I consider the policy of allowing the physicians to participate in the private market. Referring a patient to a private practice is now possible. I assume that the dedicated doctor does not participate in the private market; only the profit-maximizing doctors do. I use a Nash bargaining solution to describe the outcome in the private sector. First, the disagreement point is one where a physician and the patient fail to agree on a referral. In this case, the patient's utility is $\alpha B(q)$, while the physician's utility is $C(q^r) - C(q)$. An agreement is a pair (p,q), where p is the price the consumer pays the physician and q the quality the physician has to provide. A Nash bargaining solution maximizes the joint surplus (above the disagreement) and specifies a price to split the surplus in some ratio, say s, $0 < s < 1$, between the two parties.

The quality in the Nash bargaining solution maximizes the joint surplus $\alpha B(q) - C(q)$; and the quality level is the first-best $q(\alpha)$. The physician and patient will be able to agree on a referral when their joint surplus is higher than in the disagreement point. This is the condition: $\alpha B(q(\alpha)) - C(q(\alpha)) \geq \alpha B(\underline{q}) - C(\underline{q}) - C(q^r)$, which is satisfied whenever α is above a certain threshold, say $\hat{\alpha}$. If α is below this threshold, the physician and patient cannot agree on a referral with mutual benefits; they stay within the public system.

In summary, if the consumer is matched with a dedicated physician, she receives treatment at quality q^r in the public sector. Otherwise, a referral is recommended by the physician if her valuation parameter α is sufficiently high. In that case, she receives the efficient quality $q(\alpha)$, and pays the physician a price, say $p(\alpha)$. If a referral is not recommended, the patient stays within the public system, receives the minimum quality $\underline{q}$ but pays nothing. I now write the welfare index in this regime:

$$\mu\left\{E(\alpha)B(q^r) - C(q^r)\right\} + (1-\mu)\left\{\int_{\underline{\alpha}}^{\hat{\alpha}}\left[\alpha B(\underline{q}) - C(q^r)\right]f(\alpha)d\alpha + \right.$$

$$\left.\int_{\hat{\alpha}}^{\overline{\alpha}}\left[\alpha B(q(\alpha)) - p(\alpha)\right]f(\alpha)d\alpha\right\}.$$

The first term refers to the welfare generated by the dedicated physicians. The second term is divided into two parts: the first refers to the net benefit provided by the profit-maximizing doctors in the public sector, while the second is the surplus from a referral to the private sector. To compare the welfare index under the referral regime with that in (1), I write the price in the private market as a markup over cost: $p(\alpha) \equiv C(q(\alpha)) + m(\alpha)$. So now I rewrite the welfare index as:

$$\mu\left\{E(\alpha)B(q^r) - C(q^r)\right\} +$$

$$(1-\mu)\left\{\int_{\underline{\alpha}}^{\hat{\alpha}}\left[\alpha B(\underline{q}) - C(q^r)\right]f(\alpha)d\alpha + \int_{\hat{\alpha}}^{\overline{\alpha}}\left[\alpha B(q(\alpha)) - C(q(\alpha)) - m(\alpha)\right]f(\alpha)d\alpha\right\},$$

which is further simplifed to

$$\mu E(\alpha)B(q^r)+(1-\mu)E(\alpha)B(\underline{q})-C(q^r)+$$

$$(1-\mu)\int_{\hat{\alpha}}^{\bar{\alpha}}\{[\alpha B(q(\alpha))-C(q(\alpha))-m(\alpha)]-[\alpha B(\underline{q})-C(q^r)]\}f(\alpha)d\alpha.$$

The welfare index in (2) embeds a number of properties. By the definition of Nash bargaining, $\alpha B(q(\alpha)) - C(q(\alpha)) - m(\alpha) > \alpha B(\underline{q})$, for all $\alpha > \hat{\alpha}$, so for the same recommended quality q^r, the welfare index is higher than (1) when the physicians are allowed to refer patients to the private market. There are two sources of the gain. The first is due to the efficiency of quality (as a function of the valuation parameter α) since the physician and the patient can contract on quality and price in the private market. Second, there is some cost saving: the last term $(1-\mu)(1-F(\hat{\alpha}))C(q^r)$ is the government's cost saving from not paying those physicians who refer patients.

The private market allows an opportunity for the profit-maximizing doctors to supply higher quality to those consumers with higher valuations. It also alleviates the incentive problem in the public sector. A profit-maximizing doctor forgoes the rent $C(q^r) - C(\underline{q})$in order to refer a patient to the private practice. The private sector serves as an efficient supply and a sorting mechanism.

Although the welfare index must be improved when referral is allowed, it does not imply that all consumers must be better off. This depends on whether the government actually changes the recommended quality level once referral is permitted. There are two counteracting effects for the government to consider. First, private market referral leads to some cost savings. The government's payment $C(q^r)$ is more likely given to a dedicated doctor. Hence there is a tendency for the government to raise the recommended quality.

Second, private market referral leads to qualities that are efficient relative to consumers' valuations. The recommended quality in the public sector applies to all consumers who seek services there; consumers with different valuations receive the same quality.

Some of the rigidity can be avoided when referrals are possible; consumers in the private sector receives qualities that are more in-line with their valuations. To exploit this market flexibility, the government may want to steer more consumers to the private sector. The way to achieve this is to lower the recommended quality level in the public sector; more consumers will then find the private sector attractive.

If the recommended quality actually increases because the cost savings are so significant, then all consumers become better off compared to the regime where referrals are impossible. Clearly those who stay behind in the public sector are better off. Those who decide to accept the referral must also find it beneficial too. If the recommended quality falls because the government wants to steer more consumers to the private market, then those consumers who stay behind in the public sector become worse off.

4.2.5 Robustness and Adverse Effects

The gain in welfare when the profit-maximizing physicians are allowed to refer patients is based on the trading opportunity in the private sector. This is a fundamental point. I have used the Nash bargaining solution, which implies an efficient allocation from a referral. This allows me to make the point in a succinct fashion, but the precise efficiency property in the private market is unimportant. What is important is that the opportunity for trading in the private sector does not require the patient or the physician to forgo any surplus in the public sector. For the Nash bargaining solution, this is captured by the disagreement point being defined as the allocation the patient and the physician would have achieved if referral were impossible.

The welfare property can be supported even when there is asymmetric information between the patient and the physician, or when some other market frictions exist. These obstacles will reduce the potential gain, as well as the volume of trade in the private market. However, in any equilibrium, if the physician succeeds in making a referral, both the physician and the patient must expect to gain from the decision. For referral to be a rational outcome, each party must expect to be better off than staying within the public sector. Market frictions affect how much each party gains, and how often referrals occur. Nevertheless, as long as there is some likelihood of a successful referral, the expected welfare must increase.

Adverse and unintended effects must be considered when the private market is opened.[1] I have assumed that when referral to the private market is allowed, this is the only new activity for profit-maximizing doctors. In practice, new activities or effects may occur. First, this may affect the behaviors of the dedicated doctors. The government enjoys the sincere behaviors of the dedicated doctors in the public system. If these doctors find the higher income from the private market attractive, they may abandon the public service. Morale may be affected, and incentive costs may increase.

Private market participation may further adversely affect profit-maximizing doctors' behavior in the public sector. They may decide to lower quality at the public sector even more. This let them devote more time and energy to private patients. Moreover, a lower quality at the public sector may make patients more willing to accept a referral suggestion. This deterioration of quality in the public sector typically affects the less well-off consumers.

There may be some countervailing incentives, however. A physician may have an incentive to perform well in the public system in order to signal to the private market his effort or ability (Gonzalez, 2004). The asymmetric information may come from physicians having different abilities; their performance may reveal some information. Signaling however may lead to other distortions. For example, in order to enhance his performance in the public sector, patients may be treated excessively. There is no reason, however, to expect any signaling incentive to exactly counteract incentives for other adverse reactions.

The supply of physicians in the public sector may be affected when the private market expands. Some physicians may choose to leave the public sector altogether. The departure of the profit maximizing physicians poses less of a problem for the public system. In any case, they are supposed to take up the profit opportunities there. The departure of the dedicated physicians is a definite negative for the public sector.

1 See Biglaiser and Ma (2003) for more detailed discussions.

Adverse effects due to the expansion of the private market may be mitigated by regulations. If the magnitude of the adverse effect is positively related to the profit opportunities in the referral market, then price ceiling regulation may be considered. This regulation simply says that a physician's referral fee cannot exceed a certain amount. In practice, this may be implemented by a limit on physician income. For example, a regulation may forbid a physician working in the public system to have more than 30% of his income derived from private work. The lower the ceiling, the less a physician can exploit the profit opportunities in the private sector. There is then less incentive to engage in activities that will further reduce quality in the public system.

4.3 Altruism and Physician Agency

In this section, I study the physician agency model. I begin with the optimal mechanism for physicians in the private sector. Then I study the impact of the public sector offering low-powered incentives to hire some of these physicians. The public and private sectors are in effect sorting physicians with different degrees of altruism. Recall that the altruistic physician's utility is given by $R - C(q) + \beta B(q)$. Firms are supposed to be competitive; they offer incentive mechanisms to physicians in order maximize consumers' expected utility, subject to incentive and nonnegative profit constraints for the physicians.

Recall that the mechanism $(R(\alpha,\beta),q(\alpha,\beta))$ must satisfy the truth-telling (or incentive compatibility) and nonnegative profit constraints. The incentive constraints are:

$$(\alpha,\beta) = \arg\max_{\alpha',\beta'} R(\alpha',\beta') - C(q(\alpha',\beta')) + \beta B(q(\alpha',\beta')).$$

The right-hand side expression is the utility a physician who has altruism parameter β and who is matched with a consumer with valuation parameter α. The incentive constraint says that the physician can do no better than reporting truthfully this information. The maximum or indirect utility is

$$\max_{\alpha',\beta'} R(\alpha',\beta') - C(q(\alpha',\beta')) + \beta B(q(\alpha',\beta')).$$

Because the physician's preferences are not directly dependent on the patient's valuation parameter α, the indirect utility function is only a function of β. Let the indirect utility be $U(\beta)$. Again, because the physician's preferences are independent of α, any incentive mechanism is generically independent of α, the patient's valuation.[2] In other words, the physician's private information about α cannot be extracted directly. So from now on I write a mechanism as $(R(\beta),q(\beta))$. I can now write the indirect utility function as $U(\beta) = R(\beta) - C(q(\beta)) + \beta B(q(\beta))$ By the Envelope Theorem, $U'(\beta) = C(q(\beta))$. Furthermore, because U is the maximum of affine linear functions of β, it is convex. So $U''(\beta) = B'(q(\beta))q'(\beta) \geq 0$; this implies that any quantity q in an incentive compatible mechanism must be nondecreasing in β. In fact incentive compatibility is equivalent to U being convex, and q being nondecreasing.

Next, I consider the nonnegative profit condition. Let $\pi(\beta) \equiv R(\beta) - C(q(\beta))$. I use the definition of the indirect utility function to obtain $\pi(\beta) \equiv U(\beta) - \beta B(q(\beta)) = U(\beta) - \beta U'(\beta)$. Differentiating the profit function, I get

$$\frac{d\pi}{d\beta} = \frac{d}{d\beta}\left[U(\beta) - \beta U'(\beta)\right] = -\beta U''(\beta) = -\beta B'(q(\beta))q'(\beta) \leq 0. \qquad (3)$$

2 The requirement for the genericity is due to the following. The indirect utility function U is the maximum of affine linear functions of β. So it must be convex in β, and hence almost everywhere differentiable. If $B(q(\alpha,\beta))$ were not independent of α, the differentiability of U would be violated.

The profit decreases as the degree of altruism increases, at a rate determined by the quantity increase rate; see (3). Both the quantity and profit schedules have to be compatible, the former being increasing while the latter decreasing, with (3) relating the rates they can change. If the quantity increases too rapidly, then the profit falls to zero and the nonnegative profit constraint binds. This then forces the quantity to become constant. Actually, if the profit constraint binds

at $\hat{\beta}$, then it binds for all $\beta > \hat{\beta}$, and as a result the quantity becomes constant too. Altruism and nonnegative profits may lead to a choice of a set of quantities that are insensitive to the variation of the altruism parameter when β is sufficiently high.

4.4 The Optimal Mechanism for Altruistic Physicians

I assume that firms in the private sector operate in a competitive environment. So they choose mechanisms to maximize consumers' expected utility. The objective function is

$$W \equiv \iint [\alpha B(q(\beta)) - R(\beta)]\, h(\alpha, \beta)\, d\alpha\, d\beta ,$$

where h is the joint density of α and β. Because the physician's private information about α cannot be extracted, I rewrite this objective fucntion as follows:

$$W \equiv \int [\alpha_m(\beta) B(q(\beta)) - R(\beta)]\, g(\beta)\, d\beta ,$$

where $\alpha_m(\beta)$ is the conditional mean of α given β. The optimal mechanism is one that maximizes this objective function subject to the incentive and nonnegative profit constraints. I will skip over the derivation of the optimal mechanism; see Chone and Ma (2004) for the technical details. It is, however, important to mention a key step. The optimal mechanism must consider a pooling regime. For a range of higher values of β the quantity becomes constant (no longer strictly increasing in β), and profit zero. For lower values of β, the nonnegative profit constraint does not bind, and quantity is strictly monotone increasing in β. So the objective function for the optimal mechanism can be written in two parts. For low values of β, it is a separating regime; physicians with different degrees of altruism receive different payments and provide different quantities. For high values of β, it is a pooling regime; physicians with different degrees of altruism receive the same payment and provide the same quantity. Moreover, physicians in the pooling regime earn zero profit.

I will use the following assumption: $\alpha_m(\beta) + \beta + G(\beta)/g(\beta)$ is continuous and nondecreasing in β. This assumption guarantees that the optimal quantity is indeed nondecreasing. Furthermore, I assume that the unconditional mean of α is less than $\alpha_m(\underline{\beta}) + \underline{\beta}$, which guarantees that the pooling regime does not extend to the entire support of β so that some separation remains.

The optimal quantity is depicted in Figure 1. The quantity is continuous, strictly increasing for β between $\underline{\beta}$ and $\hat{\beta}$ but constant between $\hat{\beta}$ and $\overline{\beta}$. Let the pooling quantity be denoted by $q(\hat{\beta})$. The physician profit is illustrated in Figure 2. The profit is continuous, strictly decreasing for β between $\underline{\beta}$ and $\hat{\beta}$, but remains at 0 between $\hat{\beta}$ and $\overline{\beta}$. For pooling regime, the indirect utility $U(\beta)$ is $\beta B(q(\hat{\beta}))$ since profit is zero.

Let me provide some intuition behind the properties of the optimal mechanism. The information about α cannot be extracted directly; it can only be inferred from the information of β. revealed by the physician. The conditional expectation $\alpha_m(\beta)$ then becomes a critical part in the determination of the optimal quantity. The hazard rate $G(\beta)/g(\beta)$ takes into account the information rent available to the physician due to his superior information. Furthermore, when the physician's nonnegative profit constraint does not bind, part of the information rent is derived from the quantity, which is distorted upward.

For small values of β, profits contribute more to the indirect utility. As quantity begins to increase along with β, the altruistic part $\beta B(q)$ contributes more to the indirect utility. Pooling must occur for high values of β. If it was separating for all values of β, then the quantity would be strictly increasing. Now reducing the quantity for the highest value of β would lead to a second-order loss in terms of quantity efficiency but a first-order gain in terms of information-rent saving. This implies that pooling must be part of the optimal schedule. The assumption that α is less than $\alpha_m(\underline{\beta}) + \underline{\beta}$ guarantees that pooling at the lowest value of β is not as attractive as extracting the information. In this case, the quantity schedule is the one in Figure 1.

Finally, the indirect utility is $U \equiv \pi(\beta) + \beta B(q(\beta))$. Let me assume here that the benefit function B is positive.[3] Figure 3 presents the (convex) indirect utility, the physician's payoff when both profit and altruistic components are taken into account. Here, although physicians who have lower values of β enjoy more profits, their maximum utility is actually lower, since there is less satisfaction from servicing their patients with the lower quantities. Furthermore, for higher values of β, the indirect utility is linear in β since profit is zero: $U = \beta B(q(\hat{\beta}))$.

3 The benefit function is an ordinal measure; its sign has no relevance for the design of the optimal mechanism. For the derivation of the optimal mechanism, only the marginal benefit B' is relevant. Nevertheless, the sign of the benefit function does a.ect the sign of the indirect utility, although, in a symmetric fashion, only the sign of U' (which is positive) has a bearing on the optimal mechanism.

4.4.1 Low-powered Incentives in the Public Sector

I have presented the optimal mechanism in the private sector. Now a government intends to expand its public health services, and must hire some physicians from the private market. I maintain the assumption that the incentive system in the public sector is low-powered. The payment mechanism is simply a fixed wage, in exchange for the physician providing some quantity. Let the quantity chosen by the government be $\tilde{q}$. The compensation for providing this quantity is the physician's cost $C(\tilde{q})$. Under this scheme, a physician with altruism parameter β obtains a utility $\beta B(\tilde{q})$.

When the public sector offers this package, how do physicians respond? A physician with altruism parameter β compares $U(\beta) = \pi(\beta) + \beta B(q(\beta))$ in the private sector with $\beta B(\tilde{q})$ in the public sectors. In Figure 4, I have drawn these two utilities as functions of β. The solid line is the same as in Figure 3 (where the indirect utility from the optimal mechanism is depicted), while the dotted line is $\beta B(\tilde{q})$. When $\tilde{q}$ is greater than $q(\hat{\beta})$, the utility function for the public sector intersects that for the private sector from below. Those physicians with a degree of altruism higher than $\tilde{\beta}$ will find it more attractive to work in the public sector.

The simple policy of asking physicians to provide in the public sector a fixed (and higher) quantity attracts those who are more altruistic. A policy of the public sector prescribing a higher level of health care than in the private sector is unusual; in most countries, the opposite is true. This is due to the restriction that physicians are paid only their costs of services. A low-powered incentive may also involve a salary, or a lump-sum payment, say T. In this case the utility from joining the public service becomes $T + \beta B(\tilde{q})$. Various combinations of T and $\tilde{q}$ can be considered. The salary component has the effect of raising the level of the dotted line in Figure 4. It is then possible for the government's quantity prescription to be reduced below the highest in the private sector. Figure 5 illustrates an example of this policy. There, the requested quantity in the public sector is lower than the highest in the private sector; in fact, the salary is not sufficiently high to attract those physicians who value consumer benefits the most.

The point is that quantity standards and salary schemes in the public sector attract different types of physicians. Depending on the specific kind of health services and providers, the government's offer of salary and standard of care may vary. Each combination sets up a different partition of physicians with different degrees of altruism across the public and private sectors. The impact does not stop there. When a subset of physicians leave for public service, firms in the private sector face a different population of physicians. For example, if the scheme in Figure 4 is used by the government, then physicians with lower degrees of altruism remain in the private sector. Formally, this is equivalent to a change in the distribution of the physician's altruism parameter. The support of β will have a subset at the top truncated, and the optimal mechanism will have to adjust for this change.

It is beyond the scope of this paper to study fully the equilibrium between the public and private sectors when each offers quantity and compensation schemes. Nevertheless, I speculate that private firms will be able to pay less information rent when the government hires some physicians. The incentive constraints may become more relaxed because of the smaller physician population.

4.5 Concluding Remarks

I have presented two models with heterogenous physician behaviors. In the first, some physicians provide health care with good qualities despite the lack of incentives. They work in the public sector. Other physicians behave in strategic ways. When the strategic physicians are allowed to refer patients from the public sector to their private practices, the government may save costs. Welfare increases as a result. I also consider adverse reactions when referrals are possible. They tend to reduce the welfare gain, but public policies that limit the profits to be made by strategic physicians may remedy these adverse effects.

In the second model, physicians possess private information about their own degree of altruism towards patients' benefits as well as

patients' valuation of health care. The optimal mechanism offered by competitive firms in the private sector is presented. I then consider how low-powered incentive schemes in the public sector may attract some of the physicians. I illustrate how the public sector may select those physicians with higher degrees of altruism. In both models I deviate from the conventional wisdom by assuming a variety of behaviors and preferences. I argue that the narrow optimizing behavior framework is inappropriate for the health sector. This is a significant departure. It incorporates a certain degree of realism. I do not in any way suggest that optimizing behaviors are unimportant, but argue that they should be only part of a spectrum of behaviors to be considered.

Figure 1. Optimal quantity

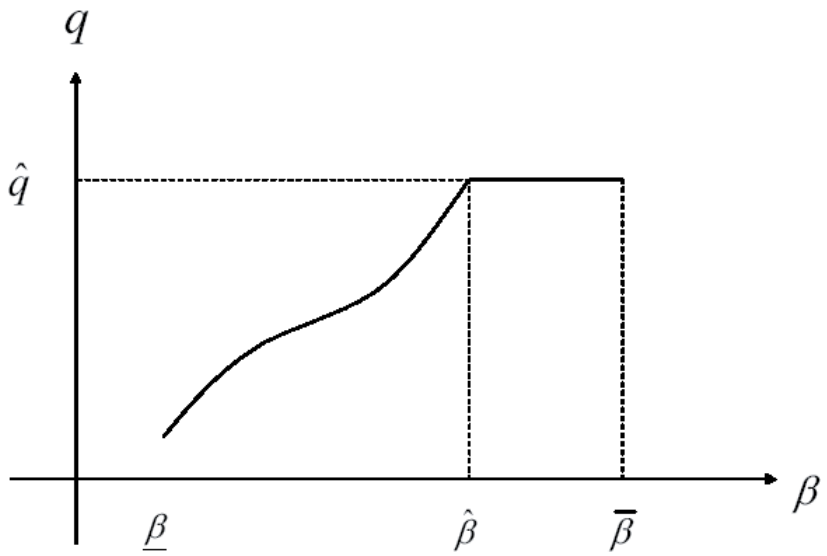


Figure 2. Physician profit

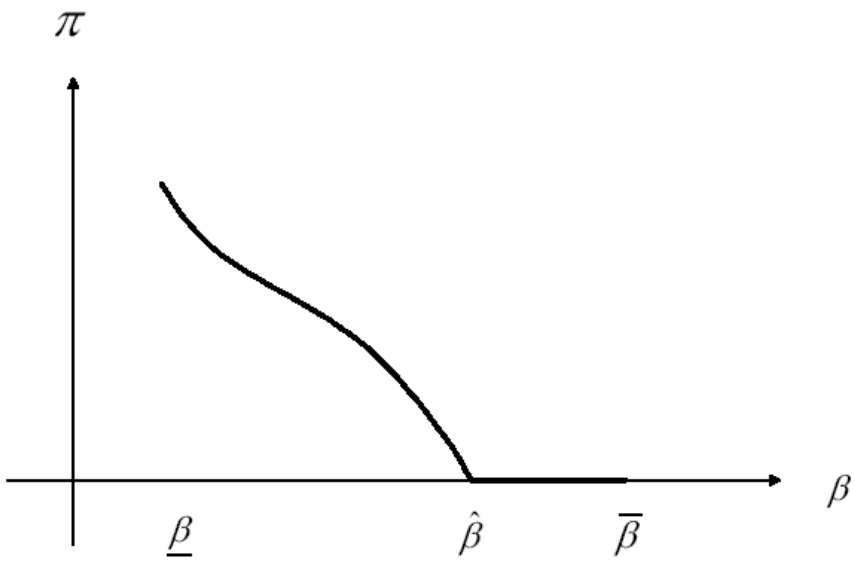


Figure 3. Indirect utility

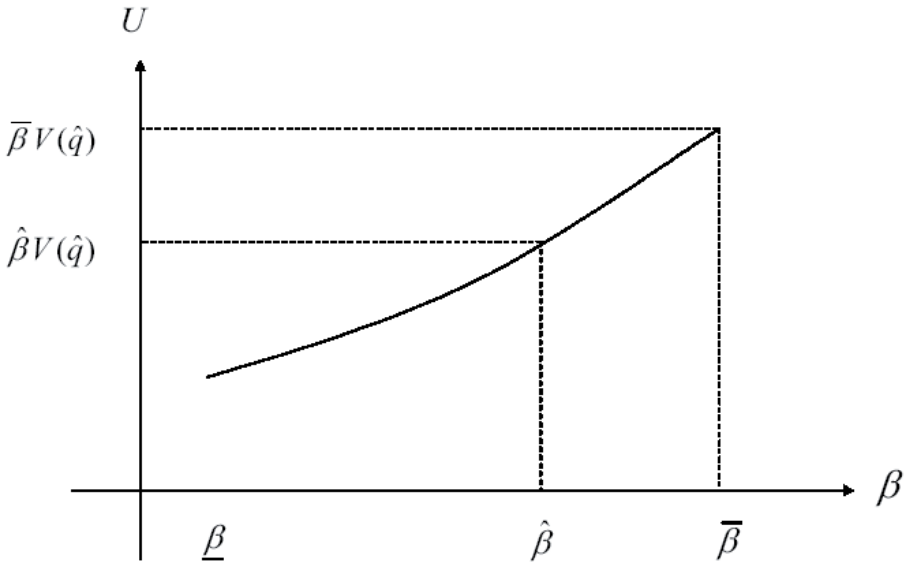


Figure 4. Indirect utilities for different sectors

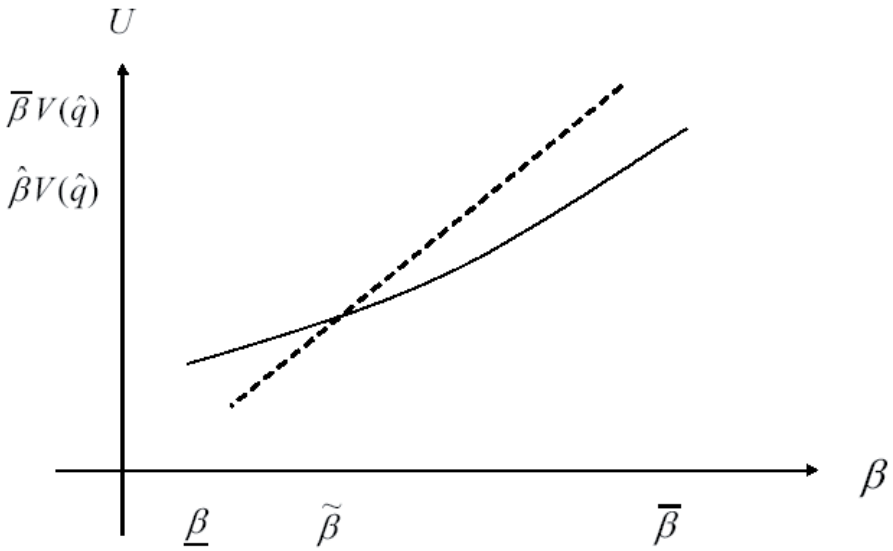
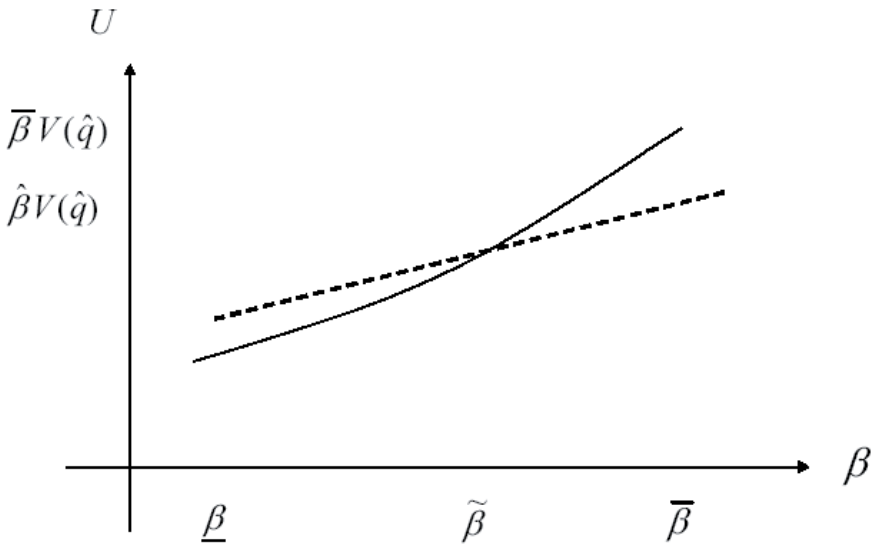


Figure 5. Indirect utility with salary in public sector



References

[1] Alger, I, and Ma, C.A. (2003), "Moral Hazard, Insurance, and Some Collusion," Journal of Economic Behavior and Organization, 50, pp. 225-247.

[2] Biglaiser, G. and Ma C.A. (2003), "Moonlighting: Public Service and Private Practice," Boston University working paper.

[3] Brekke, K.A., Kverndokk, S., and Nyborg, K. (2003) "An Economic Model of Moral Motivation," Journal of Public Economics, 87, pp 1967-1983.

[4] Barros, P. and Olivella, P. (2002), "Waiting Lists and Patient Selection," CEPR working paper.

[5] Besley, T. and Ghatak, M (2003) "Competition and Incentives with Motivated Agents," London School of Economics working paper.

[6] Besley, T., Hall, J., and Preston, I. (1998), "Private and Public Health Insurance in the UK," Journal of Public Economics, 42, pp. 491-497.

[7] Che, Y. (1995), "Revolving Doors and the Optimal Tolerance for Agency Collusion," Rand Journal of Economics, 26, pp. 378-397.

[8] Choné, P. and Ma, C.A. (2004) "Asymmetric Information from Physician Agency: Optimal Payment and Healthcare Quantity," Boston University working paper.

[9] Gonzalez, P. (2004), "Should Physicians' Dual Practice be Limited? An Incentive Approach," Health Economics, 13, pp 505-524.

[10] Frank, R. (1996), "What Price the Moral High Ground?" Southern Economic Journal, 63, pp. 1-17.

[11] Holmstrom, B., and Milgrom, P. (1994), "The Firm as an Incentive Scheme," American Economic Review, 84, pp. 972-991.

[12] Iversen, T. (1997), "The E.ect of a Private Sector on the Waiting Time in a National Health Service," Journal of Health Economics, 16, pp. 381-396.

[13] Jack, W. (2004) "Purchasing Health Care Services from Providers with Unknown Altruism," Georgetown University working paper.

[14] Lewis, T. and Sappington, D. (1989), "Countervailing Incentives in Agency Problems," Journal of Economic Theory, 49, pp294-313.

[15] Ma, C.A. (1994), "Health Care Payment Systems: Cost and Quality Incentives," Journal of Economics & Management Strategy, 3, pp. 93-112.

[16] Ma, C.A. (2003), "Public Rationing and Private Cost Incentives," Journal of Public Economics, 88, pp. 333-352.

[17] Rabin, M. (1993), "Incorporating Fairness into Game Theory and Economics," American Economic Review, Vol. 83, pp. 1281-1302.

[18] Rickman, N. and McGuire, A. (1999), Regulating Providers' Reimbursement in a Mixed Market for Health Care," Scottish Journal of Political Economy, 46, pp. 53-71.

Discussion I

Hans Keiding

Department of Economics, University of Copenhagen

The paper by Ma deals with some problems – mainly such that are related to incentives – arising from the co-existence in the health care sector of public and private providers, in the present case physicians. Comparisons of the functioning of public and private enterprise is a recurrent theme in the health economics literature; in *production of services,* the public sector enjoys the advantage of easier access to capital while the private sector may have advantages in efficiency of operation. For the *patients*, public health care provision entails universal enrolment and consequently avoidance of adverse selection problems otherwise encountered, whereas private health care provision opens up for a greater freedom of choice. For *overall performance*, public health care makes it easier to set and enforce standards (as it is done e.g. by NICE in the UK), whereas a private health care sector can be regulated in a more flexible way relying on competition.

In the present paper, the distinction between public and private health care provision is to be found in the incentive mechanisms and the resulting allocation of providers according to the degree of altruism. The two models of the paper both deal with physicians who are randomly matched to patients. In *model 1,* providers may be altruistic or selfish; in any case they can observe patients' preferences, so if they are private profit maximizers, they can in principle perform first degree price discrimination, extracting all surplus. The result of the model is that allowing for selfish providers to behave in this way is actually welfare superior to a situation where all physicians receive the same fixed remuneration.

In *model 2,* providers come with different degree of altruism, which however cannot be observed by employers, public or private, so that

remuneration of services can only depend on the amount delivered. This gives rise to an agency problem, where the optimal contract sets a limit to the amount of services to be delivered, since physicians wanting to provide better service would then run into losses; a public health care service offering fixed salaries and stipulating high quality service would then attract the more altruistic physicians, whereas the less altrustic providers will be perfectly satisfied with the optimal contracts.

Common to the two models is the vision of a public sector using what is called *low-powered* incentives as contrasted with a private sector where incentive contracts prevail. This does not seem unreasonable; what is perhaps less intuitive is the vision underlying both models of a public sector providing high-quality medical care as compared with a private sector where it is possible to get a more humble service at a lower cost. This does not fit well with the European experience of the last decade; in eastern Europe, former all-embracing public health care has run hopelessly out of funds and provide only the most rudimentary services, if they exist at all. And also in western Europe, public health care has had a protracted record of cost containment efforts with resulting increase in waiting time and medical error. Even for the US environment, it is no obvious whether this vision of public health care (Medicare and Medicaid) versus private health care reflects the actual situation.

Be this as it may, it is interesting to look into welfare comparisons of different arrangements with respect to the split between public and private health care, and it is thought-provoking that the introduction of public health care provision in model 1 will lead to higher welfare. To some extent this counterintuitive result comes from the well-known fact that first degree price discrimination – which can be exercised by a monopolist knowing the demand of the consumer – is welfare optimal. Therefore, transferring from a situation where everybody is treated in the same way to one, where at least some consumers can get an individual treatment, points towards an increase in welfare.

What is not taken into account in the model is the fact that allowing for some private provision of health care will change the situation also in the public sector. If all provision is public, the model determines an equilibrium provision where average satisfaction equals cost. Once the private sector is introduced, the previous equilibrium level of provision in the public sector must be changed, depending on whether the private sector attracts patients with high or low quality parameter. This effect is not taken into account in the comparison, and presumably it would not change the result, but a formal proof would have been reassuring in this respect.

In model 2, the point of departure is the private rather than the public sector. The author studies incentive contracts for physicians who care for the quality of service that they provide. Since this degree of caring for the quality of the outcome (which in the paper is called altruism) is not observable (and anyway might well be irrelevant for the employer, public or private) it does not enter into the contract. What will happen is therefore that the physicians will select quality delivered (and associated payoff) so that it fits as well as possible with their preferences for income and quality of work. If the latter weighs heavily, income will be small, and since it is restricted to be nonnegative, there is an upper limit to the quality that physicians can afford to deliver. In this case, a public sector offering fixed remuneration – presumably high enough to cover the cost – will be a preferred alternative to the quality-aware physicians.

As it was mentioned already, what happens in the model, namely that physicians who insist on providing high quality service are forced out of private and into public health care, is a far cry from reality, where exactly the reverse happens every day – stories abound about highly competent doctors leaving public health care not because of notoriously low pay hey consider to be right for the patients.

One of the reasons that the models proposed by Ma do not display situations which look like those of contemporary European health care systems is that allocation in the models are chosen as welfare optimizers; budget considerations, on the other hand, are largely

absent. In reality, budget considerations in the health care organizations tend to overrule all other considerations, and certainly the abstract considerations of overall welfare. Consequently, models of the public/private interrelations in health care will have to incorporate such overall budget constraints in order to capture essential aspects of what is going on.

Discussion II

Klaus Kultti

Department of Economics, University of Helsinki

The organisation of health services is an economically significant matter and it is likely to be even more significant in the future. The rising expenditures put pressure into the restructuring of the health sector in many countries, and it is an active playing ground for politicians. Thus, it looks like it would be of importance to understand how the health sector works; what are the incentives and constraints of the participants.

The starting point of Albert Ma's work is the observation that public and private health care sectors coexist in many countries. The incentive structures in the two sectors are different: The private sector resembles any other service provided in the markets while the public sector features low-powered incentives.

To understand the interplay of the private and public health care sectors and their co-existenca Ma postulates that there are two kind of physicians; some of them he calls altruistic while the others are standard profit maximising agents. The former are medical professionals who follow the orders and standards given by the regulator or some supervisory body. The latter ones only care about profit. It is assumed that the quality of the physicians' services is not verifiable, and thus they cannot be paid on the basis of the quality of the services they provide.

Ma determines how the health sector with heterogenous physicians works. In particular, he determines the optimal compensation schemes for both kind of physicians, and shows that there is an equilibrium in which altruistic physicians work in the public sector and receive low

powered incentives, and the profit maximising physicians work in the private sector.

Except for the assumption about altruistic preferences Ma's work fits well into the mainstream of economic modelling. His analysis is technically fluent, and the model succinctly parameterised. It also seems very likely that there are physicians with different degrees of altruism. Still, I should like to offer criticism to the approach Ma has chosen from the theoretical standpoint.

In behavioural economics new categories are introduced quite liberally. Altruistic preferences are the prime example while things like spitefulness, fairness, sacrifice and myopic behaviour are also common. When one enriches the standard economic models with this kind of possibilities it seems that there are no limits to what one can explain. But once one can explain anything the explanatory power of the theory is zero. Roughly put, the best theory is not one that explains everything but the one that explains exactly half of the things; namely the things that are true.

My view is that in an effort to understand economically interesting phenomena and institutions one should make at least a reasonable effort to understand them within the current standard approach to the behaviour of economic agents. There are at least two advantages in doing so. First, the results are understandable within the currently accepted core of economics and comparable to previous results. Secondly, the limitations of the current theoretical framework become clear. This, as a guiding principle I suggest the following ideas as one (out of many) plausible alternative to Ma's assumption about the physicians' altruistic preferences.

There are two things to be explained: 1. The existence of the public sector and the private sector in health care. 2. Low-powered incentive schemes in the public sector and high-powered incentive schemes in the private sector.

Instead of sorting the physicians to the public and private sector according to their characteristics we could think that two sectors is an optimal solution to the pervasive problems of incomplete information in the physicians' services: The suppliers have much better information than the demanders of the services. The public sector then establishes some kind of benchmark or minimum quality that is guaranteed. This prevents the private sector from unravelling down in quality, and gives an incentive to good physicians to go to the private sector to provide better than standard service. What prevents everybody to go to the private sector is the fact that there are differences in the ability of physicians. Also in many countries, including Finland, most research is conducted in the public sector. This would keep research orientated physicians in the public sector.

As to the low-powered and high-powered incentives, it looks to me that the work of Holmström and Milgrom (1991) on multitask principal-agent problems seems particularly relevant in the health sector. One could ask in what circumstances the compensation schemes in the private and the public sectors are optimal. Ma's answer is that they are optimal when physicians are heterogeneous as to their altruism, and the society wants to sort different types.

But we know that if an agent is to perform several duties where the measurement of the duties is not accurate low-powered incentive schemes are optimal. A general practitioner in the public sector is to provide both quality and quantity, and the former is much more difficult to measure than the latter. But quality is important as the GP has to decide whether to refer a patient to an expert or not; and here both H1 and H2 type errors may be very costly.

One can think that there are two ways to organise the health care. The public system where physicians are given low-powered incentives because there is plenty of anonymity in the system: the GP, for instance, does not know the expert to whom s/he refers patients. In the private health care system physicians know the experts to whom they refer patients. There reputation effects take care of the discipline needed.

Finally, I think that in contrast to many other fields of economics here a researcher can get valuable information by just asking physicians what they think the true state of affairs is. My extremely limited query indicates that Finnish physicians believe altruistic aspects to be important. In their opinion this is not, however, manifested in the functioning of the everyday Finnish health care system but in the physicians' decisions to go and work in the developing countries and organisations like *Médecins Sans Frontières*; they indisputably forgo economic rewards unlike physicians who decide to work in the public sector.

References

B. Holmström and Milgrom P. Multi-Task Principal-Agent Problems: Incentive Contracts, Asset Ownership and Job Design, Journal of Law, Economics and Organization, Vol 7, Special Issue.

Economists and the Quest for Regressive Health Care Financing: Conclusions in Search of Arguments

5

Robert G. Evans

Department of Economics, University of British Columbia

5.1 Kenneth Arrow and the "Welfare Burden" of Health Insurance

Kenneth Arrow's 1963 paper, "Uncertainty and the Welfare Economics of Medical Care" is generally celebrated among health economists as a seminal work in the field. It has often been reprinted, and in 2001 was the subject of an entire special issue of the *Journal of Health Politics, Policy and Law* (Peterson, 2001). The Arrow Prize, given for the best student paper presented at the biennial meetings of the International Health Economics Association, reflects the profession's assessment of the paper's significance.

Yet there is another later paper (Arrow, 1976), rarely cited, that is arguably more significant in its implications for economists trying to advance our understanding of health care systems. In that paper, Arrow attempts to demonstrate rigourously, on the basis of conventional economic theory, that the "optimal" pattern of financing of health care – the distribution of its costs across the population – must necessarily include some level of user charges, of direct payment by recipients proportionate to the care they use. "Free" (to the user) care is allocatively inefficient.

The importance of this paper lies not so much in its conclusions as in its explicit and transparent presentation of the assumptions necessary to reach them. Those assumptions have traveled far beyond this one

somewhat obscure paper; they are in fact essential to the whole corpus of economic analysis that can be included under the general rubric of the "welfare burden" generated by third party coverage for health care costs, public or private (e.g. Pauly, 1969; Feldstein, 1973; Manning *et al*., 1987). Without the assumptions spelled out by Arrow, this literature falls apart. The conclusions, and the policies and practices, that draw on it have no intellectual foundation, or at least none in economic analysis. And empirical estimates of the alleged welfare burden of alternative insurance arrangements are meaningless.

A "welfare burden" or loss of aggregate social well-being is postulated to arise because health insurance lowers the out-of-pocket cost to users of these services, relative to their opportunity cost in terms of other commodities foregone. The result is "over-consumption" of health care, and a corresponding reduction of other goods available. People would be happier with less health care, and more of other things, but insurance coverage distorts their choices and makes them collectively worse off.

It is important to be clear that "over-consumption" in this framework has nothing to do with ineffective or harmful care. It is defined solely as consumption of services for which an individual user would have been unwilling (which includes unable) to pay a price equal to or greater than the marginal cost of their production. The relationship of health care to health nowhere enters the analysis.

The assertion that "free" care generates a welfare burden rests purely on *a priori* reasoning. It is thus impervious to evidence that in a system of insurance coverage without user charges, health care use is closely associated with measures of need, and most is used by a small proportion of the population who are quite or very sick (e.g. Finkelstein, 2001; Roos *et al*., 2004). No matter how serious the patient's condition and need for care, any provision of services would by definition be "overused" if the recipient would have been "unwilling" to pay for it. Caring for patients in such circumstances allegedly generates a welfare burden, and the world would be a better place without it.

But health status and needs for health care are uncertain, and uncertainty *per se* is for most individuals a "bad," a source of disutility or distress from which they will pay for relief. That is why people buy insurance. There is thus presumed to be a fundamental trade-off between the costs of too much insurance coverage (over-consumption of care) and the costs of too little (individual exposure to excessive risk), with the optimal level of coverage requiring some balance between the two. Some insurance is good, complete coverage is too much.

This conceptual framework has been extremely influential and is deeply embedded in the intellectual structure of modern health economics (particularly in North America). Nyman (1999, p. 681) for example, refers to Pauly's (1968) early outline of the welfare burden framework (see also Pauly, 1969) as "…one of the most influential papers in the health economics literature", and cites a number of subsequent empirical applications in the American literature. His assessment seems valid, though a corresponding influence on the actual design of health care financing systems, over the intervening thirty years, is hard to detect outside the United States, and even there the picture is unclear.

The obvious disjuncture should perhaps have been troubling to academic health economists, but there is little evidence that it has. Victor Fuchs, in his presidential address to the American Economic Association (1996) reports on an informal "grab sample" survey of health economists, economic theorists, and physicians. He found that an overwhelming majority of each group, from 77% of physicians to 93% of economic theorists, agreed with the statement "Third party payment results in patients using services whose costs exceed their benefits …". Indeed Fuchs himself classified this as a positive statement, a statement of fact (true or false, and testable against evidence) rather than a normative statement of values, despite its obvious distributional content. (Costs and benefits *to whom, measured how*?)

It is therefore somewhat unfortunate that, as Arrow's 1976 paper makes clear, the fundamental assumptions of this intellectual framework characterize a world of total fantasy. Relative to our own world, they are simply nonsense.

5.2 A World of Clones: Fantasy and Finance

Arrow is quite open: *"I ignore distributional considerations and assume a single person in the economy*." (Arrow, 1976, p. 4.). Perhaps sensitive to the difficulties of describing insurance markets in a single-person economy, however, he rapidly revises: "*To avoid distributional considerations I assume that all individuals have identical endowments and identical utility functions. I further assume a very large population*." (*Ibid*., p. 5). His "individuals" (?) also have identical risks of illness or injury; they differ only in that some, randomly selected, actually do suffer an adverse event, and others do not. And since the analysis is static, in a timeless world, "today" and "tomorrow" are also identical. Today's health misfortune does not affect tomorrow's probabilities (no chronic illness). Tomorrow is always a whole new day.

It follows that the choice of financing mechanism *cannot* have any "distributional implications". That was, after all, Arrow's intent in choosing his assumptions. Whether financing is all third party, or all self-pay, or any intermediate mix, the choice will affect all individuals equally. After the fact (*ex post* in the traditional jargon of economics), at the end of the period of analysis, those who suffer illness and purchase health care will be financially worse off than the more fortunate, if the financing system requires them to pay some or all of its cost out of pocket. But looking forwards (*ex ante*), before anyone can know who will be unlucky, all are equally placed. The choice of financing mechanisms at the social level – how much insurance coverage – will involve the trade-off above, but each individual faces *exactly the same trade-off*, and therefore the "optimal" financing mix is exactly the same for each. What is best for one, is best for all.

These assumptions permit Arrow – and all those who postulate a welfare burden/risk reduction trade-off – to focus solely on how the choice of financing mechanism will affect the mix of commodities produced and used in this society, as between health care and "other goods". As Joan Robinson has said (somewhere) people disappear into the background; only the commodities have speaking parts.

A further set of assumptions, not spelled out by Arrow, is necessary to guarantee that that in Arroworld the prices for health care faced by the consumer-clones are equal to the marginal resource costs, the opportunity costs, of its production. These are equally fantastical: health care must be produced by private, strictly for-profit firms, supplying perfectly competitive markets, with free entry and exit of firms and with "well-behaved" factor supply and production functions. We will not deal here with the distributional significance of these assumptions, except to note that they make it impossible to express questions about the relative incomes of providers of care, questions that are central to the management of all real-world public payment systems. It is however notable that Kessel's (1958) classic paper "Price Discrimination in Medicine", once widely cited but now largely forgotten, did explore the structural features of American medical markets that had restricted competition and protected physician incomes. These have vanished from Arroworld.

An obvious implication of this framework is that there will be no political disagreement among citizens over how to finance health care. Costs will not be spread equally over the whole population in Arroworld; those who become ill will bear a somewhat larger share of total costs. But they will agree to this in advance, while they are still behind the Rawlsian "veil of ignorance" because on average everybody is better off with this distribution than with full insurance coverage. Everyone prefers to run a certain risk of incurring out-of-pocket costs rather than to accept the lower level of well-being associated with too much health care, relative to the "other things" they could have instead. And since everyone has an equal income anyway – regardless of illness status – they are all equally capable of bearing whatever payment burden may be imposed on them by the luck of the draw.

Choice of the optimal financing mix, in Arroworld, will require technical information as to the elasticity of demand for care of the representative individual, and his/her degree of distaste for risk. But each individual will have the same interest in having these parameters correctly identified, and in using that information to design the optimal balance of collective payment and user pay. Any disagreements will be solely over these technical issues, and therefore resolvable through data and analysis; the debate will not be clouded and prolonged by embedded conflicts of economic interest masquerading as "scientific" disagreements. The economists who spend their time investigating these questions will indeed be, in Keynes' description, "… humble, useful people, like dentists." and not at all ideologically (and financially) inspired advocates for clashing values and objectives…

There goes the alarm clock. It's time to wake up and go to work.

5.3 Conversations of the Deaf: Enduring Conflicts, Enduring Interests

The most obvious characteristic of debates over public financing decisions, at least in all high-income countries, is precisely that they are *not* settled by fact and argument. A fairly stable consensus has emerged in most countries over the last half-century, but it is exposed to constant political challenge, particularly from the right-wing. Moreover the nature of those challenges has changed little, if at all, over the decades.

Public discussions of financing issues are haunted by "zombies" – ideas and arguments that are intellectually dead but can never be permanently laid to rest. These intellectual zombies are impervious to evidence or logic; no matter how many times they are refuted, they are always dug up again, often refurbished with new labels and accompanying claims of "new ideas" or "thinking outside the box", and sent marching back into the public arena to spread confusion and disinformation. Zombies cannot be killed because they serve economic interests that are very much alive, and indeed immortal (Evans et al., 1994; Barer et al., 1998).

The reason for these unquiet graves is not far to seek. Outside Arroworld, people are not clones. They vary enormously in their economic resources, as represented memorably by Pen's (1971) "parade of dwarfs, and a few giants". Moreover the international collaborative Luxembourg Income Study finds that these inequalities have been increasing in recent years in nearly all high-income countries (Smeeding, 2002).

[Rather chillingly, there is also evidence that the "surge in economic inequality in many countries around the world in the last few decades" … "stacks the deck of democracy in favor of the richest citizens, and as a result, everyone else is more likely to conclude that politics is simply not a game worth playing." (Solt, 2004). Phillips (2002), in a longer-term historical analysis of the United States, makes the same point. This trend may have a bearing on the regressive tendencies in so many recent proposals for "reform" in health care financing – see below.]

There is also a large literature documenting the enormous variations in individual risk status and in experience of illness; references are provided by Reid et al., (2003) who demonstrate the central role of co-morbidities in this interpersonal variation. Accordingly the choice of financing mechanisms is fundamentally political. That choice has major implications for the incidence of total health care costs – how the total national bill for health care is distributed across the population, and how closely individuals' shares of that bill is relate to their economic resources. In this on-going political debate, conclusions that are valid only in Arrow's fantasy world of clones have no explanatory value, but they can be, and are, used for propaganda purposes to advance particular economic interests.

In these debates, the question addressed by Arrow (and fundamental to the welfare burden story) drops out of sight. No one other than a handful of economists addresses (or likely even understands) the concept of "allocative efficiency" or "opportunity costs" in the economist's meaning of the terms. Rather the primary public contest is between cost containers and "under-fundists", where both parties

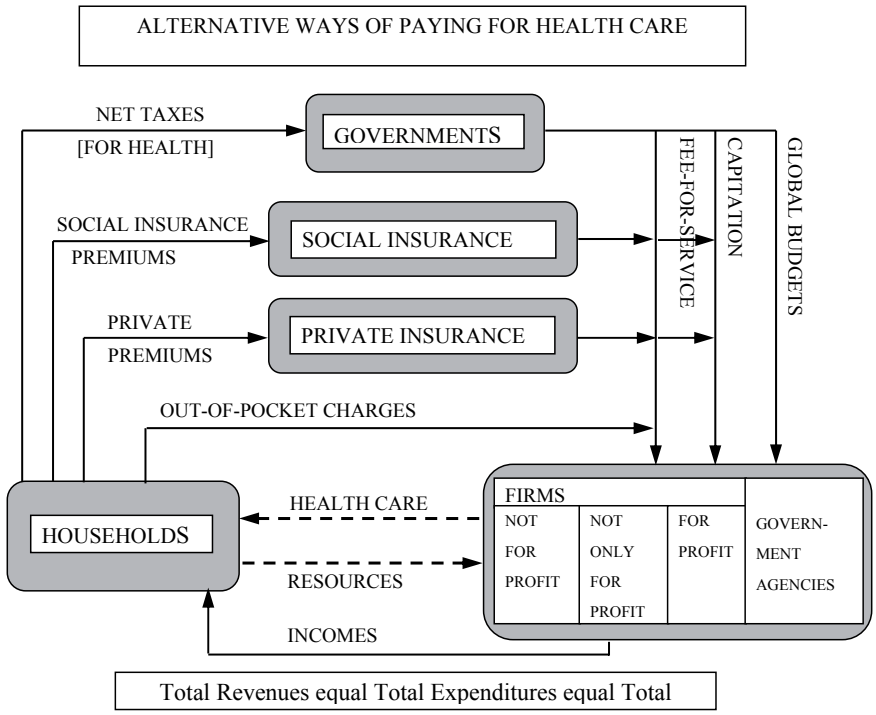
are focused on financial measures of cost, respectively trying to hold them down or to push them up.

Everyone agrees that the point of providing health care is to protect or advance someone's health; no one (not even Americans, and not even economists on their days off) believes that, in a highly unequal world, access to care should be rationed solely by willingness/ability to pay. (It happens, particularly in the United States as well as, obviously, outside the high-income world, but no one seriously tries to support it as a moral principle.) And (almost) everyone seems to accept, on the basis of minimal evidence and sometimes in the face of counter-evidence, that more health care would be a good thing. A dissident collection of clinical epidemiologists and other health care researchers keep trying to focus attention on "evidence-based medicine" or more generally evidence-based health care – do what works, not merely what pays – but their impact on high level public debates over health policy seems in most countries to be minimal at best.

5.4 Accounting for Health Finance: Follow the Money

The basic accounting structure of health care systems is laid out in Figure I, a simplified version of the National Accounts framework with the foreign sector omitted. "Real" resources of human time, skills, and effort, and services of capital and natural resources, are owned directly or indirectly by people (grouped into households) and are supplied to "firms" – the provider organizations that transform these resources into various types of health care and supply them back to individuals or communities of people. Provider organizations pay for these resources through the various forms of income that flow back to persons. Providers in turn are funded on various terms either by the payment organizations that assemble the necessary revenue from households, or by households themselves through direct payments.

Figure 1.



Equation 1 presents the corresponding identity linking Total Revenue collected for, Total Expenditure on, and Total Income earned from health care. T, SI, PI, and C (measured in currency) represent the channels through which finances flow from households – taxes, social insurance premiums, private insurance premiums, and direct charges to users of service. P and Q are vectors of the amounts of the various different types of health care services provided and used (Q), and their corresponding prices (P), while Z and W are vectors of the amounts of the various resources used up – such as person-hours of a particular skill type or square meters of building space – and of their corresponding rates of reimbursement.

$$T + SI + PI + C \equiv P \times Q \equiv W \times Z \qquad \ldots (1)$$

In Arroworld the choice of financing channels has no implications for the incidence of health care costs. Clones all pay the same amount, whether through taxes, or social insurance premiums, or private insurance premiums. As described above, the extent of user payment will affect incidence *ex post*, but not *ex ante*, because at the beginning of the analytic period, everyone's risk and expected use is identical. (It is also presumed to influence the amount that each identical individual pays, insofar as it is presumed that total outlays will be larger or smaller depending on the split between self-pay, and public or private insurance coverage.)

That is why, if insurance is privately provided by competitive commercial firms, everyone will pay the same premium, a premium just equal to the amount of tax they would have to pay if the same level of coverage were financed through income taxation, or compulsory social insurance premiums. In this world Equation 1 holds as an identity not only for the population as a whole, but also for each individual in that population (with the substitution of individual expected values for C and Q).

But in our world, the selection of channels matters enormously. Shifting the mix generates a direct, and potentially very large, transfer of wealth from some individuals to others. That is why financing choices are endlessly controversial, and the notion of an "optimal" mix that is best for everyone is a chimaera, a logically impossibility. Its analysis in the literature of academic economics is at best an irrelevance and at worst a source of confusion and mischief.

5.5 Who Pays? Who Gets? – And How to Shift the Burden

Figure 2 provides an extreme but real example, using data from a random sample of 4% of the population of the province of Manitoba (total sample, just over 40,000) assembled and analysed by Mustard *et al.* (1998a,b). They match actual individual-level utilization and estimated corresponding public expenditure for all hospital and physicians' services and long-term care (from the administrative

records of the universal public insurance programs, in which there are no user fees or extra-billing of patients), with corresponding family-level incomes reported to the Canadian Census. This unique data set enables one to estimate, by income decile, the amounts of public health care expenditure generated and the corresponding contributions through income and sales taxes paid to both the provincial and the federal government.

Figure 2a. *Expenditures on Publicly Financed Health Care, by Income Decile, Manitoba, 1994*

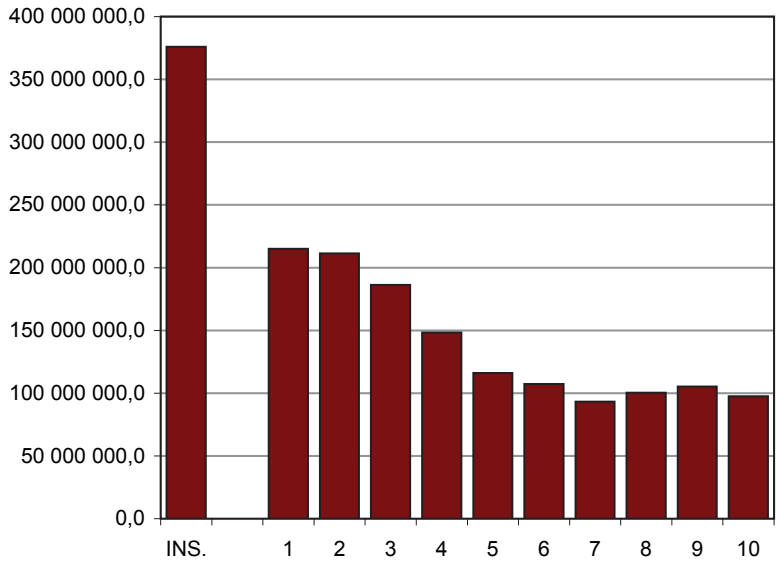


Figure 2b. *Tax Contribution to Health Care, by Income Decile, Manitoba, 1994*

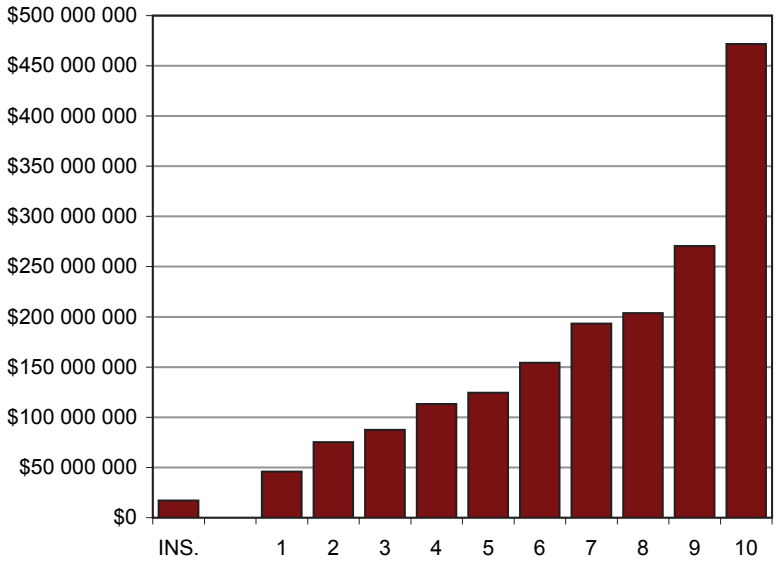


Figure 2c. *Net Transfer to/from Income Decile, Public Financing of Health Care, Manitoba, 1994*

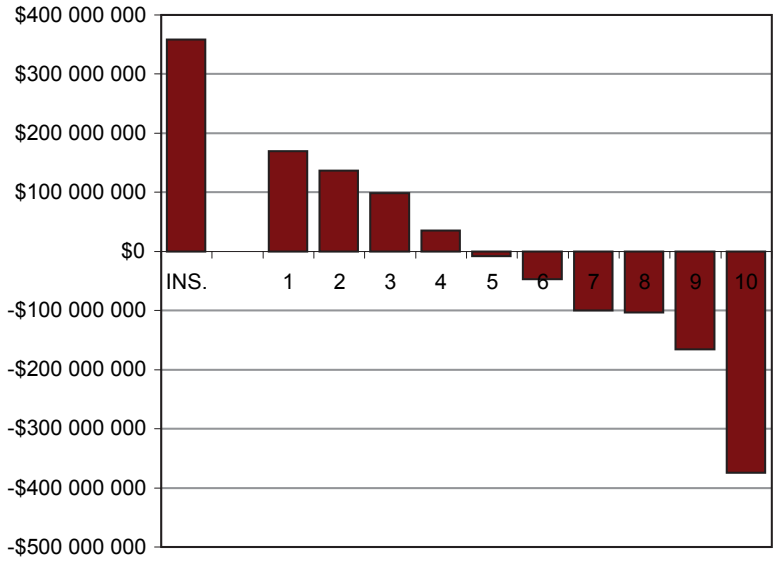
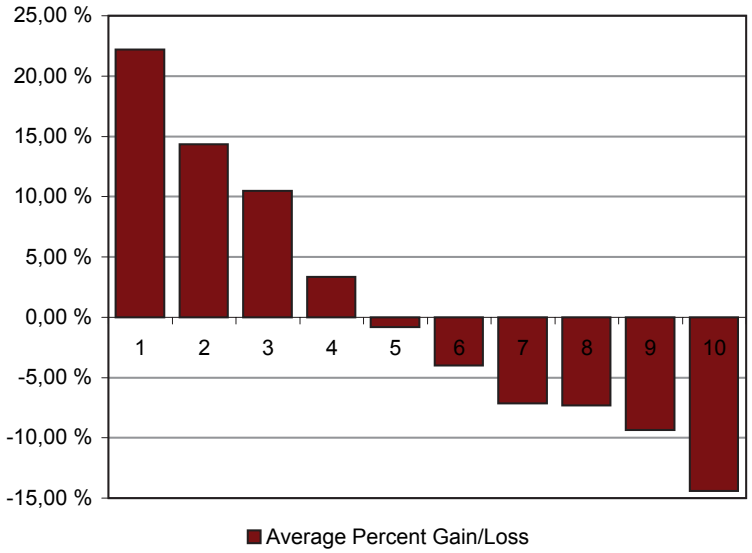


Figure 2d. *Net Transfer by Income Decile, as Share of Consumable Income, Manitoba, 1994*



The distributional impact of these programs emerges with striking clarity. Panel 2a indicates the distribution of total program expenditures (in 1994 CAD) across income deciles, with highest average per capita costs in the lowest decile and a steady fall down to the middle of the income range. For the upper-income half of the population average outlays vary little with income. It must be recalled, however, that within each decile there are very large variations in expenditures on behalf of different individuals (see Figure 4 below). Moreover, as shown, a small group of permanently institutionalized people (INS, not included in the income deciles) account for a remarkably large share of total expenditure.

Panel 2b displays the estimated distribution across deciles of tax liability, under the tax structure then in effect, to raise revenues sufficient to cover these expenditures. The concentration of burden on the top decile reflects both the degree of progressivity in the Manitoba and Canadian tax systems at that time, and the concentration

of income in the highest decile. Panel 2c indicates the distribution of net contributions, showing the extent of the net transfer of income from those in the higher (and particularly the highest) deciles to support the costs of those with lower incomes. Net costs incurred to support those permanently institutionalized are particularly notable, and are very similar to the net contribution of those in the highest income decile. Panel 2d shows the proportion of consumable income (disposable income less consumption taxes) associated with these transfers. (This measure is meaningless for the permanently institutionalized, with very high costs and almost no income.)

These data are now a decade old, but they still serve to illustrate the very significant redistributive impact of a comprehensive and relatively expensive public program financed by a progressive tax system. A shift in the channels of payment, in the absence of any change in the distribution of service utilization, would change this pattern in a perfectly predictable way. An (obviously hypothetical) shift to full self-payment with a tax cut proportional across income deciles would collapse all the bars in all panels to the horizontal axis with a corresponding transfer of wealth up the income distribution. It would also, of course, redistribute wealth dramatically within each decile, from the ill to the healthy.

More economically and politically realistic, Figure 3 indicates the impact of an across-the-board 20% cut in the level of income taxes absorbed by health care, that is exactly offset by the imposition of either a poll tax – usually referred to in this context as a "compulsory health (or social) insurance premium" – or a coinsurance charge such that users of health care would be required to pay some proportion of the expenditures made for their care. It is assumed that those in the lowest income decile, and those permanently institutionalized, would be exempt from either of these charges.

Figure 3a. *Net Transfer by Income Decile, in $ per Family, from 20% cut in Income Tax and Off-setting Poll Tax (INS and Dec. 1 exempt), Manitoba, 1994*

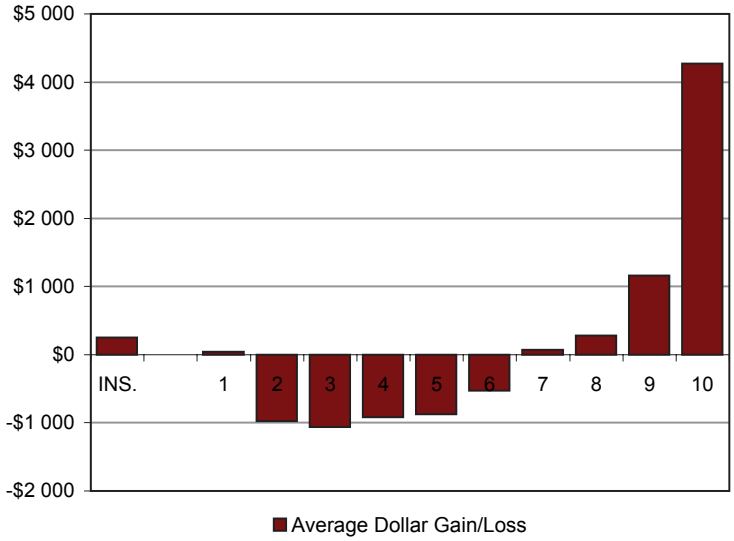
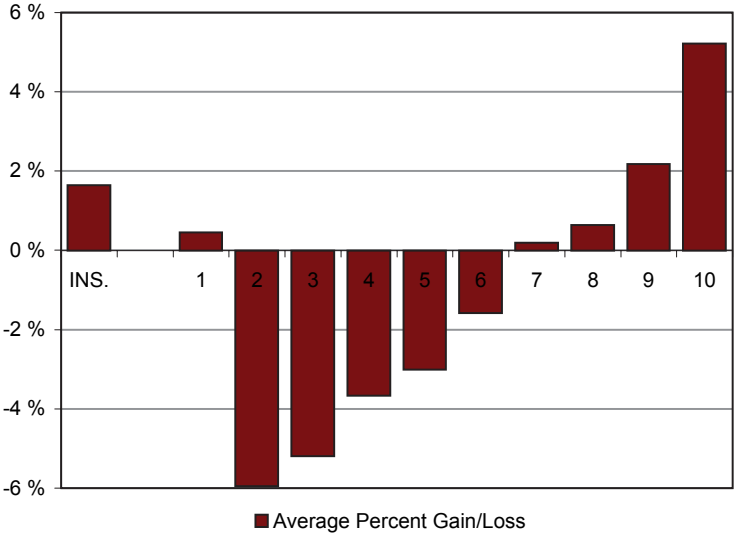


Figure 3b. *Net Transfer by Income Decile, as Percent of Consumable Income, from 20% cut in Income Tax and Off-setting Poll Tax (INS and Dec. 1 exempt), Manitoba, 1994*



Panel 3a shows the redistributive effects across the income spectrum of a shift from income to poll taxes, with individuals grouped by census families. Families in the top decile are the big winners, gaining an average of just over $4000 per year each (in after tax, 1994 CAD). This reflects the high concentration of income at the top end, as well as the larger impact of a proportionate tax cut at the high end of a progressive tax schedule. Panel 3b shows the corresponding change in percentage of consumable income. While the average losses are roughly equal in dollar terms from the second decile to the sixth – at or close to $1000 per family – they have a much larger proportionate impact farther down the income distribution. Families in the second lowest decile would lose 6% of their consumable income, while those in the highest decile would gain just over 5%.

Offsetting the tax cuts with a coinsurance charge for health care use has similar qualitative effects, as illustrated in Panels 3c and 3d. But the heavier concentration of health care use among lower income groups implies that the losses will be greater there – nearly $2,000 per family in the second and third deciles, while the top decile gain just under $5,000. This translates into a loss of consumable income of over 10% in the high-using second decile – the lowest income group to whom the charges are applied – while the gain for the top decile now reaches 6%. The point of approximate break-even moves down from the seventh to the sixth decile; families in the seventh decile break even, on average, with the poll tax but are slightly better off (than under full tax financing) with the coinsurance charge.

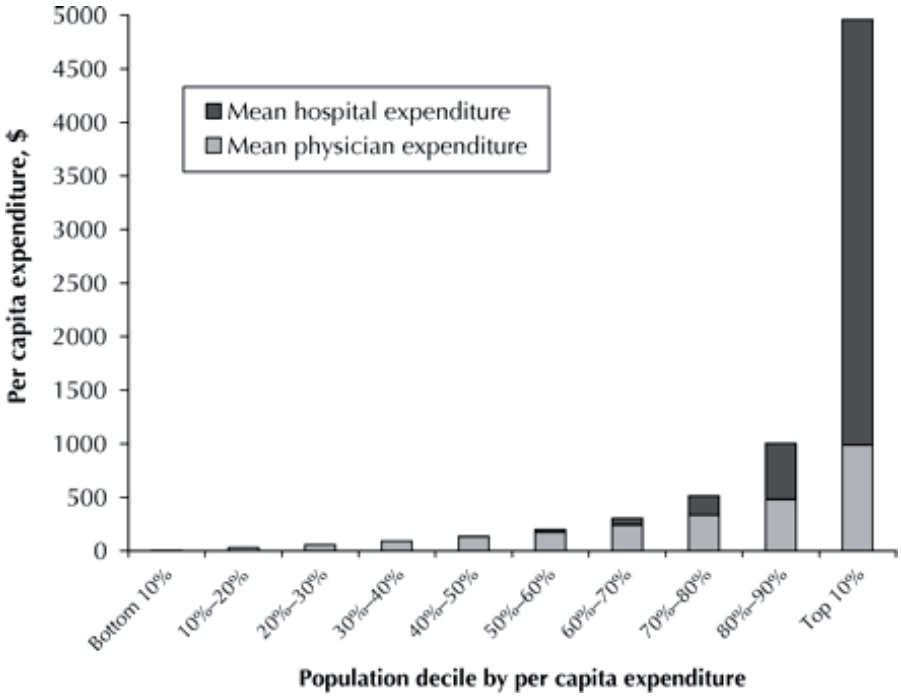
These data are from a uniquely rich data set, but for a small Canadian province at a particular point in time. The quantitative effects would be different in jurisdictions with different systems of taxation and health care delivery, as will be discussed below. And these figures themselves would shift if one changed the postulated tax cuts and exemption patterns. In particular the extent of the impact on people with very low incomes may be mitigated by discounting either the poll tax or the coinsurance rates for those in the lower deciles. (Such mitigation implies that the gains for those at the top of the income distribution would come primarily at the expense of those in the

middle.) But the qualitative results, the directions of effect, are fully general. Debates over health care financing are first and foremost debates about the distribution of income, net of taxes and health care expenditures.

5.6 Taxing Incomes or Taxing Illness? Small Numbers, Large Needs

There is, however, an obvious and very important difference in incidence between the poll tax and the coinsurance requirement, beyond their effects across income classes. The coinsurance charges will impose very different burdens on different people *within* each income class, from zero for the incorrigibly healthy to very heavy costs, perhaps greater than their incomes, on the very ill. As noted above, care use and corresponding illness expense are heavily concentrated, in every health care system, on a relatively small proportion of the population, typically elderly, poor, and chronically ill, often with multiple co-morbidities. Figure 4, from Forget *et al.* (2002), provides a dramatic illustration, showing the distribution of actual expenditures over a three-year period within the Manitoba population (unfortunately not cross-classified by income). The incidence of liability for any form of proportionate user charge would follow a similar pattern, while user charges capped at a relatively low level per person would exempt most of the expenditures in the system.

Figure 4. *Average Physician and Hospital Expenditures per capita, Residents of Manitoba Ordered by Individual Level of Expenditure, 1997-1999 (Forget et al., 2002)*



A second difference, much emphasized by some economists, is that by linking financial liability to the level of care use, the coinsurance charge creates a financial incentive to limit use. This is, however, a very peculiar argument in the context of the broader policy debate.

In the first place, while there is empirical evidence in support of the presumption, drawn from elementary economic theory, that when people (particularly lower income people) must pay all or part of the costs of their own care, they will use less. This idea forms the core of Arrow's argument above, and of all prior and subsequent analyses using the concept of a "welfare burden". But there is in fact *no* evidence that these individual responses aggregate to a collective response, i.e. that *total* use of services will fall in response to a shift in financing mix. To assume such an aggregate response on the basis

of individual responses is an example of the well-known "fallacy of composition", a logical error against which all first-year economics students are (supposed to be) warned. A variety of forms of evidence, including that from comparisons across national systems, supports the view that aggregate health care expenditures have little or nothing to do with the level of user charges, at least within the range of current experience. Rose's Law (1985) reminds us that the causes of cases are not in general the same as the causes of rates, and that Law seems to be operative for health care financing.

More important still, however, there is no evidence that individuals respond to out-of-pocket charges by reducing their use of less necessary or effective care. Indeed the evidence is that they do not. This is of no significance in Arroworld, where the value of commodities is be judged solely by the user's willingness to pay. Lower use means a lower welfare burden: whether the services given up were desperately needed to preserve life and limb, or minor conveniences, or entirely ineffective, or even actively harmful, is simply irrelevant. And since everyone has the same income, and the same tastes, the question of differential ability to pay does not arise.

But in the real world of human concerns and public policy debate, people care a very great deal about the relationship between health care and health, and the existence of unmet needs for care. There is a broad public consensus that people should get the care they need, irrespective of their ability/willingness to pay for it. Thus advocates for user fees must typically confront a powerful counter-argument, that such fees will discourage people with lower incomes from seeking needed care, and will thus threaten or damage their health.

Figure 3 brings out clearly the inevitably regressive nature of the two hypothetical policies represented there. A shift in the mix of funding sources toward greater reliance on either compulsory social insurance premiums (poll taxes) or direct charges to users, and away from income or even sales taxes, inevitably transfers net income in favour of those at the high end of the income distribution. (The regressive effect of private insurance is even greater, since a competitive marketplace

forces private insurers to charge higher premiums to those at greater risk of illness. On average, people with lower incomes have greater health needs.)

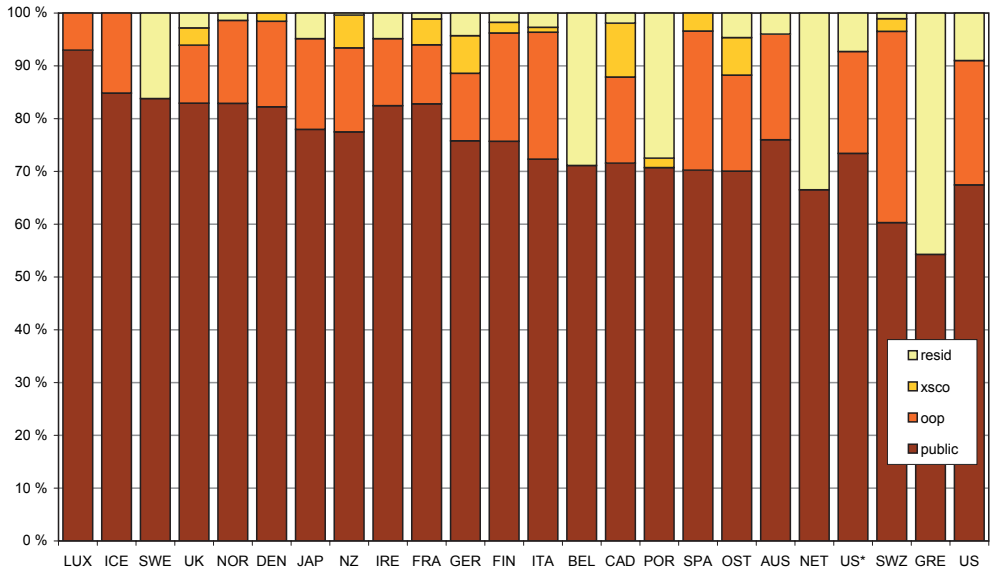
It is not therefore surprising that advocacy of such shifts comes primarily from the representatives of people at the top end of the income distribution. They have the most to gain, and tend to have privileged access to the media and the policy-making process. (The appeal of their message may well be enhanced by the fact that the senior members of public bureaucracies, with the greatest responsibility for financing policies, are themselves in the income brackets that would benefit most from shifts in mix of the sort represented in Figure 3. So are most of their friends and associates. This is not to suggest corruption, but rather participation in "communities of shared understanding".) It would be naïve, therefore, to expect controversy over the financing mix ever to disappear,

Figure 4, however, brings out the important additional dimension, hidden in Figures 2 and 3, of the extreme inequality of patterns of health care use among individuals, independent of the distribution of incomes. A coinsurance charge – the requirement that individuals pay user fees proportionate to their use of care – would have very powerful redistributional effects within as well as across income classes, effects significantly greater than those across classes. It is hard to find any justification on equity grounds for such a "horizontal" redistribution. Furthermore, insofar as such charges do limit access and use by those with lower incomes, to the benefit of those at higher levels, they weaken the link between use and need, and strengthen that between use and ability to pay. These two effects are the Achilles heel of user charges, in societies where (unlike Arroworld) most of the population seem to believe that the purpose of a health care system is to meet the needs of the ill or injured, not to respond to the willingness to pay of people with more money.

Chipping Away at the Consensus: Private Financing and the Protection of Privilege

Accordingly it appears that at present the endless debates and conflicts over the financing mix take place, at least in high-income countries, against the background of a quite stable consensus among the population as a whole in support of a preponderant role for public financing. That consensus is reflected in the patterns of financing mix shown in Figure 5.

Figure 5.



In almost all OECD countries, health care is primarily financed either through taxation or through social insurance. In some countries, notably the United States but including among others Canada, some part of the public financing flows "through the fiscal back door" in the form of public tax concessions (tax-expenditure subsidies) to employer-purchased private coverage rather than appearing openly in government expenditure budgets. In the United States this indirect public expenditure by all levels of government is projected to reach $209.9 bn. USD in 2004 (Sheils and Hogan, 1999; Sheils and Haught, 2004). Adding to this some additional smaller forms of *de facto* public reimbursement raises the public share of outlays from its officially reported level of about 45% to a more realistic 60% (Fox and Fronstin, 2000; Woolhandler and Himmelstein, 2002).

Outside the United States, the competitive private commercial insurance markets that are the primary focus of analysis in the academic health economics literature are virtually non-existent. There are a few countries with significant pseudo-private insurance systems, such as France, Australia, or Switzerland, but all are heavily regulated. They are in effect manipulated or managed to try to prevent them from behaving in the manner dictated by competitive private markets – risk-rating their premiums and excluding unprofitable customers. But wherever they operate, private insurers cover a proportion of health care costs that is much lower than the proportion of the population with such coverage. Private firms cannot afford to insure unhealthy people, because those are the ones who need and use care most.

The overwhelming predominance of public financing, however, still leaves plenty of scope for disagreement at the margin over financing sources, inspiring *inter alia* an apparently endless supply of academic papers and health economics conferences on "the public-private mix". But it is only in these purely academic settings that analyses based (explicitly or more commonly implicitly) on the assumptions of Arroworld can receive a serious hearing. The scale and relative transparency of the potential redistributive effects of changes in the financing mix, changes in both financial burden and access to health care, have ensured that the "distributional considerations" that Arrow

was careful to avoid or ignore have remained front and centre in the public debates.

Accordingly those promoting regressive changes in the financing mix have typically sought to mitigate the impact of their preferred policies by exempting the lowest income groups from their impact – as in Figure 3 – or by building some degree of income sensitivity into either the poll tax/premium structure or the user charge structure. A popular approach that has resurfaced periodically for over thirty years as a "new idea" is to integrate user charges in some way with the income tax system, so that the financial liability associated with a given level of health care use will vary with the user's tax bracket. Still more complex structures have been developed, such as the variants on the recently prominent "Medical Savings Accounts". These complexities serve to blur the inevitable redistributional effects.

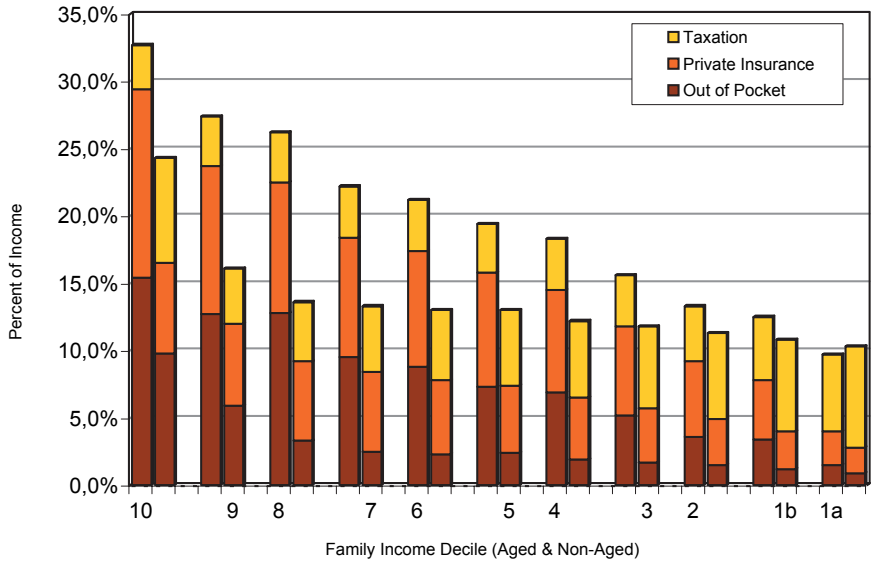
Mitigation of regressive effects at the lower end of the income distribution, however, leaves untouched the effects at the top end. There is no getting around the fact that emerges so clearly in Figures 2 and 3 – there is in most countries a lot of money in the upper reaches of the income distribution. Financing systems that draw heavily on income taxation, whether progressive or proportionate or even somewhat regressive, transfer substantial amounts of wealth from people at the top end. The more expensive the health care system, the larger the transfer. Any shift that weakens the linkage between income and burden incidence leaves more money, potentially a lot more money, in the hands of the highest income groups. User charges are especially advantageous, because needs for care are negatively correlated with income. Such charges also serve to improve access for the better off, insofar as they limit access at the lower end of the income distribution.

The Private Insurance Trade-Off: Greater Inequity But Higher Costs

The American experience, however, demonstrates that private commercial insurance can be as regressive, and thus as advantageous

for the wealthy, as out-of-pocket charges. Figure 6 is drawn with data from Rasell *et al*. (1993). It shows that for the non-institutionalized American population, out-of-pocket payments for all forms of health care take up a steadily falling share of family incomes as family incomes rise – as one would expect. (Absolute outlays, however, rise with rising incomes.) But private insurance premiums have a regressive pattern very similar to that of out-of-pocket payments – consistent with the expected behaviour of a competitive private insurance market. Theory would lead one to expect that commercial insurers would establish a premium structure, across the population, corresponding to expected outlays within each income class.

Figure 6. *Share of Income Spent on Health Care, United States (1987), by Family Income Decile and Payment Form*



The proportion of income paid for health care through public channels, however, rises as income rises, indicating that at least in the late 1980s the American tax system was mildly progressive. (It might be different today.) On the other hand one might also have expected that

since virtually the whole of the American population over the age of sixty-five is covered by the national Medicare insurance system, the overall burden of health care costs, public and private, would be distributed progressively for this group. It is not; on the contrary the burden distribution is quite markedly regressive even for the over sixty-five population. This presumably reflects the extent of costs for services excluded from the public program, as well as the structure of user charges, coinsurance and deductibles, built into the American version of Medicare. These do not appear to have been very effective in controlling cost escalation, but they do ensure considerable regressivity in program financing (see also Figure 7 below).

The regressivity of private insurance is enhanced if it is supported by tax expenditure subsidies. When employer-paid premiums are exempt from taxation in the hands of the employee, this subsidy is of greatest value to those in the highest tax brackets. Sheils and Haught (2004) estimate that in 2004 this public subsidy will be worth, on average, $2780 to families in the $100,000 and up income bracket. This group comprised 14% of the American population, but received 26.7% of the subsidy. Families earning under $10,000 received an average subsidy worth $102; the average subsidy value rises monotonically with family income.

The disadvantage of significant reliance on private insurance financing, however, is that it is extraordinarily expensive. Private coverage imposes very large administrative costs, both for the services of the insurers themselves and for providers who must deal with them. In the United States the extra costs – relative to a universal public health insurance program – are estimated for 1999 at $209 bn. or about 17% of total health care expenditures (Woolhandler *et al*., 2003). A simple projection suggests that by 2004 these extra administrative costs will be over $300 bn. or more than $1000 for every member of the U.S. population.

In effect the private insurance system imposes an "inequality-efficiency trade-off" – it permits Americans to have greater inequality of access and a highly regressive burden distribution, in return for a

very bureaucratic and inefficient payment system. These extra costs are generated as private insurers seek to acquire the most profitable aggregate pool of insurees and to offer the most profitable menu of contractual forms. They must identify the risk status of more and more finely divided subpopulations and recruit the profitable while rejecting the unprofitable, and must devise contractual forms to limit their liability. Providers, in turn, must invest in systems to determine patient coverage status and reimbursement eligibility, and to negotiate with insurers for payment.

But one does not have to introduce the extravagances of a private health insurance system to advance the regressivity agenda, and indeed no country other than the United States has chosen to do so. Any shift in the tax structure – to a less progressive income tax schedule, or to more reliance on consumption taxes, or best of all the introduction or expansion of social insurance premiums perhaps linked roughly to incomes but with a ceiling somewhere in the middle income range – will achieve the primary objective of substantial benefits for those at the top of the income schedule, at the expense of those further down. Or one can simply advocate for "small" user charges.

5.7 Marketing Regressive Financing; Mangling Economic Theory

In pursuit of the regressivity objective, advocates of "reform" in health care financing have to develop arguments for more general benefit. This is the effect, if not the purpose, of models based on Arrow's approach, implicitly assuming a society of clones. With that assumption one can maintain the pretense that economic theory provides an objective, quasi-scientific justification for the "...out-of-pocket payments [that] tend to be a highly regressive means of financing health care..." (van Doorslaer *et al*., 1993, p.42).

But of course economic theory does no such thing, and never could. As Hume pointed out more than two hundred years ago, it is logically impossible to derive normative propositions about how the world

should be, from positive propositions as to how it is – *ought* from *is*. To the extent that it is a science, or at least an academic discipline, economics is the activity of generating and testing positive propositions (valid or otherwise) about how (real or hypothetical) economies work. How they *should* work, in Arrow's terms what is "optimal" cannot be derived from the analysis alone. Such conclusions require the addition of some set of values, some rules for ranking alternative states of the world.

Arrow implicitly postulated that people *should* get what they value most, *as reflected solely by their willingness to pay* – which though apparently self-evident to economists is a substantive and often controversial moral judgement – and then explicitly assumed that in Arroworld everyone wanted the same things and had the same resources. In that world any policy has the same (*ex ante*) impact on everyone. In the real world all policies make some better off and some worse off so that policy choices always favour some interests over others – an inherently political, value-laden decision.

Noting wrong with that – it happens and must happen every day in every society. But a discipline whose participants claim to avoid interpersonal comparisons and to reach scientific, value-free conclusions cannot then be used to claim that some particular set of arrangements is "optimal" – better than any other for everyone. The best that can be done is to map as accurately as possible the distribution of effects associated with different policies. To pronounce professionally on real-world policy questions while deliberately ignoring the incidence of benefits and burdens is indefensible. When these effects are transparent and large, such behaviour suggests covert advocacy for a particular set of economic interests – professional malpractice, if the term had any meaning in economics (see, e.g. Morgan, 2000; Pear, 2003).

The urge to extend professional jurisdiction, in the face of Hume's fundamental logical constraint, has led to a remarkable intellectual dodge, the concept of *potential* Pareto improvement. It has been argued by some economists that one can declare a policy change

unambiguously good if the gainers could fully compensate the losers, and still be better off – *even if no compensation is ever paid.* (Alternatively the losers could not compensate the gainers for foregoing the policy, or both.) This bizarre notion has been memorably skewered by Reinhardt (1992) as the "uncompensated-punch-in-the-nose" test.

> *A is willing, for reasons unspecified, to pay up to $400 for the privilege of punching B in the nose. B is willing to be punched in return for a payment of $200 or more. The economist declares that aggregate social welfare will be improved if A punches B; and B accepts in return for a payment of something between $200 and $400. The punch is delivered; the payment is not. But the economist assures B that there has nonetheless been an aggregate improvement in social welfare, because A could have compensated B and still been better off. At this point, B may punch the economist.*

5.8 Forget the Economists, We Need More Health Care! But Do We?

Another class of arguments for general benefits from shifting the financing mix takes off from the widespread concern over the constant pressure, in all modern health care systems, for increases in health care expenditure. The pressures for expansion arise from an assumed relationship – typically untested – between increased expenditure and an increased quantity or quality of care services, and a corresponding and again rarely tested improvement in the health of some, at least, of the population. Needs are ever expanding, and more money is needed to meet those needs. But these casual assumptions are false.

Examples of effective, beneficial, and often life saving or function preserving health care are legion in modern health care systems, and progress in medical technique continues at a rapid rate. It does not follow, however, that general expansions in output, much less in expenditure, necessarily result in improvements in anyone's

health. Increased spending may be absorbed in increased incomes for resource suppliers; there may be a significant degree of just plain technical inefficiency – unnecessarily costly, resource-intensive ways of producing particular services; or the services themselves may, despite the best intentions of their providers, be ineffective or even harmful to health.

Concerns about the extent of ineffective care, inefficiently provided, have a long history, and well-documented examples abound. More recently evidence has begun to emerge that, at least in the United States, greater expenditure on physicians' services and hospital care is *at the aggregate level* actually associated with *lower* quality care and *poorer* health outcomes for the population served (e.g. Fisher et al. 2003a, 2003b; Baicker and Chandra, 2004; see also the collected papers in *Health Affairs*, 2004).

This evidence should be strengthening long-standing concerns for improving the efficiency and effectiveness of modern health care delivery systems – for getting more "value for money". Certainly the greater value is there to be gotten.

Pharmaceuticals provide a number of recent examples of widely prescribed and quite expensive drugs that have turned out, on randomized controlled trial, to be no better than much cheaper off-patent alternatives, or to have dangerous side-effects. In 2004 the non-steroidal anti-inflammatory *Vioxx* for treatment of arthritis – new and much more expensive than older alternatives such as ASA or ibuprofen – was taken off the market because of findings of an associated risk of stroke. These findings raise questions about the whole class of new "coxibs" (FitzGerald, 2004). In 2002 long-term use of a widely-prescribed form of hormone-replacement therapy was found to increase the risk of heart disease and stroke (Rossouw *et al*., 2002); and later that year hypertension treatment with (off-patent and very inexpensive) diuretics was found to be as effective as treatment with much more expensive (patented) ACE inhibitors or calcium channel blockers (Furberg *et al*., 2002).

These concerns do not arise – indeed have no meaning – in Arroworld. Although the word "efficiency" may occur frequently and centrally in the discussion of welfare burdens, its meaning is totally different from, indeed bears no relation to, more normal usage. In Arroworld, health care and all other commodities are always produced "efficiently" in the normal sense by for-profit firms in competitive markets. Survival in those markets requires them to employ least-cost technology and avoid any waste of resources. The economist's concept of "allocative inefficiency" that is Arrow's concern, means rather that the "wrong" mix of commodities is being produced, that the same resources could be used to produce a different mix that is "better", more highly valued, as evidenced by the fact that people would be willing to pay more for it.

As for the effectiveness of health care services, the question does not arise. Care is, in Arroworld, provided solely in response to consumers' (not patients') willingness to pay. The relationship of health care to health is undefined, because the concept of health itself does not enter the discussion. Once adopted, the assumptions of Arroworld provide no language in which to describe or express concern over ineffective or harmful care, or technically inefficient, wasteful production. Such thoughts are literally unthinkable.

5.9 To Send a Message to Doctors, Write it on a Cheque

In the real world, however, there is a long tradition of keen interest in the relationship between "value for money" and the structure of the right-hand side of Figure 1. How providers of health care are organized, with what objectives and constraints, and particularly how they are paid, are seen as exerting a major influence on health system performance, on how care is provided, to whom, and for what. It is taken as self-evident that providers can and do exercise a substantial degree of influence over the patterns and costs of care used by patients, and will exercise this discretion in different ways depending upon the

terms on which they have access to funding – the upper right-hand side of Figure 1.

How providers respond to different funding incentives, however, will depend upon how they are organized and regulated, and the mix of motives behind their behaviour. The unusual forms of provider organizations that populate the health care industries and the extent of explicit and implicit regulation binding the markets in which they operate, have their origins and justifications in concerns to elicit socially desired provider behaviour. A minimal set of differentially motivated types is represented on the lower right side of Figure 1. This perspective makes a fundamental break from the "supply side" of Arroworld and its intellectual derivatives, in which utilization decisions are made solely by users of care, in response to prices per unit charged by profit-maximizing firms in perfectly competitive markets.

From this perspective, attempts to improve the efficiency, effectiveness, and appropriateness of care delivery are best focused on the objectives of the firms providing care, and the terms on which they have access to funding. As Aneurin Bevan famously remarked: "If you want to send a message to doctors, write it on a cheque." Policies of "structural reform" in the organization of health care delivery hold out the promise of limiting the cost of health care without compromising the health of the populations served.

Perhaps the archetypical illustration of this line of argument has been provided by the capitation-reimbursed prepaid group practices in the United States, "re-branded" in the 1970s as health maintenance organizations (HMOs). For more than fifty years these organizations have served their enrolled populations at costs, and particularly at rates of hospital utilization, that are significantly below those of the surrounding fee-for-service delivery system. Comparative studies, including trials with randomized assignment of enrollees, have shown that these cost advantages are real; they do not arise from patient selection or under-treatment. How providers are funded, does matter.

These observations have been discussed in the academic literature over the last fifty years, though studiously ignored by the analysts of Arroworld and their concerns about welfare burdens. Yet they had surprisingly little impact on public policies toward health care prior to the 1980s, and in many countries (Canada, for example) they still have little. The academics and "policy wonks" know the stories well, but the high-level public and political debates have been focused elsewhere. This disjuncture underlies a long-standing interest, beginning about thirty-five years ago, in whether structural reform on the upper right hand side of Figure 1 might be promoted by linking it to competitive market forces focused on the upper left-hand side.

Could Competitive Markets Drive Structural Reform? Enthoven's Long Quest

The best-known advocate for such a linkage is Alain Enthoven, who for nearly thirty years has been proposing and modifying suggestions for creating a competitive market for financing organizations (e.g. Enthoven, 1980; Enthoven and Kronick, 1988a,b). The core idea is that households should have the opportunity – and the financial incentive – to select among insurers (public or private) on the basis of their premiums for a comprehensive package of coverage of "all necessary care".

Insurers that were most successful in "managing" the care process – negotiating with providers to purchase for their enrollees a package of effective care, efficiently provided, would earn profits and expand their market share. In the process, much would be learned about "what works and what does not", and most importantly, what was learned would, under the discipline of the competitive market, actually be applied. (Kindig (1998) has taken this idea to its logical next step, suggesting that such organizations go beyond purchasing health care, to being paid on the basis of their success in improving the health of their enrollees, by whatever strategies are most effective.)

Much can be, and has been, said and written about the feasibility and risks of this class of strategies. They have the considerable virtue that, as Enthoven emphasized, they do not assume or rely on fully informed "consumers" to determine, by their price-sensitive choices of particular diagnostic and therapeutic interventions from different providers, the mix and volume of health care produced, or its mode of production. But all strategies built around individual choices and market incentives, if they are not simply to replicate market outcomes – them as has, gets; and them as hasn't and gets sick, has had it – require enormous amounts of information to build in compensation for individual differences in risk and resources.

They also require highly sophisticated and complex monitoring systems to detect and discourage the multitude of forms of opportunistic behaviour that market systems encourage and reward. The current state of "managed care" in the United States contrasted with that of "total commissioning practices" in the United Kingdom is at least suggestive that the idea of organizations paid by capitation to provide a comprehensive package of care to a designated population may be on the right track, but that embedding such providers in a competitive private market is a recipe for failure. Arroworld, where everybody has perfect information about everything (except who will become ill next year), has no need for concepts of patient trust or professional obligation. In the real world, however, these appear to be essential for system survival.

Forward to the Past: Patients at Risk Leads to More Regressive Financing (Again)

Be that as it may, the key notion in the Enthoven approach was that households must select among care managers on the basis of their relative costliness. But once households are placed at financial risk, a policy that begins as a means of improving health care delivery can easily become a means of generating a more regressive pattern of financing. The Netherlands, where Enthoven's ideas appear to have been particularly influential, provide a recent and clear illustration of this process (Helderman *et al*., forthcoming, 2005).

In early 2002 the left- and right-wing parties in the then-governing "purple coalition": …strongly disagreed about the method of premium setting in the national insurance scheme. The Labour Party adhered to a largely income-related contribution and a relatively small flat rate premium as already present in the sickness fund scheme. The conservative Liberal Party, however, favoured a fully community-rated premium [i.e. unrelated to income] with tax compensation for those on low incomes.

This disagreement had no significance for the operations of the health care system *per se*; it was simply an unusually naked struggle over the incidence of system costs. The Liberal Party were trying to use the process of "reform" to reduce the share of costs borne by their higher-income supporters through a manoeuvre equivalent to substituting a poll tax for an income tax. The Labour Party were resisting this regressive shifting of costs farther down the income spectrum – to *their* members.

The two-part premium system (income-related plus fund-specific flat charge) was originally introduced with the presumption that it would encourage competition among insurers in promoting efficiency and containing costs. The larger the flat-rate component, the greater the alleged incentive for insurers to compete in the effective management of the care of their enrollees. The more regressive burden distribution would "buy" more efficient or effective health care – the equity/efficiency trade-off so firmly embedded in the minds of a certain school of economists. "Distributional considerations" could be dealt with by subsidizing the poor, which still leaves plenty of scope for moving costs from the higher to the middle and lower bands of the income spectrum.

What happened, however, was rather different. Between 1995 and 2002 the maximum spread among flat-rate premiums charged by different sickness funds rose from three percent to over fifty percent. (Helderman *et al*., op. cit.). That a spread of this magnitude has arisen from differences in fund effectiveness in managing health care, is not credible. The "reform" has enabled the better-off to buy

a substantially enhanced service package within the "universal" public system, even as it increases the regressivity of contributions compared with a strictly income-related system.

The Liberal/Labour disagreement in 2002 suggests that such a result should not be a complete surprise. A regressive agenda was hidden behind the promise of greater efficiency through market competition. Managing health care delivery and containing costs is an extremely difficult undertaking – as managers of every system in the world are acutely aware. A competitive, profit-oriented management is usually better advised to concentrate on patient selection. Accordingly for-profit managed care firms have been endlessly inventive in finding ways to recruit a less costly enrollee mix. This activity has been most highly developed in the United States, but Australia, where private firms are not permitted to discriminate by risk in premium setting, also offers examples that are at least entertaining.

Proposals to promote right-hand side reform by linking in and modifying the left-hand side financing systems, to place enrollees at some financial risk, at least start from the recognition that current patterns of health care production and use offer considerable opportunities for improvement. Significant savings are there to be extracted, if only one could find better forms of organization and incentives for performance. ("Better" is, of course, judged in terms of the financial cost of achieving some health outcome, or at least of producing a basket of services; it bears no relation to Arrow's concept of "optimality".)

This may be a genuine objective, as in Enthoven's early analysis, which becomes modified into the promotion of more regressive financing (and/or "progressive" access) as the difficulties of genuine system reform become apparent. Alternatively the regressive agenda may in fact be covert behind all proposals for "market-based" reform – as in the recent Dutch example. But at least the rhetoric of greater value for money is prominent.

Value for Money? No, Just More (Private) Money: The Rhetoric of Unsustainability

This rhetoric contrasts with another and recently popular line of argument for promoting a more regressive incidence of health care costs – "unsustainability" – that turns the welfare burden analysis completely inside out. Advocates of a shift to more private financing in universal public systems argue that health care is being *under-used*, not over-used. Health care is underfunded because of excessively tight public controls on spending. This is an ancient argument among providers of care, for whom any increase in expenditure is by accounting necessity an increase in aggregate incomes.

But the argument has a new twist. Public programs do not fund health care adequately, not because they will not, but because they cannot. Public sector resources are insufficient, or will be in the near future, to support an adequate level of health care provision. Expansion or even maintenance of current levels of public financing is "not sustainable"; if current and emerging needs are to be met additional revenues must be raised through private sources. There is no alternative.

In terms of Equation 1 above, the "welfare burden" argument supports advocacy of more out-of-pocket payment, with the objective of shifting real resources from the production of health care to that of more highly valued "other things" – raise C so as to lower Q and transfer some of the Z to other production. The "unsustainability" argument, by contrast, finds private third-party insurance coverage quite acceptable, even desirable. The objective is to increase the flow of financing into health care while shifting it from less to more regressive sources. A shift from public tax to private insurance financing would achieve precisely this – raising Q and Z while shifting the financing mix toward a higher proportion of C and PI.

It is not always clear whether the objective of increased financing necessarily implies an increase of real resources into health care – and away from the production of "other things" – or merely an increase in the rate of remuneration of those supplying resources for

health care production. In terms of Equation 1 above, some of those who claim underfunding advocate increases in Z to support increases in Q, while others may simply be seeking increases in W and P. But either interpretation is directly inconsistent with the presumptions of the "welfare burden" argument, in which care is over-provided, and rates of remuneration cannot be changed because they are fixed in an imaginary perfectly competitive market for resources.

Those advancing the "unsustainability" claim are often a bit fuzzy as to the precise criteria for "too much" or "too little" health care. Superficially, they seem to be adopting a strictly needs-based standard, and rejecting the criterion of willingness to pay that underlies the economic concept of allocative efficiency. More private money would permit an expanded health care system to meet more needs. On this criterion, however, the "unsustainability" argument is vulnerable in that it must bypass – simply ignore – the whole question of "value for money" and the extensive evidence of inappropriate care, inefficiently provided. It appears as a rather mindless assertion that "more would be better".

In the hands of some proponents, however, the argument seems to undergo a subtle shift such that the willingness to pay criterion enters through the back door. If some better-off individuals would be willing to pay privately for more care, of perceived higher quality, more rapidly accessible, then they should be permitted to do so, whether through private insurance or out-of-pocket. Care is allegedly being underprovided (in quantity and/or quality), relative to the willingness-to-pay of the better off, because of public sector constraints.

But the better-off should not be required to support a similar standard, through public financing, for the rest of the population. This would be "unsustainable" in effect *overprovision* (defined relative not to their needs but to their unwillingness or inability to pay). The broader issue of alleged global underfunding relative to needs blurs into the narrower one of preserving privilege for those able to pay.

There is unquestionably evidence of localized underprovision or overly delayed provision of particular services in all health care systems. Modern health care is a huge and highly complex enterprise, and there is no question but that management techniques have not kept pace with this complexity – partly because of covert and sometimes overt internal opposition by providers. But the "unsustainability" argument converts such local observations into a claim of systemic underprovision, and that in turn into a claim of global undersupply of resources rather than failures of management. These logical progressions are not made explicitly, and lack any marshalling of evidence to support what is otherwise a logical *non sequitur*. Indeed claims of systemic underprovision have been impervious to direct evidence that localized shortages are *not* the consequence of systemic underfunding.

The main leap of faith, however, in this form of advocacy for more regressive funding, is that the public sectors of modern, high income economies are simply incapable of mobilizing sufficient resources to meet their populations' needs (or possibly wants) for health care, either now or in the not too distant future. More money is needed, governments cannot raise it, therefore private funding *must* be expanded. QED.

Just why the necessary resources could not be raised through public sources – assuming that they were in fact needed – is never made clear. At this point the argument becomes very obscure and seems to involve a lot of rhetorical hand-waving. (In the sermon notes of the legendary Scottish preacher expounding free will versus determinism: "This point very doubtful. Shout loudly.") Apparently modern societies can no longer afford expanding public health care systems, but can perfectly well afford more rapid expansion if it is financed from private sources – user fees and private insurance. But from a society-wide perspective, why is the one sustainable and the other not?

Those alleging the unsustainability of universal public systems sometimes suggest that in some mystical way a given expansion in

health care spending will reduce economic growth and general welfare if financed through the public sector, but will promote them if financed privately. No evidence is offered. The most plausible interpretation, however, is the simplest – the incidence of the additional costs will be more regressive. Expansion through private channels will bear less heavily on the higher income members of the population than would an equivalent expansion through public sources. And it will offer them privileged access to services, often even to those financed primarily through the public sector. The welfare burden analysis, implicitly assuming a society of clones, could hardly be less relevant. But economists willing to adopt the Arrow assumptions can be recruited in support of a larger role for user charges, which then provide a market for private insurers, and a more regressive financing mix overall.

Apocalyptic Demography, Limitless Technology, and Empty Treasuries? – Here Come the Zombies Again

That this is the real thrust is suggested by the speciousness of the arguments offered in support of unsustainability. The claim is very widespread that demographic trends, aging populations, will simultaneously increase health care costs and reduce the working population available to support them. This may well be a significant problem for pension funds and more generally for retirement policy. But studies of the impact of population aging on health care costs have consistently shown them to be real but small, well within the normal rates of growth of modern economies, and nowhere near the magnitude of effects of trends and policies within health care systems themselves. Nor are the international variations in health care costs among high-income societies linked in any way to the relative ages of their populations. The issue is no longer an interesting question for research; what is interesting is that the scenarios of "apocalyptic demography" are completely imperviousness to the overwhelming counter-evidence. They are classic examples of the "zombies" referred to above.

Yet another specious argument, of the "everybody benefits" form, has been the claim that users of public systems would actually benefit

from the existence of a second, privately financed tier to relieve the otherwise "unsustainable" pressure on those systems. By not using the publicly funded systems, those in the upper tier free up capacity to improve access for those in the lower tier. And indeed this might be true, if those in the upper tier were purchasing services from Mars. But if the same providers are offering services to patients in both tiers, and being paid higher fees by those in the upper tier, one does not have to be an economist to predict that they will manage access so as to encourage "those who can afford it" to seek private care. And of course they do, in every two-tier system in the world. If access were equalized, why would anyone "go private"?

Yet again, it is argued by right-wing governments in several of the Canadian provinces that health care costs are escalating so fast that they are absorbing a larger and larger share of public budgets and crowding out other worthwhile forms of public spending. This is clearly unsustainable, and more private funding is essential to preserve other public services while supporting the continuing expansion of the health care sector.

What they do not say, is first that each of these provinces has put through major cuts in income tax rates over the last five years – cuts that were proportionately more beneficial for taxpayers in the higher income brackets. Yet even with these tax cuts, spending on the universal public health insurance programs now takes up roughly the same share of provincial revenues – as opposed to expenditures – as in the early 1980s (Evans, 2004, 2005). Right-wing governments have been cutting both income taxes and public (non-health) expenditures, but public reaction has made it difficult for them to cut health spending as well.

It is of course undeniable that sufficiently aggressive tax cutting can make public programs of all sorts – health care or otherwise – "unsustainable". That appears to be most clear in the current strategy of the Republican Party in the United States. Its' significance for other high income countries is fortunately much less clear. But what is clear is that claims of fiscal "unsustainability" of modern health care systems, and corresponding efforts to shift the financing balance of

those systems, are rooted not in fiscal exigency but in redistributional objectives.

The point is put with particular bluntness and vigour by Vladeck (2000) with respect to the Medicare system in the United States:

> *"The "crisis" in financing Medicare and Social Security has largely been manufactured by individuals and institutions with a political agenda to shrink or abolish those programs, and we should stop letting ourselves fall victim to their propaganda."*

But his point has applicability wherever claims of unsustainability are made.

That is not to say that vigilant cost containment in all health care systems, public or private, is unnecessary. It is now thirty years since Aaron Wildavsky (1977) formulated the "Law of Medical Money", which states in effect that a medical system will absorb as much money as it can acquire, irrespective of any other considerations, and that there is no way of containing that urge other than by strictly controlling the supply of money. Thirty years of experience have repeatedly confirmed this Law. But industrialized countries other than the United States have for the last quarter century worked out administrative mechanisms that, while far from perfect, at least manage to place reasonable constraints on the growth of their health care systems.

Indeed the outstanding exception, the United States, is also the country with the highest level of private *and of public* funding for health care (relative to national income). But it also has one of the most regressive financing systems among OECD countries. Private funding buys regressivity, not sustainability – and that is presumably the dominant political objective. Reinhardt's (2001) comment on the United States Congress is worth quoting:

> *"That no one in the U.S. Congress shows much interest in the glaring inefficiencies that could easily be addressed within the current Medicare program [in the U.S., covering only those 65 and over] speaks volumes about the true, but hidden, agenda that actually drives the quest for privatizing ... Crisply put, the objective is to shift responsibility for health spending on older persons from the general taxpayer onto the older people themselves..." (p.201).*

Variations on a Common Theme: Distributional Tensions in the ECuity Project Findings

The European situation is much more diverse, with most countries having struck a very different balance between social solidarity and the pressures for regressivity and privilege. While public financing is everywhere predominant, the mix of public sources and the extent of private financing can be structured to yield very different patterns of incidence. Figure 7, drawn from Tables 2 and 6 in Wagstaff *et al.* (1999), shows for a number of European countries the relationship between the progressivity or regressivity of the overall burden distribution, and particular components of the financing mix. (Regrettably, the ECuity Project has not been in a position to update this extremely significant comparative data in this paper and its companion, van Doorslaer *et al*. (1999).)

Figure 7a. *Relation between Private (Direct Payments) Financing and Progressivity of Total Health Expenditures, Selected OECD Countries and Years*

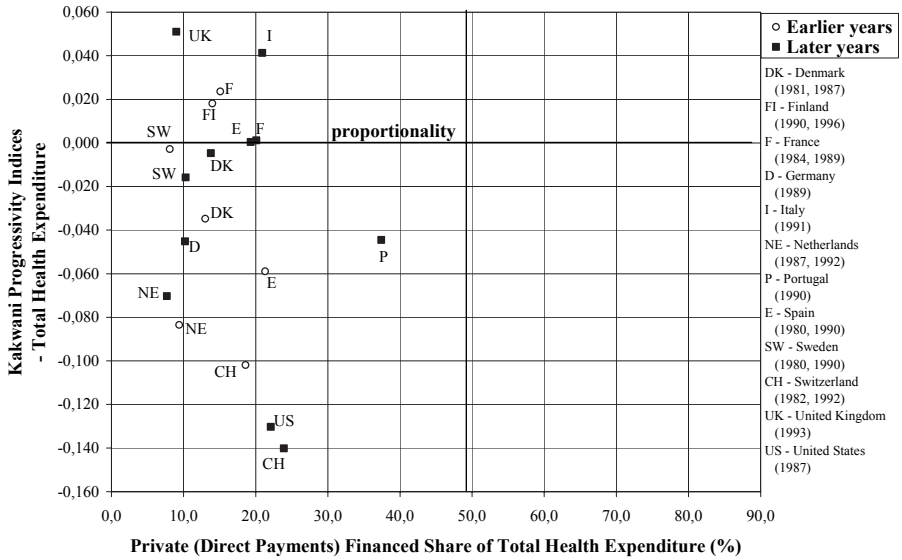


Figure 7b. *Relation between Private (Total) Financing and Progressivity of Total Health Expenditures, Selected OECD Countries and Years*

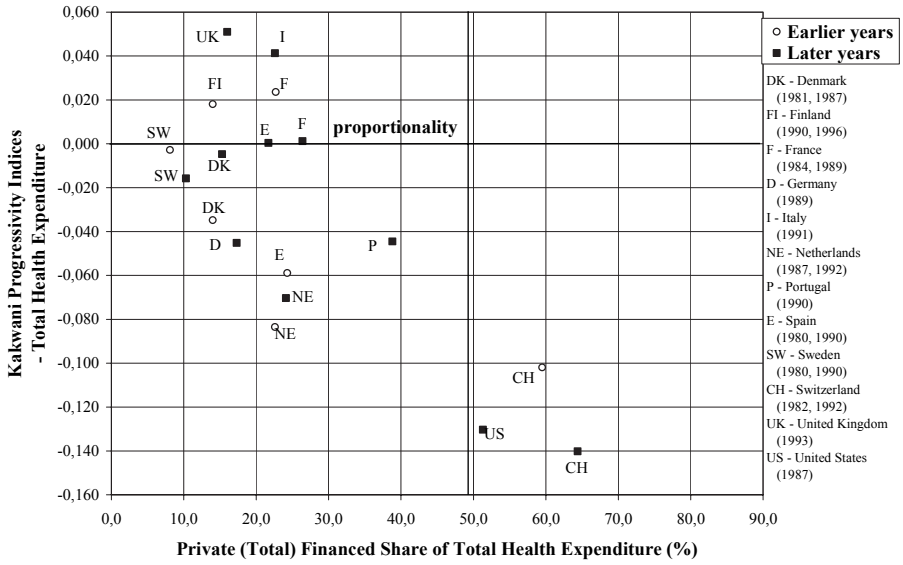
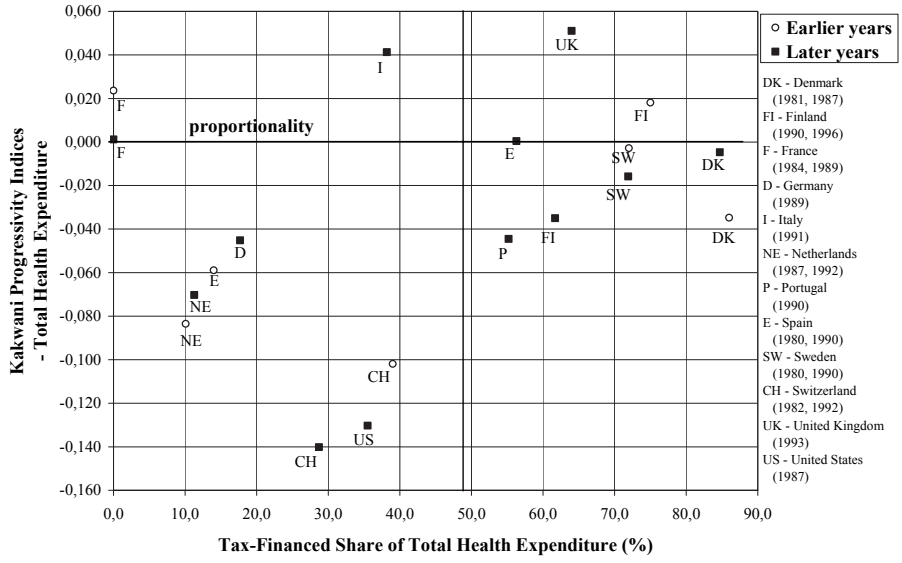


Figure 7c. *Relation between Tax Financing and Progressivity of Total Health Expenditures, Selected OECD Countries and Years*



Each of the panels of Figure 7 displays Kakwani indices (Kakwani, 1977) calculated for the various countries in the ECuity Project study, at different points in time. This index summarizes the progressivity or regressivity of the distribution across income classes of total health care expenditures – through all the channels on the left-hand side of Figure 1. It takes a value of zero if families at each income level pay on average a roughly equal proportion of their incomes for health care. A positive index indicates that the proportion of family income paid for health care, through all channels, rises with family income; a negative value indicates that the share of income paid for health care falls, on average, with rising income. The size of the index indicates the strength of the corresponding relationship.

Most of the countries in the study, at least at that time, were clustered relatively closely around the proportionality line, with the United States, Switzerland (CH) and to a lesser extent the Netherlands as outliers on the regressive side. There were no countries with correspondingly strong progressive financing systems, the most progressive – the UK and Italy – had Kakwani values that were only half the size, in absolute value, as those in the US and Switzerland.

Figure 7a plots these index values against the proportion of health expenditures financed from out-of-pocket payments. While out-of-pocket payment is the most regressive form of financing, bearing most heavily on those at lower incomes, it turns out that there is relatively little variation across countries in the extent of reliance on this form of financing. Differences across countries in the regressivity or progressivity of financing systems as a whole do not appear to arise primarily from this source.

On the other hand when one plots the measure of progressivity or regressivity against total private funding, as in Figure 7b, quite a strong relationship emerges. It would appear that it is differences in the extent of private insurance coverage that primarily determine the distribution of financing burden. In most of these countries private insurance coverage plays a very small role in financing health care – as observed in the more extensive OECD data shown in Figure 5. But where such coverage exists on any significant scale, it is a

powerful mechanism for maintaining a regressive distribution of the overall expenditure burden.

This is not a new observation. In reporting results from the fist round of the ECuity Project, van Doorslaer et al. (1993) observed that:

> *"The two countries with predominantly private financing systems – Switzerland and the US – have the most regressive structures overall. This is scarcely surprising in view of just how regressive private insurance and out-of-pocket payments are when used to finance such a large proportion of health care expenditures for such a large proportion of the population. The group of countries with the next most regressive systems are the countries operating the so-called social insurance model,...countries which...rely mainly on tax-finance... have the least regressive financing systems. (p.44)"*.

The later papers confirm this observation.

This pattern presumably arises because there is such a large variation in health expenditures *within* income classes, as illustrated in Figure 4. Heavy reliance on out-of-pocket financing implies a correspondingly large and capricious variation in the distribution of economic burdens within as well as across income classes, a variation that will also undermine any objective of equality of access. This appears to be politically unacceptable in high-income countries, even (for the majority of the population) in the United States. Accordingly the regressivity agenda must be pursued through private insurance coverage, with advocacy of out-of-pocket payment not as an end in itself – as the "welfare burden" argument would imply, but as a way of providing something for private insurance to cover.

But there are other ways to advance that agenda. Figure 7c plots the Kakwani Indices against the proportion of health care financing drawn from general taxation. There is a clear relationship, as previously observed by van Doorslaer et al. above. More financing

from general taxation tends to be associated with a less regressive incidence of health care costs. But social insurance financing can be quite progressive, as in France, or quite regressive, as in Germany, depending upon how closely the premiums are linked to income. A ceiling on premiums can result in a relatively regressive system, limiting the liabilities of those with higher incomes.

Still better for the wealthy is the opportunity to opt out of a "universal" social insurance system entirely, as in Germany or the Netherlands. This enables the better-off to secure preferred access as well as more regressive financing. On the other hand the ability to "go private" in a system financed primarily from general taxation, as in the United Kingdom, serves to secure better access, but does not exempt the wealthy from contributing, in proportion to their incomes, to the costs of the public system.

The extent of private financing shows up consistently as a major source of system regressivity, underlining the fact that the" public/ private" debate is always and inevitably about "Who Pays?" But the source tables in Wagstaff et al. (1999) reveal some subtleties. Income taxation, at least at the time of these analyses, tended to be *more* progressive in the United States, Switzerland, and Germany than in the Nordic countries, although the inequality of incomes is significantly greater (see also Figure 6).

This suggests a possible long-term political trade-off; where the tax system is less progressive and incomes are at least relatively more equal there may be less resistance to taxation as the primary source of health care finance. A fully tax-financed health care system in the United States, or more plausibly, say, 80% tax financing, would bring about a great deal more redistribution down the income spectrum that it does in Sweden. Correspondingly the political resistance is more intense.

There is another dimension to the contest over incidence. Most of the countries included in Figure 7 have more or less regressive financing systems, taking a higher proportion of the incomes of people with

lower incomes. Some are close to proportional, others are quite highly regressive. Only a few are progressive and those not by much. It is important to note, however, that since health expenditures on behalf of lower income people are typically much larger in relation to their incomes than those for higher income people, a mildly regressive financing system that supports universal access and public payment will still transfer significant resources from higher to lower income people (van Doorslaer *et al.,* 1999). This transfer is accentuated by the fact that health status is negatively correlated with income.

Moreover the actual amount of redistribution that takes place will vary according to the scale of the health care system itself. The more expensive the system, relative to the total income of the country, the larger the amount of income that will be redistributed for any given pattern of incidence. Thus Switzerland and the United States, with the most regressive financing systems, are also the countries spending the largest share of their national incomes on health care. The usual interpretation, from a political economy perspective, is that the politically dominant economic interests in these two countries favour private financing sources because they are relatively regressive, but privately financed systems are also least effective in controlling health care cost escalation.

Undermining Cost Containment While Advancing Inequality: A Dual Role for Arroworld Analyses

But the political causality may run the other way as well. Because these systems are so expensive, a less regressive financing system would transfer relatively more money down the income distribution than it does in, say, the United Kingdom or Sweden. High system costs provide a greater incentive among the politically influential classes to preserve or promote regressive financing structures. If this mechanism is at work, it might help to explain some of the rising political pressures for regressivity that can be observed in several of the OECD countries. As health costs rise, there is more to gain by pushing a larger share of them onto someone else. Certainly advocacy

for the regressivity agenda is concentrated primarily at the upper end of the income distribution.

In advancing this agenda, the myths of Arroworld are ready to hand. It is not difficult to find an economist willing, even eager, to lend professional support to the proposition that a shift to a more regressive financing mix will moderate the rate of cost escalation. After all one of the icons of the field, indeed one of the leading economic theorists of his generation, has long ago established the precedent, followed by a number of prominent health economists, for "ignor[ing] distributional considerations" and the absence of supporting evidence at the aggregate national level, indeed the significant counter-evidence, has no impact on strongly-held theories.

The linkage between escalating health care costs and increased pressure for more regressive financing mechanisms opens a broader interpretation of the incidence of health care costs. As illustrated in Figure 1, total health care expenditures are always and inevitably equal to total income drawn from the provision of health care. It follows that policies to improve the efficiency and effectiveness of health care delivery must necessarily threaten the incomes of some of those employed in or profiting from the provision of care.

Reductions in inpatient hospital use, for example, long overdue on clinical grounds, have reduced the availability of low-skilled jobs in this sector. The discovery that a popular form of hormone replacement therapy was actually a threat to health, not a benefit, sent the shares of Wyeth Corporation down 40% in a week. The similar finding for Merck's coxib Vioxx dropped its stock price by one-third. And if doctors could ever be convinced to absorb and act upon the findings of comparative tests of different anti-hypertension drugs, costs of this therapy could be deflated overnight and pharmaceutical stock prices would go with them.

It does not matter whether the intervention is dramatically effective, useless, or downright dangerous – so long as it is being paid for, someone is being paid. And those who are paid can be counted on to

resist, in a variety of ways, efforts to mitigate the escalation of health care costs that threaten their particular incomes.

Again the fantasies of Arroworld are at hand to conceal the significance of these fundamental accounting realities. In Arroworld, everyone is identical so that a shift in the mix of products produced – more (or less) health care and less (or more) of other things – has no distributional consequences. By hypothesis everybody is paid for (its'?) labour and other resource inputs at exactly their "opportunity cost". This means that if the market for one form of output shrinks, the labour etc. freed up can be instantly and costlessly redeployed into producing the alternative goods. Losing a job is no hardship, because there is always another, equally well-paid and satisfying, available next door – albeit in a different industry, producing those unspecified but more highly valued "other goods". There are no years of professional training or experience to be written off, no early retirements for redundancy. (Actually, people seem to live forever in this timeless world.).

But back in the real world, there are deeply embedded and bitterly defended interests in every health care system. Policy choices have an enormous impact on who gets paid, for what, and how much. This is the back-drop to Reinhardt's comments on the US Congress, quoted above. There are massive opportunities for system rationalization and cost containment, in the American system more than any other. But these huge pots of gold are all guarded by particularly ferocious dragons that are perfectly well aware of the distributional consequences of improvements in system efficiency or effectiveness. At present, it appears that the American government has decided to abandon efforts to contain its hypertrophied health care system, and to concentrate on trying to shift the burden of financing Medicare from taxpayers to the elderly themselves.

Exactly the same is now happening in the private sector. The American system of employment-based private health insurance (albeit with heavy government subsidy) experienced several years of relative cost stability after 1992 – sufficient to get past what

appeared in 1994 to be a serious threat of national health insurance. Now "managed care" appears to have collapsed and costs (and health care incomes) are again escalating uncontrollably. The reaction has been to begin pushing costs back onto employees, and particularly to cut supplementary benefits for future retiring employees. This group appears not to be aware of the extent to which their future coverage has already been eroded, and the medium-term future may be quite interesting when they find out. But the key point for our purposes is the way in which an unmanageable and politically too potent collection of industries is indirectly driving the regressive agenda through the pressure of uncontrollable cost escalation.

The United States is the extreme example by far – as in so many other ways it is "not a country like the others". But a similar process may be underway in other OECD countries with respect to pharmaceutical reimbursement. There are now a number of examples of otherwise universal, publicly-financed systems dealing with uncontrolled pharmaceutical cost escalation by transferring costs to patients. Failed cost control efforts tend to evolve – often with the enthusiastic encouragement of providers – into regressive cost shifting (Evans, 1990).

Any discussion of incidence, whether of expenditures or of incomes, inevitably focuses on conflicts of interest. At least in the short run there are typically no policies that do not threaten someone's interests. Yet at the same time there *are* obviously opportunities for very general, if not necessarily universal, benefits from choices in health policy. One could get pretty general agreement that inappropriate prescribing of drugs that are a threat to the health of some or all to whom they are offered, is not a social *desideratum*. Nor is keeping patients in hospital beds far longer than is necessary. These services are not "goods" in the economist's sense of the word; they are threats to health no matter who gets paid for them.

The criterion of health benefit is a much more politically potent tool, in principle at least, for achieving (almost) consensus on what should be done and what should not than are the fantasies of Arroworld. But the

process of winning through to any such operational consensus is often gravely, sometimes fatally, hampered by the noise and confusion and outright disinformation generated pursuant to the conflicts over the incidence of corresponding burdens and benefits, the very real conflicts of economic interests. In this conflict, economists have absolutely no legitimate role in pretending to offer objective, professionally-based determinations of "optimal" arrangements (though they will certainly find interested parties willing to pay them to do so).

But we may play a more useful role by mapping the terrain and providing plausible assessments of the likely consequences of different policies – particularly their distributional impacts – and unmasking the distributional agenda lying behind so many proposals for "reform". One should always keep front and centre the old Roman question: "Cui bono?" or the slightly more up to date: "Cherchez l'interêt." Careful economic analysis can be very powerful for this purpose, so long as it is grounded in the institutional and behavioural realities of working health care systems.

Analysis that has been "lobotomized" by the assumptions of Arroworld, however, – and this includes the whole "welfare burden" tradition – is thereby rendered incapable even of formulating the questions, let alone identifying the interests, that drive most of the real-world debates over health care policy. Conclusions drawn from such crippled analyses become merely part of the propaganda weaponry of those promoting a regressivity agenda. They have not, outside the United States, been a markedly effective weapon. But they can be quite a significant distraction, absorbing a great deal of professional and sometimes public attention that could better be focused on serious efforts to improve health care system performance. How many more times must economists, in particular, go through the same sterile "public-private" debates – conversations of the deaf – whose terms have not changed significantly in thirty years? It might help if we could agree that any analysis of health care financing that "ignore[s] distributional considerations" should itself be ignored.

As for Arrow himself, we may read him two ways. He may have been the brilliant theorist, living on the island of Laputa totally out of touch with the world around him. More charitably, however, we could read him as posting a warning to those faced with prescriptions for "optimal" financing arrangements on the basis of what is essentially intermediate microeconomic theory for undergraduates, by spelling out clearly the essential, and absurd, underlying assumptions.

References

Arrow, K. (1963) "Uncertainty and the Welfare Economics of Medical Care", *American Economic Review* Vol. 53, no. 4, (December) pp. 941-73.

Arrow, K.J. (1976) "Welfare Analysis of Changes in Health Coinsurance Rates" in R. Rosett, ed. *The Role of Health Insurance in the Health Services Sector* New York: National Bureau of Economic Research, pp. 3-23. (See also R-1281-OEO Santa Monica, Ca.: The Rand Corp., 1973.)

Baicker, K. and A. Chandra (2004) "Medicare Spending, The PhysicianWorkforce, And Beneficiaries' Quality Of Care" *Health Affairs* Web Exclusive (April 7) pp. 184-197.

Barer M.L., R.G. Evans, C. Hertzman and M. Johri (1998) "Lies, Damned Lies, and Health Care Zombies: Discredited Ideas that Will Not Die" HPRU 1998:05D (March), Centre for Health Services and Policy Research, UBC, Vancouver, and HPI Discussion Paper 10, Health Policy Institute, The University of Texas-Houston, Health Science Center.

Enthoven, A.C. (1980) Health Plan: the only practical solution to the soaring costs of medical care Reading, Mass.: Addison-Wesley.

Enthoven, A.C. and R. Kronick (1988a) "A Consumer Choice Health Care Plan for the 1990s: Part I" *New England Journal of Medicine*, Vol. 320, no. 1, pp. 29-37.

Enthoven, A.C. and R. Kronick (1988b) "A Consumer Choice Health Care Plan for the 1990s: Part II" *New England Journal of Medicine*, Vol. 320, no. 2, pp. 94-101.

Evans, R.G. (1990) "Tension, Compression, and Shear: Directions, Stresses, and Outcomes of Health Care Cost Control" *Journal of Health Politics, Policy and Law*, Vol. 15, no. 1, (Spring) pp. 101-128.

Evans, R.G., M.L. Barer, G.L. Stoddart and V. Bhatia (1994) *Who Are the Zombie Masters, and What Do They Want?* Toronto: The Premier's Council on Health, Well-being and Social Justice, June.

Evans, R.G. (2004) "Financing Health Care: Options, Consequences and Objectives" in G.P. Marchildon, T. McIntosh and P.-G. Forest (eds.) *The Fiscal Sustainability of Health Care in Canada: The Romanow Papers, Volume I* Toronto: University of Toronto Press, pp. 139-197.

Evans, R.G. (2005) "Political Wolves and Economic Sheep: The Sustainability of Public Health Insurance in Canada" in Alan Maynard, ed. *The Public-Private Mix for Health* London: The Nuffield Trust, forthcoming.

Feldstein, M.S., (1973) "The Welfare Loss of Excess Health Insurance" *Journal of Political Economy* Vol. 81, no. 2 Part 1 (March-April) pp. 251-280.

Finkelstein, M.M. (2001) "Do factors other than need determine utilization of physicians' services in Ontario?" *Canadian Medical Association Journal* Vol. 165, no. 5 (September 4) pp. 565 - 570.

Fisher, E.S. *et al.* (2003a) "The Implications of Regional Variations in Medicare Spending: Part 1. The Content, Quality and Accessibility of Care" *Annals of Internal Medicine* Vol. 138, no. 4 (Feb. 18, 2003) pp. 273-287.

Fisher, E.S. *et al*., The Implications of Regional Variations in Medicare Spending: Part 2. Health Outcomes and Satisfaction with Care" *Annals of Internal Medicine* Vol. 138, no. 4 (Feb. 18, 2003) pp. 288-298.

FitzGerald, G.A. (2004) "Coxibs and Cardiovascular Disease" *New England Journal of Medicine* Vol. 351, no.17, (October 21) pp 1709-1711.

Forget, E., R. Deber and L.L. Roos (2002) "Medical Savings Accounts; Will They reduce Costs?" *Canadian Medical Association Journal* Vol. 167, no. 2 (July 23) pp. 143-147.

Fox, D.M. and P. Fronstin (2000) "Public Spending For Health Care Approaches 60 Percent" *Health Affairs* , Vol. 19. no. 2 (March/April) pp. 271-273.

Fuchs, V.R. (1996) "Economics, Values, and Health Care Reform" *American Economic Review* Vo. 86, no. 1 (March) pp. 1-24.

Furberg, C.D. and the ALLHAT Investigators (2002) "Major Outcomes in High-Risk Hypertensive Patients Randomized to Angiotensin-Converting Enzyme Inhibitor or Calcium Channel Blocker vs Diuretic" *Journal of the American Medical Association* Vol. 288, no. 23 (December 18) pp. 2981-2997.

Health Affairs Web Exclusive, (2004) October 7.

Helderman, J.-K., F.T. Schut, T.E.D. van der Grinten and W.P.M. van de Ven (2005) "Market-oriented health care reforms and policy learning in the Netherlands" *Journal of Health Politics, Policy and Law* Vol. 30, no. 1 (forthcoming).

Kakwani, N.C. (1977) "Measurement of Tax Progressivity: An International Comparison" Economic Journal Vol. 87, issue 345 (March) pp. 71-80.

Kessel, R.A. (1958) "Price Discrimination in Medicine" Journal of Law and Economics Vol. 1 (October) pp. 20-53.

Kindig, D.A. (1998) *Purchasing Population Health. Paying for Results* Ann Arbor: University of Michigan Press.

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* Vol. 77 no. 3 (June) pp. 251-277.

Morgan, S.G., M.L. Barer and R.G. Evans (2000) "Health economists Meet the Fourth Tempter: Drug Dependency and Scientific Discourse" (Editorial) *Health Economics* Vol. 9, pp. 659-67.

Mustard, C.A. et al. (1998a) "Use of Insured Health Care Services in Relation to Income in a Canadian Province" in M.L. Barer, T.E. Getzen and G.L. Stoddart (eds.) *Health, Health Care and Health Economics: Perspectives on Distribution*, Chichester: John Wiley, 1998, pp. 157-78.

Mustard, C.A., M.L. Barer, R.G. Evans and J. Horne (1998b) Paying Taxes and Using Health Care Services: The Distributional Consequences of Tax Financed Universal Health Insurance in a Canadian Province. CSLS (Centre for the Study of Living Standards) conference on The State of Living Standards and the Quality of Life in Canada, October 30-31, Ottawa. (www.csls.ca/events/oct98/must1.pdf)

Nyman, J. A. (1999) "The economics of moral hazard revisited" Journal of Health Economics Vol. 18, no. 6 (December) pp.681-828.

Pauly, M.V. (1998) "The economics of moral hazard: comment" *American Economic Review* Vol. 58, no. 3, part 1 (June) pp. 531-537.

Pauly, M.V. (1969) "A Measure of the Welfare Cost of Health Insurance" *Health Services Research* Vol. 4, no. 4 (Winter) pp. 281-292.

Pear, R. (2003) "Drug Companies Increase Spending to Lobby Congress and Governments" *New York Times* May 31.

Pen, J. (1971) Income Distribution: Facts, Theories, Policies New York: Praeger, 1971.

Peterson, M.A., ed. (2001) "Kenneth Arrow and the Changing Economics of Health Care" *Journal of Health Politics, Policy and Law* (Special Issue) Vol. 26, no. 5 (October).

Phillips, K. (2002) *Wealth and Democracy* New York: Broadway Books.

Rasell, E., J. Bernstein, and K. Tang (1993) *The Impact of Health Care Financing on Family Budgets*. Briefing Paper (April). Washington, DC: Economic Policy Institute.

Reinhardt, U.E. (1992) "Reflections on the Meaning of Efficiency: Can Efficiency Be Separated from Equity?" *Yale Law & Policy Review* Vol. 10, no. 2, pp. 302-315.

Reinhardt, U.E. (2001) "Commentary: On the Apocalypse of the Retiring Baby Boom" in "Northern Lights: Perspectives on Canadian Gerontological Research", Special Supplement, *Canadian Journal of Aging"* Vol. 20, supplement 1, (Summer) 2001, pp. 192-204.

Reid, R.J., R.G. Evans, M.L. Barer, S.B. Sheps, K. Kerluke, K.M. McGrail, C. Hertzman and N. Pagliccia (2003) "Conspicuous consumption: Characterizing high users of physician services in e Canadian province" *Journal of Health Services Research and Policy*, Vol. 8, no. 4 (October), pp. 215-224.

Roos, N.P., E, Forget, R. Wall and L. MacWilliam (2004) "Does universal comprehensive insurance encourage unnecessary use? Evidence from Manitoba says 'no'" *Canadian Medical Association Journal* Vol. 170, no. 2 (January 20) pp. 209-214.

Rose, G. (1985) "Sick Individuals and Sick Populations" *International Journal of Epidemiology* Vol. 14, no. 1, pp. 32-38.

Rossouw, J.E. and the Women's Health Initiative trial investigators (2002) "Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women" *Journal of the American Medical Association* Vol. 288, no. 3 (July 17) pp. 321-333.

Sheils, J., and P. Hogan (1999) "Cost of Tax-Exempt Health Benefits in 1998" *Health Affairs* Vol. 18, no. 2 (March-April) pp. 176-81.

Sheils, J. and R. Haught (2004) "The Cost Of Tax-Exempt Health Benefits In 2004" *Health Affairs* Web Exclusive, February 25, 2004.

Smeeding, T. (2002) "Globalization, Inequality, and the Richh Countries of the G-20: Evidence from the Luxembourg Income Study (LIS)" Luxembourg Income Study Working Paper Series, Working Paper No. 320 (July).

Solt, F. (2004) "Economic Inequality and Democratic Political

Engagement" Luxembourg Income Study Working Paper Series, Working Paper No. 385 (July).

van Doorslaer, E., A. Wagstaff, and F. Rutten, eds. (1993) *Equity in the Finance and Delivery of Health Care: An International Perspective* New York: Oxford University Press.

Vladeck, B. C. (2000) "Mending Medicare", *Issues in Science and Technology* (Spring) p. 17.

Wagstaff, A. *et al*. (1999) "Equity in the finance of health caree in twelve OECD countries" *Journal of Health Economics* Vol. 18, no. 3, pp. 291-314.

Wildavsky, A. (1977) "Doing Better and Feeling Worse: The Political Pathology of Health Policy" *Daedalus* Vol. 106, no. 1, pp.105-24.

Woolhandler, S. and D.U. Himmelstein (2002) "Paying for National Health Insurance – and Not Getting It" *Health Affairs* Vol. 21, no. 4 (July-August) pp. 88-98.

Woolhandler, S., T. Campbell and D.U. Himmelstein (2003) "Costs of Health Care Administration in the United States and Canada" *New England Journal of Medicine*, Vol. 349, no. 8 (August 21) pp. 768-775.

Discussion

Pekka Rissanen

Tampere School of Public Health, University of Tampere

Typically, it is a pleasure to read Professor Evans's texts. This piece of work did not disappoint, it did not differ in style from the previous ones. Especially, the comparisons of pure academic "value free" economics to more down-to-earth or real life phenomena in the economy and society made enjoyable reading.

In his paper, Evans analyses diverse arguments, which have the common feature that they are aiming at changing current health care financing mechanisms. And moreover, they aim at changes towards more regressive incidence of this financing.

These arguments stem from two distinct points of departure. The first line of argumentation is built on the question of the allocative efficiency of health care. It is argued that the financing mechanisms, especially low or zero user charges will lead to excess use of health care, and therefore its welfare burden is too large. The second point of departure lies in the opposite direction: it is argued that more resources are needed for health care in order to secure its ability to meet the rising needs of the population.

The solution to the efficiency argument according to economics text books is an increase in health care consumers' own financial contribution, which then leads to decreased use of health care and increased use of other commodities (which obviously yield higher utility than health care). However, the problem in this solution is that it changes not only efficiency of allocation, but also the distribution in incidence of financing. Moreover, in real life, such shifts in the financial burden tend to hit the low income population groups hardest.

Of course, this is contrary (in most countries) to the prevailing ethical principles of equality.

Professor Evans presents an amazing amount of evidence concerning potential distributive issues and real life observations, most of these naturally from Canadian data sources. Most of the data are from ten or so years ago, but the evidence convincingly supports the idea of considerable distributional consequences of any changes in health care financing mechanisms. However, these kinds of phenomena do not change rapidly.

I have only few comments to add on Professor Evans's paper. Firstly, I make an addition to the Canadian figures concerning the role of the financing mix. Secondly, Evans handles income groups as if there are no dynamic changes in the composition of such groups. I think, there is at least one methodological concern in such a treatment. Thirdly, I think, the influence of the global economy is an important topic especially for small economies, and perhaps also for their health care systems.

From financial burden point of view, as Evans writes, the form of financing mechanism has an essential role to play. In Finland about fifteen years ago, the government made a remarkable change in taxation system, where the proportion of indirect taxes in total government income increased and the share of direct (income) taxes decreased. Until that point in time, about 35 % of total health care funding came from the central government. As a consequence, one would have expected that such a change would lead to a less progressive (and therefore more favourable to the rich) system. However, such development was only rather modest. One reason was that at the same time, the role of government in total health care funding decreased and role of local authorities, municipalities increased – local authorities collect most of their income by progressive local income taxes.

Professor Evans talks about the health insurance system, and tax-based health care systems certainly contain a number of similar features. However, it is essential to note, that not only co-payments

matter. At the general population level, choice of taxation system, i.e. share of direct / indirect taxes may influence the incidence of health financing even more dramatically.

As a matter of fact – and this will lead to my second concern – according to the results of Klavus and Häkkinen (1998), the influence of households' direct health care payments was less regressive in 1994 than in 1990, even though between these years, dramatic co-payments were imposed on various health care services and drugs. Indeed, at the macroeconomic level, the households' proportion of total health care expenditure grew from about 13% to 20% between these measurement points.

Klavus and Häkkinen explain this unexpected development by methodological artefacts. The data in 1990 contained a cross-sectional sample, and a similar sample was drawn in 1994, i.e. no closed cohort follow-up was made. In 1994, the composition of income deciles was different from that in 1990. Between these measurement points, a deep economic recession took place in Finland. This hit young adults hardest, who had difficulties getting jobs after completing their studies. At the same time, there were only very modest cuts in old age pensions, and therefore, retired people climbed to higher income deciles (from the lowest), while the youngest found themselves at the lower end of the distribution. The poorest income deciles suddenly contained the healthiest part of the population!

This really highlights a strong need for true follow-up datasets in these kinds of assessments. Therefore, although I really think that professor Evans's evidence is reliable, his findings would have been even more convincing if these had been confirmed by, e.g., some closed cohort follow-up data.

My third point refers to the small open economies, such as the countries in Northern Europe. The economic capacity in all of these is strong but small from the perspective of international markets, and consequently, they are rather dependent on their international environments. From this departure, I would like to stress two issues.

As Professor Evans writes, collecting more money and resources to health care sector has often been taken as one argument in favour of higher user co-payments. Indeed, this has been one important argument also in the Finnish health policy debate. Not so much has been discussed about the potential consequences of increased health care budget.

At present, in Finland, there exists an under-supply of medical doctors, and the same situation will soon emerge in the labour markets of nurses and other health care professionals. Thus, we have two options from which to choose how to use the extra money in the health care budget: we can buy resources from abroad or we can pay higher wages to the existing resources. But, where are these countries or economies, from which we could buy health care professionals? Where are there countries with an excess supply of doctors or nurses? At least all the European countries seem to face problems of ageing populations, where a remarkable proportion of doctors and nurses will retire within the next ten years, at the same time as the need for health care services will grow. The professionals from the developing countries might compose a potential source for recruitment, but how ethical is the policy that might emerge? If the supply of health care professionals is restricted all that may remain of higher co-payments is that we simply redistribute incomes from users (in many cases the poorest) to health care professionals.

However, there is another argument in favour of higher co-payments from the foreign affairs point of view. In addition to international competition for health care resources, there is international competition of many other resources as well, such as IT-professionals etc. Many economists believe that in small open economies we are forced to compete for many kinds of resources and one instrument in this competition is the level of (income) taxes. In order to be competitive, it is necessary for the economy to reduce its level of taxation, and in that pursuit, it is necessary to decrease public spending. This is one reason for higher user charges, a means to keep up the present level of welfare services.

How solid is this argument? I have some doubts, but during recent years we have observed a flow of business and enterprises to China or other countries with low costs of production. At least it widens the scope of present national (and fragmented!) argumentation in health economics towards a more international one. Therefore I think, it would be interesting to read, perhaps from the writings to come from Professor Evans, how he would respond to this final point.

A Missing Piece in the Trade-offs between Efficiency and Equity in Health Care Financing

6

William C. Hsiao

Harvard School of Public Health

6.1 Introduction

Heated debates on health care financing rage in many OECD nations. Rapid rising health expenditures strain government and household budgets that lead societies to argue as how to best finance their health care. Canadians and British debate whether to relieve pressure on government budgets by shifting more to private insurance financing. Germans debate over how much to rely on patients' copayment to hold down the social insurance contribution rates. The American has its perennial crisis as how to finance its 45 million uninsured and how to control its double digit health insurance premium inflation rates. In essence, all of these debates come to one question: How best to finance health care?

Health care financing involves trade-offs between equity and efficiency. Inefficiencies that are often cited included excess burden produced from tax payments and moral hazard produced from health insurance. Health care financing methods influence both equity and cost containment, but the influence on the latter is often overlooked. Cost containment effects inherent in any health financing method influence efficiency of a national health system because they impact allocative and productive efficiency and thus the total amount of resources we have to spend for health. Consequently, certain financing methods may have superior features that enhance both equity and efficiency that can minimize the trade-off between them. The focus

of this paper is to analyze the different degrees of trade-off of several major financing methods.

This paper is organized into five parts. We begin by defining health care financing and summarizing the special features health care. Next, we examine the varying degree of influences of health care financing methods in managing national health expenditure. Third, we clarify the concepts of efficiency and equity. The debates on the trade-offs have been confused by the different meanings of efficiency. Consequently often these debates generated more heat than light for public policy. Meanwhile, equity is a value laden term where different societies hold different values, influenced by their history, culture and political philosophy. There is no agreement on a common concept of equity in health. This paper clarifies the different concepts. Fourth, we compare the trade-offs made by selected nations in terms of equity and efficiency when they employ different health care financing methods. Finally, we offer our conclusions.

6.2 Background

As economists, we are asked to evaluate and advise government's role in health care financing. Should we use general tax revenues or private insurance premium to fund health care? Should we have National Health Service or National Health Insurance? Do we require patients to pay directly for health services that they demand? Do we compel people to save for their post-retirement health expenses? The list could be endless. The answers to these questions, however, can only be answered when we are clear what a nation wants to achieve with its national health system.

Health care financing refers to the approach by which money is mobilized to fund health sector activities, and to which health services money is allocated. Financing plays critical roles in determining how much money is available for health care, who bears the financial burden, who has access to health services, health status of the population, how risks are pooled, and whether health care costs can

be controlled. There are four methods of financing, each of which taps a different source for funds and creates a political economy for it. The four methods of financing are: general revenue, social and private insurance, and patients' out-of-pocket payment. Each method has different equity and efficiency implications. For a full explanation of these methods, see W.C. Hsiao (1998)

Health has several unique features that make health care financing unusually complex. First, unlike other basic goods for life such as food, housing and education that are needed by everyone, serious illnesses do not afflict everyone. Uncertainty of illnesses makes risk pooling desirable and advances our economic welfare (Arrow, 1963). Insurance, however, creates moral hazard which is likely to lead to greater demand for services. Second, resource allocation becomes more complicated because it requires trade-offs between health improvements and financial risk protection against catastrophic expenses. Third, health is a basic necessity of life and a key contributor to our well being. All OECD nations accepted health as a human right. While the rhetoric usually far exceeds the actions, nonetheless every nation, except the USA, assures their citizens in varying degrees of "equal" access to health care. Fourth, our preferences for medical services depend on our "state of the world." When we are well and face with a low probability of become ill, we have a particular preference for medical services that influences our willingness to prepay. But when we become ill we are in a different state of world, our preference may change. As a result, health financing suffers from serious incomplete contracts in which sick patients frequently complaint being short changed by National Health Service or insurance. Finally, many people look upon life as priceless and demand heroic measures to prolong their lives when they become seriously ill. In this era of rapid advancements in new expansive medical technology, these heroic measures would bankrupt most households or put a severe financial strain on government or insurance programs.

Besides the unique features of health and health care, health care financing faces additional challenges because of the serious market

failures in the insurance and health service delivery markets. These failures are due to asymmetry of information, imperfect agency relationships, incomplete contracts, and monopoly. Insurance creates moral hazard and raises the issue of what would be an optimal insurance that offers "reasonable" protection while minimizes inefficiency. Adverse selection becomes a serious problem under voluntary insurance because of asymmetry of information between the buyer and seller. Moreover, competing insurance organizations practice risk selection and it poses a serious barrier to establishing a competitive insurance market.

6.3 Health care financing method and cost containment

The national health system like a household or a firm faces a budget constraint. The financing method creates an overall budget constraint on the complex health sector where more than a dozen of imperfect and monopolistic markets interact and that influences total expenditure and its inflation rates. The effectiveness of the budget constraint, its tightness and whether the budget is set prospectively influence the production function of health services and therefore the expenditures. Like a household, the health system adopts different production functions to produce the outputs, depending on the budget constraint placed on the system. If the budget is tight and revenue effectively capped, the system is pressured to adopt more cost-effective production functions, in turn, influence national health expenditure.

The financing method chosen by a nation imposes a "hard," "soft" or "open" budget over the health sector. Moreover, the financing method also places the control of the financial resources into the hands of different actors who allocate how the funds will be used as well as decide on the payment methods by which the providers are incentivized.

We term a *"hard-budget" system* as a system where the budget is preset effectively (such as not allowing for deficit or retrospective adjustment) and thus significantly affects the quantity and quality of services supplied. A society may decide on an aggregate prospective budget for health through negotiations or through a political process. The UK system would be a clear example of a closed budget system in which the budget is determined through a political negotiation process conducted in the Cabinet. Another example is the global budget negotiated between Germany's Sickness Fund Association and Physician and Hospital Associations. The US Balanced Budget Act of 1997 also preset a budget for the Medicare program through a legislative process.

In comparison, we term a "*soft-budget" system* as a system where the budget is preset but allowing for significant retrospective adjustments or a budget that is based on the actual costs incurred. For examples, Blue Cross had previously used the actual cost incurred as the basis to pay hospitals. The US Defense Department often follows this practice in purchasing armaments.

We term an *"open-budget" system* as a system where the budget is established largely by market forces, relying on the demand of patients (willingness and ability to pay) to set the budget constraint on the system. However, patients' demand can't set an effective budget on the providers because individual patients and providers can't predict the charges in advance for most major illnesses. The uncertainty of diagnosis and treatments leave the charges indeterminable in advance. The patient can't make rational choices based on price when charges for a patient are largely determined retrospectively by whatever the diagnosis and treatment have been given.

6.3.1 A Diagrammatic Model of budget constraints and national health expenditure

The relationship between the system's budget constraint and the supply and demand of health services can be illustrated by a six-part diagram. Figure A analyzes their relationships. It illustrates

the interrelationships between health status, price/quality adjusted quantity demanded and supplied, and total expenditure derived by the different budget setting approach. Quadrant I shows a simple supply and demand relationship, while Quadrant II illustrates the relationship between quantity of health services and health status. Quadrant III maps the total cost (calculated as C_1xQ_1) and price, and Quadrant IV shows the plausible relationship between total spending for health services and population's health status. Under a free market system without health insurance, the demand and supply at P_1Q_1 will determine the total expenditure for health services. Under a break-even assumption, the revenue needed to fund the expenditure is also determined by supply and demand. This is shown as E_1 in Quadrant IV.

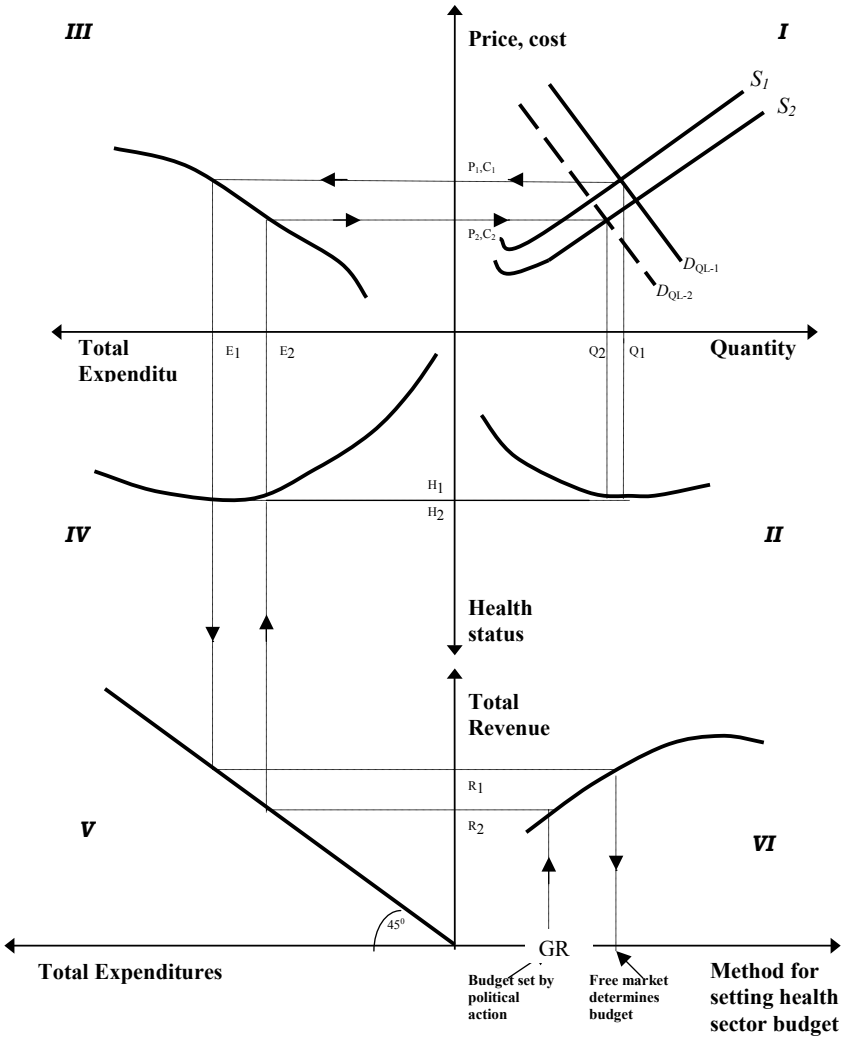
In a "hard-budget" system, the budget is set prospectively at point GR in Quadrant VI. This prospective budget constrains the quantity of services which can be produced to Q_2, and the marginal cost declines to C_2. Without a price change for patients, demand could exceed supply. We assume that suppliers reduce the amenities of services (i.e. quality of health care) or increase bedside rationing until demand shifts downwards to a new equilibrium point. In some cases, the supply and demand maybe equilibrated by waiting time.

In an "open-budget" system, the health services provision market determines the expenditures. No "open-budget" country relies totally on patients' direct out-of-pocket payment to finance health care. Most these countries, such as US and Singapore, use a combination of private insurance, social insurance, public subsidized health services, and out-of-pocket payments. Since the health services provision market is dominated by the providers, the expenditure inflation rate may be affected by increases in the supply of practitioners and hospitals beds. The quantity of services and charges produced in the supply and demand interactions in the provision market (shown in Quadrant I) determines the national health expenditure.

Managed Competition (Enthoven, 1987), attempts to equalize the market power between the insurance purchasers and insurers, and

transform the insurers into managed care plans to "manage" the quantity, quality and charges of providers. These market actions take place in Quadrant I where the managed care plans try to shift the production function of the providers (the supply curve) and the demand curve of the patients through various incentives (payments) and by imposing private regulations on modality of medical treatments. Nonetheless, managed competition is still an open-budget system where the providers can shift charges. There is no effective prospective budget imposed on the total national health expenditure. The market actions in Quadrant I determine the total national health expenditure.

Figure A. *Relationship between health status, supply, demand, and the budget*



Empirical Evidence

We have constructed an econometric model to test the hypothesis derived from our concept of different kind of budget constrains

imposed by the various financing methods. We tested whether the share of total national health expenditure funded by different methods has a differential impact in determining health expenditure inflation. Previous studies (Gerdtham and Johnsson, 1999) have included the share of public financing in their regression models to explain health expenditure growth and they found negative associations. In our research we provided a theoretical framework. We summarize our study and its result here. For the full paper, see Hsiao & Yip (2004).

To empirically test the health financing method of a country as a measure of the nation's budget constraint, we employ equation [1]:

$$\ln(HE_{pc})_{it} = \beta_0 + \beta_1 \ln(GDP_{pc})_{it} + \delta_1\left(\frac{GTax}{THE}\right)_{it} + \delta_2\left(\frac{SocIns}{THE}\right)_{it} + \delta_3\left(\frac{FP\,\mathrm{Pr}iv}{THE}\right)_{it} + \beta_2\begin{Bmatrix}\mathrm{Pr}\,oviderPayment\\ Macro-organizatin\\ \mathrm{Re}\,gulation\\ Persuasion\end{Bmatrix} + \beta_3 X_{it} + \alpha_i + \tau_t + \varepsilon_{it} \quad [1]$$

where $\left(\frac{GTax}{THE}\right)_{it}$ is the share of total health expenditure financed by general tax revenue. Similarly, $\left(\frac{SocIns}{THE}\right)_{it}$, and $\left(\frac{FP\,\mathrm{Pr}iv}{THE}\right)_{it}$ are the share of total health expenditure financed by social insurance contribution and for-profit private health insurance premium, respectively. To avoid perfect colinearity, we omitted the share of out-of-pocket payments in our estimation as a comparison group.

In this specification, δ_1, δ_2 and δ_3 measure the relationship between general tax revenue, social insurance and for-profit private insurance financing and health expenditure, relative to out-of-pocket payment, respectively. Based on our hypothesis, we predict that $\delta_2 < \delta_1 < 0 < \delta_3$.

Our primary source of data is the OECD health database (2000). Of the 29 countries, we first excluded Turkey, Greece, and Mexico because they are at a different stage of economic development compared to other member countries. Similarly, we excluded the transitional economies, such as Poland, Hungary and the Czech Republic because their social and economic systems are undergoing

major transformations. Our analytical sample consists of 23 countries for years 1971 to 1998/9.

In choosing between the fixed and the random effect models for panel data estimation, the primary criterion is whether there exists unobservables that may be correlated with the budget constraint variables and therefore bias their coefficient estimates. To the extent that we do not have direct measures on provider payment, macro-organization, regulation and persuasion, we are in favor of the fixed effect model. That is, to the extent that countries that are more conscious of controlling health expenditure are more likely to adopt a close-budget system and other provider payment and organizational forms that are more effective in controlling cost, failing to control for these other structural components of the health system could bias the coefficient estimates of the budget constraint variables. We estimated both the random effect and the fixed effect models and performed a Hausman specification test. As expected, the Hausman specification test rejects the random effect model in favor of the fixed effect model ($\chi^2 = 365.9$).

Table A presents the estimation results for equation 1. Overall, the results are consistent with our theoretical prediction. We found that social insurance financing, characterized by prospective and hard budget constraints, is associated with the slowest growth of health expenditure, followed by general tax revenue financing, when compared to out-of-pocket payment. In contrast, for-profit private insurance financing is associated with a more rapid increase in health expenditure. The results in Table A imply that an increase in government tax revenue financing share by 10 percentage-points (relative to out-of-pocket payment, the omitted group in the regression) is associated with a 2.5 percent lower growth in real health expenditure per capita. Similarly, a 10 percentage-points increase in the social insurance share is associated with 4.3 percent smaller increase in real health expenditure per capita. On the other hand, an increase in for-profit private insurance financing by 10 percentage-points is associated with a 24 percentage-points greater increase in health expenditure.

Table A. The Financing Method Approach: Panel Data Estimation Results

	Fixed Effect	Random Effect
ln(GDPpc)	0.7977	1.1367
	(16.39)**	(35.69)**
Tax Rev. Share	-0.0025	-0.0047
	(2.49)*	(6.90)**
Soc. Ins. Share	-0.0043	-0.0064
	(4.07)**	(9.79)**
For-profit Ins. Share	0.0238	0.0156
	(7.90)**	(9.20)**
% Pop. <1	0.0994	0.0431
	(3.53)**	(1.35)
% Pop. 65-74	0.0167	0.0147
	(2.50)*	(1.65)
% Pop. 75-79	-0.0132	-0.0322
	(0.79)	(1.40)
% Pop. >80	-0.0528	0.0741
	(2.68)**	(4.08)**
% Pop. Female	-0.0295	-0.0169
	(1.50)	(1.35)
Doctors/Pop.	-0.0085	0.0804
	(0.62)	(9.91)**
Beds/Pop.	0.0213	0.0085
	(6.75)**	(4.67)**
Cancer Mortality	-0.0005	0.0007
	(1.23)	(2.27)*
IHD Mortality	0.0008	0.0002
	(4.43)**	(1.40)
Cardiovascular Mortality	-0.0002	-0.0005
	(0.54)	(2.47)*
Infant Mortality	-0.001	-0.0205
	(0.43)	(7.38)**
Alcohol Consumption	-0.0064	0.0102
	(2.22)*	(5.36)**
% Pop. W/ Ins. Coverage	0.003	0.0036
	(3.13)**	(4.53)**
Constant	-0.0226	-3.7858
	(0.02)	(5.71)**
Observations	537	537
Year dummies	Yes	Yes
R-squared	0.96	

Absolute value of t statistics in parentheses
* significant at 5%; ** significant at 1%

6.4 Equity and Efficiency

What can standard normative economics say about health care financing? Public economics offers an approach to examine the role of the government and market in health care financing. But its usefulness has been severely limited. On the revenue side, we can examine the excess burden and the incidence of tax, social insurance contributions and self-pay. However, public economics doesn't guide us on how much to trade-off between equity and efficiency and how progressive the taxes should be. These are normative issues that have to be decided by a nation's own social values as revealed through a political process. On resource allocations, public economics guides us to distinguish between public, merit and private goods as well as recognizing the externalities of certain private consumptions, but the benefit of a good is usually measured as a single dimensional factor such as health gains or protection against impoverishment. Furthermore, public economics has limited application on how to trade-off between spending for health gains versus medical rescue (such as emergency services), or health improvement versus risk protection. These trade-offs can be decided through the market place when people purchase private insurance, however people's ability to pay differs greatly. Alternatively, the trade-off can be decided through a political process where the majority indicates what services to include or exclude under National Health Service or compulsory social insurance.

Health care financing has to be concerned with distributive justice in health cost and benefits. Arthur Okun (1975) argued that the most difficult and controversial aspects of public policy and public economics involve the balance and trade-off of equity and efficiency. Not everyone can have all the health care that they want or need (Evans, 1984). All nations have to make choices on how much to spend on health, how to obtain the greatest amount of benefits from the resources spent, who shall bear the cost burden and who shall receive the benefits. While we are clear about the general concept of trade-off between equity and efficiency, the specifics are much more complicated. First, we must clarify the concepts and meanings of efficiency and equity because they vary and each have many layers of consideration in their application.

Efficiency

Economists have several definitions of efficiency that carry very different meanings and often do not make clear which meaning they are concerned with. In discussions, US economists frequently present efficiency in the social welfare context that defines efficiency as the optimal condition that maximizes total utility. Besides implying the source of funding to finance health care should come from non-tax sources to minimize excess burden, this definition also implies that consumers must be given choices of insurance plans and health services so individual's preferences can be revealed through the market place and maximize the utility. Using this definition of efficiency, Enthoven (1978) declared choice is the final objective in health care financing and delivery. Mark Pauly (2001) also argued that the US health system is efficient because it offers people more choices. Under this criterion, universal insurance and tax financed national health services would be inefficient because they do not offer free choice of insurance and compulsory tax or social insurance contribution incur excess burden. Furthermore, NHS may be inefficient when they do not offer patients free choices of health service providers. Public economics literature follows similar theory and often refers to efficiency as defined by Pareto (Harvey Rosen, 2003). Pareto Optimality involves voluntary exchange to maximize individual utility, but it does not take into account the initial wealth and income of individuals.

On the other hand, when European economists discuss efficiency, they often refer to a different social welfare function – allocate limited resources efficiently to produce the optimal health status under circumstances where the equity parameters have already been established. It's clear that the social welfare functions of the US and European economists often differ and maximizing a population's health status is not the same as maximizing utility. Meanwhile, an average citizen thinks of efficiency in term of productive efficiency – i.e. minimizes the cost to produce a given quantity of a good. Once we become sensitive about the different meanings of efficiency and how they are used in our discourse, the confusion in public debates becomes clear.

Equity

While public discussions already were confused by different meanings of efficiency, equity complicates it further. Equity is a complex concept that derives from various theories of social and distributive justice, including utilitarianism, libertarianism, egalitarianism and communitarianism (Roberts et al. 2003). Each theory of justice generates a different set of social and individual obligations in respect to the distribution of health, health care resources, and health-related financial risk protection. Hence, nations have given different interpretations to the meaning of equity in their policies (Wagstaff et al, 2003).

Arguments made by John Rawls in his *Theory of Justice*, (Rawls, 1971) have had significance impact on the concept of equity in health. Norman Daniels has persuasively argued that health care should be treated as a primary social good whose distribution is subjected to the Rawlsian principles (Daniels 1985). According to the Difference Principle, inequalities in health care and the other social determinants of health are inequitable if they fail to enhance the welfare of the least advantaged in society. Thus, equity means the public resource is to be allocated by following the *maxmin* principle to improve the position of the worse- off members of a nation. Amartya Sen has offered a similar account of the value of health, focusing on its relationship to capabilities (Sen 1993; Sen 2002). The capabilities approach evaluates a person's well-being and advantage in terms of her entire possible set of life functions, her ability to attain valued states of being or to achieve valuable actions. For Daniels and Sen, it seems that equity of health – not merely opportunity to attain health – is valued (Sen 2002). Both Daniels and Sen offer a broad scope of social justice beyond the health care system for our obligations to decrease inequities in distribution of health. Meeting the goals set by their theories would require a wide array of socioeconomic interventions that are beyond the range of the health system.

Various theories of distributive justice suggest that health equity encompasses three fundamental questions. First, not all differences

in health are unjust: when do health differences become inequities? Secondly, do the principles of distribution apply to health outcomes, access to health care or health care usage? Third, does equity require equal treatment for equal need or a minimum standard of health treatment according to need?

In every country there are differences in health between individuals and across socioeconomic groups. Yet not all of these disparities are the source of ethical concern. When are health disparities unfair and inequitable? According to Arneson and Cohen (Reference), society is obliged to equalize an individual's opportunity of welfare, but not obligated to compensate individual for voluntary choices. Roemer builds on this to explicitly argue that society should only be held to account for those health outcomes that arise from social circumstances (Roemer 1993; Roemer 1995). Margaret Whitehead and Goran Dahlgren propose that health inequalities are inequitable when they are "avoidable, unnecessary, and unfair" (Dahlgren and Whitehead 1991).

Different theories of distributive justice offer another metric for identification of inequitable inequalities, often inseparable from an assessment of what should be equalized. Egalitarian theories are divided on whether the focus should be on health *outcomes, opportunity* to attain health status or equalization of *resources.* The most modest of egalitarians agree that society should strive to distribute resources and burdens in a way that is ambition sensitive and endowment insensitive, thereby compensating individual for their "brute bad luck" (e.g. bad genes) but not for the outcomes of their chosen actions "option luck" (e.g. bad lifestyle choices) (Dworkin 1981; Dworkin 1981).

Brian Barry's definition of equity focused on the comparative treatment of individuals of similar needs (Barry 1990 [1965]). Appeal to equal use of health services as a benchmark for equity is also problematic. Individuals with equal need and access may differ in their health care usage due to variations in their information base and personal and cultural preferences. As a result, the health systems literature tends to

focus on equity of access. Equity of access requires equalization of financial and physical (e.g. travel time) obstacles for individuals with similar health needs.

In practice, equity can mean ensuring that individuals do not fall below some specified minimum level. Daniels proposes that society's obligations to provide health care and the social determinants of health cease once the threshold of normal species functioning is obtained. All health systems of advanced nations currently recognize a minimal standard or basic benefits package. People with greater resources may have access to superior services when they can afford to pay and willing to pay. Even the United States provides minimal care to the poor via its Medicaid system and legally mandated emergency care for all. Singapore subsidizes specified levels of public hospital services with modest coinsurance. Germany and the Netherlands provide a basic benefits package and allow richer citizens to purchase private insurance of self-pay for superior services, generating a two-tiered system. Other countries, such as Canada limit the development of a two-tiered system by prohibiting private insurance for publicly-covered services.

Equity in financial risk protection has seldom being discussed in literature. Financial risk varies by people's ability to pay for health care when uncertain major illnesses strike. One standard used in measuring risk protection involves whether a household will be impoverished by a major medical expense. OECD nations have set poverty level of income as the standard for impoverished households. The common approach to measure equity in financial risk protection involves ascertain by socioeconomic groups as the proportion of households in a particular group has fallen below poverty line when they paid for their health expenses. However, this concept and measurement approach does not take into the value of insurance given to everyone that's imbedded in National Health Service or national health insurance, regardless a person had a catastrophic expenditure. John Nyman (2003) has argued that for measuring the value of these programs and ascertaining their equity, we need to measure the value of insurance for all households.

In addition, equity criteria for health and risk protection differ. Yet, in health care financing, we have to consider equity for both health and risk protection. Finally, health care financing concerns who pays and who benefits. Equity applies to both. The most meaningful way to examine equity and the trade-off between equity and efficiency involves the examination of equity of the net benefits.

6.5 Equity in practice and their trade-offs

We compare the efficiency and equity of several types of health systems under two different concepts of social welfare function. First, we consider the *system productive efficiency* as the social welfare function where a country's national health expenditure (in ppp terms) approximates the amount of resource spent to produce that nation's health welfare and the expenditures reflect the productive efficiency. While levels of cost are important, the equitable distributions of these costs also have to be considered. Welfare outcome consists of several dimensions: level of health status and its equitable distribution, risk protection and its distribution, level of access and equality of access for equal need. For the second concept of social welfare function, we use maximizing total utility as welfare criterion.

We select five nations for comparison because they have different health systems and use different methods of health care financing. UK and Finland use centralized budget to finance health care with a National Health Service to deliver health care, Canada finances health care through a National Health Insurance system with a mixture of public and private provision of services, Germany finances health care through a social insurance system with a mixture of public and private provision of services, and the USA relies on a pluralistic system of private insurance, social insurance and self-pay to finance health care with a mostly private provision of services.

6.5.1 Social welfare function A: System productive efficiency

Health gains

Health status measures such as infant mortality rate or life expectancy have often been used as welfare indicators.[1] However, these indicators are insufficient measures to inform health policy as to the impacts of health care interventions, particularly for advanced nations. Health status is produced from many other factors other than prevention and health services and they include genetics, environmental and socioeconomic factors and lifestyle. Also, those indicators do not measure the quality of life. Nonetheless, these measurements give a first approximation of welfare. Table B compares the relative health status among the five nations. Finland and Germany have the lowest infant morality rates while the USA has the highest. Canada has the highest life expectancy rate while the USA has the lowest.

OECD has estimated the national health expenditures for its member nations and we present the expenditures per capita in Table B. Finland spent the least per capita along with UK while the USA spent more than twice of the low spending nations.

Table B. National health expenditure and health status of selected nations

Country	2001 Total health expenditure/ capita, ppp basis	National health expenditure as % of GDP	Rank of expenditures (from lowest to highest)	2000 Infant mortality rate	1999 Life Expectancy
Finland	1,841	7.0%	1	5.3	79.0
UK	1,992	7.6	2	3.8	77.4
Canada	2,792	9.7	3	4.4	77.7
Germany	2,808	10.7	4	5.6	77.4
USA	4,887	13.9	5	6.9	76.7

Source: OECD Data, 2003.

1 Economic theory would suggest that health status is not a sufficient indicator of welfare because health contributes to human capital development and economic growth. Economic studies of developing nations have found empirical evidence to support this hypothesis. This paper does not include this consideration in our definition of welfare because it's not clear how much health care contributes to health in advanced economies.

(b) Equity distribution of the cost burden

The comparative distribution of the costs (expenditures) has been analyzed by van Doorslear and Wagstaff (1999). Their equity principle was based on people's ability to pay. They used the Kakwani index to compute the income redistribution resulted from sources of funds used to finance health care. We use the results produced by their equity index for the selected nations to rank the equitable distribution of the health care costs and they are shown in Table C. Unfortunately, Canada was not included in their study. UK and Finland ranked first being most equitable, while the USA was the most inequitable.

Risk Protection

Finland and UK offer free of nearly free essential health services to its citizens. Thus they are universally protected from financial risks arise from catastrophic medical expenses. Germany compels its citizens who earns less than €40,500 per year to be insured under its social insurance program which provides them with risk protection. It's possible that a few affluent Germans who may be impoverished by catastrophic medical expenses because they did not voluntarily enrolled in the social insurance program or purchased private insurance. Canadian NHI insures all citizens for all essential health services except drugs and dental care. It's possible that some Canadian middle class households may become impoverished from large drug and/or dental expenses. In contrast, 15% of Americans under age 65 are uninsured, and another 20% have inadequate insurance coverage for catastrophic medical expenses. Most of the uninsured and under-insured are low income households (Reference). One study found that the financial burden from out-of-pocket medical expenses was the second most frequently cited cause for households who filed bankruptcy in the USA (Reference).

Absolute Access and Equitable Access to health care

Access to health care is another reasonable indicator used as a welfare measure. Access is a multi-dimensional concept. The equity aspect of access is measured by equal access according to need.

Schoen and Doty (2003) found that when access is measure by ability to see specialists on a timely basis and patients' perception of quality of health services, UK, Canada and USA came out to be similar. As for financial barrier to access, it's well established that the USA has the worst condition. While access vary among nations depending on the measure used, evidence seem to indicate that overall access is somewhat similar among the nations studied once we exclude financial access as a factor.

When we use waiting time as a measure for access, UK has the highest percentage of population waiting for six months or more for elective surgery and Canada ranked second while the USA being the lowest (Schoen and Doty, 2003). The lengthy waiting time has become a major political issue in UK and Canada. It shows that when government use rational approach such as cost/effectiveness method based on health improvement (measured in QALY) to allocate limited resources, it must satisfy the demands of a political market. When politics overly restrain the amount spend for health and significantly affected personal quality such as waiting time, the public reacts through a political process to remedy it.

Equity in access in Europe and USA has been studied by van Doorslear and Wagstaff (JHE, 2000.) They defined equity as "equal utilization according to need." Using self assessed health status as the indicator for need, they found that West Germany to be most equitable (but result is statistically insignificant), followed by UK (statistically insignificant), Finland with the USA being the last. Using their definition of access, Schoen and Doty (2003) ranked UK as being most equitable, followed by Canada with the USA being the last.

Table C. Ranking of Countries by various health system outcome variables

Country	Ranking on expenditure per capita or as % of GDP	Equitable distribution of cost burden	Access[a]	Equity in access[a]	Equity in access[b]
Finland	1	1	nis	nis	3
UK	2	1	1	1	2
Canada	3	nis	1	2	nis
Germany	4	nis	nis	nis	1
USA	5	2	2	3	3

[a] Schoen and Doty.

[b] van Doorslear and Wagstaff.

Note: Ranking of 1 means the best and 5 means the worst. nis denotes country not included in the study cited.

Empirical studies offer several conclusions. The private insurance method of financing as practiced in the USA yields the worst possible results. Compare to the other selected nations, the USA spends the largest amount of resources, the cost burden shared most inequitably while produce the worst and most inequitable benefits in terms of lower health status, lack of adequate risk protection for at least 50% of its population, problematic access to health care for the average citizen and most inequitable access. However, many economists argue that the USA tries to optimize a different social welfare function, total utility. When that measure is used, the USA may come out in a favorable position because the US system gives people choice.

6.5.2 Social welfare function B: Maximize total utility

Health care financing methods can reduce consumer surplus in four ways under National Health Insurance (NHI) or national health services (NHS) as compared to a voluntary private insurance or patient direct out-of-pocket financing systems. First part comes from the well established fact that tax or compulsory social insurance contribution create excess burden, so there is an efficiency loss. The second part comes from compulsory insurance rather free choice of insurance

that can affect individual utility, the third part comes from free choice of providers if a NHI or NHS restrict free choice of providers and the fourth part comes from moral hazard induced by insurance. On the other hand, there are efficiency gains from health care financing methods when they can influence the productive efficiency of health care.

Our econometric study found that NHI or NHS may impose a more effective budget constraint over the health systems and produce health services more efficiently than those under health systems that rely on private insurance and patient out-of-pocket payments. One area of greater productive efficiency is administrative expenses under NHI or NHS. Studies found that under the US private health insurance system, close to 25% of the premium income was spent for the administrative costs of insurance plans and additional administrative work imposed on the hospitals and doctors by the varied requirements of different insurance plans (Reference.) Meanwhile, only 5-8% of the NHI or NHS expenditures were used for administrative expenses (Reference.) Thus we can conclude that while NHI or NHS restricts choice on insurance, they reduce the national health expenditure by at least 15% by unifying and simplifying the administrative procedures. In addition, NHI or NHS also prevents providers from shifting their charges from one payer to another as they can do under multiple private insurance financing. Hence, hospitals and clinics face a tighter budget constrain and they have to give more attention to operational efficiency. The production cost could be reduced by another 15-20% as illustrated by the managed care "revolution" in the US which showed that the market constraints imposed by the managed care plans could improve hospital operational efficiency by 15-20% (Reference.) We can deduce that NHI or NHS may reduce the national health cost by 20% or more and accordingly reduces the individual tax or social insurance contribution rates. Then the consumers would have more discretion income to spend on other goods that produce consumer surplus. Moreover, more people would be insured because the premium is lower and affordable than what otherwise would be under private insurance or out-of-pocket financing systems. There is a gain in consumer surplus for this group of people.

While there would be a loss of efficiency from the lack of free choice of providers, such a loss would be experienced under NHS or NHI or voluntary private insurance, it depends of how the health care delivery system is designed. Restrict free choice of providers is not an inherent feature of NHS or NHI. As a matter of fact, the Canadian NHI offers free choice of providers while the US managed care plans restrict free choice of providers.

As for the loss of efficiency from moral hazard, it depends on the benefit design. Any form of health insurance (or free care) induces an efficiency loss, regardless it's public or private plan. The Rand Study was able to measure the insurance effect on demand for services and the elasticity is around 0.21. Thus when design benefit, the trade-off between equity and efficiency arises. But for a nation, this trade-off has to be considered in a larger context, the overall health care financing method to be used to influence the productive efficiency of health care.

It has not been possible to measure the gains and loss of consumer surplus from various health care financing methods and other factors to derive a net balance. Nonetheless, we have shown that the selection of an appropriate financing method that can improve the productive efficiency of a health system can at least reduce the trade-off between equity and efficiency. The US pluralistic system of financing, relying mostly on private insurance seems to have put its health system on a less efficient production frontier. Consequently, it sacrificed equity while lost efficiency.

6.6 Conclusion

Health care financing methods have differential impacts on the productive efficiency of a health system as reflected in the level of national health expenditures and its inflation rate. The pluralistic private insurance financing approach as practiced by the USA incurred the largest health expenditures and the highest average expenditure inflation rate over the past four decades. So with one

concept of social welfare function – system productive efficiency – the US system is the worst among the five nations we have compared. Paradoxically, the equity of the US health system is also the worst, using various criteria of equity. The answer seems to lay in the health care financing approach adopted by the US that offers an environment for less efficient production of health care and health.

When we use another concept of social welfare function – maximize total utility – to assess its efficiency, the US financing method still does not seem to produce better results than the other countries. Comparing NHS, NHI and compulsory social insurance systems with the US health system, they incurred greater losses from excess burden, moral hazard and less free choice of insurance. On the other hand, these systems generated additional consumer surplus from two sources: (a) their financing methods hold down national health expenditures so people who would have purchased voluntary health insurance before now have more income for non-health goods and (b) these systems provided health insurance to these people who preferred health insurance but were not able to afford it before. Meanwhile, the US system is the most inequitable one, using any generally accepted criteria for equity. Thus, it seems that a pluralistic system of health care financing sacrificed equity while didn't gain any efficiency in terms of utility.

References

Barry, B. (1990 [1965]). Political argument. Berkeley, University of California Press.

Dahlgren, G. and M. Whitehead (1991). Policies and Strategies to Promote Social Equality in Health. Stockholm, Institute of Future Studies.

Daniels, N. (1985). Just Health Care. New York, Cambridge University Press.

Dworkin, R. (1981). "What is equality? Part I: Equality of welfare." Philosophy and Public Affairs 10(3): 185-246.

Dworkin, R. (1981). "What is equality? Part II: Equality of resources." Philosophy and Public Affairs 10(4): 283-345.

Enthoven, Alain C. (1987). Consumer choice health plan (CCHP): an approach to national health insurance (NHI) based on regulated competition in the private sector. Standford, Standford University, Graduate School of Business.

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

Okun, Arthur M. (1975). *Equality and efficiency, the big tradeoff*. Washington, The Brookings Institution.

Phelps, Charles E. (1983). *Taxing Health Insurance: How Much is Enough?* Rand Documents, www.rand.org/heath/Cites/org1980-1984.html.

Rawls, J. (1971). A Theory of Justice. Cambridge, Mass., Belknap Press.

Roberts, Marc J. et al. (2004). Getting Health Reform Right: A Guide to Improving Performance and Equity. Oxford, Oxford University Press.

Roemer, J. E. (1993). "A pragmatic theory of responsibility for the egalitarian planner." Philosophy and Public Affairs 22(2): 146-166.

Roemer, J. E. (1995). "Equality and responsiblity." Boston Review.

Rosen, Harvey S. (2002). *Public Finance*, 6th ed. Boston, McGraw-Hill/Irwin.

Sen, A. (1993). Capability and well-being. The Quality of Life. M. Nussbaum and A. Sen. Oxford, Clarendon Press**:** 30-53.

Sen, A. (2002). "Why health equity." Health Economics 11: 659-666.

Schoen, Cathy and Doty, Michelle (2004). "Inequities in access to medical care in five countries: finding from the 2001 Commonwealth Fund International Health Policy Survey." Health Policy 67:309-322. (Accepted in 2003).

van Doorslaer, Eddy and Wagstaff, Adam et al. (2000). "Equity in the delivery of health care in Europe and the US." Journal of Health Economics, 19 issue 5.

van Doorslaer, Eddy, Wagstaff, Adam, et al. (1999). "The redistributive effect of health care finance in twelve OECD countries." Journal of Health Economics, vol. 18, issue 3.

Discussion

Unto Häkkinen

STAKES

The paper introduces different approaches and concepts that are used to analyse efficiency and equity of health care systems both theoretically and empirically. The author starts by introducing a diagrammatic model in order to analyse the effects of budget constraints on national health expenditure. The model includes simple production functions of health (health status as a function of quantity and expenditure). Although earlier in the paper the author has mentioned the unique features of a health care system (such as asymmetry of information, imperfect agency relationships, incomplete contracts, monopoly), these features have not been included in the model. In addition, the simple model does not take into account that health care financing methods as such may affect the shape of the health production function i.e. functions given in quadrants IV and II (in Figure A) might be quite different in various financing systems. According to the model given in the figure, health status should be higher in countries with open budget systems, since there is a positive "plausible", though mechanical (see Evans and Stoddart 1990) relationship between health spending and level of health. This contradicts the empirical evidence given later in the paper. The fact that in developed countries there is no clear relation between level of health and total health spending may be due to some indirect effects of financing such as a hard budget system giving more income to buy more non-health goods that affects health etc.

The author constructs an econometric model according to the hypothesis based on different budgets constraints. However, the theoretical foundations for the empirical model are not given, which makes it difficult to evaluate the empirical model although the results as such are plausible. In addition, the econometric study

includes other unclear issues. It is not clear why the author arrives at the hypothesis that social insurance financing contains better health expenditure than tax financing (which more typically has a hard-budget). As usual, one should be very cautious about the results because of the well-known problems in international comparisons, such as the definition of health care, measurement of prices, lack of PPPs for health expenditure, comparability of many measures such as hospital beds and small numbers of observations. In addition to the financing method variables, the empirical model includes other ad hoc variables. The justification for including many disease-specific mortality variables is unclear. The correlation of the two mortality measures (IHD and cardiovascular mortality) may be high. Also the magnitudes and signs of many coefficients vary considerable between the two specifications and some have even "wrong signs", an example being the negative sign of the variables for population over 80 and alcohol consumption. Finally, the stability of results that include many ad hoc variables in addition to alternative econometric specifications are not discussed.

In the third part the author discusses the different concepts of efficiency and equity and clarifies the differences between US health economists (welfarists) and European (also some Canadian) health economists (extra-welfarists). In addition, he discusses different theories on distributive justice. However, this discussion fails to mention the important extra-welfarist literature on equity-efficiency trade-offs in health (Wagstaff 1991, Williams and Cookson 2000) that are applied to cost utility analyses but can also be used in macro level analyses. Earlier discussion on this topic can also be found in the Finnish health economics literature (Sintonen 1981, Häkkinen 1992).

In the fourth part, the author gives practical examples of trade-offs, selecting five nations for a comparison. Here the paper includes a misunderstanding of the Finnish system by claiming that it uses a centralised budget to finance health care. In fact, the Finnish health care system is one of the most decentralised systems in the world, as most important economic decisions in health care are made by

the 432 munipalities, which decide annually the amount of money to be allocated to health care as well as dividing resources among the differing sectors within the area of health (Häkkinen 2005).

The main message from the last part of the paper is that the US system yields the worst possible results with respect to the many dimensions evaluated. This conclusion is well known also in Finnish discussions and is as such plausible although it is based on analysis of only five countries. But added to this conclusion, as well more generally to the whole paper, one can argue that we are still missing pieces of both theoretical and empirical research on the trade-offs between efficiency and equity in healthcare in order to give policy guidance to countries other than the USA.

References:

Evans RG Stoddart G. Producing health, consuming health care. Soc Sci Med 1990:31:12:1347–1363.

Häkkinen U. Terveyspalvelujen käyttö, terveydentila ja sosioekonominen tasa-arvo Suomessa. (Health care utilization, health and socioeconomic equality in Finland) Sosiaali- ja terveyshallitus. Tutkimuksia 20, Helsinki 1992.

Häkkinen U. The impact of changes in Finland's health care systems. Health Econ 2005:14:S1(September):S101-S118.

Sintonen H. An approach to economic evaluation of actions for health. SVT XXXII:74. Helsinki 1981.

Wagstaff A. QALYs and the equity-efficiency trade off. J Health Econ 1991:10:21–41.

Williams A. Cookson R. Equity in health. In Culyer A, Newhouse JP eds. Handbook of health economics. Volume 1B Amsterdam Elsevier 2000. 1863–1910.

VATT-JULKAISUJA / VATT PUBLICATIONS

VATT-JULKAISUJA SARJASSA AIEMMIN ILMESTYNEET JULKAISUT

27:5. 1990-luvun talouskriisi. Suomen Akatemian tutkimusohjelma. Jaakko Kiander: Laman opetukset. Helsinki. 2001.

27:6. 1990s economic crisis. The Research Programme of the Academy of Finland: Down from the heavens, up from the ashes. The Finnish economic crisis of the 1990s in the light of economic and social research. Jorma Kalela – Jaakko Kiander – Ullamaija Kivikuru – Heikki A. Loikkanen – Jussi Simpura (eds.). Helsinki 2001.

27:7. 1990s economic crisis. The Research Programme of the Academy of Finland: Final Report. Eds. Jaakko Kiander – Sari Virtanen, Helsinki 2002.

28:1. Hyvinvointivaltio 2000-luvun kynnyksellä. VATT-vuosikirja 1999. Toimittaneet Reino Hjerppe – Seija Ilmakunnas – Iikko B. Voipio. Helsinki 1999.

28:2. Hyvinvointivaltio 2000-luvun kynnyksellä. VATT-vuosikirjan 1999 artikkelit. Toimittaneet Reino Hjerppe – Seija Ilmakunnas – Pirkko Valppu – Iikko B. Voipio. Helsinki 1999.

28:3. The Finnish Welfare State at the Turn of the Millennium. VATT Yearbook 1999. Helsinki 2000.

29. Social Capital: Global and Local Perspectives. Jouko Kajanoja – Jussi Simpura (eds.). Helsinki 2000.

30. Työmarkkinat testissä. Seija Ilmakunnas (toim.). Helsinki 2001.

31. Yhteistyöllä ja kilpaillen – Peruspalvelut 2000 - tutkimusraportti. Kauko Aronen – Maija-Liisa Järviö – Kalevi Luoma – Tarmo Räty. Helsinki 2001.

32. Towards Higher Employment, The Role of Labour Market Institutions. Seija Ilmakunnas – Erkki Koskela (eds.). Helsinki 2002.

33. Tutkimus ja talouspolitiikka – Kirjoituksia taloudellisesta kasvusta. Toimittaneet Jouko Kajanoja – Jaakko Kiander – Iikko B. Voipio. Helsinki 2002.

34. Kansankodin tuolla puolen – Hyvinvointivaltion tilivelvollisuuden tarkastelua. Seppo Lindblom. Helsinki 2002.

35. Parempi työllisyys, pienempi työttömyys – Kuinka se tehdään? Heikki Räisänen (toim.) Helsinki 2003.

36. Julkinen talous globalisoituvassa maailmantaloudessa. Reino Hjerppe (toim.) Helsinki 2003.

37. Kunnalliset palvelut – Terveyden- ja vanhustenhuollon tuottavuus. Reino Hjerppe – Aki Kangasharju – Reijo Vuorento (toim.). Helsinki 2003.

38. Mitä on työvoimapolitiikka? Elisabet Heinonen – Kari Hämäläinen – Heikki Räisänen – Matti Sihto – Juha Tuomala. Helsinki 2004.

39. Kohti globaalivastuuta, Maailmanhallinta – politiikkaa, taloutta ja demokratiaa. Seppo Tiihonen – Paula Tiihonen. Helsinki 2004.

40. Hyvinvointi ja työmarkkinoiden eriytyminen. Reino Hjerppe – Heikki Räisänen (toim.). Helsinki 2004.

41. Verotus, talouspolitiikka ja kansantalous – Taxation, Economic Policy and the Economy. Vesa Kanniainen – Seppo Kari (toim.). Helsinki 2005.

42. Suomi 10 vuotta Euroopan unionissa – Taloudelliset vaikutukset. Jaakko Kiander – Antti Romppanen (toim.). Helsinki 2005.

43. Suomi 2025 – Kestävän kasvun haasteet. Reino Hjerppe – Juha Honkatukia (toim.). Helsinki 2005.

44. Kasvumallin muutos ja veropolitiikan haasteet. Reino Hjerppe – Seppo Kari – Jaakko Kiander (toim.). Helsinki 2006.