uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Olga Milliken
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Economics)
GRADE / DEGREE

Department of Economics
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Three Essays in Health Economics and Public Policy

TITRE DE LA THÈSE / TITLE OF THESIS

Vicky Barham
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

Rose Ann Devlin
CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Marie Allard

Louis Hotte

Stefan Dodds

Jean-François Tremblay

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

THREE ESSAYS IN HEALTH ECONOMICS AND PUBLIC POLICY

by

Olga V. Milliken

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Economics

Department of Economics
Faculty of Social Sciences
University of Ottawa

September 2008

Your file Votre référence
ISBN: 978-0-494-48407-4
Our file Notre référence
ISBN: 978-0-494-48407-4

This dissertation is dedicated
to the memory of my mother,
Ludmila V. Lavrova

ABSTRACT

Essay One: Genetic Health Risks: The Case for Universal Public Health Insurance. This paper examines the appropriate role for the public sector in providing genetic and health insurance when health risks are genetically determined at conception. The *ex ante* efficient outcome (when individuals are ignorant about their health risks) is characterized. It is demonstrated that this outcome cannot be achieved by private health insurance markets or by a government which cannot commit to a once-and-for-all transfer policy. In contrast, the desired outcome is attained through public provision of universal health insurance and of genetic testing, coupled with a public pension scheme.

Essay Two: Fee-for-Service vs. Capitation: Anything You Can Do - I Can Do Better (and Cheaper). This paper recasts the analysis of optimal physician remuneration - generally presented as a contest between prospective (capitation) and retrospective (fee-for-service) schemes - as a problem in price theory. This approach abstracts from problems of asymmetric information and concentrates on the design of the price mechanism. It demonstrates that when the demand for health care is price-inelastic, the appropriately designed fee-for-service and capitation schemes both lead to Pareto efficient outcomes. When a patient's demand for care is uncertain, or the risk of poor health outcomes depends on the preventive care provision, standard arguments concerning risk bearing are used to

prove that paying physicians on a fee-for-service basis can deliver socially-optimal outcomes at a lower cost than if they are paid under a capitation scheme.

Essay Three: Comparative Efficiency Assessment of Primary Care Models Using Data Envelopment Analysis. This paper compares the productive efficiency of four models of primary care service delivery in Ontario, Canada, using the methodology of Data Envelopment Analysis. Particular care is taken to include quality of service in the output measure. The influence of the delivery model on productive efficiency is disentangled from patient characteristics using regression analysis. The traditional fee-for-service arrangement ranks highest and the Community Health Centre model (which involves a multidisciplinary team of health care professionals paid on a salary basis) the lowest in efficiency scoring. The reliance of input measures on the costs of running a practice and on the number of patients favours the fee-for-service model.

CONTENTS

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisors, Professor Rose Anne Devlin and Professor Vicky Barham. They have advised, guided, and inspired me throughout the extended journey of writing this dissertation. I met Professor Barham when I was her teaching assistant for an Introductory Microeconomics course. The first essay of my dissertation was conceived while writing a research paper for a public economics course she taught. Professor Barham introduced me to Professor Devlin when they were searching for a student who would be interested in pursuing Health Economics, and thus I found my supervisors. I want to thank Professor Barham for her optimism and inspiration over the course of the past five years. She taught me how to see a forest rather than trees (which is no mean feat for a girl used to picking mushrooms in forests), and how to apply theory to the study of a practical policy problem. Moreover, she held me to a high standard of thoroughness and academic rigor. Professor Devlin taught me the fine art of empirical research: teasing out relevant information from a mountain of data and distilling it into a meaningful economic analysis. Vicky and Rose Anne introduced me to a style of academic writing that I will endeavor to pursue throughout my career.

Professor Devlin facilitated my entry to the "Comparison of Primary Care Models in Ontario" project at the C. T. Lamont Centre for Research in Primary

Health Care at the Élisabeth Bruyère Research Institute where I was able to collect data and conduct an economic analysis which became the third essay of my dissertation. This essay could not have been written without the research environment provided by Dr. William Hogg, Professor in the Department of Family Medicine at the University of Ottawa, Director of the Centre, and the principle investigator of this project. I benefited enormously from the multidisciplinary team of researchers at the Centre: Dr. Laura Muldoon, Dr. Grant Russell, Simone Dahrouge, and Dr. Meltem Tuna.

Without the welcoming and unwavering support of the administrative personnel at the Department of Economics, my life as a Ph.D. student, would have been much more difficult. I would like to thank Diane Ritchot, Bernadette Benoit, Irène Paré and Luce Laviolette for their assistance.

I benefited from insightful and invaluable comments regarding my dissertation from Professor Marie Allard, Professor Louis Hotte, Professor Jean-François Tremblay, and Professor Mike Hoy. Many thanks go to Professor Nguyen Van Quyen for advising me as a supervisor of the Ph.D. Program and for teaching me valuable mathematical skills.

I am grateful to my family who have been my stability, my mother-in-law, Dorothy Milliken for her listening ear, invaluable practical assistance, and for providing numerous hours of child care for her granddaughter Ekaterina during the last several months of my work on this dissertation. I am indebted to my family in Russia, my aunts and my cousin Natasha who have always believed in me, and

supported and nurtured my academic endeavours. And most of all, I would like to thank my dear husband Scot for his love, inspiration and reassurance when I did not see "the light at the end of the tunnel". Thank you for listening and being patient, helping me to develop my writing style, and encouraging me to excel.

Ottawa, Canada, September 2008

CHAPTER 1

Introduction

Throughout the world, publicly funded health systems appear to be in an unfortunate state. Health care costs are going up due to developments in technology, including new pharmaceuticals and equipment, and due to increased longevity and the rising demand of an aging population. This, coupled with a lack of financing, undermines quality of care, timely access, and the effectiveness of medical treatments. More funding would appear to be a potential remedy, but governments argue that they simply cannot afford to allocate more resources to health care than is presently the case. Widespread frustration with backlogs has led many policy makers - and voters - to advocate a reduction in public sector involvement in the provision and financing of health care, and therefore a shifting of more of the financial burden to private individuals.

Advances in genetic testing technologies add a new spin to the raison d'être of public sector's involvement in the provision of health care and of health insurance. These new developments mark a considerable change in the ability to detect early in life (or even before a child is conceived) health risks and to identify individual predispositions to many diseases. Increased knowledge about an individual's health risks is a double-edged sword: on the one hand, it may facilitate the socially desirable prevention of the onset of diseases while, on the other hand, and in the absence of a regulated insurance market, it creates an underclass of individuals

who will be uninsured due to high health risks or because their insurance premiums are prohibitively high. Some countries, but not Canada, have outlawed the use of genetic information for insurance purposes. However, the complete suppression of the family health history to protect an individual against genetic risk discrimination may not be feasible. In a publicly funded health care system with universal access to health care at an equalized price the problems of risk differentiation, underinsurance, and prohibitive premiums are not present. However, in this system, which struggles to allocate limited health care dollars, the reasoning for financing the costs of providing a panoply of genetic tests and of further treatments must be justified.

In addition to contributing to the justification for the public sector's role in the provision of health insurance, this thesis addresses key issues affecting the effective delivery of health care services. Although physician services in Canada account only for 13 per cent of total health care expenditures[1], policy-makers pay close attention to how providers of primary health services organize their practices and how they are remunerated for the services provided. The existing literature has focused primarily on the strengths and weaknesses of a traditional fee-for-service payment mechanism which pays physicians per unit of service, and of a capitation scheme in which physicians receive a monthly fee for each patient in their rosters. Since neither of these two payment schemes is without flaws (at least in the way they are set up in practice), a blend of the two is often presented as a better

[1] Compared to hospital expenditures of 30 per cent of the total health care spending in 2006 (CIHI 2007).

alternative. It seems, however, that policy-makers have been concentrating on the differences in outcomes that these two payment mechanisms bring about instead of focusing on the nature and the appropriate design of the two. Therefore, it must be asked whether incentives can be introduced via the pricing of the two mechanisms that yield desirable health care outcomes.

Relatedly, it is evident that primary care can be delivered in a variety of institutional settings – which vary with respect to whether physicians are part of a group or are solo practitioners, and how they promote and handle multidisciplinary groups as well as those practices where care is delivered only by physicians and nurses. In many countries, efforts have been made to influence the institutional environment in which family physicians work. What is not clear is whether each of these models are equally effective at delivering primary health care to the diverse populations served by publicly-funded health care systems. A systematic study of these models is paramount. Furthermore, it is important to understand how the characteristics of the model may affect its performance. In Canada and, particularly, in the province of Ontario there is a lot of room for a comprehensive examination of primary care service delivery.

This dissertation examines the above health policy issues in three essays. Essay One studies theoretically the implications of an early (at birth or at conception) identification of individual health risks for social and individual welfare. This is examined in the environment where genetic testing and genetic insurance are available on the market. The paper recognizes that the developments in new genetic

technologies and in genetic testing make it impossible to suppress completely the information about an individual's risk type, because one cannot insure against having parents and/or grandparents with a certain gene pool. Since individuals would like (but are unable) to purchase insurance against falling into a high-risk category, competitive insurance markets are incapable of implementing ex ante efficient outcomes: government involvement is necessary. At first inspection, however, it is not evident that public sector involvement extends as far as actual public provision of health and genetic insurance: an ex post transfer to compensate those in the high-risk category for incurring high health insurance premiums would seem to solve the problem. However, strategic manipulation of savings on the part of individuals renders government policy time inconsistent, and means that cash transfers will not be effective at implementing the first best. Instead, it is shown that a policy consisting of public universal health insurance and of a social pension ensures redistribution via the expenditure arm of the public sector, rather than through taxation. The analysis developed in this paper demonstrates that it may be desirable to limit the government's capacity to use observable information about an individual's type for taxation purposes. Also, the result that a universal health insurance program (or mandating health insurance purchases) cannot by itself ensure an efficient redistribution of resources in the economy highlights the importance of making joint decisions regarding different social programs.

The two other essays of the dissertation contribute to our understanding of key issues with respect to the reform of primary health care provision. They ex-

amine how an incentive payment structure and other organizational characteristics of physician practices influence efficiency in the delivery of primary health care. Essay Two, inspired by work in price theory and, in particular, the theory of clubs, compares two compensation schemes for physician's services: fee-for-service and capitation. The application of these approaches to the analysis of physician compensation mechanisms is novel; it abstracts from the difficulties created by asymmetries of information in order to focus on the role of the price mechanism. The paper starts with the observation that in an environment where both patients and physicians are heterogeneous, efficient provision of health care services may require a specific matching of patients with physicians. Thus, one role of the payment scheme is to ensure that the right provider is matched with each patient, and the other is to guarantee the provision of necessary medical services. The theoretical approach chosen in this paper underscores the similarities rather than the differences between the outcomes that can be achieved under fee-for-service and capitation payment schemes, and highlights the fact that fees must reflect the heterogeneity of patient medical needs. The analysis shows that if the demand for physicians' services is exogenously determined and if it is certain for an individual physician, then both the fee-for-service and capitation mechanisms result in the same health care outcomes delivered at identical costs. In addition, the paper considers the performance of the two payment schemes under demand uncertainty, and when preventive measures taken by a physician decrease this uncertainty. The approach of the paper permits an examination of how differences with respect

to attitudes towards risk between an individual physician and the public sector influence the cost of providing primary care services under the two payment mechanisms. It is demonstrated that under some circumstances physicians who are paid using the fee-for-service scheme are able to deliver the required services at a lower cost than their peers compensated on the capitation basis.

Essay Three of the dissertation undertakes an efficiency comparison of four distinct models of primary health care service delivery in the province of Ontario, Canada, employing the methodology of Data Envelopment Analysis (DEA). These models are: traditional fee-for-service practices (FFS), health service organizations (HSOs), family health networks (FHNs), and community health centres (CHCs). A typical FFS practice consists of a solo physician – the owner – who is paid by the provincial government a relatively flat fee per unit of medical service provided. In HSO practices, which are solo or group, physicians are compensated on a capitation basis for each patient in their rosters. In Family Health Networks, physicians provide comprehensive care for their patients 24 hours a day, seven days a week and are paid under a blended scheme, which combines capitation payments with incentives for preventive care and other services. Community health centres are interdisciplinary teams, employing various health care professionals and social workers to address disadvantaged populations' needs. They pay physicians straight salaries.

The empirical analysis of this paper draws on the data collected as part of a multidisciplinary project funded by the Ontario Ministry of Health and Long-Term

Care. This project made it possible to undertake a comprehensive study of the performance of these models. This essay pioneers an economic evaluation of these models in Ontario. Another unique feature of this paper is that it incorporates several quality-of-care and performance indicators and is able to assess their cumulative influence on the efficiency of primary care practices. Previous studies have not employed a wide spectrum of clinically approved performance measures. The DEA technique allows for the combining of several measures of primary care outcomes into a single output for each primary care practice. The assessment of this output relative to the inputs used to deliver it results in an efficiency score for each primary care practice. The paper conducts a sensitivity analysis to evaluate the importance of the input measures to the primary care model's efficiency performance. Once efficiency scores are calculated for different input scenarios, a regression analysis is carried out to disentangle the "model" factor from patient characteristics and other environmental variables which help explain the variation in efficiency scores across the models.

The paper concludes that the CHC model is at least as good as others in terms of the quality of care provided, and for some indicators it outperforms the other models. However, once the costs of running a primary care practice are taken into account, the FFS model is the leader in the efficiency ranking, with CHC practices being, on average, the worst performers. Several explanations are offered for the results, including the possibility of a non-linear relationship between performance indicators and costs – it may be relatively inexpensive to achieve a low

level of performance, but very costly to push the quality of care beyond a certain threshold. It may also be the case that CHCs are too small relative to the large fixed costs necessary to operate a multidisciplinary centre, and therefore a larger patient "roster" is needed to bring the average cost of operating a centre down. The finding that HSO and FHN models have lower efficiency ranking than the FFS model is interesting since they have been introduced to replace the traditional FFS model, where physicians are arguably not concerned with the costs of providing medical services. This result, in some sense, confirms the theoretical findings of the second essay, i.e., that capitation is more expensive than the fee-for-service payment mechanism due to the risk premium that is required to compensate a risk-averse physician for facing income uncertainty brought in by a perspective nature of capitation.

Many additional questions emerge from the research presented in this dissertation. In particular, both Essays Two and Three highlight the importance of taking account of the heterogeneity of both physicians and patients. This suggests that the quest for one optimal compensation scheme, or one optimal model of delivery of primary care delivery may be ill-conceived: it may in fact be preferable to offer a menu of options, which offer meaningful choices to both physicians and patients. In terms of future empirical work, it is evident that the insights would be much more compelling if it were possible to capture more fully all of the costs incurred by each patient – including the costs of hospital care, and of specialist care. There is good reason to anticipate that different models of primary care provision may

differ systematically with respect to the way in which they offload service provision to other actors in the health care system, and a truly adequate comparison of different institutional reforms must take all of these costs into account. This of course leads back directly to the concerns of Essay One, and the reasons for public sector involvement in the financing of health care. Over the next few years, health economics researchers need to help policy makers understand better the relationship between dollars spent on health care, dollars spent on other programs (such as pensions), and the actual production of health. Who is going to decide how to allocate the public sector budget? Political economy models will be particularly useful in gaining insight into these matters.

REFERENCES

Canadian Institute for Health Information (2007) *Health Care in Canada*, available at cihi.ca

CHAPTER 2

Genetic Health Risks: The Case for Universal Public Health Insurance

2.1 Introduction

Medical researchers worldwide are currently expending enormous effort searching for genetic markers for specific diseases and health conditions. Genetic testing is already widespread: hundreds of genetic disorders can be detected through pre-natal screening, and in most industrialized countries programs are in place for the universal testing of newborns for a number of genetic and metabolic diseases (Hanley 2005). In countries with publicly-funded health care systems, the growth in the panoply of available genetic tests has been accompanied by a corresponding growth in concern regarding the cost of testing and of treatment, and of the potential cost burden this may represent for the health care system (Miller *et al.* 2002). This discussion is often embedded in a broader reflection on the appropriate role of the public sector in the provision of health care.

This paper examines the role of the public sector in providing health insurance when health risks are genetically determined at conception. It is assumed that health risks are known, thus abstracting from the specific problems that arise due to either moral hazard or adverse selection. While it is now well-established that private insurance markets will generally not allocate resources efficiently when information is asymmetric, the case for publicly-provided health care (or, equivalently, publicly-provided health insurance) is less compelling when there are no

asymmetries of information. Below, the *ex ante* efficient outcome is characterized, and it is shown that competitive private insurance companies will not offer individuals health insurance on the terms required to attain this outcome. This is because, as Doherty and Thistle (1996) point out, in an environment where differences in individual's genetic risks are publicly known, one can acquire insurance against the risk that a certain disease predisposition will translate into an actual disease, but it is impossible to insure against having the 'wrong' parents with defective genes. As hinted by Cutler (2002) and Cutler and Zeckhauser (2000), in this sort of economic environment the *ex ante* efficient outcome cannot be attained unless individuals can purchase antenatal insurance - before they know who their parents are - an option that is not of any practical interest.

In principle, however, the outcome that would prevail if genetic (antenatal) insurance were available can be obtained by implementing a system of lump-sum taxes and transfers from low-risk to high-risk citizens, and then letting each purchase insurance in the private market. For universal public health insurance to be optimal, it must be the case that government policy is time-inconsistent, therefore precluding it from implementing a redistributive policy of this kind: the lesson drawn from models such as those developed by Bruce and Waldman (1991) and Coate (1995) is that when a social-welfare-maximizing government cannot commit to denying assistance to individuals who have not purchased adequate protection for themselves, then a cash transfer policy is time inconsistent and welfare reducing as compared to the provision of universal public health insurance. Typically,

infirm (or high-risk) individuals have an incentive to not purchase insurance, since a greater degree of neediness in the disaster state will increase the transfer which they receive. Strikingly, the paper demonstrates that, even if the government were to choose the transfer policy required to implement the *ex ante* efficient outcome, and even if infirm (or high-risk) individuals purchase full genetic and health insurance, the equilibrium is not efficient. This is because all individuals - healthy and infirm strategically adjust their savings to minimize their tax burden. In contrast, public provision of health care (and genetic testing), coupled with publicly provided pensions which discourage the government from redistributing in period two, make it possible to implement the desired outcome as an equilibrium.

The analysis developed below shows how the incentives for individuals to undergo genetic testing and to purchase genetic insurance are affected by redistributive policy. When the government can commit to a once-and-for-all income transfer at the beginning of the first period, then individuals undergo genetic testing, and purchase full insurance, whenever it is socially efficient for them to do so. However, when governments cannot prevent themselves from redistributing wealth in the second period then, much in the spirit of Hoel and Iverson (2002), individuals will typically not choose to undergo genetic testing - even when it is socially optimal - if they are required to bear the full cost of these procedures. When individuals do choose to submit to genetic testing, their expected utility is independent of whether or not they purchase genetic insurance. This result can be compared to Tabarrok (1994) who underscores the inefficiencies that may arise in the insur-

ance market if individuals are able to undergo genetic testing, but who suggests that the socially optimal outcome can be attained by obliging those undergoing genetic testing to also purchase genetic insurance. Our analysis strongly suggests that whenever genetic testing is socially optimal, it should also be publicly funded.

The remainder of the paper is organized as follows. Section 2.2 describes the *ex ante* efficient outcome - when individuals are behind the veil of ignorance. It also examines whether individuals undertake genetic testing when it is socially optimal. Section 2.3 shows that this outcome could be implemented by a Benthamite government which can commit to a once-and-for-all redistributive policy, but is not attainable if the government cannot resist redistributing wealth from rich to poor in the second period. In contrast, the *ex ante* efficient outcome is attained when genetic testing, health care and pensions to finance consumptions in period two are publicly provided. Section 2.4 provides a discussion and concludes.

2.2 Efficient Outcomes with Genetically Determined Health Risks

Consider a two-period economy in which individuals are born either healthy or infirm. Regardless of their risk type, all individuals are healthy in period one and during this period they can learn their type by undergoing a genetic test, at a cost c.[1] Otherwise, risk types are publicly - and costlessly - revealed at the beginning of period two.

In period two, infirm-type individuals face two states of nature: 'good' and 'disaster'. The probability that an infirm-type individual experiences ill health

[1] Later this assumption is relaxed to allow for observable health risk heterogeneity since birth.

is π. In the disaster state he/she becomes ill and incurs loss L , which may be interpreted as the cost of medical treatment. In contrast, healthy-type individuals remain healthy in period two.

Assume that, at birth, all individuals are equally likely to turn out to be infirm, and that this probability is equal to γ, $0 < \gamma < 1$. An individual who undergoes a genetic test and learns that he or she is infirm can, in period one, undertake preventive medical treatment, at fixed cost M. Individuals who undertake preventive medical treatment reduce the risk of the disaster state to $\pi - \alpha$. The results of the genetic test are public information.

Individual utility depends on consumption, C, in each period and, for simplicity, it is assumed that there is no discounting of future consumption. Consequently, $U = U(C_1) + U(C_2)$. In period one, all individuals have income y, and they finance consumption in period two by savings in period one. Assume that there exists a competitive insurance market. In period one, before undergoing genetic testing, individuals can purchase genetic insurance G at actuarially-fair premium γ. In period two, health care insurance coverage Z against loss L is sold to infirm-type individuals who did not undergo genetic testing at price π per unit of coverage and to those who have undergone genetic testing and preventive medical treatment, at price $\pi - \alpha$ per unit of coverage.

Given that all individuals are equally likely to be infirm, and that they have identical preferences and income, in this setting they will all either choose to undergo genetic testing, or no one will. So, assume for the meantime that everyone

chooses to submit to genetic testing, and that individuals who discover that they are infirm choose to undertake preventive medical treatment. In this case, the optimization problem to be solved by a given individual can be expressed as

$$\begin{aligned}\max_{G} EU \;=\; & \gamma[U\left(y-c-M-\widehat{s}_I+(1-\gamma)G\right) \\ & +(\pi-\alpha)U\left(\widehat{s}_I-L+(1-\pi+\alpha)\widehat{Z}\right) \\ & +(1-\pi+\alpha)U\left(\widehat{s}_I-(\pi-\alpha)\widehat{Z}\right)] \\ & +(1-\gamma)\left[U\left(y-c-\gamma G-\widehat{s}_H\right)+U(\widehat{s}_H)\right],\end{aligned} \tag{2.1}$$

subject to $\widehat{s}_I$ and $\widehat{s}_H$ (period-one savings by individuals whom genetic testing reveals that they are infirm-type and healthy-type, respectively) being the solutions to

$$\begin{aligned}&\max_{s_I} U\left(y-c-M-s_I+(1-\gamma)G\right) \\ &+(\pi-\alpha)U\left(s_I-L+(1-\pi+\alpha)\widehat{Z}\right) \\ &+(1-\pi+\alpha)U\left(s_I-(\pi-\alpha)\widehat{Z}\right),\end{aligned} \tag{2.2}$$

$$\max_{s_H} U\left(y-c-\gamma G-s_H\right)+U(s_H), \tag{2.3}$$

respectively. The maximization problems in (2.2) and (2.3) are subject to $\widehat{Z}$ solving

$$\begin{aligned}&\max_{Z}\ (\pi-\alpha)U\left(s_I-L+(1-\pi+\alpha)Z\right) \\ &+(1-\pi+\alpha)U\left(s_I-(\pi-\alpha)Z\right).\end{aligned} \tag{2.4}$$

After straightforward calculations (see Appendix A), (2.1)-(2.4) yield

$$\widehat{Z} = L, \tag{2.5}$$

$$\widehat{s}_I = \frac{y-c-\gamma[M+(\pi-\alpha)L]}{2}+(\pi-\alpha)L, \tag{2.6}$$

$$\widehat{s}_H = \frac{y-c-\gamma[M+(\pi-\alpha)L]}{2}, \tag{2.7}$$

$$\widehat{G} = M+(\pi-\alpha)L, \tag{2.8}$$

$$\widehat{U}_H = \widehat{U}_I = 2U\left(\frac{y-c-\gamma[M+(\pi-\alpha)L]}{2}\right). \tag{2.9}$$

In contrast, if individuals do not undertake genetic testing, they can still purchase genetic insurance to provide protection against learning that they are infirm in period two, but since they cannot undertake preventive treatment they will still have to pay π per unit of health insurance purchased in period two. The optimization problem therefore becomes

$$\begin{aligned}\max_{G,s} EU = {} & U(y-\gamma G-s)+ \\ & \gamma\left[\begin{array}{l}\pi U\left(s+G-L+(1-\pi)\widehat{\widehat{Z}}\right) \\ +(1-\pi)U\left(s+G-\pi\widehat{\widehat{Z}}\right)\end{array}\right] \\ & +(1-\gamma)U(s),\end{aligned} \tag{2.10}$$

subject to $\widehat{\widehat{Z}}$ being the solution to

$$\begin{aligned}\max_{Z} \; & \pi U(s+G-L+(1-\pi)Z) \\ & +(1-\pi)U(s+G-\pi Z).\end{aligned} \tag{2.11}$$

Optimization problems in (2.10)-(2.11) result in

$$\begin{aligned}
\hat{\hat{Z}} &= L, \\
\hat{\hat{G}} &= \pi L, \\
\hat{\hat{s}} &= \frac{y-\gamma\pi L}{2}, \\
\hat{\hat{U}}_H &= \hat{\hat{U}}_I = 2U\left(\frac{y-\gamma\pi L}{2}\right)
\end{aligned}$$

(see Appendix B for details). This implies that individuals will purchase genetic insurance - even if they do not undergo genetic testing - in order to perfectly smooth their consumption over their life-cycle, regardless of whether or not they will be revealed to be healthy or infirm. In this case, the genetic insurance is effectively protecting them against the subsequent revelation that they are infirm and will therefore need to purchase health care insurance. Notice that, comparing the solutions for the scenario where all consumers undergo genetic testing with that where they do not test in period one, agents choose to test if and only if it is socially efficient for them to do so, i.e.,

$$\begin{aligned}
\frac{y-c-\gamma[M+(\pi-\alpha)L]}{2} &> \frac{y-\gamma\pi L}{2} \\
-c-\gamma[M-\alpha L] &> 0 \\
\gamma\alpha L &> c+\gamma M,
\end{aligned}$$

which simply means that the expected increase in consumption due to the reduced likelihood of loss exceeds the cost of undergoing testing and the expected cost of preventive health care. This result can be contrasted with that of Tabarrok (1994), who suggests that individuals would need to be compelled to purchase

genetic insurance before undergoing testing. Here, where all individuals are *a priori* identical, no such regulation need be implemented: if it is desirable for individuals to seek testing, they will do so. Moreover, it should be observed that in this extremely simple setting there are no issues - à la Tabarrok (1994) or Polborn *et al.* (2006) - with respect to a wish to suppress information regarding the results of the genetic test on the part of the consumer (or even a desire to hide the results from oneself). In period two, the only individuals who purchase health insurance are of the infirm type, and therefore even if insurance companies cannot observe the agents' types, from the fact that an individual seeks to purchase health insurance in period two, the insurer can infer that this person is of the infirm type. In contrast, it is clearly important to the consumer that the purchase of preventive medical care can be observed by the insurance company, so that period two health care premiums will be adjusted accordingly.

But of course, as observed by Doherty and Thistle (1996), it is not really plausible to suppose that all individuals face the same risk of becoming infirm. In practice, even though individuals may not know for sure whether they are the healthy or infirm type, there is often information (typically observable and verifiable) about the healthiness of their parents (and grandparents) which means that all agents do not in fact face the same risk of being infirm in period two. Therefore consider a simple generalization of our genetic testing model, and assume that agents may be at either high ($\overline{\gamma}$) or low ($\underline{\gamma}$) risk of becoming infirm in period two. Proportion λ of the population is high-risk, and proportion $(1-\lambda)$ is low

risk. As above, it is assumed that genetic testing can be undertaken at cost c, that preventive medical treatment can be taken in period one at cost M, and that this medical treatment reduces the probability of the disaster state to $\pi-\alpha$. Also assume that the fact that an individual has undertaken preventive medical treatment is observable, and the fact that they are infirm- or healthy-type is costlessly revealed at the beginning of period two. For simplicity, it is assumed that all infirm-type individuals who have not undertaken preventive medical treatment in period one face the same probability π of experiencing the disaster state in period two.

Before investigating the competitive market equilibrium, it is useful to first, characterize the *ex ante* efficient outcome, that is, the outcome which all individuals would choose behind the veil of ignorance (i.e., before they learned their risk type) but given that they were able to purchase insurance at actuarially-fair rates. Considering a situation in which individuals choose to undergo testing (and to undertake preventive care), the representative individual's decision problem can be expressed as

$$\max_{\substack{s_{\overline{\gamma}I},G_{\overline{\gamma}},G_{\lambda},Z_{\overline{\gamma}} \\ s_{\underline{\gamma}I},G_{\underline{\gamma}},Z_{\underline{\gamma}}}} EU =$$

$$\lambda\left[\begin{array}{c} \overline{\gamma}\left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}I}\right)+ \\ (\pi-\alpha)U(s_{\overline{\gamma}I}-L+(1-\pi+\alpha)Z_{\overline{\gamma}})+ \\ (1-\pi+\alpha)U(s_{\overline{\gamma}I}-(\pi-\alpha)Z_{\overline{\gamma}}) \end{array}\right] \\ +(1-\overline{\gamma})\left[U\left(y-c-\overline{\gamma}G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}H}\right)+U(s_{\overline{\gamma}H})\right] \end{array}\right]$$

$$+(1-\lambda)\left[\begin{array}{c} \underline{\gamma}\left[\begin{array}{c} U\left(y-c-M+(1-\underline{\gamma})G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}I}\right)+ \\ (\pi-\alpha)U(s_{\underline{\gamma}I}-L+(1-\pi+\alpha)Z_{\underline{\gamma}})+ \\ (1-\pi+\alpha)U(s_{\underline{\gamma}I}-(\pi-\alpha)Z_{\underline{\gamma}}) \end{array}\right] \\ +(1-\underline{\gamma})\left[U\left(y-c-\underline{\gamma}G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}H}\right)+U(s_{\underline{\gamma}H})\right] \end{array}\right],$$

where $G_{\overline{\gamma}}, G_{\underline{\gamma}}$ denote a genetic insurance coverage (as discussed above) that a high-risk/low-risk type would want to purchase to protect themselves against the risk of facing a high insurance premium, if they turn to be infirm, whereas G_{λ} denotes a second genetic insurance policy that protects agents against being at high-risk of turning out to be infirm. As above it is assumed that insurance is offered at actuarially fair rates. The solution to this problem is $Z^*_{\overline{\gamma}} = Z^*_{\underline{\gamma}} = L$, $G^*_{\overline{\gamma}} = G^*_{\underline{\gamma}} = M + (\pi-\alpha)L$, $G^*_{\lambda} = M + (\pi-\alpha)L(\overline{\gamma}-\underline{\gamma})$, $s^*_{\overline{\gamma}H} = s^*_{\underline{\gamma}H} = \frac{1}{2}\left(y-c-(M+(\pi-\alpha)L)(\underline{\gamma}+\lambda(\overline{\gamma}-\underline{\gamma})\right)$, $s^*_{\overline{\gamma}I} = s^*_{\underline{\gamma}I} = \frac{1}{2}\left(y-c-(M+(\pi-\alpha)L)(\underline{\gamma}+\lambda(\overline{\gamma}-\underline{\gamma})\right)+(\pi-\alpha)L$, and $U^*_{\underline{\gamma}H} = U^*_{\underline{\gamma}I} = U^*_{\overline{\gamma}H} = U^*_{\overline{\gamma}I} = 2U\left(\frac{1}{2}\left(y-c-(M+(\pi-\alpha)L)(\underline{\gamma}+\lambda(\overline{\gamma}-\underline{\gamma})\right)\right)$ (see Appendix C for proof).

It is now straightforward to establish that the competitive market equilibrium will not be *ex ante* efficient. Recall that the risk-class that each agent belongs to is observable, i.e., whether or not someone is at high or low risk of becoming infirm in period two is public knowledge. In this setting it is possible for competitive private insurers to offer premium insurance to each risk type, i.e. $G_{\overline{\gamma}}, G_{\underline{\gamma}}$. In contrast, it is not possible for them to sell insurance against being a $\overline{\gamma}$-risk individual rather than a $\underline{\gamma}$-risk person. What this means is that in the competitive insurance market high-risk individuals pay higher premiums for genetic insurance

than do low-risk individuals, and therefore experience lower life-time utility than do low-risk individuals. Therefore, the *ex ante* efficient outcome - in which lifetime utility is independent of type - is not achieved. In an environment where types are costlessly observable, it is not possible for individuals to insure against being born the high-risk type, as suggested by Cutler (2002), Cutler and Zeckhauser (2000), and Tabarrok (1994).

2.3 Different Routes to Rome: Cash Transfers and In Kind Provision

The previous section demonstrates that the competitive private market cannot replicate the *ex ante* efficient outcome. It might, however, be wondered whether appropriately designed public policy might alleviate this problem. One possible strategy would be for the government to provide a cash transfer to high-risk individuals - financed by a tax on low-risk citizens - which would enable them to cover the higher expenses they incur in purchasing genetic insurance as compared to the expenditures of low-risk individuals[2]; however, whether or not this policy is effective or not clearly depends upon whether or not high-risk agents choose to take out full insurance. This problem is examined here.

The sequence of events is as follows. In period one, before individuals decide on savings, the government has the opportunity to redistribute from low-risk to high–risk individuals (cash transfer T). Individuals then have the opportunity to acquire genetic insurance. They subsequently can undergo genetic testing at cost c, enabling them to learn their type (infirm or healthy), and they can undertake

[2] It is assumed here that purchases of genetic insurance are socially optimal.

preventive treatment at cost M if they are infirm. Regardless of whether they undergo genetic testing in period one, individual types are costlessly revealed at the end of period one. Individuals decide on savings to finance consumption in period two after the types are revealed. At the beginning of period two, before the disaster strikes proportion π of infirm-type individuals, every infirm person decides upon the amount of health insurance coverage, Z.

It is trivial to verify that when the government can commit to a once-and-for-all income transfer T from low-risk to high-risk individuals at the beginning of period one, then both low-risk and high-risk individuals purchase a full genetic insurance coverage to smooth their consumptions between the 'infirm' and the 'healthy' states. Also, infirm-type individuals purchase a full health insurance coverage, thus, fully smoothing their consumption between the disaster and non-disaster states. Moreover, in this setting, it is straightforward to show that a government which maximizes a Benthamite social welfare function[3] will implement a system of income transfers which replicates the *ex ante* Pareto efficient outcome. For completeness, this fact is recorded in the proposition below.

Proposition 1 *When a Benthamite government can commit to a once-and-for-all transfer policy, the optimal transfer policy results in an allocation of resources which is ex ante Pareto efficient.*

Proof. See Appendix D. ■

[3] Here, a Benthamite social welfare function is defined as a sum of utilities of all individuals in the economy with an equal weight assigned to each individual utility.

As stressed by the literature on time consistency (Boadway 1997), what is not self-evident is whether a government which seeks to maximize a Benthamite social welfare function can credibly claim to be committed to a 'once-and-for-all' transfer policy. To this end, it is useful to examine the government's behaviour should individuals fail to fully insure. Specifically, if infirm-type individuals choose to take out insurance $Z < L$, then in period 2 after the disaster occurs and proportion $\pi - \alpha$ of them have experienced the disaster state, the government solves the optimization problem

$$\begin{aligned}\max_{\tau_i} W_d = & \sum_{i=1}^{\lambda\overline{\gamma}(\pi-\alpha)N} U(s_{\overline{\gamma}Ii} + (1-\pi+\alpha)Z_i - L + \tau_{\overline{\gamma}i}) \\ & + \sum_{i=\lambda\overline{\gamma}(\pi-\alpha)N+1}^{\lambda\overline{\gamma}N} U\left(s_{\overline{\gamma}Ii} - (\pi-\alpha)Z_i + \tau^I_{\overline{\gamma}i}\right) \\ & + \sum_{i=\lambda\overline{\gamma}N+1}^{\lambda N} U\left(s_{\overline{\gamma}Hi} + \tau^H_{\overline{\gamma}i}\right) + \sum_{i=\lambda N+1}^{(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha))N} U\left(s_{\underline{\gamma}Ii} - L + (1-\pi+\alpha)Z_i + \tau_{\underline{\gamma}i}\right) \\ & + \sum_{i=(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha))N+1}^{(\lambda+(1-\lambda)\underline{\gamma})N} U\left(s_{\underline{\gamma}Ii} - (\pi-\alpha)Z_i + \tau^I_{\underline{\gamma}i}\right) \\ & + \sum_{i=(\lambda+(1-\lambda)\underline{\gamma})N+1}^{N} U\left(s_{\underline{\gamma}Hi} + \tau^H_{\underline{\gamma}i}\right),\end{aligned} \tag{2.12}$$

subject to the government's budget constraint

$$\sum_{i=1}^{N} \tau_i = 0, \tag{2.13}$$

where τ_i is the disaster assistance provided to infirm individuals who fall ill or a tax imposed on infirm-type individuals who do not fall ill or on healthy type individuals. Individuals are ordered from $i = 1$ to N according to their type:

If $i \leq \lambda\overline{\gamma}(\pi - \alpha)N$, then individual i is of high-risk, infirm and sick

in period two.

If $\lambda\overline{\gamma}(\pi-\alpha)N+1 \leq i \leq \lambda\overline{\gamma}N$, then individual i is of high-risk, infirm and non-sick in period two.

If $\lambda\overline{\gamma}N+1 \leq i \leq \lambda N$, then individual i is of high-risk and healthy in period two.

If $\lambda N+1 \leq i \leq (\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha))N$, then individual i is of low-risk, infirm and sick in period two.

If $(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha))N+1 \leq i \leq (\lambda+(1-\lambda)\underline{\gamma})N$, then individual i is of low-risk, infirm and non-sick in period two.

If $(\lambda+(1-\lambda)\underline{\gamma})N+1 \leq i \leq N$, then individual is of low-risk and healthy in period two.

The first order condition with respect to τ_i is

$$\frac{\partial W_d}{\partial \tau_i} = U_i'(\bullet)+\mu = 0, \quad \forall\, \tau_i, \tag{2.14}$$

where μ is the shadow price of the budget constraint. This condition requires that consumption in period two is equalized for all citizens.

Anticipating that disaster insurance will indeed be provided in the case that they failed to fully insure, the insurance decision of infirm-type individuals is straightforward to examine.

Proposition 2 *If the government maximizes a Benthamite social welfare function and cannot commit to a once-and-for-all transfer policy, then expected utility of individual i is independent of the amount of insurance purchased; in particular,*

therefore, it is optimal for infirm individuals to purchase no health insurance, i.e., $\widetilde{Z} = 0$.

Proof. (So as to avoid visual clutter, the indices i and $\overline{\gamma}$ in the following proof are suppressed.) In period two there are two types of infirm individuals, high-risk individual who became infirm and low-risk individual who became infirm. First, consider the problem of a representative high-risk infirm individual

$$\max_{Z} EU_I = \begin{array}{l} (\pi-\alpha)U(s_I - L + (1-\pi+\alpha)Z + \widetilde{\tau}) \\ +(1-\pi+\alpha)U(s_I - (\pi-\alpha)Z + \widetilde{\tau}^I), \end{array}$$

where $\widetilde{\tau}, \widetilde{\tau}^I$ are the solutions to the optimization problem in (2.12)-(2.13). Consequently,

$$\begin{aligned} \frac{\partial EU_I}{\partial Z} &= (\pi-\alpha)U'(s_I - L + (1-\pi+\alpha)Z + \widetilde{\tau})\left((1-\pi+\alpha) + \frac{d\widetilde{\tau}}{dZ}\right) \\ &\quad -(1-\pi+\alpha)U'(s_I - (\pi-\alpha)Z + \widetilde{\tau}^I)\left((\pi-\alpha) - \frac{d\widetilde{\tau}^I}{dZ}\right). \end{aligned}$$

It can be shown that $\frac{d\widetilde{\tau}}{dZ} = -\frac{N-1}{N}(1-(\pi-\alpha))$ and $\frac{d\widetilde{\tau}^I}{dZ} = \frac{N-1}{N}(\pi-\alpha)$ (see Appendix E for details). Since second period consumptions are equalized, the above first order condition is equal to zero. Moreover, the optimization problem for a low-risk individual who becomes infirm-type yields exactly the same solution. Thus, the expected utility of an infirm-type individual is independent of the amount of health insurance that they purchase, and an optimal solution is to choose $\widetilde{Z} = 0$.

■

The result that the expected utility of infirm individuals is independent of the amount of health insurance, Z, which they purchase – and therefore that $\widetilde{Z} = 0$ is

only one of many possible solutions to Infirm's optimisation problem – is somewhat surprising. One might expect infirm individuals to have an incentive to systematically underinsure - thereby making themselves eligible for an enhanced government handout in the disaster state. This result is driven, however, by the fact that aggregate income is effectively certain in period two, and a Benthamite government will therefore always ultimately equalize the distribution of this aggregate income across all citizens. Consequently, infirm individuals' consumption in period two is independent of the amount of insurance which they purchase. For convenience in what follows it is assumed that $\widetilde{Z} = 0$.[4]

Turning now to the determination of savings at the end of period one observe that, if individuals have undertaken genetic testing at the beginning of period one, then uncertainty regarding their type is resolved: they know, for sure, whether they are healthy or infirm, and whether they were originally low-risk or high-risk individuals is no longer of any relevance to their savings decisions. In this case, therefore, the optimization problem solved by a representative healthy individual is

$$\max_{s_{H_i}} U_H = U(Y_i - s_{Hi}) + U\left(s_{Hi} + \widetilde{\tau}_i^H\right),$$

where $Y_i = y - c - \overline{\gamma} G_{\overline{\gamma}i} + T_i$ if the individual is of high-risk, or $Y_i = y - c - \underline{\gamma} G_{\underline{\gamma}i} + T_i$ if the individual is of low-risk; T_i designates the tax levied by the government in period one, and $\widetilde{\tau}_i^H$ - which is the solution to (2.12)-(2.13) and depends therefore on the savings decisions of all individuals in period one - denotes the period-two

[4] The assumption that individuals purchase no insurance merely simplifies the exposition, and does not alter the results.

transfer (which, in the case of healthy individuals, will in fact be a tax). Similarly, if the individual i is infirm, applying the condition that second period consumptions are equalized, the level of their period one savings, s_{Ii}, can be found by solving

$$\max_{s_{Ii}} U_I = U(Y_i - s_{Ii}) + U\left(s_{Ii} - L + \widetilde{\tau}_i\right),$$

where $Y_i = y - c - M + (1-\overline{\gamma})G_{\overline{\gamma}i} + T_{\overline{\gamma}}$ if the individual is of high-risk, and $Y_i = y - c - M + (1-\underline{\gamma})G_{\underline{\gamma}i} + T_{\underline{\gamma}}$ if the individual is of low risk. At an optimum, it must therefore be true that

$$\frac{dU_H}{ds_{Hi}} = -U'(Y_i - s_{Hi}) + U'\left(s_{Hi} + \widetilde{\tau}_i^H\right)\left[1 + \frac{d\widetilde{\tau}_i^H}{ds_{Hi}}\right] = 0, \tag{2.15}$$

$$\frac{dU_I}{ds_{Ii}} = -U'(Y_i - s_{Ii}) + U'\left(s_{Ii} - L + \widetilde{\tau}_i\right)\left[1 + \frac{d\widetilde{\tau}_i}{ds_{Ii}}\right] = 0. \tag{2.16}$$

Let $\widetilde{s}_{Hi}$ and $\widetilde{s}_{Ii}$ be the solutions to (2.15) and (2.16) respectively.

Proposition 3 *Both high and low-risk individuals have the same first period consumptions.*

Proof. Totally differentiating the first order conditions for $\widetilde{\tau}_i^H, \widetilde{\tau}$ and $\widetilde{\tau}_i^I$ (equation (2.14)) and applying Cramer's Rule one can show that $\frac{d\widetilde{\tau}_i}{ds_{Ii}} = \frac{d\widetilde{\tau}_i^H}{ds_{Hi}} = -\frac{N-1}{N}$ (see Appendix F for details). Therefore, (2.15) and (2.16) yield

$$\frac{U'(Y_i - \widetilde{s}_{Hi})}{U'\left(\widetilde{s}_{Hi} - \widetilde{\tau}_i\right)} = \frac{U'(Y_i - \widetilde{s}_{Ii})}{U'\left(\widetilde{s}_{Ii} - \widetilde{\tau}_i\right)} = \frac{1}{N}, \ \forall i, \tag{2.17}$$

which, since consumption is equalized in period two, also implies that all individuals must achieve the same consumption levels in period one. ■

Observe from (2.17) that there is only imperfect consumption smoothing, which implies that the *ex ante* outcome is not attained: this is because both healthy and infirm individuals overconsume in period one, in order to reduce the burden of redistributive taxation in period two. Moreover, the inefficiency clearly becomes more acute as the size of the population increases: individuals effectively consume their entire endowment during period one, leaving themselves in relative penury in period two.

Consider now the decision problem of a high-risk individual with respect to the purchase of genetic (premium) insurance.

Proposition 4 *In the case in which individuals undertake genetic testing, expected utility is independent of the amount of genetic insurance purchased. In particular, it is optimal to purchase no genetic insurance, i.e.* $\widetilde{G}_{\overline{\gamma}} = \widetilde{G}_{\underline{\gamma}}$.

Proof. A representative high-risk individual solves

$$\max_{G_{\overline{\gamma}i}} U_i = \overline{\gamma}\left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}i}+T_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Ii}\right)+ \\ U(\widetilde{s}_{\overline{\gamma}I}-L+\widetilde{\tau}_i) \end{array}\right]$$
$$+(1-\overline{\gamma})\left[U\left(y-c-\overline{\gamma}G_{\overline{\gamma}i}+T_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Hi}\right)+U(\widetilde{s}_{\overline{\gamma}Hi}+\widetilde{\tau}_{Hi})\right].$$

The first order condition yields

$$\frac{\partial U_i}{\partial G_{\overline{\gamma}i}} = \overline{\gamma}\left(1-\overline{\gamma}-\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dG_{\overline{\gamma}i}}\right)U'(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}i}+T_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Ii})$$
$$+\overline{\gamma}U'(\widetilde{s}_{\overline{\gamma}Ii}-L+\widetilde{\tau}_i)\left(1+\frac{d\widetilde{\tau}_i}{d\widetilde{s}_{\overline{\gamma}Ii}}\right)\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dG_{\overline{\gamma}i}}$$
$$+\overline{\gamma}U'(\widetilde{s}_{\overline{\gamma}Ii}-L+\widetilde{\tau}_i)\sum_{j\neq i}\frac{d\widetilde{\tau}_i}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}}$$
$$-(1-\overline{\gamma})\left(\overline{\gamma}+\frac{d\widetilde{s}_{\overline{\gamma}Hi}}{dG_{\overline{\gamma}i}}\right)U'(y-c-\overline{\gamma}G_{\overline{\gamma}}+T_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Hi})$$
$$+(1-\overline{\gamma})U'(\widetilde{s}_{\overline{\gamma}Hi}+\widetilde{\tau}_{Hi})\left(1+\frac{d\widetilde{\tau}_i^H}{d\widetilde{s}_{\overline{\gamma}Hi}}\right)\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dG_{\overline{\gamma}i}}$$
$$+(1-\overline{\gamma})U'(\widetilde{s}_{\overline{\gamma}Hi}+\widetilde{\tau}_{Hi})\sum_{j\neq i}\frac{d\widetilde{\tau}_i^H}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}}.$$

The application of the first-order conditions (2.15) and (2.16) (the Envelope Theorem) to the above yields

$$\frac{\partial U_i}{\partial G_{\overline{\gamma}i}} = \overline{\gamma}(1-\overline{\gamma})U'(I1)+\overline{\gamma}U'(I2)\sum_{j\neq i}\frac{d\widetilde{\tau}_i}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}}$$
$$-(1-\overline{\gamma})\overline{\gamma}U'(H1)+(1-\overline{\gamma})U'(H2)\sum_{j\neq i}\frac{d\widetilde{\tau}_i^H}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}},$$

where $I1, I2$ $(H1, H2)$ denote the consumption of an infirm (healthy) individual in periods one and two, respectively. The application of the conditions that second period consumptions are equalized and (2.17) to the above yields

$$\frac{\partial U_i}{\partial G_{\overline{\gamma}i}} = U'(I2)\left(\overline{\gamma}\sum_{j\neq i}\frac{d\widetilde{\tau}_i}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}}+(1-\overline{\gamma})\sum_{j\neq i}\frac{d\widetilde{\tau}_i^H}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}}\right)=0,$$

because $\frac{d\widetilde{s}_j}{dG_{\overline{\gamma}i}}=0$ for $j\neq i$ (see Appendix G for details). This implies that expected utility is independent of $G_{\overline{\gamma}i}$. ■

Given that individuals may not choose to purchase genetic insurance, it is clearly of interest to determine whether they will in fact choose to undergo genetic

testing (and, in the event that they turn out to be infirm, undergo preventive medical treatment). This is answered in the next proposition.

Proposition 5 *When government redistributive policy is time-inconsistent, individuals may not choose to undergo genetic testing although it is socially desirable for them to do so.*

Proof. It is socially desirable for a high-risk individual to undergo genetic testing as long as the reduction in the expected loss due to illness in period two is at least as high as the cost of testing and preventive medical care

$$\gamma_i \alpha L \geqslant c + \gamma_i M, \ \gamma_i = \{\overline{\gamma}, \underline{\gamma}\}.$$

Whereas the social benefit of a single agent undergoing genetic testing and preventive medical treatment is equal to the expected increase in aggregate income available for the government to redistribute in period two (the term on the left) less the expected cost of testing and preventive treatment (the term on the right), when government redistributive policy is time-inconsistent the increase in aggregate income is shared equally among all N individuals. Thus, a high-risk individual undergoes genetic testing if and only if

$$\frac{\gamma_i \alpha L}{N} \geqslant c + \gamma_i M. \tag{2.18}$$

This means that individuals will not undergo genetic testing although it is socially desirable for them to do so if

$$\gamma_i \alpha L \geqslant c + \gamma_i M \geqslant \frac{\gamma_i \alpha L}{N}. \tag{2.19}$$

■

It is straightforward to check that, since $\overline{\gamma} > \underline{\gamma}$, if (2.18) is satisfied for low-risk individuals, then it is also satisfied for high-risk persons. More interestingly, however, the reverse is not true, and therefore it is possible that high-risk individuals choose to undergo testing (and preventive treatment), whereas low-risk individuals do not, although it would in both cases be desirable for them to do so.

Turning now to the optimal first-period tax/transfer policy of the government, notice that the government can only observe that individuals are of low-risk or of high-risk. All other types (infirm, healthy, sick or not sick) are revealed after the first-period transfer is distributed.

Proposition 6 *Expected social welfare is independent of the first-period government transfer.*

Proof. See Appendix H. ■

Proposition 6 is intriguing. Intuitively, it seems plausible to predict that the optimal policy for the government would be to choose a zero transfer policy in period one or, if a positive level of transfer were optimal, to anticipate that this transfer would be chosen strategically to influence the savings decisions of healthy- and infirm-type individuals. It is slightly startling to discover that, even when the government is not able to resist offering disaster assistance at the end of period two, the level of social welfare is in fact independent of the government's first period redistributive policy. In particular, it is consistent with optimal policy to choose the same redistributive transfer policy as would be selected in the full

commitment case, even though infirm individuals may rationally choose to forego health insurance, and will therefore require — and be offered — disaster relief. Even more startling is that even if the government chooses the first-best level of cash transfer, and infirm-type individuals choose to fully insure, the *ex ante* optimum will still not be achieved (see equation (2.17)) .

The equilibrium path deviates from that required to achieve the *ex ante* optimum because the tax burden associated with the government's redistributive program is sensitive to changes in period one savings. Even if the government provides the *ex ante* optimal transfer at the beginning of period one, even if individuals choose to take out genetic insurance, to undertake genetic testing and preventive medical treatment, and even if all infirm individuals choose to fully insure, the equilibrium will still not be *ex ante* efficient because agents' strategic behaviour leads to the distortion of their savings decision. Observe, in particular, that if $\frac{d\tilde{\tau}_H}{ds_H}, \frac{d\tilde{\tau}_I}{ds_I} = 0$, then equations (2.15) and (2.16) show that there will be full consumption-smoothing between the first and second periods, just as is required for the implementation of the *ex ante* efficient outcome. This fact underlies the next proposition.

Proposition 7 *The ex ante efficient outcome can be implemented by a Benthamite government which provides health care to sick individuals and public pension of* $\frac{1}{2}\left(y - c - (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma})\right)$ *to all individuals in period two, and finances the provision of these services as well as of socially desirable genetic testing and preventive measures by imposing an equal per capita tax,*

$T^* = \frac{1}{2}\left(y + c + (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma})\right)$ *in period one.*

Proof. Since the government provides medical care and covers expenses of the genetic test and preventive measures, individuals do not face any expenses related to health care, and losses (L) of sick individuals are covered fully. Given a tax T^*, individual i's consumption in period one is $\frac{1}{2}\left(y - c - (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma})\right) - s_i$, where s_i is private saving. Observe that consumptions in period two can be financed by public pensions and private saving, s_i. One can demonstrate that $s_i = 0$, $\forall i \in [1, N]$. Individual i solves

$$\max_{s_i} U_i = U(y - T^* - s_i) + U(T^* - c - (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma}) + s_i),$$

subject to

$$T^* = \frac{1}{2}\left(y + c + (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma})\right).$$

The first order condition yields at the optimum

$$U'(y - T^* - s_i) = U'(T^* - c - (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma}) + s_i).$$

Thus, $s_i^* = 0$. ■

Similar to Bruce and Waldman (1991) and Coate (1995), the problem of time-consistent government policy is resolved by implementing an appropriate program which undertakes in-kind redistribution (health care and pension). Observe that a public health insurance program alone is not sufficient to obtain the *ex ante* efficient outcome, since individuals are still motivated to overconsume in period one, anticipating a bail-out in period two. Offering public pension alone does not

lead to the ex ante efficient outcome either, since in the economy with multiple health risks individuals will not be able to insure against their type on a private competitive market. Also, they may not undertake genetic testing when it is socially desirable to do so. Thus, a combination of two expenditure programs is necessary. Also, note that the solution in Proposition 7 does not require for the government to use any information about individual's type. This result is interesting since often the inability of the government to observe an individual's type restrains the tax policy from reaching an optimal outcome (Mirrlees 1971, Diamond and Mirrlees 1971, Stiglitz 1982, Boadway *et al.* 1996).

2.4 Public Policy Options and Concluding Remarks

This paper examines *ex ante* efficiency and its importance for health insurance markets. The analysis of section 2.3 shows that in settings in which individuals have information about either their health status, or about the likelihood that they will turn out to have a poor health status, a competitive insurance industry will not be able to offer insurance products (including genetic insurance) which implement the *ex ante* efficient outcome. In the absence of corrective public policy, individuals who are born infirm, or who are born knowing that they are at higher risk of becoming infirm, are born unlucky: through no fault of their own, they will experience lower life-time utility than individuals who are born healthy, or who are born knowing that they have a greater likelihood of being healthy. Only when individuals are born ignorant of their health status, but with an identical risk of becoming healthy or infirm, is it possible for markets to provide consumers

with insurance products that enable them to achieve the same lifetime utility level as they would obtain if they could purchase insurance behind the veil of ignorance.

Is there any way to achieve the *ex ante* efficient outcome in an economy with multiple risk types? As discussed in section 2.3, the first-best outcome cannot be obtained by providing high-risk individuals with compensatory cash transfers, even if the government were able to costlessly verify to which risk class individuals belonged. However, somewhat surprisingly, this is not because high-risk and/or infirm-type individuals fail to fully insure themselves, but because the savings decision is distorted: knowing that a Benthamite-type government will not be able to resist the temptation of providing disaster assistance, both high-risk and low-risk individuals overconsume in period one. Moreover, to the extent that the government cannot prevent itself from implementing a redistributive tax and transfer policy at the end of the second period, the problems that were identified in section 2.3 also mean that the first-best outcome cannot be achieved when genetic testing is costly.

At least one mechanism can solve the problem of the time inconsistency of government redistributive policy when it cannot commit to a once-and-for-all transfer. That mechanism — widely adopted in many industrialized economies —has two parts. The first element is public universal health insurance, i.e., a publicly funded comprehensive health insurance system, that covers the costs of genetic testing and of preventive medical care for individuals who are identified as being at high-risk of becoming infirm, and also covers the loss associated with the disaster state for

infirm-type individuals who subsequently fall ill. The costs of providing universal health care are appropriately covered by imposing an equal per capita tax on all citizens in each period. The second element is a public pension provided in period two and financed by a tax in period one. In effect, the *ex ante* optimal outcome can be attained through a combination of in-kind provision of services to sick individuals as well as a public pension for everyone in period two and a uniform tax on all citizens in period one (when incomes are equal).

Strikingly, the *ex ante* optimal outcome cannot be achieved solely through mandatory insurance. Recall that, along the equilibrium path, individuals may choose to purchase genetic insurance, as well as disaster insurance. The reason that the equilibrium outcome is not the desired outcome is not due to problems with the purchase of insurance, but to inadequate savings: a policy which makes the purchase of genetic or disaster insurance mandatory does not influence the savings decision, and hence does not influence the equilibrium outcome. Mandatory insurance - which covers genetic testing and preventive treatment as well as providing disaster insurance - will prove effective only if coupled with a disincentive on the government part to redistribute wealth in period two. This can be achieved through a publicly provided pension. This paper makes it clear that as in the real world, the decisions with regard to different social programs are interconnected.

Another lesson to be drawn from the results in this paper is that the failure of income tax systems in many countries to differentiate individual tax burdens on the basis of what appear to be observable differences between taxpayers - e.g.,

health status - is, in fact, desirable. Governments can use expenditure policy to redistribute resources between those who are healthy and those who are infirm, but by restricting the public sector's capacity to redistribute income through the tax system, this mitigates the distortion of other decisions - in our case, the savings decision - thus achieving a higher overall level of welfare.

2.5 Appendices

Appendix A

By backward induction, first, solve

$$\max_{Z} \ (\pi-\alpha)U\left(s_I - L + (1-\pi+\alpha)Z\right) \tag{2.20}$$
$$+(1-\pi+\alpha)U\left(s_I-(\pi-\alpha)Z\right),$$

to yield $\widehat{Z} = L$. Second, given $\widehat{Z} = L$, solve

$$\max_{s_I} U\left(y-c-M-s_I+(1-\gamma)G\right) \tag{2.21}$$
$$+(\pi-\alpha)U\left(s_I - L + (1-\pi+\alpha)\widehat{Z}\right) \tag{2.22}$$
$$+(1-\pi+\alpha)U\left(s_I-(\pi-\alpha)\widehat{Z}\right),$$

to yield

$$\widehat{s}_I = \frac{1}{2}\left(y-c-M+(1-\gamma)G+(\pi-\alpha)L\right), \tag{2.23}$$

and

$$\max_{s_H} U\left(y-c-\gamma G-s_H\right)+U(s_H), \tag{2.24}$$

to yield

$$\widehat{s}_H = \frac{1}{2}\left(y-c-\gamma G\right). \tag{2.25}$$

Finally, given (2.23) and (2.25) solve

$$\begin{aligned}\max_{G} EU &= \gamma[U(y-c-M-\widehat{s}_I+(1-\gamma)G) \\ &\quad +U(\widehat{s}_I-(\pi-\alpha)L)] \\ &\quad +(1-\gamma)[U(y-c-\gamma G-\widehat{s}_H)+U(\widehat{s}_H)].\end{aligned}$$

The first order condition at the optimum yields

$$G = M + \widehat{s}_I - \widehat{s}_H,$$

thus,

$$\widehat{G} = M + (\pi-\alpha)L,$$

$$\widehat{s}_I = \frac{y-c-\gamma[M+(\pi-\alpha)L]}{2} + (\pi-\alpha)L,$$

and

$$\widehat{s}_H = \frac{y-c-\gamma[M+(\pi-\alpha)L]}{2}.$$

Note that the lifetime utility of infirm-type and healthy-type individuals are identical at

$$\widehat{U}_I = \widehat{U}_H = 2U\left(\frac{y-c-\gamma[M+(\pi-\alpha)L]}{2}\right).$$

Appendix B

By backward induction, first, solve

$$\begin{aligned}\max_{Z} \ &\pi U(s+G-L+(1-\pi)Z) \\ &+(1-\pi)U(s+G-\pi Z),\end{aligned} \tag{2.26}$$

which yields $\widehat{\widehat{Z}} = L$. Second, given that $\widehat{\widehat{Z}} = L$ solve

$$\max_{s,G} EU = U(y-\gamma G-s) \tag{2.27}$$
$$+\gamma U(s+G-\pi L)+(1-\gamma)U(s).$$

The first order conditions at the optimum are

$$\frac{\partial EU}{\partial G} = -U'(y-\gamma G-s)+U'(s+G-\pi L)=0,$$

$$\frac{\partial EU}{\partial s} = -U'(y-\gamma G-s)+\gamma U'(s+G-\pi L)$$
$$+(1-\gamma)U'(s)=0$$

$$G(1+\gamma) = y-2s+\pi L. \tag{2.28}$$

Since genetic insurance protects against realization of the 'infirm' state it must be true that consumption in the 'healthy' and the 'infirm' states are equalized, i.e.,

$$U'(s+G-\pi L) = U'(s). \tag{2.29}$$

Substituting (2.28) into (2.29) yields

$$\hat{\hat{s}} = \frac{y-\gamma\pi L}{2}. \tag{2.30}$$

Subsequently, given (2.30) equation (2.28) yields

$$\hat{\hat{G}} = \pi L.$$

Thus, the utility of a representative individual who does not undertake genetic testing is

$$\hat{\hat{U}} = 2U\left(\frac{y-\gamma\pi L}{2}\right).$$

Appendix C

Behind the veil of ignorance an individual solves

$$\max_{\substack{s_{\overline{\gamma}I},G_{\overline{\gamma}},G_{\lambda},Z_{\overline{\gamma}} \\ s_{\underline{\gamma}I},G_{\underline{\gamma}},Z_{\underline{\gamma}}}} EU =$$

$$\lambda\left[\begin{array}{c} \overline{\gamma}\left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}I}\right) \\ +(\pi-\alpha)U(s_{\overline{\gamma}I}-L+(1-\pi+\alpha)Z_{\overline{\gamma}}) \\ +(1-\pi+\alpha)U(s_{\overline{\gamma}I}-(\pi-\alpha)Z_{\overline{\gamma}}) \end{array}\right] \\ +(1-\overline{\gamma})\left[U\left(y-c-\overline{\gamma}G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}H}\right)+U(s_{\overline{\gamma}H})\right] \end{array}\right]$$

$$+(1-\lambda)\left[\begin{array}{c} \underline{\gamma}\left[\begin{array}{c} U\left(y-c-M+(1-\underline{\gamma})G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}I}\right)+ \\ (\pi-\alpha)U(s_{\underline{\gamma}I}-L+(1-\pi+\alpha)Z_{\underline{\gamma}})+ \\ (1-\pi+\alpha)U(s_{\underline{\gamma}I}-(\pi-\alpha)Z_{\underline{\gamma}}) \end{array}\right] \\ +(1-\underline{\gamma})\left[U\left(y-c-\underline{\gamma}G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}H}\right)+U(s_{\underline{\gamma}H})\right] \end{array}\right].$$

The first order condition for $Z_{\overline{\gamma}}$ is

$$\begin{aligned} \frac{\partial EU}{\partial Z_{\overline{\gamma}}} &= \lambda\overline{\gamma}[(\pi-\alpha)(1-\pi+\alpha)U'(s_{\overline{\gamma}I}-L+(1-\pi+\alpha)Z_{\overline{\gamma}}) \\ &\quad -(\pi-\alpha)(1-\pi+\alpha)U'(s_{\overline{\gamma}I}-(\pi-\alpha)Z_{\overline{\gamma}})], \end{aligned}$$

so that at the optimum,

$$U'(s_{\overline{\gamma}I}-L+(1-\pi+\alpha)Z_{\overline{\gamma}})=U'(s_{\overline{\gamma}I}-(\pi-\alpha)Z_{\overline{\gamma}}), \tag{2.31}$$

which yields $Z^*_{\overline{\gamma}}=L$. Similarly, one can show that $Z^*_{\underline{\gamma}}=L$.

Using (2.31) and the fact that $Z^*_{\overline{\gamma}}=Z^*_{\underline{\gamma}}=L$, it is obtained that

$$\frac{\partial EU}{\partial s_{\overline{\gamma}I}}=\lambda\overline{\gamma}\left[\begin{array}{c} -U'\left(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}I}\right) \\ +U'(s_{\overline{\gamma}I}-(\pi-\alpha)L) \end{array}\right]. \tag{2.32}$$

Similarly,

$$\frac{\partial EU}{\partial s_{\underline{\gamma}I}} = (1-\lambda)\underline{\gamma}\left[\begin{array}{c} -U'\left(y-c-M+(1-\underline{\gamma})G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}I}\right) \\ +U(s_{\underline{\gamma}I}-(\pi-\alpha)L) \end{array}\right]. \tag{2.33}$$

First order conditions with respect to $s_{\overline{\gamma}H}$, $G_{\overline{\gamma}}$, G_{λ} are

$$\frac{\partial EU}{\partial s_{\overline{\gamma}H}} = \lambda(1-\overline{\gamma})\left[-U'\left(y-c-\overline{\gamma}G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}H}\right)+U'(s_{\overline{\gamma}H})\right], \tag{2.34}$$

$$\frac{\partial EU}{\partial G_{\overline{\gamma}}} = \lambda\left[\begin{array}{c} \overline{\gamma}(1-\overline{\gamma})U'\left(\begin{array}{c} y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}} \\ +(1-\lambda)G_{\lambda}-s_{\overline{\gamma}I} \end{array}\right) \\ -\overline{\gamma}(1-\overline{\gamma})U'\left(y-c-\overline{\gamma}G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}H}\right) \end{array}\right], \tag{2.35}$$

$$\begin{aligned} \frac{\partial EU}{\partial G_{\lambda}} &= \lambda\left[\begin{array}{c} \overline{\gamma}(1-\lambda)U'\left(\begin{array}{c} y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}} \\ +(1-\lambda)G_{\lambda}-s_{\overline{\gamma}I} \end{array}\right) \\ +(1-\overline{\gamma})(1-\lambda)U'\left(\begin{array}{c} y-c-\overline{\gamma}G_{\overline{\gamma}} \\ +(1-\lambda)G_{\lambda}-s_{\overline{\gamma}H} \end{array}\right) \end{array}\right] \\ &\quad +(1-\lambda)\left[\begin{array}{c} -\underline{\gamma}\lambda U'\left(y-c-M+(1-\underline{\gamma})G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}I}\right) \\ -\lambda(1-\underline{\gamma})U'\left(y-c-\underline{\gamma}G_{\underline{\gamma}}-\lambda G_{\lambda}-s_{\underline{\gamma}H}\right). \end{array}\right]. \end{aligned} \tag{2.36}$$

Combining (2.32), (2.34) and (2.35) one can show that

$$\begin{aligned} U'(s_{\overline{\gamma}H}) &= U'\left(y-c-\overline{\gamma}G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}H}\right) \\ &= U'\left(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}}+(1-\lambda)G_{\lambda}-s_{\overline{\gamma}I}\right) \\ &= U'(s_{\overline{\gamma}I}-(\pi-\alpha)L). \end{aligned} \tag{2.37}$$

Similarly, it is straightforward to show that

$$\begin{aligned} U'(s_{\underline{\gamma}H}) &= U'\left(y - c - \underline{\gamma}G_{\underline{\gamma}} - \lambda G_{\lambda} - s_{\underline{\gamma}H}\right) \\ &= U'\left(y - c - M + (1-\underline{\gamma})G_{\underline{\gamma}} - \lambda G_{\lambda} - s_{\underline{\gamma}I}\right) \\ &= U'(s_{\underline{\gamma}I} - (\pi - \alpha)L). \end{aligned} \tag{2.38}$$

Using (2.37) and (2.38), at the optimum (2.36) yields

$$\begin{aligned} &\lambda(1-\lambda)U'\left(y - c - M + (1-\overline{\gamma})G_{\overline{\gamma}} + (1-\lambda)G_{\lambda} - s_{\overline{\gamma}I}\right) \\ &= (1-\lambda)\lambda U'\left(y - c - M + (1-\underline{\gamma})G_{\underline{\gamma}} - \lambda G_{\lambda} - s_{\underline{\gamma}I}\right). \end{aligned} \tag{2.39}$$

Equation (2.39) can be expressed as

$$G_{\lambda} = (1-\underline{\gamma})G_{\underline{\gamma}} - (1-\overline{\gamma})G_{\overline{\gamma}} + s_{\overline{\gamma}I} - s_{\underline{\gamma}I}$$

Further, at the optimum (2.35) yields

$$G_{\overline{\gamma}} = M + s_{\overline{\gamma}I} - s_{\overline{\gamma}H}, \tag{2.40}$$

and it is similarly straightforward to show that

$$G_{\underline{\gamma}} = M + s_{\underline{\gamma}I} - s_{\underline{\gamma}H}.$$

Given (2.40), at the optimum, (2.32) and (2.34) become, respectively,

$$y - c - M + (1-\overline{\gamma})\left(M + s_{\overline{\gamma}I} - s_{\overline{\gamma}H}\right) + (1-\lambda)G_{\lambda} - s_{\overline{\gamma}I} = s_{\overline{\gamma}I} - (\pi - \alpha)L), \tag{2.41}$$

$$y - c - \overline{\gamma}\left(M + s_{\overline{\gamma}I} - s_{\overline{\gamma}H}\right) + (1-\lambda)G_{\lambda} - s_{\overline{\gamma}H} = s_{\overline{\gamma}H}. \tag{2.42}$$

Further, subtracting (2.42) from (2.41) yields

$$s_{\overline{\gamma}I} - s_{\overline{\gamma}H} = (\pi - \alpha)L. \tag{2.43}$$

Using (2.43), (2.40) yields

$$G^*_{\overline{\gamma}} = M + (\pi - \alpha)L. \tag{2.44}$$

Similarly, it can be shown that $s_{\underline{\gamma}I} - s_{\underline{\gamma}H} = (\pi - \alpha)L$, and therefore it is obtained that

$$G^*_{\underline{\gamma}} = M + (\pi - \alpha)L. \tag{2.45}$$

Next, given (2.44) at the optimum, (2.32) yields

$$2s_{\overline{\gamma}I} = y - c - \overline{\gamma}M + (2 - \overline{\gamma})(\pi - \alpha)L + (1 - \lambda)G_\lambda. \tag{2.46}$$

Similarly, given (2.45), at the optimum (2.33) yields

$$2s_{\underline{\gamma}I} = y - c - \underline{\gamma}M + (2 - \underline{\gamma})(\pi - \alpha)L - \lambda G_\lambda. \tag{2.47}$$

Using (2.44), (2.45), (2.46), and (2.47), it can be shown that

$$G^*_\lambda = (\overline{\gamma} - \underline{\gamma})(M + (\pi - \alpha)L). \tag{2.48}$$

Given (2.48), (2.46) and (2.47) it is obtained that

$$s^*_{\overline{\gamma}I} = s^*_{\underline{\gamma}I} = \frac{1}{2}\left(y - c - (\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma}))(M + (\pi - \alpha)L)\right) + (\pi - \alpha)L. \tag{2.49}$$

Further, using (2.43) and (2.49) it can be shown that

$$s^*_{\overline{\gamma}H} = s^*_{\underline{\gamma}H} = \frac{1}{2}\left(y - c - (\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma}))(M + (\pi - \alpha)L)\right). \tag{2.50}$$

Lastly, given (2.37) and (2.38), and using (2.49) and (2.50) it can be demonstrated that

$$\begin{aligned} U^*_{\underline{\gamma}H} &= U^*_{\underline{\gamma}I} = U^*_{\overline{\gamma}H} = U^*_{\overline{\gamma}I} \\ &= 2U\left(\frac{1}{2}\left(y - c - (M + (\pi - \alpha)L)(\underline{\gamma} + \lambda(\overline{\gamma} - \underline{\gamma}))\right)\right). \end{aligned}$$

Appendix D: Proof of Proposition 1

At the beginning of period one the only realized and observable difference among individuals is their risk-type (high or low). Thus, the government redistributes from low-risk to high risk individuals by choosing lump-sum taxes/transfers $T_{\overline{\gamma}}$, $T_{\underline{\gamma}}$, respectively, by solving

$$\max_{T_{\overline{\gamma}},T_{\underline{\gamma}}} W = \sum_{i=1}^{\lambda N} \left[\begin{array}{c} \overline{\gamma} \left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})G^*_{\overline{\gamma}i}-s^*_{\overline{\gamma}Ii}+T_{\overline{\gamma}}\right)+ \\ (\pi-\alpha)U(s^*_{\overline{\gamma}Ii}-L+(1-\pi+\alpha)Z^*_{\overline{\gamma}i})+ \\ (1-\pi+\alpha)U(s^*_{\overline{\gamma}Ii}-(\pi-\alpha)Z^*_{\overline{\gamma}i}) \end{array} \right] \\ +(1-\overline{\gamma})\left[U\left(y-c-\overline{\gamma}G^*_{\overline{\gamma}i}-s^*_{\overline{\gamma}Hi}+T_{\overline{\gamma}}\right)+U(s^*_{\overline{\gamma}Hi})\right] \end{array} \right]$$

$$+\sum_{i=\lambda N+1}^{N} \left[\begin{array}{c} \underline{\gamma} \left[\begin{array}{c} U\left(y-c-M+(1-\underline{\gamma})G^*_{\underline{\gamma}i}-s^*_{\underline{\gamma}Ii}+T_{\underline{\gamma}}\right)+ \\ (\pi-\alpha)U(s^*_{\underline{\gamma}Ii}-L+(1-\pi+\alpha)Z^*_{\underline{\gamma}i})+ \\ (1-\pi+\alpha)U(s^*_{\underline{\gamma}Ii}-(\pi-\alpha)Z^*_{\underline{\gamma}i}) \end{array} \right] \\ +(1-\underline{\gamma})\left[U\left(y-c-\underline{\gamma}G^*_{\underline{\gamma}i}-s^*_{\underline{\gamma}Hi}+T_{\underline{\gamma}}\right)+U(s^*_{\underline{\gamma}Hi})\right] \end{array} \right],$$

subject to the budget constraint

$$\lambda T_{\overline{\gamma}} + (1-\lambda)T_{\underline{\gamma}} = 0, \tag{2.51}$$

and subject to $s^*_{\overline{\gamma}Ii}, G^*_{\overline{\gamma}i}, Z^*_{\overline{\gamma}i}$ (or $s^*_{\underline{\gamma}Ii}, G^*_{\underline{\gamma}i}, Z^*_{\underline{\gamma}i}$) being the solution to the optimisation problems of a representative high-risk (or a low-risk) individual, that is,

$$\max_{s_{\overline{\gamma}I},G_{\overline{\gamma}},Z_{\overline{\gamma}}} EU = \overline{\gamma} \left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})G_{\overline{\gamma}}-s_{\overline{\gamma}I}+T_{\overline{\gamma}}\right)+ \\ (\pi-\alpha)U(s_{\overline{\gamma}I}-L+(1-\pi+\alpha)Z_{\overline{\gamma}})+ \\ (1-\pi+\alpha)U(s_{\overline{\gamma}I}-(\pi-\alpha)Z_{\overline{\gamma}}) \end{array} \right]$$

$$+(1-\overline{\gamma})\left[U\left(y-c-\overline{\gamma}G_{\overline{\gamma}}-s_{\overline{\gamma}H}+T_{\overline{\gamma}}\right)+U(s_{\overline{\gamma}H})\right].$$

It is easy to demonstrate that if the government is committed to a once-and-for-all transfer policy, then infirm-type individuals always purchase full health insurance $Z^*_{\overline{\gamma}} = Z^*_{\underline{\gamma}} = L$; and both high-risk and low-risk individuals purchase full genetic insurance coverage $G^*_{\overline{\gamma}} = G^*_{\underline{\gamma}} = M + (\pi - \alpha)L$, which protects them if they were to become an infirm-type, requiring treatment at cost M and health insurance (albeit at a lower premium than if they were not to undertake the treatment).

Consequently, high–risk individuals have a life-time utility

$$2U\left(\frac{1}{2}(y - c - \overline{\gamma}(M + (\pi - \alpha)L) + T_{\overline{\gamma}}\right),$$

and low-risk individuals have life-time utility

$$2U\left(\frac{1}{2}(y - c - \underline{\gamma}(M + (\pi - \alpha)L) + T_{\underline{\gamma}}\right).$$

This implies that the government's optimisation problem can be re-written as

$$\max_{T_{\overline{\gamma}}, T_{\underline{\gamma}}} W = \sum_{i=1}^{\lambda N} 2U\left(\frac{1}{2}(y - c - \overline{\gamma}(M + (\pi - \alpha)L) + T_{\overline{\gamma}}\right) + \sum_{i=\lambda N+1}^{N} 2U\left(\frac{1}{2}(y - c - \underline{\gamma}(M + (\pi - \alpha)L) + T_{\underline{\gamma}})\right),$$

subject to (2.51).

The first order conditions are

$$\sum_{i=1}^{\lambda N} 2U'\left(\frac{1}{2}(y - c - \overline{\gamma}(M + (\pi - \alpha)L) + T_{\overline{\gamma}}\right) + \rho\lambda = 0, \forall i \in [1, ..., \lambda N],$$

$$\sum_{i=\lambda N+1}^{N} 2U'\left(\frac{1}{2}(y - c - \underline{\gamma}(M + (\pi - \alpha)L) + T_{\underline{\gamma}})\right) + \rho(1-\lambda) = 0, \forall i \in [\lambda N+1, ..., N],$$

and (2.51), where ρ is the shadow price of the government budget constraint, and N is the number of individuals in the economy.

From the above first-order conditions it follows that government chooses a tax/transfer policy to equalize the utilities across the population with the optimal choice of transfer $T^*_{\overline{\gamma}} = (1-\lambda)(\overline{\gamma}-\underline{\gamma})(M+(\pi-\alpha)L)$ and tax $T^*_{\underline{\gamma}} = -\lambda(\overline{\gamma}-\underline{\gamma})(M+(\pi-\alpha)L)$. Thus, as described in Section 2.2, this implements the *ex ante* efficient allocation.

Appendix E

Fully differentiating (2.14) and (2.13) with respect to Z_i and $\tau_i, \forall i \in [1, N]$ yields

$$\begin{pmatrix} U'' & 0 & \dots & 0 & 0 & 1 \\ 0 & U'' & 0 & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & U'' & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} d\widetilde{\tau}_1 \\ d\widetilde{\tau}_2 \\ \dots \\ d\widetilde{\tau}_N \\ d\mu \end{pmatrix}$$

$$= \begin{pmatrix} K_i U'' & 0 & \dots & 0 & 0 \\ 0 & K_i U'' & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & K_i U'' \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} dZ_1 \\ dZ_2 \\ \dots \\ dZ_{N\left(\lambda\overline{\gamma}+(1-\lambda)\underline{\gamma}\right)} \end{pmatrix},$$

where $K_i = -1+(\pi-\alpha)\ \forall i \in [1, \lambda\overline{\gamma}(\pi-\alpha)N]$ or $\forall i \in [\lambda N+1, \left(\lambda+(1-\lambda)\underline{\gamma}\right)(\pi-\alpha)N]$ and $K_i = (\pi-\alpha)\ \forall i \in [\lambda\overline{\gamma}(\pi-\alpha)N+1, \lambda\overline{\gamma}N]\ \forall i \in [\left(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha)\right)N]$. Note that because second period utilities are equalized the index i in U_J can be suppressed.

Using Cramer's rule, observe that for infirm individuals who fall sick, i.e., $\forall i \in [1, \lambda\overline{\gamma}(\pi-\alpha)N]$ or $\forall i \in [\lambda N+1, \left(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha)\right)N]$,

$$\frac{d\widetilde{\tau}_i}{dZ_i} = \frac{1-(\pi-\alpha)(N-1)(U'')^{N-1}}{-N(U'')^{N-1}} = -\frac{N-1}{N}(1-(\pi-\alpha));$$

and that for infirm individuals who do not fall sick, i.e., $\forall i \in [\lambda\overline{\gamma}(\pi-\alpha)N+1, \lambda\overline{\gamma}N]$ or $\forall i \in [\left(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha)\right)N+1, \left(\lambda+(1-\lambda)\underline{\gamma}\right)N]$,

$$\frac{d\widetilde{\tau}_i}{dZ_i} = \frac{-(\pi-\alpha)(N-1)(U'')^{N-1}}{-N(U'')^{N-1}} = \frac{N-1}{N}(\pi-\alpha).$$

Appendix F

Fully differentiating (2.14) and (2.13) with respect to s_i and $\tau_i, \forall i \in [1, N]$ yields

$$\begin{pmatrix} U'' & 0 & \dots & 0 & 0 & 1 \\ 0 & U'' & 0 & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & U'' & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} d\widetilde{\tau}_1 \\ d\widetilde{\tau}_2 \\ \dots \\ d\widetilde{\tau}_N \\ d\mu \end{pmatrix}$$

$$= \begin{pmatrix} -U'' & 0 & \dots & 0 & 0 \\ 0 & -U'' & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & -U'' \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} ds_1 \\ ds_2 \\ \dots \\ ds_N \end{pmatrix}.$$

Note that because second period utilities are equalized the index i in U'' can be suppressed. Using Cramer's rule note that

$$\frac{d\widetilde{\tau}_i}{ds_i} = \frac{(N-1)\,(U'')^{N-1}}{-N(U'')^{N-1}} = -\frac{(N-1)}{N}.$$

Similarly,

$$\frac{d\widetilde{\tau}_i}{ds_j} = \frac{-\,(U'')^{N-1}}{-N(U'')^{N-1}} = \frac{1}{N},\ \forall i \neq j.$$

Appendix G

Fully differentiating (2.15) and (2.16) with respect to s_i and G_i yields

$$\begin{pmatrix} U''_{11}+\frac{1}{N}U''_{21} & 0 & \ldots & 0 & 0 \\ 0 & U''_{12}+\frac{1}{N}U''_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & U''_{1N}+\frac{1}{N}U''_{2N} \end{pmatrix} \begin{pmatrix} d\widetilde{s}_1 \\ d\widetilde{s}_2 \\ \ldots \\ d\widetilde{s}_N \end{pmatrix}$$

$$= \begin{pmatrix} (1-\gamma_i)U''_{11} & 0 & \ldots & 0 & 0 \\ 0 & (1-\gamma_i)U''_{12} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & (1-\gamma_i)U''_{1N} \end{pmatrix} \begin{pmatrix} dG_1 \\ dG_2 \\ \ldots \\ dG_N \end{pmatrix},$$

where $\gamma_i \in \{\overline{\gamma}, \underline{\gamma}\}$. Applying Cramer's rule, one obtains that $\frac{d\widetilde{s}_i}{dG_j} = 0$.

Appendix H: Proof of Proposition 6

The optimal redistributive policy of the government in period one is found by

solving

$$\max_{T_{\overline{\gamma}},T_{\underline{\gamma}}} W = \sum_{i=1}^{\lambda\overline{\gamma}(\pi-\alpha)N} \left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})\widetilde{G}_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Ii}+T_{\overline{\gamma}}\right) \\ +U\left(\widetilde{s}_{\overline{\gamma}Ii}-L+(1-\pi+\alpha)\widetilde{Z}_{\overline{\gamma}i}+\widetilde{\tau}_{\overline{\gamma}i}\right) \end{array} \right]$$
$$+ \sum_{i=\lambda\overline{\gamma}(\pi-\alpha)N+1}^{\lambda\overline{\gamma}N} \left[\begin{array}{c} U\left(y-c-M+(1-\overline{\gamma})\widetilde{G}_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Ii}+T_{\overline{\gamma}}\right) \\ +U\left(\widetilde{s}_{\overline{\gamma}Ii}-(\pi-\alpha)\widetilde{Z}_i+\widetilde{\tau}^I_{\overline{\gamma}i}\right) \end{array} \right]$$
$$+ \sum_{i=\lambda\overline{\gamma}N+1}^{\lambda N} \left[\begin{array}{c} U\left(y-c-\overline{\gamma}\widetilde{G}_{\overline{\gamma}i}-\widetilde{s}_{\overline{\gamma}Hi}+T_{\overline{\gamma}}\right) \\ +U\left(\widetilde{s}_{\overline{\gamma}Hi}+\tau^H_{\overline{\gamma}i}\right) \end{array} \right]$$
$$+ \sum_{i=\lambda N+1}^{\lambda N+(1-\lambda)\underline{\gamma}(\pi-\alpha)N} \left[\begin{array}{c} U\left(y-c-M+(1-\underline{\gamma})\widetilde{G}_{\underline{\gamma}i}-\widetilde{s}_{\underline{\gamma}Ii}+T_{\underline{\gamma}}\right) \\ +U\left(\widetilde{s}_{\underline{\gamma}Ii}-L+(1-\pi+\alpha)\widetilde{Z}_{\underline{\gamma}i}+\widetilde{\tau}_{\underline{\gamma}i}\right) \end{array} \right]$$
$$+ \sum_{i=(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha))N+1}^{(\lambda+(1-\lambda)\underline{\gamma})N} \left[\begin{array}{c} U\left(y-c-M+(1-\underline{\gamma})\widetilde{G}_{\underline{\gamma}i}-\widetilde{s}_{\underline{\gamma}Ii}+T_{\underline{\gamma}}\right) \\ +U\left(\widetilde{s}_{\underline{\gamma}Ii}-(\pi-\alpha)\widetilde{Z}_{\underline{\gamma}i}+\widetilde{\tau}^I_{\underline{\gamma}i}\right) \end{array} \right]$$
$$+ \sum_{i=(\lambda+(1-\lambda)\underline{\gamma})N+1}^{N} \left[\begin{array}{c} U\left(y-c+(1-\underline{\gamma})\widetilde{G}_{\underline{\gamma}i}-\widetilde{s}_{\underline{\gamma}Hi}+T_{\underline{\gamma}}\right) \\ +U\left(\widetilde{s}_{\underline{\gamma}Hi}+\widetilde{\tau}^H_{\underline{\gamma}i}\right) \end{array} \right],$$

subject to

$$\lambda T_{\overline{\gamma}} + (1-\lambda)T_{\underline{\gamma}} = 0,$$

where $T_{\overline{\gamma}}$ and $T_{\underline{\gamma}}$ are the first-period transfers (taxes) to a low-risk individual and high-risk individual, respectively.

Using $T_{\underline{\gamma}} = -\frac{\lambda}{1-\lambda}T_{\overline{\gamma}}$ from the budget constraint and applying the results derived above, i.e., $\widetilde{G}_{\overline{\gamma}i} = \widetilde{G}_{\underline{\gamma}i} = 0$ and $\widetilde{Z}_i = 0$, $\forall\ i \in [1, N]$, the first order condition

with respect to $T_{\underline{\gamma}}$ becomes

$$\frac{dW}{dT_{\overline{\gamma}}} = \sum_{i=1}^{\lambda\overline{\gamma}(\pi-\alpha)N} \left[\begin{array}{c} U'\left(y-c-M-\widetilde{s}_{\overline{\gamma}Ii}+T_{\overline{\gamma}}\right)\left(-\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dT_{\overline{\gamma}}}+1\right) \\ +U'\left(\widetilde{s}_{\overline{\gamma}Ii}-L+\widetilde{\tau}_{\overline{\gamma}i}\right)\left(1+\frac{d\widetilde{\tau}_{\overline{\gamma}i}}{d\widetilde{s}_{\overline{\gamma}Ii}}\right)\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dT_{\overline{\gamma}}} \\ +U'\left(\widetilde{s}_{\overline{\gamma}Ii}-L+\widetilde{\tau}_{\overline{\gamma}i}\right)\sum_{j=1,j\neq i}^{N}\frac{d\widetilde{\tau}_i}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dT_{\overline{\gamma}}} \end{array} \right]$$

$$+ \sum_{i=\lambda\overline{\gamma}(\pi-\alpha)N+1}^{\lambda\overline{\gamma}N} \left[\begin{array}{c} U'\left(y-c-M-\widetilde{s}_{\overline{\gamma}Ii}+T_{\overline{\gamma}}\right)\left(-\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dT_{\overline{\gamma}}}+1\right) \\ +U'\left(\widetilde{s}_{\overline{\gamma}Ii}+\widetilde{\tau}^{I}_{\overline{\gamma}i}\right)\left(1+\frac{d\widetilde{\tau}^{I}_{\overline{\gamma}i}}{d\widetilde{s}_{\overline{\gamma}Ii}}\right)\frac{d\widetilde{s}_{\overline{\gamma}Ii}}{dT_{\overline{\gamma}}} \\ +U'\left(\widetilde{s}_{\overline{\gamma}Ii}+\widetilde{\tau}^{I}_{\overline{\gamma}i}\right)\sum_{j=1,j\neq i}^{N}\frac{d\widetilde{\tau}^{I}_{\overline{\gamma}i}}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dT_{\overline{\gamma}}} \end{array} \right]$$

$$+ \sum_{i=\lambda\overline{\gamma}N+1}^{\lambda N} \left[\begin{array}{c} U'\left(y-c-\widetilde{s}_{\overline{\gamma}Hi}+T_{\overline{\gamma}}\right)\left(-\frac{d\widetilde{s}_{\overline{\gamma}Hi}}{dT_{\overline{\gamma}}}+1\right) \\ +U'\left(\widetilde{s}_{\overline{\gamma}Hi}+\tau^{H}_{\overline{\gamma}i}\right)\left(1+\frac{d\widetilde{\tau}^{H}_{\overline{\gamma}i}}{d\widetilde{s}_{\overline{\gamma}Hi}}\right)\frac{d\widetilde{s}_{\overline{\gamma}Hi}}{dT_{\overline{\gamma}}} \\ +U'\left(\widetilde{s}_{\overline{\gamma}Hi}+\tau^{H}_{\overline{\gamma}i}\right)\sum_{j=1,j\neq i}^{N}\frac{d\widetilde{\tau}^{H}_{\overline{\gamma}i}}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dT_{\overline{\gamma}}} \end{array} \right]$$

$$+ \sum_{i=\lambda N+1}^{\lambda N+(1-\lambda)\underline{\gamma}(\pi-\alpha)N} \left[\begin{array}{c} -\frac{\lambda}{1-\lambda}U'\left(y-c-M-\widetilde{s}_{\underline{\gamma}Ii}-\frac{\lambda}{1-\lambda}T_{\overline{\gamma}}\right)\left(-\frac{d\widetilde{s}_{\underline{\gamma}Ii}}{dT_{\underline{\gamma}}}+1\right) \\ -\frac{\lambda}{1-\lambda}U'\left(\widetilde{s}_{\underline{\gamma}Ii}-L+\widetilde{\tau}_{\underline{\gamma}i}\right)\left(1+\frac{d\widetilde{\tau}_{\underline{\gamma}i}}{\widetilde{s}_{\underline{\gamma}Ii}}\right)\frac{d\widetilde{s}_{\underline{\gamma}Ii}}{dT_{\underline{\gamma}}} \\ -\frac{\lambda}{1-\lambda}U'\left(\widetilde{s}_{\underline{\gamma}Ii}-L+\widetilde{\tau}_{\underline{\gamma}i}\right)\sum_{j=1,j\neq i}^{N}\frac{d\widetilde{\tau}_{\underline{\gamma}i}}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dT_{\overline{\gamma}}} \end{array} \right]$$

$$+ \sum_{i=(\lambda+(1-\lambda)\underline{\gamma}(\pi-\alpha))N+1}^{(\lambda+(1-\lambda)\underline{\gamma})N} \left[\begin{array}{c} -\frac{\lambda}{1-\lambda}U'\left(y-c-M-\widetilde{s}_{\underline{\gamma}Ii}-\frac{\lambda}{1-\lambda}T_{\overline{\gamma}}\right)\left(-\frac{d\widetilde{s}_{\underline{\gamma}Ii}}{dT_{\underline{\gamma}}}+1\right) \\ -\frac{\lambda}{1-\lambda}U'\left(\widetilde{s}_{\underline{\gamma}Ii}+\widetilde{\tau}^{I}_{\underline{\gamma}i}\right)\left(1+\frac{d\widetilde{\tau}^{I}_{\underline{\gamma}i}}{\widetilde{s}_{\underline{\gamma}Ii}}\right)\frac{d\widetilde{s}_{\underline{\gamma}Ii}}{dT_{\underline{\gamma}}} \\ -\frac{\lambda}{1-\lambda}U'\left(\widetilde{s}_{\underline{\gamma}Ii}+\widetilde{\tau}^{I}_{\underline{\gamma}i}\right)\sum_{j=1,j\neq i}^{N}\frac{d\widetilde{\tau}^{I}_{\underline{\gamma}i}}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dT_{\overline{\gamma}}} \end{array} \right]$$

$$+ \sum_{i=(\lambda+(1-\lambda)\underline{\gamma})N+1}^{N} \left[\begin{array}{c} -\frac{\lambda}{1-\lambda}U'\left(y-c-\widetilde{s}_{\underline{\gamma}Hi}-\frac{\lambda}{1-\lambda}T_{\overline{\gamma}}\right)\left(-\frac{d\widetilde{s}_{\underline{\gamma}Hi}}{dT_{\underline{\gamma}}}+1\right) \\ -\frac{\lambda}{1-\lambda}U'\left(\widetilde{s}_{\underline{\gamma}Hi}+\widetilde{\tau}^{H}_{\underline{\gamma}i}\right)\left(1+\frac{d\widetilde{\tau}^{H}_{\underline{\gamma}i}}{\widetilde{s}_{\underline{\gamma}Hi}}\right)\frac{d\widetilde{s}_{\underline{\gamma}Hi}}{dT_{\underline{\gamma}}} \\ -\frac{\lambda}{1-\lambda}U'\left(\widetilde{s}_{\underline{\gamma}Hi}+\widetilde{\tau}^{H}_{\underline{\gamma}i}\right)\sum_{j=1,j\neq i}^{N}\frac{d\widetilde{\tau}^{H}_{\underline{\gamma}i}}{d\widetilde{s}_j}\frac{d\widetilde{s}_j}{dT_{\overline{\gamma}}} \end{array} \right].$$

Using the result that second period consumptions are equalized across indi-

viduals, that first period consumptions are equalized across the population, that $1+\frac{d\tilde{\tau}_{\overline{\gamma}i}}{d\tilde{s}_{\overline{\gamma}Ii}}=1+\frac{d\tilde{\tau}^{I}_{\overline{\gamma}i}}{d\tilde{s}_{\overline{\gamma}Ii}}=1+\frac{d\tau^{H}_{\overline{\gamma}i}}{d\tilde{s}_{\overline{\gamma}Hi}}=\frac{1}{N}\ \forall i$,that $\frac{d\tilde{\tau}_{i}}{d\tilde{s}_{j}}=\frac{1}{N}\ \forall i\neq j,\ i\in[1,N],\ j\in[1,N]$, and that $\frac{d\tilde{s}_{i}}{dT_{\overline{\gamma}}}=\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}\ \forall i\in[1,N]$ (where (1) and (2) denote consumptions in period one and two respectively), the first order condition can be rewritten as

$$\begin{aligned}
\frac{dW}{dT_{\overline{\gamma}}} &= \sum_{i=1}^{\lambda N}\left[\begin{array}{c} U'(1)\left(-\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}+1\right) \\ +U'(2)\frac{1}{N}\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)} \\ +U'(2)\frac{1}{N}\sum_{j=1,j\neq i}^{N}\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}\end{array}\right] \\
&\quad+\sum_{i=\lambda N+1}^{N}\left[\begin{array}{c} -\frac{\lambda}{1-\lambda}\left(-\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}+1\right)U'(1) \\ -\left(\frac{\lambda}{1-\lambda}\right)U'(2)\frac{1}{N}\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)} \\ -\frac{\lambda}{1-\lambda}U'(2)\frac{1}{N}\sum_{j=1,j\neq i}^{N}\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}\end{array}\right] \\
&= \lambda N\left[\begin{array}{c}\left(-\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}+1\right)U'(1) \\ +\frac{1}{N}\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}U'(2) \\ +U'(2)\frac{1}{N}(N-1)\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}\end{array}\right] \\
&\quad+(1-\lambda)N\left[\begin{array}{c} -\frac{\lambda}{1-\lambda}\left(-\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}+1\right)U'(1) \\ -\left(\frac{\lambda}{1-\lambda}\right)\frac{1}{N}\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}U'(2) \\ -\frac{\lambda}{1-\lambda}U'(2)\frac{1}{N}(N-1)\frac{U''(1)}{U''(1)+\frac{1}{N^2}U''(2)}\end{array}\right] \\
&= 0.
\end{aligned}$$

Similarly it can be shown that $\frac{dW}{dT_{\underline{\gamma}}}=0\ \forall T_{\underline{\gamma}}$. Thus, social welfare is independent of the first period redistributive policy.

REFERENCES

Boadway, R. (1997) Public economics and the theory of public policy, *Canadian Journal of Economics,* Vol. 30 pp. 753-772.

Boadway, R., N. Marceau, and M. Marchand (1996) Investment in education and the time inconsistency of redistributive tax policy, *Economica,* Vol.63(250) pp. 171-189.

Bruce, W., and M. Waldman (1991) Transfers in kind: how they can be efficient and non-paternalistic, *American Economic Review,* Vol. 81 pp. 1345-1351.

Coate, S. (1995) Altruism, the Samaritan's Dilemma, and the government transfer policy, *American Economic Review,* Vol. 85(1) pp. 46-57.

Cutler, D. M. (2002) Health care and the public sector, National Bureau of Economic Research, *NBER working paper 8802.*

Cutler, D. M., and R. J. Zeckhauser (2000) "The anatomy of health insurance" In *Handbook of Health Economics* by Culyer A. J., and J. P. Newhouse, Eds., North-Holland: Amsterdam, Vol. 1 pp. 564-629.

Diamond, P. A., and J. A. Mirrlees (1971) Optimal taxation and public production 1: Production efficiency and 2: Tax rules, *American Economic Review,* Vol.61 pp. 8-27 and 261-278.

Doherty, N. A., and P. D. Thistle (1996) Adverse selection with endogenous information in insurance markets, *Journal of Public Economics* Vol.63 pp. 83-102.

Hanley, W. B. (2005) Newborn screening in Canada: Are we out of step? *Paediatrics and Child Health,* Vol.10(4) pp. 203-207.

Hoel, M., and T. Iversen (2002) Genetic testing when there is a mix of compulsory and voluntary health insurance, *Journal of Health Economics,* Vol. 21 pp. 253-270.

Polborn, M. K., M. Hoy, and A. Sadanand (2006) Advantageous effects of regulatory adverse selection in the life insurance market, *The Economic Journal,* Vol.116 pp. 327-354.

Miller, F., J. Hurley, S. Morgan, R. Goeree, P. Collins, G. Blackhouse, M. Giacomini, and B. O'Brien (2002) Predictive genetic tests and health care costs: final report prepared for the Ontario Ministry of Health and Long Term Care, available at www.health.gov.on.ca/english/public/pub/ministry_reports/geneticsrep02/chepa_rep.pdf

Mirrlees, J. A. (1971) An exploration in the theory of optimum income taxation, *Review of Economic Studies,* Vol.38 pp. 175-208.

Stiglitz, J. E. (1982) Self-selection and Pareto efficient taxation, *Journal of Public Economics*, Vol.17 pp. 213-240.

Tabarrok, A. (1994) Genetic testing: an economic and contractarian analysis, *Journal of Health Economics,* Vol.13 pp. 75-91.

CHAPTER 3

Fee-For-Service vs. Capitation: Anything You Can Do - I Can Do Better (and Cheaper)

3.1 Introduction

A classic concern of the literature on health economics is the design of the payment scheme for physicians. Should they be paid on a fee-for-service basis – as has traditionally been the norm – retrospectively reimbursing physicians for service provided according to a pre-determined fee schedule? Or is a prospective payment scheme – such as those increasingly used by both private health maintenance organizations and public health insurers — which pays physicians a fixed fee per patient on the physician's roster on the understanding that the physician must provide all necessary health care services to the rostered patients, a more cost-effective approach to securing medical services?

Many researchers have compared these two payment schemes (see Léger 2008 and McGuire 2000 for extensive overviews). The problem is generally viewed as an application of the principles of mechanism design, where the principal (the health maintenance organization, or public insurer) designs a compensation scheme to elicit the desired behaviour by the agent (the physician), and where it is assumed that the agent's actions are only imperfectly observable. In a nutshell, it is argued that fee-for-service does not provide incentives for cost-minimization in the

selection of medical services and may even induce demand for unnecessary services (Evans 1974, Pauly 1980, McGuire 2000). In contrast, capitation provides a profit-maximizing physician with powerful incentives to conserve service usage. It is also argued that capitation may encourage the provision of preventive services if they reduce future costs of service provision, although there is no clear evidence to support this theoretical proposition (Shimmura 1988).

Of course, the actual cost of servicing a patient is uncertain at the time of receiving the capitation payment. And since, on average, healthy patients make fewer visits to their physician than patients with complicated health conditions, this may induce physicians to roster healthy patients rather than patients with complicated health conditions. This type of selection is known as cream-skimming (Pauly 1984, Ellis 1998). Indeed, opponents of capitation argue that the prospective nature of capitation leads to quality stinting in general and, in the extreme, to the denial of services to high-cost patients (Pauly 1984, Allen and Gertler 1991, Newhouse 1996, Ellis 1998). Capitation may also encourage excessive referrals, thereby shifting medical costs to secondary care providers, i.e., specialists (Shimmura 1988, Stearns, Wolfe and Kindig 1992). Broadly speaking, the consensus emerging from this literature is that it is desirable to create a blend of the prospective and retrospective schemes in order to mitigate problems associated with each pure system (Ellis and McGuire 1986, 1990, Ma 1994, Chalkley and Malcomson 1998, Eggleston 2000, 2005).

This paper attacks the problem of the design of physician payment schemes

from a different perspective. In particular, rather than assuming that physicians and patients are essentially homogeneous, a key premise of this analysis is that some physicians enjoy more than others the challenge of solving complex health problems and are perhaps better at doing so, and that some patients have complex medical needs whereas other patients require only straightforward care. A key role of the payment scheme is therefore to help match physicians with patients, and ensure that appropriate care is provided. To this end, the paper views the problem of physician remuneration as nothing more - nor less - than a problem in pricing theory, and the approach is very much in the spirit of discussions of pricing in club economies (Wooders 1978, Scotchmer 1994, Barham and Wooders 1998). As in the literature on club economies, the model abstracts from the challenges created by asymmetric information, in order to focus on the role of the price mechanism: the notions of fee-for-service and capitation equilibria developed here are closely analogous to the concepts of competitive equilibrium and admission-price equilibrium studied in these papers.

An important benefit of the approach taken here is that it allows a careful assessment of the relative strengths and weaknesses of fee-for-service versus capitation in an environment of heterogenous agents when the demand for medical services is certain, as well as under uncertainty. Perhaps not surprisingly to those familiar with the club theory literature, the analysis suggests that there may be fewer intrinsic differences between the outcomes that can be obtained under fee-for-service or capitation than is commonly perceived: if policy-makers were to take

a more sophisticated approach to the design of the pricing scheme under either fee-for-service or capitation, then this may reduce or indeed entirely eliminate the differences in the economic outcomes obtained under these payment mechanisms. The theoretical analysis of the paper also makes it possible to examine how differences in the tolerance of individual physicians and of the public health insurer to risk affect the overall cost of obtaining health care services under retrospective and prospective payment schemes. In particular, the model suggests that it is reasonable to expect that an appropriately-designed fee-for-service system may, under some circumstances, be able to deliver desired health care outcomes at lower cost than can a capitation-based system.

The next section of this paper describes the model and Pareto efficient allocations in the economy. Section 3.3 studies fee-for-service and capitation equilibria and establishes the equivalence between these two types of equilibria in an environment in which the demand for medical services is exogenously determined. Section 3.4 analyzes fee-for-service and capitation equilibria when there is uncertainty with respect to the health care needs of any individual patient. Section 3.5 extends the analysis to allow for preventive care. Section 3.6 concludes.

3.2 The Model

3.2.1 Agents, practices and states of the economy

Consider an economy populated by medical service providers, i.e., physicians, and users of their services, i.e., patients. There are two goods in the economy: medical services and yogurt. Each patient has an equal endowment of y units of yogurt and

no medical services. Each physician's endowment of yogurt is zero. Physicians, however, can provide medical services to patients in exchange for yogurt. In this and the next section it is assumed that all patients consume one (and only one) medical assessment.

Patients differ with respect to the complexity of their medical problems. There are H patients in the economy, indexed by h. There are h_1 type-1 patients with complex problems (further - complex patients) and h_2 type-2 patients with simple problems, e.g., a sore throat or an ear infection (further - simple patients). $h_1 + h_2 = H$, and patients are ordered such that if $h \leq h_1$ then the patient is a type-1 patient, whereas if $h > h_1$ then the patient is a type-2 patient. Patient h's utility function is as follows

$$U_h = u(M_h) + \delta_{hv}$$

where $u(M_h)$ is continuous, monotonically increasing, quasi-concave, and measures the utility of the patient from consumption of yogurt, M_h. δ_{hv} denotes the utility of patient h from medical services of type v, $v = l$ if a complex assessment is provided, and $v = s$ if the patient receives a straightforward assessment; δ_{hv} is a discrete function, taking on different values depending on h and v, and equals zero if the patient does not visit a physician.

A complex patient benefits more from a long visit than from a short one. A simple patient's benefit from a short visit is identical to her benefit from a long

visit. Formally,

$$\delta_{hl} > \delta_{hs}, \text{ for } h \leq h_1 \text{ and} \tag{3.1}$$

$$\delta_{hl} = \delta_{hs} \text{ for } h > h_1. \tag{3.2}$$

There are K physicians in the economy, indexed by k. Physicians differ with respect to their opportunity cost of performing complex assessments. This paper refers to physicians who have a low disutility of performing complex procedures as high-productivity or type-1 physicians, and to those with high disutility of dealing with complex patients as low-productivity or type-2. There are k_1 high-productivity physicians and k_2 low-productivity physicians: $k_1 + k_2 = K$, and physicians are ordered such that if $k \leq k_1$ then the physician is a high-productivity physician, whereas if $k > k_1$ then the physician is a low-productivity physician. The utility function of physician k of type j $(j = 1, 2)$ is

$$U^k = V(M^k) - C(\gamma_j l^k + \rho s^k)$$

where $V(M^k)$ denotes the utility from income, M^k, which the physician derives from the yogurt received for providing straightforward and complex assessments; V is continuous, monotonically increasing and concave. $C(I)$ denotes the cost of service provision, where $I = \gamma_j l + \rho s$ is the caseload (or the number of assessments in productivity units)[1]. l^k denotes the number of complex assessments (long visits) provided by physician k, s^k denotes the number of straightforward assessments (short visits) provided by physician k, and γ_j is interpreted as the time required

[1] One may consider I as the total time spent treating patients.

for a type-j physician to perform a complex assessment, whereas ρ is interpreted as the time required to perform a straightforward assessment. C is continuous, monotonically increasing and convex in the caseload, I. Note that a physician who does not provide medical services, and thus, is not allocated any income (yogurt), obtains a utility of zero; this is interpreted as the physician's reservation utility level.

Given that the time required to perform complex assessments is less for type-1 physicians than for type-2 physicians, and assuming that straightforward assessments can be performed more quickly than complex assessments, the following inequality must hold

$$\rho < \gamma_1 < \gamma_2 \tag{3.3}$$

Next practices are constructed to match physicians with patients. A practice is characterized by one physician of type j ($j = 1, 2$) and an associated set of patients of type i ($i = 1, 2$), who use the physician's medical services. A profile of practice ϕ^k is an ordered pair $\phi^k = (\phi_1^k, \phi_2^k)$ where ϕ_1^k and ϕ_2^k are the number of type-1 and type-2 patients in practice ϕ^k. A practice structure $\aleph$ for the entire population is a set of ordered pairs $\{(\phi_1^1, \phi_2^1), ..., (\phi_1^K, \phi_2^K)\}$ such that $\sum_{k=1}^{K} \phi_i^k = h_i$, $(i = 1, 2)$, where h_i is the total number of patients of type i. Also, $\sum_{i=1}^{2} \sum_{k=1}^{K} \phi_i^k = H$. In other words, a practice structure ensures that all patients have a physician and that all physicians are assigned to practices.

This paper focuses on two price mechanisms (fee-for-service and capitation) with a uniform schedule of prices set exogenously for the whole patient popu-

lation. With this purpose in mind, all allocations considered in this paper are restricted to equal treatment allocations, i.e., those with equalized consumptions for all patients receiving comparable assessments, and for physicians with an identical practice profile. Define, therefore, (T_l, T_s) to be the transfer of yogurt received by a physician, where T_l is defined as a transfer for each complex assessment, and T_s as a transfer for each straightforward assessment. For convenience, in the exposition below it is assumed that the yogurt foregone by a patient who receives an assessment of a given length is equal to the transfer received by the physician who provides the service. However, if the institutional details of publicly funded health care systems were to be captured in greater detail, it would be possible to define τ_h as the health tax paid by a patient of type h, and the Ministry of Health budget constraint as $\sum_{h=1}^{H} \tau_h = \sum_{k=1}^{H} T_l l^k + T_s s^k$.

An allocation for a practice k is characterized by a list $(((T_l, T_s), l_i^k, s_i^k, \phi_1^k, \phi_2^k); i = 1, 2)$ where l_i^k denotes the total number of long visits by patients of type i in practice k, and s_i^k denotes the total number of short visits by patients of type i in practice k. Note that this notation is general enough to permit patients of a given type to have either a long or a short visit. An allocation relative to a practice structure $\aleph$ is denoted by $(((T_l, T_s), l_i^k, s_i^k, \phi_1^k, \phi_2^k); i = 1, 2, k = 1, ..., K)$ and consists of an allocation for each practice ϕ^k in the practice structure $\aleph$. A state of the economy, Ω, consists of a practice structure $\aleph$ of H and an allocation relative to $\aleph$, thus

$$\Omega = (\aleph, (((T_l, T_s), l_i^k, s_i^k, \phi_1^k, \phi_2^k); i = 1, 2, k = 1, ..., K)).$$

The state of the economy, Ω, is *feasible* if (i) all patients are assigned to practices, i.e., for each type i, $\sum_{k=1}^{K} \phi_i^k = h_i$; and (ii) $\sum_{i=1}^{2} \phi_i^k T_i^k \leq (\phi_1^k + \phi_2^k)y$. Condition (i) ensures that all patients receive medical care (either a short visit or a long visit) and condition (ii) guarantees that the total transfers of yogurt from patients to physicians within the practice do not exceed the amount of yogurt available to patients. Note that although the supply of medical services in each practice will be an endogenously determined variable, there is no assumed capacity constraint with regard to medical service production, i.e., it is physically possible for one physician to see all patients.

3.2.2 Pareto Efficient Allocations

The set of Pareto efficient allocations of yogurt and medical services is a subset of the set of feasible allocations. A Pareto efficient allocation is described by a state of the economy Ω such that $\nexists \aleph'$ and a feasible allocation relative to $\aleph', (((T_l', T_s'), l_i'^k, s_i'^k, \phi_1'^k, \phi_2'^k); i = 1,2; k = 1, ..., K)$, such that all patients and physicians are at least as well off under $\aleph'$ as they were under $\aleph$, and at least one agent is strictly better off.

Observe that in a Pareto efficient allocation both patients and physicians must be better off then under autarchy. Patient h's participation constraint can be expressed as

$$U^h = u(y - T_v) + \delta_{hv} > u(y), \forall h, \forall v. \tag{3.4}$$

Condition (3.4) indicates that patients are better off forgoing T_v units of yogurt and receiving a medical assessment of some sort than not seeing a physician at

all. For example, a complex patient is better off with a straightforward assessment rather than not seeing a physician at all. Observe that this constraint implicitly defines an upper limit on the amount of yogurt that the patient is willing to give up exchange for an assessment of a given duration.

Physician k's participation constraint is

$$U^k = V(T_s s^k + T_l l^k) - C(\gamma_j l^k + \rho s^k) > 0. \tag{3.5}$$

For a physician it must be true that the total benefit of the income received in exchange for servicing a given caseload outweighs the cost of providing these services. Note that due to the concavity of the utility income function, V, physicians will require high compensation if they are to be induced to carry a heavy caseload.

For there to be a nontrivial economic problem in this economy it must be the case that there exists at least one feasible state of the economy, Ω, for which conditions (3.4) and (3.5) can be simultaneously satisfied. If this is not true then there is an irresolvable allocation problem: in the absence of coercive measures, there is no way of dividing the available supply of yogurt and of allocating physicians and patients to practices which leaves all parties better off than they are under autarky. Since there is no physical capacity constraint on the physicians, this is equivalent to requiring that there is 'enough yogurt'. In what follows, it is assumed that this is true.

A related issue is whether complex patients should receive long visits. Due to the separability of the patient's utility function in income and medical services, the marginal benefit to the complex patient of receiving a long visit rather than a short

one is $\delta_{hl} - \delta_{hs}$. This is the upper bound on the amount of additional yogurt that a complex patient would be prepared to exchange for a long visit rather than a short one. As well, the additional cost to any physician of supplying a long visit rather than a short one is $C(\gamma_j l + \rho s) - C(\gamma_j(l-1) + \rho(s+1))$. It is then immediate that if there does not exist a feasible state of the economy, Ω, in which all complex patients receive long visits and for which $\delta_{hl} - \delta_{hs} > C(\gamma_j l^k + \rho s^k) - C(\gamma_j(l^k - 1) + \rho(s^k + 1))$ for $\forall\ h \leq h_1$ and all physicians who supply long visits in Ω, then there are no potential 'gains from trade' with respect to complex assessments rather than simple ones.

Remark 8 *If there is 'enough yogurt' then there exists a feasible state of the economy Ω in which complex patients prefer to receive long visits, and in which the physician's participation constraint is satisfied.*

Proof. For any pair (T_l, T_s) complex patient h, $\forall\ h \leq h_1$, prefers a long assessment if

$$\begin{aligned} u(y - T_l)) + \delta_{hl} \quad &> \quad u(y - T_s) + \delta_{hs} \\ &\iff \delta_{hl} - \delta_{hs} > u(y - T_s) - u(y - T_l). \end{aligned} \tag{3.6}$$

Observe that as $y \to \infty$, the right-hand side of (3.6) approaches zero, whereas the left-hand side is a positive constant. Consequently, if there is 'enough yogurt' then there always exists a pair of transfers of yogurt (T_l, T_s) from the patient to the physician that will make it worthwhile for the physician to provide a complex assessment, and which leaves the complex patient better off than if she were to only receive a simple assessment. ■

Corollary 9 *Suppose that there is 'enough yogurt'. Then at a Pareto efficient allocation all complex patients receive complex assessments, and all simple patients receive straightforward assessments.*

Proof. The first part of the corollary follows directly from equation (3.6). Suppose that initially, a complex patient was receiving a straightforward assessment. A Pareto improvement can now be constructed: consider an alternative allocation where the transfer of yogurt is $\widetilde{T} > T_s$ and where $\widetilde{T}$ is chosen so that the treating physician is indifferent between providing the complex assessment and the straightforward assessment; at this alternative allocation, the complex patient gains $\delta_{hl} - \delta_{hs}$ in additional utility from the complex assessment, and, if there is 'enough yogurt', this gain outweighs the loss in utility from reduced yogurt consumption. Now, suppose that initially a simple patient is receiving a complex assessment. Since there is no benefit to the patient from the longer assessment, and there is a greater cost to the physician when providing a complex assessment rather than a straightforward one, it is then immediate that - with no change in the structure of transfers - a Pareto improvement occurs when a simple assessment is provided. ■

Given that Pareto efficient allocations all involve providing straightforward assessments to simple patients, and complex assessments to complex patients, the next issue to resolve is how patients are allocated across practices. In what follows, therefore, the subscript i for l_i^k, s_i^k is suppressed because $l_2^k = 0, s_1^k = 0$ in all Pareto efficient allocations. Moreover, in this section, where each patient visits only once,

$l^k = \phi_1^k$ and $s^k = \phi_2^k$.

Turning now to the question of whether or not complex assessments are provided only by high-productivity physicians, the proposition below indicates that this is not generally true: depending on the relative proportions of complex and simple patients in the economy, complex assessments may in fact be provided by low-productivity physicians as well as by high productivity ones. What is true, however, is that at least one type of physician typically has a specialized caseload of either complex or straightforward assessments.

Proposition 10 *A Pareto efficient allocation of medical services and yogurt is described by a state of the economy* $\Omega^e = (\aleph, (((T_l, T_s), l^k, s^k, \phi_1^k, \phi_2^k); k = 1, ..., K))$, *where* $l^k = \phi_1^k, s^k = \phi_2^k$ *and at least one type of physician has a specialized caseload: either (i) all complex patients receive complex assessments with high-productivity physicians, and high-productivity physicians may provide some straightforward assessments to simple patients, whereas the low-productivity physician is specialized in the provision of straightforward assessments or (ii) high-productivity physicians provide complex assessments to complex patients and no assessments to simple patients and some complex patients may receive complex assessments with low-productivity physicians.*

Proof. Consider first the situation in which some complex patients receive complex assessments with low-productivity physicians and there exists at least one high-productivity physician who provides at least one straightforward assessment

to a simple patient. The high-productivity physician obtains the payoff

$$V(T_s s^k + T_l l^k) - C(\gamma_1 l^k + \rho s^k),$$

whereas the low-productivity physician who sees both simple and complex patients obtains the payoff

$$V(T_s s^{k\prime} + T_l l^{k\prime}) - C(\gamma_2 l^{k\prime} + \rho s^{k\prime}),$$

where $s^k, s^{k\prime}, l^k, l^{k\prime}$ denote the number of simple and complex assessments provided by the high-productivity and low-productivity physicians. In what follows, it is assumed that patients are divisible. Observe that, for any (l^k, s^k), a high-productivity physician is willing to exchange one simple patient for $\frac{\rho}{\gamma_1}$ complex patients; similarly, for any $(l^{k'}, s^{k'})$, a low-productivity physician is willing to exchange $\frac{\rho}{\gamma_2}$ complex patients for one simple patient. Now consider a trade in caseload between the high-productivity and low-productivity physician: reallocate simple patients from the high-productivity physician in exchange for up to $\frac{\rho}{\gamma_1}$ complex patients; assume that although there is an exchange of patients, there is no change in physician transfers. The high-productivity physician is indifferent between the new allocation and the initial one, and the same is true of the complex and simple patients: they do not care about the skill level (or productivity) of the doctor, but only about the complexity (length) of the assessment. However, the low-productivity physician is now strictly better off because his costs are lowered, i.e., $C(\gamma_2 l^{k'} + \rho s^{k'}) > C(\gamma_2(l^{k'} - \frac{\rho}{\gamma_1}) + \rho(s^{k'} + 1))$[2]. Consequently, the new allocation is a Pareto improvement with respect to the initial one.

[2] $C(\gamma_2(l^{k\prime} - \frac{\rho}{\gamma_1}) + \rho(s^{k\prime} + 1)) = C(\gamma_2 l^{k\prime} + \rho s^{k\prime} - \frac{\gamma_2}{\gamma_1}\rho + \rho))$. Since $\gamma_2 > \gamma_1$, then $-\frac{\gamma_2 \rho}{\gamma_1} + \rho < 0$. It follows that $C(\gamma_2 l^{k\prime} + \rho s^{k\prime}) > C(\gamma_2(l^{k\prime} - \frac{\rho}{\gamma_1}) + \rho(s^{k\prime} + 1))$.

In contrast, if high-productivity physicians provide assessments only to complex patients (and no straightforward assessments to simple patients), then there may exist Pareto efficient allocations where low-productivity physicians provide long assessments to some complex patients. To show that this is possible, assume initially that all complex patients are assigned to practices with high-productivity physicians, and that all simple patients are assigned to practices with low-productivity physicians. For convenience, assume that all practices with high-productivity physicians are identical (with $\widehat{l}^k$ complex patients), and all practices with low-productivity physicians are identical too (with $\widehat{s}^{k'}$ simple patients). The possibility of a Pareto improvement now depends upon whether or not high-productivity physicians have too high a caseload relative to low-productivity ones. Observe that it is Pareto improving to transfer a complex patient from a high-productivity doctor to a low-productivity one if there exists a Δ such that

$$V(T_l\widehat{l}^k - \Delta) - C(\gamma_1(\widehat{l}^k - 1)) > V(T_l\widehat{l}^k) - C(\gamma_1\widehat{l}^k) \tag{3.7}$$

and

$$V(T_s\widehat{s}^{k'} + \Delta) - C(\gamma_2 + \rho\widehat{s}^{k'}) > V(T_s\widehat{s}^{k'}) - C(\rho\widehat{s}^{k'}). \tag{3.8}$$

In particular, inequalities (3.7) and (3.8) are satisfied for $\Delta = T_l$, which means that for a high-productivity physician with caseload $\widehat{l}^k$, the marginal net benefit of giving up one complex patient and income T_l received in exchange of the complex assessment provided to this patient, is negative , i.e., $V_1'T_l - C_1'\gamma_1 < 0$; and for a low-productivity physician, the marginal benefit of adding one complex patient to the caseload of $\widehat{s}^{k'}$ simple patients and receiving a transfer T_l for the complex as-

sessment provided to this patient, is positive, i.e., $V_2'T_l - C_1'\gamma_2 > 0$. In other words, if there are too many complex patients, then at a Pareto efficient outcome some may be treated by low-productivity physicians, whereas if there are few complex patients, then not only will they all be treated by high-productivity physicians but some simple patients may also be seen by high-productivity physicians. ■

Corollary 11 *At Pareto efficient allocations where low-productivity physicians are specialized, each high-productivity physician provides h_1/k_1 complex assessments; at Pareto efficient allocations where high-productivity physicians are specialized, each low-productivity physician provides h_2/k_2 straightforward assessments.*

The partial characterization of Pareto efficient allocations provided above shows that there are two sorts of matching problems that arise: firstly, complex and simple patients need to be matched with complex and straightforward assessments and, secondly, high-productivity and low-productivity physicians need to be matched with the appropriate mix of patients. The next section examines whether fee-for-service and capitation-based payment schemes can be designed to implement equal treatment Pareto efficient outcomes.

3.3 Fixed Demand for Medical Services

This section describes an economy in which a regulator sets a schedule of fees for services provided by physicians. Physicians take these fees as given. Recall that demand for medical services is set exogenously at one visit per patient.

3.3.1 Fee-for-Service Equilibria and Pareto Efficient Allocations

The analysis of this subsection proceeds as follows. First, fee-for-service equilibria are described and, second, it is examined whether or not the equal treatment Pareto efficient allocations defined in Subsection 3.2.2 can be implemented with an appropriately designed fee-for-service payment schedule. It is assumed that there are no problems of asymmetric information: whether a patient is complex or simple, and whether a straightforward or complex assessment is provided can be observed at zero cost.

Definition 12 *A fee-for-service equilibrium* $e^f = (\Omega^f)$ *is a feasible state of the economy* $\Omega^f = (\aleph, (((p_l^f, p_s^f), l_i^k, s_i^k, \phi_1^k, \phi_2^k); i = 1, 2, k = 1, ..., K))$ *where* (p_l^f, p_s^f) *is a fee-for-service schedule for complex and straightforward assessments, such that (i) no patient wishes to change physician, (ii) no physician gains from adjusting the mix of straightforward and complex assessments which he offers, and (iii) both patient and physician are better off than under autarchy, i.e., both conditions (3.4) and (3.5) are satisfied.*

First consider the physician's decision with regard to service provision. Given the payment scheme the physician of type j maximizes his utility by choosing a number of complex and straightforward assessments to offer, l^k, s^k, i.e.,

$$\begin{aligned} \max_{l^k, s^k} U_j^k &= V(p_s s^k + p_l l^k) \qquad (3.9) \\ +C\left(\gamma_j l^k + \rho s^k\right), j &= 1 \text{ if } k \leq k_1,\ j = 2 \text{ if } k > k_2. \end{aligned}$$

The first order conditions[3] are

$$\begin{aligned} \frac{\partial U_j^k}{\partial l^k} &= p_l V'(p_s s^k + p_l l^k) - \gamma_j C'(\cdot) \leq 0 \\ l^{k*} &\geq 0; l^{k*}\left(p_l V'(p_s s^k + p_l l^k) - \gamma_j C'(\cdot)\right) = 0. \end{aligned} \tag{3.10}$$

$$\begin{aligned} \frac{\partial U_j^k}{\partial s^k} &= p_s V'(p_s s^k + p_l l^k) - \rho C'(\cdot) \leq 0 \\ s^{k*} &\geq 0; s^{k*}\left(p_s V'(p_s s^k + p_l l^k) - \rho C'(\cdot)\right) = 0. \end{aligned} \tag{3.11}$$

A high-productivity physician k (where $k \leq k_1$) provides only complex assessments if for all $s^{k*} > 0$,

$$\begin{aligned} p_s V'(p_s s^{k*} + p_l l^k) - \rho C'(\cdot) &< 0 \text{ and} \\ p_l V'(p_s s^{k*} + p_l l^k) - \gamma_1 C'(\cdot) &\geq 0. \end{aligned}$$

A high-productivity physician k (where $k \leq k_1$) provides both complex and simple assessments if there exists $(l^{k*} > 0, s^{k*} > 0)$ such that

$$\begin{aligned} p_l V'(p_s s^{k*} + p_l l^{k*}) - \gamma_1 C'(\cdot) &= 0 \text{ and} \\ p_s V'(p_s s^{k*} + p_l l^{k*}) - \rho C'(\cdot) &= 0. \end{aligned}$$

Similarly, a low-productivity physician k (where $k > k_1$) provides both complex and simple assessments if there exists $(l^{k*} > 0, s^{k*} > 0)$ such that

$$\begin{aligned} p_l V'(p_s s^{k*} + p_l l^{k*}) - \gamma_2 C'(\cdot) &= 0 \text{ and} \\ p_s V'(p_s s^{k*} + p_l l^{k*}) - \rho C'(\cdot) &= 0, \end{aligned}$$

[3] Since $V(\cdot) - C(\cdot)$ is a concave function, the second order conditions for the utility maximization are satisfied.

and the low-productivity physician specializes in the provision of straightforward assessments if for all $l_k^* > 0$

$$\begin{aligned} p_l V'(p_s s^{k*} + p_l l^{k*}) - C'(\cdot)\gamma_2 &< 0 \text{ and} \\ p_s V'(p_s s^{k*} + p_l l^{k*}) - C'(\cdot)\rho &\geq 0. \end{aligned}$$

It remains to establish whether a system of relative prices can be found which ensures that complex patients receive complex assessments, and simple patients receive straightforward assessments. Notice that if

$$\frac{p_l}{p_s} > \frac{\gamma_2}{\rho},$$

then both high and low productivity physicians will provide only complex assessments. Similarly, if

$$\frac{p_l}{p_s} < \frac{\gamma_1}{\rho},$$

then only simple assessments are provided. A physician of type j, $j = 1, 2$ is willing to provide both straightforward and complex assessments if

$$\frac{p_l}{p_s} = \frac{\gamma_j}{\rho}.$$

Notice that if the relative price of a complex versus a straightforward assessment is such that $\frac{\gamma_1}{\rho} < \frac{p_l}{p_s} < \frac{\gamma_2}{\rho}$, then a high-productivity physician offers only complex assessments and a low-productivity physician only straightforward ones.

In this section, with fixed demand for medical services and exogenously set prices, patients essentially face a 'take-it or leave-it' offer: they are given a choice of accepting a fixed quantity of services for a preset price or of having no medical

services at all. Consequently, there may exist equilibria in which the prices are 'wrong' and so, although all patients receive an assessment, only short *or* only long visits are provided.

Proposition 13 *There exist fee-for-service equilibria which are not Pareto efficient.*

Proof. The proof is by construction. Suppose the fees are set such that $p_l = p_s = p$. At price p, equations (3.10) and (3.11) become respectively

$$\frac{\partial U_j}{\partial l} = pV'(p \cdot s + p \cdot l) - \gamma_j C'(\cdot) \leq 0,$$

$$\frac{\partial U_j}{\partial s} = pV'(p \cdot s + p \cdot l) - \rho C'(\cdot) \leq 0.$$

It is evident that at any s and l for any physician of type j $(j = 1, 2)$ the inequality $pV'(p \cdot s + p \cdot l) - \gamma_j C'(\cdot) < pV'(p \cdot s + p \cdot l) - \rho C'(\cdot)$ holds, because $\gamma_j > \rho$. Therefore, all patients are served with straightforward assessments. It can be checked that the optimal number of straightforward assessments, s^* is increasing in price, p. Now choose p high enough so that all patients are served, i.e., $s^* \geq \frac{(h_1+h_2)}{(k_1+k_2)}$. Thus, an equilibrium exists in which all patients receive medical care, but because complex patients do not receive long visits this equilibrium is not Pareto efficient. ■

Proposition 13 demonstrates that an error in setting physician's fees by a public health authority, even when it leads to an equilibrium in which every patient is served, may not bring about an efficient match of heterogeneous physicians and

patients. If physicians are not paid enough to perform complex assessments, then complex patients will not receive adequate medical care.

However, as demonstrated below, if fees are set appropriately then fee-for-service equilibria are also equal treatment Pareto efficient allocations.

Consider the equal treatment Pareto efficient allocation, $\Omega^{\widehat{e}} = (\aleph, (((T_l, T_s), \widehat{l}^k, \widehat{s}^k, \widehat{\phi}_1^k, \widehat{\phi}_2^k); k = 1, ..., K))$ where $\widehat{l}^k = \widehat{\phi}_1^k, \widehat{s}^k = \widehat{\phi}_2^k$. For convenience assume that in $\Omega^{\widehat{e}}$ low-productivity physicians specialize in the provision of simple assessments, whereas high-productivity physicians provide both complex and straightforward assessments. Thus, $\forall\ k \leq k_1$, $\widehat{l}^k = h_1/k_1$, and $\sum_{k=1}^{k_1} \widehat{s}^k + \sum_{k=k_1+1}^{K} \widehat{s}^k = h_2$. In order to persuade a high-productivity physician to see both types of patients, it must be true that $\frac{p_l}{p_s} = \frac{\gamma_1}{\rho}$; if prices are set in this fashion, then the low-productivity physician will only want to see simple patients. Given this relative price, a representative high-productivity physician solves

$$\max_{l_k, s_k} V(p_s \frac{\gamma_1}{\rho} l^k + p_s s^k) - C(\gamma_1 l^k + \rho s^k).$$

$$= \max_{s_c} V(p_s s_c) - C(\rho s_c)$$

where $s_c = s^k + \frac{\gamma_1}{\rho} l^k$ can be interpreted as the total caseload. Note that at the solution to the physician's optimization problem the first-order condition is

$$p_s V'(p_s \widehat{s_c}) - \rho C'(\rho \widehat{s_c}) = 0. \tag{3.12}$$

Equation (3.12) implicitly defines $\widehat{s_c}$ and it can be verified that $d\widehat{s_c}/dp_s > 0$, i.e., that as the price per visit increases physicians are willing to increase their total

caseload. Consequently, there exists $(\widehat{p}_s, \widehat{p}_l) = (\widehat{p}_s, \frac{\gamma_1}{\rho}\widehat{p}_s)$ such that $\widehat{s}_c(\widehat{p}_s, \widehat{p}_l) = \widehat{s}^{1k} + \frac{\gamma_1}{\rho}\frac{h_1}{k_1}$.

Now consider the problem of low-productivity physicians. Recall that if they provide only straightforward assessments, at an optimum it must therefore be true that

$$\widehat{p}_s V'(p_s s^k) - \rho C'(\rho s^k) = 0. \tag{3.13}$$

Let $\widehat{s}^{2k}$ be the number of straightforward assessments that solves the low-productivity physician's optimisation problem above. Observe that for the price structure $(\widehat{p}_s, \widehat{p}_l) = (\widehat{p}_s, \frac{\gamma_1}{\rho}\widehat{p}_s)$ the first-order conditions for a high-productivity physician in (3.12) and for a low-productivity one in (3.13) are equivalent and thus,

$$\widehat{s}_c = \widehat{s}^{2k}. \tag{3.14}$$

This implies that in equilibrium the marginal cost of adding an additional simple patient (or equivalently, $\frac{\rho}{\gamma_1}$ complex patients) to the total caseload of a high-productivity physician is the same as the cost of adding one more simple patient to the caseload of a low-productivity physician. Moreover, the fact that physicians' caseloads are equalized allows an analytic solution to be established. In equilibrium each high-productivity physician serves h_1/k_1 complex patients and $\widehat{s}^{1k}$ simple patients, where

$$\widehat{s}^{1k} = \frac{h_2}{(k_1 + k_2)} - \frac{\gamma_1 h_1 k_2}{\rho k_1 (k_1 + k_2)}.$$

Each low-productivity physician serves $\widehat{s}^{2k}$ simple patients, where

$$\widehat{s}^{2k} = \frac{\rho h_2 + \gamma_1 h_1}{\rho(k_1 + k_2)}.$$

Thus, it is established that there exists a fee-for-service equilibrium e^f with the price vector $(\widehat{p}_l, \widehat{p}_s) = (\frac{\gamma_1}{\rho}\widehat{p}_s, \widehat{p}_s)$ and an allocation correspondent to this price vector such that it is an equal treatment Pareto-efficient allocation $\Omega^{\widehat{e}} = \left(\aleph, \left(\left(\left(\widehat{T}_l, \widehat{T}_s\right), \widehat{l}^k, \widehat{s}^k, \widehat{\phi}_1^k, \widehat{\phi}_2^k\right); k = 1, ..., K\right)\right)$ where $\widehat{T}_l = \widehat{p}_l, \widehat{T}_s = \widehat{p}_s$. By construction $\widehat{p}_s$ is the lowest price at which all simple patients are served and, similarly, $\widehat{p}_l$ is the lowest price at which all complex patients are served. Observe that any price structure $(\widehat{\widehat{p}}_l, \widehat{\widehat{p}}_s) = (\frac{\gamma_1}{\rho}\widehat{\widehat{p}}_s, \widehat{\widehat{p}}_s)$ where $\widehat{\widehat{p}}_s > \widehat{p}_s$ also implements the same allocation as a fee-for-service equilibrium, however, at this efficient allocation both physician types are rationed, i.e., they would like to treat more patients than there are total patients in the economy. Physicians in fee-for-service equilibria with rationing obtain higher utility than they do at the fee-for-service equilibria with price vector $(\widehat{p}_l, \widehat{p}_s)$ where they serve the desired number of patients.[4]

Now consider an equal treatment Pareto efficient allocation $\Omega^{\widetilde{e}} = \left(\left(\aleph, \left(\left(\widetilde{T}_l, \widetilde{T}_s\right), \widetilde{l}^k, \widetilde{s}^k, \widetilde{\phi}_1^k, \widetilde{\phi}_2^k\right); k = 1, ..., K\right)\right)$ where $\widetilde{s}_2^k = \widetilde{\phi}_2^k = h_2/k_2$ and $\sum_{k>k_1} \widetilde{l}^k + \sum_{k<k_1} \widetilde{l}^k = h_1$. The same procedure is applied. To persuade high-productivity physicians to specialize in complex assessments, and low-productivity physicians to see both complex and simple patients, it must be true that the relative price of complex and simple assessments is set at $\frac{p_l}{p_s} = \frac{\gamma_2}{\rho}$. A representative low-productivity physician solves

$$\max_{l^k, s^k} V(p_l l^k + p_l \frac{\rho}{\gamma_2} s^k) - C(\gamma_2 l^k + \rho s^k)$$

$$= \max_{l_c} V(p_l l_c) - C(\gamma_2 l_c),$$

[4] They provide the same number of assessments as before, but obtain more yogurt.

where $l_c = l^{2k} + \frac{h_2}{k_2}\frac{\rho}{\gamma_2}$ is the total caseload, expressed in terms of complex patients. Let $\widetilde{l}_c$ be the interior solution to the above. Similarly, a representative high productivity physician solves

$$\max_{l^{1k}} V(p_l l^{1k}) - C(\gamma_1 l^{1k}).$$

Denote by $\widetilde{l}^{1k}$ the solution to the high productivity physician's optimization problem. At a fee-for-service equilibrium, it must then be true that $k_1\widetilde{l}^{1k}+k_2\left(\widetilde{l}_c - \frac{h_2}{k_2}\frac{\rho}{\gamma_2}\right) \geq h_1$. Observe that at any $(\widetilde{p}_l, \widetilde{p}_s) = (\widetilde{p}_l, \frac{\gamma_2}{\rho}\widetilde{p}_l)$ it must be true that $\widetilde{l}^{1k} > \widetilde{l}_c$. Since both $\widetilde{l}^{1k}, \widetilde{l}_c$ are continuous and increasing functions of $(\widetilde{p}_l, \widetilde{p}_s)$ it is then trivial to establish that there exists a fee-for-service equilibrium e^f with the price vector $(\widetilde{p}_l, \widetilde{p}_s) = (\widetilde{p}_l, \frac{\gamma_2}{\rho}\widetilde{p}_l)$ and an allocation correspondent to this price vector such that it is an equal treatment Pareto-efficient allocation $\Omega^{\widetilde{e}} = \left(\aleph, \left(\left(\left(\widetilde{T}_l, \widetilde{T}_s\right), \widetilde{l}^k, \widetilde{s}^k, \widetilde{\phi}_1^k, \widetilde{\phi}_2^k\right); k = 1, ..., K\right)\right)$ where $\widetilde{T}_l = \widetilde{p}_l, \widetilde{T}_s = \widetilde{p}_s$. Note that if the relative price is held constant, but prices are increased beyond $(\widetilde{p}_l, \widetilde{p}_s)$, then these fees will implement Pareto efficient fee-for-service equilibria with rationing, in which both high and low-productivity physicians are better off than at the fee-for-service equilibrium attained with $(\widetilde{p}_l, \widetilde{p}_s)$. It should be observed that at all of these Pareto efficient equilibria, the high-productivity physician is strictly better off than the low-productivity physician.[5] These results are summarized in the proposition below.

[5] Observe that a high-productivity physician has an option of choosing the same caseload, $\widetilde{l}_c$, as a low-productivity physician has. The cost associated with serving this caseload is lower for a high-productivity physician because $\gamma_1 < \gamma_2$. Thus, it must be true that given the same income and lower cost, a high-productivity physician with $\widetilde{l}_c$ caseload is better off than a low-productivity one with the same caseload. By choosing a greater caseload, $\widetilde{l}^{k1}$, a high-productivity physician must then increase his utility even more.

Proposition 14 *Consider a price vector* $(\widehat{p}_l, \widehat{p}_s)$ *such that* $\frac{\widehat{p}_l}{\widehat{p}_s} = \frac{\gamma_1}{\rho}$ *and that at this price vector the high-productivity physician chooses to provide* h_1/k_1 *long visits, and* $\widehat{s}^{1k}$ *straightforward assessments, whereas the low-productivity physician provides* $\widehat{s}^{2k}$ *straightforward assessments, and such that* $k_1\widehat{s}^{1k} + k_2\widehat{s}^{2k} \geq h_2$. *Then this price vector implements a Pareto efficient fee-for-service equilibrium. Moreover, any price vector* $(\widehat{\widehat{p}}_l, \widehat{\widehat{p}}_s)$ *such that* $\widehat{\widehat{p}}_l > \widehat{p}_l$, $\widehat{\widehat{p}}_s > \widehat{p}_s$, *and* $\frac{\widehat{p}_l}{\widehat{p}_s} = \frac{\widehat{\widehat{p}}_l}{\widehat{\widehat{p}}_s}$, *also implements a Pareto efficient fee-for-service equilibrium with the same allocation of patients to practices. Similar results hold for price vectors* $(\widetilde{p}_l, \widetilde{p}_s) = (\frac{\gamma_2}{\rho}\widetilde{p}_s, \widetilde{p}_s)$ *which implement fee-for-service equilibria where high-productivity physicians serve* $\widetilde{l}^{1k}$ *complex patients, and low-productivity physicians serve* $\widetilde{l}^{2k}$ *complex and* h_2/k_2 *simple patients.*

The above proposition provides insight into the circumstances in which fee-for-service equilibria are Pareto efficient. What should be noted, of course, is that fee-for-service equilibria are typically not unique: there are many (indeed, infinitely many) price vectors which implement the same allocation of patients to practices. These equilibria differ from each other with respect to the distribution of income between physicians and patients.

The discussion above does not imply that one can find a price vector which will implement *any* Pareto efficient allocation. Unfortunately, in the absence of a tax system, one is generally unable to support as fee-for-service equilibria those Pareto efficient allocations where $T_l < \widetilde{p}_l, T_s < \widetilde{p}_s$: at the prices $p_l = T_l < \widetilde{p}_l$ and $p_s = T_s < \widetilde{p}_s$ physicians will not choose to serve the entire patient population.

However, the next proposition establishes that when the physician's payoff is linear in income all equal treatment Pareto efficient allocations can be implemented as fee-for-service equilibria with the help of lump-sum taxes placed on physicians.

Proposition 15 *(The Second Welfare Theorem) Assume that the physician's utility function is linear in income. Then every equal treatment Pareto efficient allocation*[6] $\Omega^e = (\aleph, ((T_l, T_s, l^k, s^k, \phi_1^k, \phi_2^k); k = 1, ..., K))$ *where* $l^k = \phi_1^k, s^k = \phi_2^k$ *can be implemented as a fee-for-service equilibrium with price schedule* (p_l^*, p_s^*) *with the help of lump-sum taxes* (τ_1, τ_2) *placed on physicians.*

Proof. When the physician's utility function is linear in income, i.e., $V = p_l l^k + p_s s^k - \tau_j$, $j = 1, 2$; the first-order conditions of a utility maximizing physician k become

$$\begin{aligned} \frac{\partial U_j^k}{\partial l^k} &= p_l - \gamma_j C'(\cdot) \leq 0 \\ l^{k*} &\geq 0, l^{k*}\left(p_l - \gamma_j C'(\cdot)\right) = 0 \end{aligned} \tag{3.15}$$

$$\begin{aligned} \frac{\partial U_j^k}{\partial s^k} &= p_s - \rho C'(\cdot) \leq 0 \\ s^{k*} &\geq 0; s^{k*}\left(p_s - \rho C'(\cdot)\right) = 0 \end{aligned} \tag{3.16}$$

Observe that full differentiation of the first-order conditions yield $\frac{dl^{k*}}{d\tau_j} = 0$, $\frac{ds^{k*}}{d\tau_j} = 0$, $j = 1, 2$; and therefore the number of patients served is independent of the vector of lump-sum taxes. Therefore, choose (p_l^*, p_s^*) such that $\sum_{k>k_1} l^k + \sum_{k<k_1} l^k \geq h_1$ and

[6] Recall that at every Pareto efficient allocation all patients are served according to their medical needs: complex patients receive long visits, and simple patients are provided with straightforward assessments.

$\sum_{k>k_1} s^k + \sum_{k<k_1} s^k \geq h_2$. This is a Pareto efficient fee-for-service equilibrium. Now, by varying (τ_1, τ_2) all other allocations of yogurt between physicians and patients can be achieved, in particular those where $T_l = p_l^* - \tau_1$ and $T_s = p_s^* - \tau_2$. ■

The result of this Proposition is significant as it shows that a combination of a thoughtfully implemented fee-for-service price mechanism and tax policy makes it possible to implement *any* desired equal treatment Pareto efficient allocation. As the linearity of physician's utility in income renders the choice of caseload independent of the lump-sum tax, redistribution from (or between) physicians via these taxes makes it possible to implement Pareto efficient allocations in which physicians receive just enough compensation to be induced to provide necessary services.

This section characterizes Pareto efficient fee-for-service equilibria as a particular match of physicians with patients and services provided to these patients. The most important lesson to be drawn here is that fee-for-service equilibria will be efficient only when physicians are appropriately compensated for the higher costs incurred in treating complex patients. In countries where physicians are paid on a fee-for-service basis, it is often the case that the public bemoans the brevity of the time spent in the physician's office. What the analysis above establishes, however, is that the practice of 10-minute visits (enough for a straightforward assessment only) is a consequence of the fee-for-service schedule, which is not responsive to patient's needs. Were physicians to be paid more for spending longer with patients with complex medical histories than they are for consultations with

generally healthy individuals, then consultation times would be adjusted.

3.3.2 Capitation Fee Equilibria and Pareto Efficiency

Many publicly-funded primary health care systems have moved away from the traditional fee-for-service compensation scheme towards capitation as this is deemed to motivate physicians to be cost-conscious. A capitation contract consists of a fixed (typically monthly) payment per patient on the physician's roster. This payment is adjusted according to some observable patient characteristics, such as age and gender. As implemented in Canada, rostered patients who choose to see a doctor other then their regular provider incur no financial penalty. However, physicians paid on a capitation basis are subject to a financial penalty (clawback) which is subtracted from the capitation payment every time a rostered patient sees a physician outside the practice. In the analysis below the clawback is set uniformly across practices and reflects the type of services provided to the patient outside the clinic where she is rostered. In Canada, physicians who roster patients can also provide fee-for-service assessments for non-rostered patients under certain restrictions. In a nutshell, what this implies is that regardless of whether physicians are paid on a fee-for-service or capitation basis, the whole patient population H is served by the same pool of K physicians.

Under capitation physicians maximize their income by choosing the number of rostered patients of each type and the number of assessments of each type. The utility of physician k of type j is

$$U_j^k = V(R_1 m_1^k + R_2 m_2^k - \lambda_1 n_1^k - \lambda_2 n_2^k + p_l l_0^k + p_s s_0^k) - C(\gamma_j(l^k + l_0^k) + \rho(s^k + s_0^k)) \quad (3.17)$$

where $R_1$and R_2 denote capitation fees per complex and per simple rostered patient, respectively; m_1^k and m_2^k denote the number of complex and simple patients, rostered by physician k; n_1^k and n_2^k denote the number of visits by rostered complex and simple patients to physicians other than physician k; λ_1 and λ_2 denote the clawback for a complex and a straightforward assessment; l^k and s^k denote the number of complex and straightforward assessments provided to rostered patients in practice k; and l_0^k and s_0^k denote the number of complex and of straightforward assessments, respectively, provided to non-rostered patients at the fees of p_l and p_s respectively.[7]

Remark 16 *When demand for medical services is fixed at one visit, the clawback is optimally set to be greater than or equal to the capitation fee.*

Proof. The proof is by contradiction. Without loss of generality, assume a physician does not serve any outside patients, i.e., $l_0 = s_0 = 0$. Suppose $\lambda_i = R_i - \Delta$, where $\Delta \geq 0$. Then, given that $n_1 = m_1 - l$ and $n_2 = m_2 - s$, $U_j = V((\lambda_1 + \Delta)m_1 + (\lambda_2 + \Delta)m_2 - \lambda_1(m_1 - l) - \lambda_1(m_2 - s)) - C(\gamma_j l + \rho s) = V(\Delta\, m_1 + \Delta\, m_2 + \lambda_1 l + \lambda_1 s) - C(\gamma_j l + \rho s)$. Note, if $l = s = 0$, $U_j = V(\Delta\, m_1 + \Delta\, m_2) > 0$. This means that without serving a single patient the physician is making a positive profit by simply rostering more patients and receiving a prospective capitation fee for each additional patient. Thus, only by choosing a clawback equal to or greater than the per-patient capitation fee can a health care regulator ensure that all rostered patients are treated by their physicians. ■

[7] To reduce visual clutter, the index k indicating a particular physician (and therefore practice), is dropped in the analysis below.

Corollary 17 *Physicians treat every patient in their rosters, i.e.,* $m_1 = l$ *and* $m_2 = s$.[8]

Given the above Remark and Corollary the optimization problem of a type-j physician becomes

$$\max_{l,s} U_j = V(R_1 l + R_2 s) - C(\gamma_j l + \rho s) \tag{3.18}$$

A capitation fee equilibrium is now defined.

Definition 18 *A capitation fee equilibrium* $e^c = (\Omega^c)$ *is a feasible state of the economy* $\Omega^c = (\aleph, (((R_1, R_2), m_i^k, n_i^k, l_i^k, s_i^k, \lambda_1, \lambda_2, p_l, p_s); i = 1, 2, k = 1, ..., K))$ *where* (R_1, R_2) *is a capitation fee schedule for complex and simple patients, such that (i) no patient wishes to be rostered in a different practice; (ii) no physician gains from adjusting the mix of straightforward and complex assessments which he offers, and (iii) both patient and physician are better of than under autarchy, i.e., both conditions (3.4) and (3.5) are satisfied.*

Proposition 19 *(Equivalence of fee-for-service and capitation equilibria). Assume that the clawback is set at a level greater than or equal to the per-patient capitation fee, i.e.,* $\lambda_1 \geq R_2, \lambda_1 \geq R_2$. *Then every Pareto efficient fee-for-service equilibrium with prices* (p_l, p_s) *is also a capitation equilibrium with capitation fees* $(R_1, R_2) = (p_l, p_s)$.

Proof. This follows directly from the observation that under the above condition for the clawback a capitated physician faces the same optimization problem

[8] Recall that each patient visits only once.

in equation (3.18) as a fee-for-service physician in equation (3.9), where $R_1 = p_l$ and $R_2 = p_s$. ■

Proposition 19 establishes rather unconventional results as compared to standard analyses of physician remuneration schemes. Whereas much has been written about the differences between prospective and retrospective payment schemes, and about optimal 'blends' of the two approaches, the analysis above suggests that the focus of much of this literature may be somewhat misdirected - at least in situations in which demand for medical services is price inelastic. The observed differences between the performance of fee-for-service and capitation schemes may have more to do with the relative lack of appropriately designed prices under *both* remuneration schemes; if prices were set appropriately, the observed differences in performance would disappear.

3.4 Uncertainty about Demand for Medical Services

In the real world an important source of uncertainty is the number of visits that a particular patient will require. One may interpret this as some degree of uncertainty about the patient's type: not all elderly patients require regular visits to a doctor (or, perhaps more accurately, many elderly patients experience episodes during which they require acute care, but most of the time are in good general health). Thus, when an elderly patient walks into a physician's office there is uncertainty about whether or not this patient will require one or more visits. Indeed, it would seem fair to argue that the number of visits which each patient will require is unknown to the physician, to the regulator, and to the patient herself.

This section extends the analysis to address uncertainty about a patient's type. As before, suppose that the patient's need for medical services is determined exogenously, and does not depend on actions taken by either the physician or the patient; specifically, assume that with probability π a complex patient requires two complex assessments and with probability $(1-\pi)$ only one assessment is needed. The patient's type is costlessly revealed only after the patient's first visit.

3.4.1 Fee-for-Service Equilibria

Despite the fact that an individual's demand for medical services is uncertain, with a large enough number of complex patients in the population the aggregate number of long visits that will be required is known with certainty. Consequently, the total number of complex assessments can be divided among the population of high-productivity physicians so that, in effect, every physician faces a certain demand for medical services.[9]

Therefore, fee-for-service physicians in an environment of uncertain individual demand simply solve the same maximization problem as under certainty: for the same payment per assessment, the solution to the physician's optimization problem will be identical to (3.10)-(3.11) in Subsection 3.3.1 where all patients require a single assessment. There is only one difference between the practice structure under certain patient demand and that under uncertainty: each state

[9] Therefore in a publicly funded health care system the tax revenue needed to finance physicians' services would be known with certainty. Taxes may or may not reflect patient's type and the amount of care received. For the purpose of this paper the exact distribution of transfers from patients in exchange for received care is irrelevant. One can simply imagine that, as is typically the case with a publicly funded health insurance, taxes collected are independent of the care received. Therefore, a patient does not face uncertainty regarding her medical bill, although she does not know how many times she will require physician's care.

of nature is characterized by a different set of patients visiting a physician twice. A uniform price for a complex assessment does not provide physicians with any incentive to distinguish between a new patient and a returning one. Thus, second-visit complex patients will be spread amongst all high-productivity physicians to fill in the optimal number of assessment slots chosen by each physician.

Remark 20 *Under uncertainty about the probability of the patient's second visit, a fee-for-service remuneration scheme does not guarantee continuity of care, in the sense that a patient seeking a second visit may not be able to see the same physician she visited for the initial consultation.*

It is important to emphasize that a fee-for-service physician does not face any risk of burnout due to excessive caseload if a high proportion of his regular patients require a second visit: a fee-for-service physician does not incur any financial penalties if he chooses not to accommodate the second visit.

Proposition 21 *A price schedule* (p_l, p_s)*, that implements an equal treatment Pareto efficient allocation with* (L, S) *complex and simple assessments as a fee-for-service equilibrium under demand certainty will also implement this efficient allocation with a total of* (L, S) *complex and simple assessments as a fee-for-service equilibrium where there is individual demand uncertainty.*

Proof. This follows immediately from Proposition 15 and the fact that a fee-for-service physician's choice of the optimal number of medical services to provide is not influenced by the uncertainty of any individual's demand for medical care.

■

What is interesting to observe about the fee-for-service equilibrium under individual demand uncertainty is that the costs to the health care system depend only on the total number of assessments that are required. As physicians do not actually have to bear any costs of uncertainty, they do not have to be compensated for risk bearing.

3.4.2 Capitation Equilibria

When the number of visits required by a complex patient is not known with certainty, the cost to the physician of servicing rostered complex patients is unknown. In the case of unexpectedly high numbers of returning patients, the physician must either incur the costs associated with providing many additional consultations, or suffer a loss in income due to the clawback if some of the rostered patients are obliged to seek care from a different physician.

As above, assume that capitated physicians can provide services to non-rostered patients, and are paid on a fee-for-service basis for their services to these patients at per-visit prices of p_l and p_s for complex and straightforward assessments respectively. One can now show that physicians who experience a high volume of follow-ups in a particular state of nature may choose to serve only a subset of their follow-ups, whereas those with a low realized volume of returning patients will be willing to offer assessments to non–rostered patients.

Proposition 22 *Suppose that the clawback fees* (λ_1, λ_2) *are weakly greater than the fees paid for services provided to non-rostered patients,* (p_l, p_s). *Then, when the demand for services is uncertain, risk-averse physicians paid under a capita-*

tion scheme require greater total remuneration than is necessary to induce them to provide equivalent services under a fee-for-service payment mechanism.

Proof. Assume the physician has a mixed caseload. Applying the principle of backward induction, first consider the physician's decision after the resolution of the demand uncertainty - which means after the physician has chosen to roster (m_1, m_2) complex and simple patients, all of whom have already received an initial assessment. Let the total number of follow-up visits required by the rostered complex patients after the resolution of uncertainty be denoted by $\widetilde{m}_1$ where $\widetilde{m}_1 \leq m_1$.

The physician must now determine whether or not to service the entire demand for long visits by complex patients, knowing that any patient who is denied care will seek services elsewhere. Also, the physician must choose how many assessments to offer to non-rostered complex patients. Consequently, the decision problem of the physician can be expressed as

$$\begin{aligned}&\max_{n_1,l_0} V(R_1 m_1 + R_2 m_2 - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\ &-C(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1))) + \rho m_2).\end{aligned}$$

The first-order conditions for n_1 and l_0 are, respectively,

$$\begin{aligned}&-p_l V'(R_1 m_1 + R_2 m_2 - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\ &+\gamma_j C'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho m_2) \leq 0\end{aligned} \tag{3.19}$$

$$n_1 \geq 0,\ n_1 \begin{pmatrix} -p_l V'(R_1 m_1 + R_2 m_2 - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\ +\gamma_j C'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho m_2) \end{pmatrix} = 0,$$

$$p_l V'(R_1 m_1 + R_2 m_2 - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \tag{3.20}$$

$$-\gamma_j C'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho m_2) \leq 0$$

$$l_0 \geq 0,\ l_0 \begin{pmatrix} p_l V'(R_1 m_1 + R_2 m_2 - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\ -\gamma_j C'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho m_2) \end{pmatrix} = 0.$$

A solution to (3.19)-(3.20) can be denoted as $(\widetilde{n}_1, \widetilde{l}_0)$. Observe that if the clawback is set high enough, physicians accommodate every rostered patient's visit, i.e., $\widetilde{n}_1 = 0$. Observe that since $\widetilde{m}_1$ will vary from one practice to another, $\widetilde{n}_1$ will generally vary from one practice to another. Also, if $p_l = \lambda_1$, then $\widetilde{n}_1$ and $\widetilde{l}_0$ are substitutes. Here, it is assumed that the physician takes care of his rostered patients first; however, if the realization of the follow-up visits is particularly low in a given state, it is payoff-maximizing (due to the lower-than-expected costs) to offer services to non-rostered patients.

Now consider the decision with respect to the number of patients to roster, m_1, m_2. The physician solves

$$\max_{m_1, m_2} EV(R_1 m_1 + R_2 m_2 - p_l\,(n_1(\widetilde{m}_1) - l_0(\widetilde{m}_1))) \tag{3.21}$$

$$-EC(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0) + \rho(m_2))$$

subject to $\widetilde{n}_1$ and $\widetilde{l}_0$ solving (3.19)-(3.20).

Notice that the uncertainty comes from the fact that $\widetilde{m}_1$ varies from 0 to m_1 in different states of the world. The first-order conditions for m_1 and m_2 are,

respectively,

$$R_1 EV'(R_1 m_1 + R_2 m_2 - p_l\,(n_1(\widetilde{m}_1) - l_0(\widetilde{m}_1))) \tag{3.22}$$

$$-\gamma_j EC'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0) + \rho m_2) \leq 0$$

$$m_1 \left(\begin{array}{c} R_1 EV'(R_1 m_1 + R_2 m_2 - p_l\,(n_1(\widetilde{m}_1) - l_0(\widetilde{m}_1))) \\ -\gamma_j EC'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0) + \rho m_2) \end{array} \right) = 0,$$

$$R_2 EV'(R_1 m_1 + R_2 m_2 - p_l\,(n_1(\widetilde{m}_1) - l_0(\widetilde{m}_1))) \tag{3.23}$$

$$-\rho EC'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0) + \rho m_2) \leq 0$$

$$m_2 \left(\begin{array}{c} R_2 EV'(R_1 m_1 + R_2 m_2 - p_l\,(n_1(\widetilde{m}_1) - l_0(\widetilde{m}_1))) \\ -\rho EC'(\gamma_j(m_1 + \widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0) + \rho m_2) \end{array} \right) = 0.$$

Now compare the choices of the capitated physician with those of a peer paid on a fee-for-service basis. Suppose that (p_l^*, p_s^*) implements a fee-for-service equilibrium with caseload (l^{k*}, s^{k*}) for each practice k and an average 'virtual roster' of $(\phi_1^{k*}, \phi_2^{k*})$ for the fee-for-service physician, that is, the average number of patients of each type seen by the physician in practice k in the fee-for-service equilibrium: $\phi_1^{k*} = \frac{l^{k*}}{1+\pi}$, $\phi_2^{k*} = s^{k*}$. Define $\widehat{R}_1 = p_l^*(1+\pi)$ and $\widehat{R}_2 = p_s^*$, that is, $\widehat{R}_1, \widehat{R}_2$ are set equal to the expected revenue per patient of each type at the fee-for-service equilibrium. Note that the first-order conditions for the fee-for-service physician, (3.10) – (3.11) can be re-written in terms of the 'virtual roster'

$$\widehat{R}_1 V'(\widehat{R}_1 \phi_1^{k*} + \widehat{R}_2 \phi_2^{k*}) - \gamma_j C'(\gamma_j \phi_1^{k*}(1+\pi) + \rho \phi_2^{k*})) = 0 \tag{3.24}$$

$$R_2 V'(\widehat{R}_1 \phi_1^{k*} + \widehat{R}_2 \phi_2^{k*}) - \rho C'(\gamma_j \phi_1^{k*}(1+\pi) + \rho \phi_2^{k*}) = 0. \tag{3.25}$$

Now observe that if $\widehat{m}_1 = \phi_1^{k*}$ and $\widehat{m}_2 = \phi_2^{k*}$, and $\widehat{Z} \equiv \gamma_j(\widehat{m}_1 + \widetilde{m}_1 - \widetilde{n}_1 + \widetilde{l}_0) + \rho\widehat{m}_2$, then $E(\widehat{Z}) = \gamma_j\phi_1^{k*}(1+\pi) + \rho\phi_2^{k*}$. Also, if $\widehat{D} \equiv \widehat{R}_1\widehat{m}_1 + \widehat{R}_2\widehat{m}_2 - p_l\widetilde{n}_1 + p_l\widetilde{l}_0$, then $E(\widehat{D}) = \widehat{R}_1\phi_1^{k*} + \widehat{R}_2\phi_2^{k*}$. Substituting these expressions into the first order conditions for the capitated physician, (3.22) and (3.23), observe that due to the concavity of V and convexity of C, $EV'(D) < V'(E(D))$ and $EC'(Z) > C'(E(Z))$. It must therefore be true that if $(m_1, m_2) = (\widehat{m}_1, \widehat{m}_2)$ then

$$\widehat{R}_1 EV'(\widehat{D}) - \gamma_j EC'(\widehat{Z}) < 0, \tag{3.26}$$

$$\widehat{R}_2 EV'(\widehat{D}) - \rho EC'(\widehat{Z}) < 0. \tag{3.27}$$

This implies that a physician paid under a capitation-based payment scheme will require greater total remuneration than a physician paid on a fee-for-service basis, if the capitated physician is to be induced to serve the $(\widehat{m}_1, \widehat{m}_2)$ roster. ■

Proposition 22 establishes that when there is demand uncertainty it is less costly to induce physicians to provide medical services when they are paid on a fee-for-service basis than when they are paid under a capitation scheme.

Corollary 23 *The level of the clawback affects the risk premium that a capitated physician requires to be induced to roster $(\widehat{m}_1, \widehat{m}_2)$ patients as compared to a fee-for-service payment scheme. The lower the clawback, the more patients a risk-averse physician will choose to roster, given the same capitation fee.*

Proof. This follows from the observation that a lower clawback decreases the expected penalty for unserviced demand, thus decreasing uncertainty with regard to the physician's income. ■

This section highlights the differences in health care outcomes under the fee-for-service and the capitation compensation mechanisms, when an individual physician faces uncertainty regarding the number of follow-up visits from the patients he sees initially. Since physicians are risk-averse, they have to be compensated for this risk-taking, resulting in higher capitation fees than those that induce physicians to service the same expected caseload when the demand is certain. In contrast, in a fee-for-service arrangement, which pays physicians retrospectively for services provided, physicians do not bear a risk of a patient returning for a follow-up visit as they are not obliged to accommodate the follow-up visit. Thus, abstracting from the value of continuous care, it is demonstrated that if the aggregate demand for follow-up visits is certain, a fee-for-service mechanism is cheaper for a regulator, since the same number of consultations can be provided at a lower cost per patient under the fee-for-service arrangement compared to the capitation one.

3.5 Preventive care

There are many circumstances in which it is argued that the likelihood of a patient requiring a subsequent follow-up visit could be greatly diminished if the physician were to take the time to provide preventive care during the initial consultation. However, providing preventive care is costly - in particular, because this requires physician's time. In this section, it is assumed that the physician can choose whether or not to apply more effort during the first encounter with the patient. Increased effort during the first visit by a complex patient decreases the probability of the second visit by q, thus, making the probability of the second

visit $\pi - q$, where $\pi > q$. Additional effort is costly for the physician, increasing the cost of a complex assessment from γ_j to $\gamma_j + \alpha$. For convenience, and in order to focus on the impact of the payment scheme on the behaviour of the physician, it is assumed that the patient is indifferent between the certain prospect of a longer visit which includes preventive care and a follow-up visit with probability $\pi - q$, or one visit for certain and a second visit with probability π.[10]

3.5.1 Fee-for-Service Contract

It is often suggested that fee-for-service physicians demonstrate little interest in preventive care, in particular as compared to capitated physicians who arguably face powerful incentives to undertake actions which reduce the likelihood of the patient returning. What this argument overlooks, however, is that the price mechanism can be used to provide physicians who are compensated on a fee-for-service basis with strong incentives to undertake preventive care when it is in fact economically desirable for them to do so.

Assume that fee-for-service physicians who undertake preventive care are offered supplementary compensation, p_r, for their preventive care services (e.g., pap smears and vaccinations). Given that the patient is indifferent between the longer visit and a higher probability of a second visit then the provision of preventive care is desirable only if the certain increase in the cost of physician effort is less than the expected cost of providing a higher number of follow-up visits. Observe that, in a large population, there is no aggregate uncertainty regarding the total num-

[10] It would be straightforward to allow for patients to strictly prefer preventive care; however, this does not add substantively to the analysis.

ber of visits required under these two scenarios. If a typical physician provides (l, s) patient visits when there is no preventive care (and therefore, on average, proportion π of a given physician's patients return for a follow-up visit), then the provision of preventive care is desirable if

$$\begin{aligned} 0 \quad &\leq \quad C(\gamma_j l + \rho s) - C((\gamma_j + \alpha)\frac{l}{1+\pi} + \gamma_j \frac{(\pi - q)l}{1+\pi} + \rho s) \qquad (3.28) \\ &\iff \quad \alpha \leq q\gamma_j. \end{aligned}$$

Proposition 24 *If it is socially desirable for physicians to undertake preventive care, then there exists price $p_r \geq 0$ such that the extra payment for preventive care induces fee-for-service physicians to undertake additional effort and supply these extra services.*

Proof. A fee-for-service physician supplies preventive care if

$$\begin{aligned} &V(p_l l + p_s s) - V((p_l + p_r)l\left(1 - \frac{q}{1+\pi}\right) + p_s s) \qquad (3.29) \\ &\leq C(\gamma_j l + \rho s) - C((\gamma_j + \alpha)\frac{l}{1+\pi} + \gamma_j \frac{(\pi - q)l}{1+\pi} + \rho s) \end{aligned}$$

Since V is increasing and continuous in p_r, there exists a payment for preventive care that will ensure that (3.29) is satisfied as an equality. Denote this price as p_r^*. Notice that p_r^* is the lowest price at which fee-for-service physicians can be persuaded to provide preventive care. Observe that the right-hand side of (3.29) must be strictly positive for preventive care to be economically desirable; consequently, only if $C(\gamma_j l + \rho s) - C((\gamma_j + \alpha)\frac{l}{1+\pi} + \gamma_j \frac{(\pi-q)l}{1+\pi} + \rho s) > 0$ when $p_r = p_r^*$ is it a Pareto-improvement to induce physicians to provide preventive care. ■

As compared to the general perception that fee-for-service physicians are not motivated to provide preventive care what the above result highlights is the fact that the price mechanism can be used to elicit the desired behaviour.

3.5.2 Capitation Contract

It is often argued that a physician paid on a capitation basis has powerful incentives for undertaking preventive care, as this will reduce the likelihood of a return visit. What this view overlooks, however, is that time spent providing preventive care reduces the physician's capacity to see additional patients; therefore, there is a trade-off between reducing the likelihood of needing to provide follow-ups (or, alternatively, of facing a clawback) versus the possible reduction in income due to rostering fewer patients since the additional time commitment of preventive care means that physicians cannot roster as many patients as before. More importantly, it seems that the argument that 'capitation means prevention, but fee-for-service does not' is somewhat misleading as one needs a defined yardstick to compare one environment against the other. Starting with the question of whether prevention is socially desirable (it may actually not always be the case!), this subsection proceeds carefully to compare the costs and the benefits of treating patients with preventive care when physicians are compensated prospectively on a capitation basis and when they are paid a fee per unit of service provided.

Proposition 25 *A capitated physician may provide preventive care when it is not socially desirable.*

Proof. Recall that it is socially desirable to implement preventive care if

$$C(\textit{no preventive care}) = C\left(\gamma_j \frac{h_1}{k_1}(1+\pi) + \rho s_k^1\right) \tag{3.30}$$

$$\geq C(\text{preventive care}) = C\left(\left(\gamma_j + \alpha\right)\frac{h_1}{k_1} + \gamma_j \frac{h_1}{k_1}(\pi - q) + \rho s_k^1\right)$$

$$\Leftrightarrow \alpha < q\gamma_j.$$

A capitated physician faces certain costs of providing preventive care - αm_1− but has an uncertain return to this effort, i.e., the benefits of the expected reduction in the number of second visits. Consequently, the more risk-averse the physician, the more likely it is that preventive care will be provided. For simplicity, assume that it is the high-productivity physician who serves a mixed caseload. In this case, a capitated high-productivity physician with caseload $(m_1, m_2) = (h_1/k_1, s_1^k)$ decides to provide preventive care if

$$\begin{aligned}
& EV(R_1 m_1 + R_2 m_2 - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\
& -EC(\gamma_j m_1 + \gamma_j(\widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho m_2) \\
\leq\ & EV(R_1 m_1 + R_2 m_2 - p_l n_1(\widehat{\widetilde{m}}_1) + p_l l_0(\widehat{\widetilde{m}}_1)) \\
& -EC\left((\gamma_j + \alpha)m_1 + \gamma_j(\widehat{\widetilde{m}}_1 - n_1(\widehat{\widetilde{m}}_1) + l_0(\widehat{\widetilde{m}}_1)) + \rho m_2\right) \\
\Leftrightarrow\ & EV(R_1 \frac{h_1}{k_1} + R_2 s_1^k - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\
& -EC(\gamma_j \frac{h_1}{k_1} + \gamma_j(\widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho s_1^k) \\
\leq\ & EV(R_1 \frac{h_1}{k_1} + R_2 s_1^k - p_l n_1(\widehat{\widetilde{m}}_1) + p_l l_0(\widehat{\widetilde{m}}_1)) \\
& -EC\left((\gamma_j + \alpha)\frac{h_1}{k_1} + \gamma_j(\widehat{\widetilde{m}}_1 - n_1(\widehat{\widetilde{m}}_1) + l_0(\widehat{\widetilde{m}}_1)) + \rho s_1^k\right).
\end{aligned}$$

The above can be re-written as

$$
\begin{aligned}
& EV(R_1\frac{h_1}{k_1} + R_2 s_1^k - p_l n_1(\widetilde{m}_1) + p_l l_0(\widetilde{m}_1)) \\
& -EV(R_1\frac{h_1}{k_1} + R_2 s_1^k - p_l n_1(\widehat{\widetilde{m}}_1) + p_l l_0(\widehat{\widetilde{m}}_1)) \\
\leq\ & EC(\gamma_j\frac{h_1}{k_1} + \gamma_j(\widetilde{m}_1 - n_1(\widetilde{m}_1) + l_0(\widetilde{m}_1)) + \rho s_1^k) \\
& -EC\left((\gamma_j+\alpha)\frac{h_1}{k_1} + \gamma_j(\widehat{\widetilde{m}}_1 - n_1(\widehat{\widetilde{m}}_1) + l_0(\widehat{\widetilde{m}}_1)) + \rho s_1^k\right)
\end{aligned}
\tag{3.31}
$$

where $\widetilde{m}_1$ ($\widehat{\widetilde{m}}_1$) is the realization of the random variable $\widetilde{M}_1(\widetilde{\widetilde{M}}_1)$, which represents the number of the follow-ups without (with) preventive services. Observe that $E(\widetilde{\widetilde{M}}_1) < E(\widetilde{M}_1)$, and therefore the left-hand side of (3.31) is always negative. In contrast, the right-hand side of this inequality can be positive or negative.

Comparing (3.31) and (3.30), it is evident that the incentives for the provision of preventive care facing a physician paid under a capitation scheme and the condition for determining whether the provision of preventive care is socially optimal do not generally coincide. In particular, risk averse physicians may provide preventive care when it is socially inefficient to do so. ■

As the capitated physician is still bearing more risk than the fee-for-service physician, however, it is still more expensive to get services provided under capitation than under fee-for-service. This is demonstrated in the next proposition.

Proposition 26 *In an environment in which physicians can provide preventive care, there subsists uncertainty with respect to their patients' need for medical care. Therefore, it is more costly to provide incentives to physicians to serve* $(\widehat{m}_1, \widehat{m}_2)$ *patients when physicians are paid under a capitation contract than it is when they are paid on a fee-for-service basis.*

Proof. Suppose (p_l^*, p^r, p_s^*) implements a fee-for-service equilibrium with caseload (l^{k*}, s^{k*}) for each practice k and an average 'virtual roster' of ϕ_1^{k*}, ϕ_2^{k*} for the fee-for-service physician, that is, the average number of patients of each type seen by the physician in practice k in the fee-for-service equilibrium: $\phi_1^{k*} = \frac{l^{k*}}{1+\pi-q}$, $\phi_2^{k*} = s^{k*}$. Define $\widehat{R}_1 = p_l^*(1+\pi-q)+p^r$ and $\widehat{R}_2 = p_s^*$, that is, $\widehat{R}_1, \widehat{R}_2$ are set equal to the expected revenue per patient of each type at the fee-for-service equilibrium. Assume that the physician paid under capitation has a roster of $\widehat{m}_1 = \phi_1^{k*}$ and $\widehat{m}_2 = \phi_2^{k*}$; this implies that the expected caseload of the capitated physician is equal to the caseload of the fee-for-service physician.

The income of a capitated physician with $\widetilde{m}_1$ return visits is $\widehat{D} \equiv \widehat{R}_1\phi_1^{k*} + \widehat{R}_2\phi_2^{k*} - \lambda_l\widetilde{n}_1(\widetilde{m}_1)+p_l\widetilde{l}_0(\widetilde{m}_1)$, where $\lambda_1 \geq p_l$. In contrast, the income of the physician paid under fee-for-service is certain, and can be expressed as $D \equiv (p_l^* + p^r)\,\phi_1^{k*} + p_l^*(\pi-q)\phi_1^{k*} + p_s^*\phi_2^{k*}$. Note that $E(\widehat{D}) = D$. Next, the first-order conditions for the fee-for-service physician, in terms of the 'virtual roster' are such that

$$\phi_1^{k*} \text{ solves } \widehat{R}_1 V'(D) - \gamma_j(1+\pi-q)C'(Z) = 0, \text{ and} \tag{3.32}$$

$$\phi_2^{k*} \text{ solves } \widehat{R}_2 V'(D) - \rho C'(Z) = 0. \tag{3.33}$$

where $Z = \gamma_j(1+\pi-q)\phi_1^{k*} + \rho\phi_2^{k*}$. In contrast, the first order conditions for the capitated physician are such that

$$\widehat{R}_1 E V'(\widehat{D}) - \gamma_j(1+\pi-q)EC'(\widehat{Z}) < 0 \text{ if } (m_1, m_2) = (\phi_1^{k*}, \phi_2^{k*}), \tag{3.34}$$

$$\widehat{R}_2 E V'(\widehat{D}) - \rho EC'(\widehat{Z}) < 0 \text{ if } (m_1, m_2) = (\phi_1^{k*}, \phi_2^{k*}). \tag{3.35}$$

Equations (3.34) and (3.35) imply that only by offering capitation fees R_1 and R_2 greater than $\widehat{R}_1$ and $\widehat{R}_2$ can the capitated physician be induced to choose a caseload $(\widehat{m}, \widehat{m}_2)$, and engage in preventive care. Consequently, preventive care can be provided more cheaply under fee-for-service than under capitation. ■

The above analysis provides useful insight into the way in which payment schemes influence the decision to provide preventive care. On the one hand it is clear that physicians paid under fee-for-service can be persuaded to provide preventive care if they are financially compensated for their efforts. On the other hand, physicians paid under capitation may over-provide these services, notably if they are very risk averse: the decision problem of the capitated physician with respect to the provision of preventive care is not directly comparable to the decision that would be taken by a risk-neutral policy-maker. Most significantly, the fact that the public sector is more able to bear risk than are private physicians means that when preventive care is socially desirable, it is cheaper to procure these services when paying physicians under a fee-for-service contract than to so under a capitation scheme.

3.6 Discussion and Conclusions

Health policy makers in many countries are struggling with the design of remuneration schemes for physicians. In many countries, efforts are being made to move from a system based primarily on fee-for-service to one based largely on capitation. This move is generally presumed to be cost-containing, since physicians are paid prospectively per patient. Researchers have drawn attention to possible

differences in the quality of services provided under capitation and fee-for-service payment schemes, and many have concluded that the optimal system is, therefore, a blended (mixed) payment system which combines elements of both fee-for-service and capitation. A mixed payment mechanism is believed to mitigate the negative results associated with traditionally designed fee–for-service and capitation schemes: induced demand for unnecessary services under a fee-for-service mechanism and incentives for cream-skimming and quality stinting - under capitation.

This paper takes a different approach, and demonstrates that — at least under certain circumstances — either of these two schemes can be designed to provide the same health service outcomes if a more sophisticated pricing scheme is implemented. The analysis stresses the principle that payment rates should reflect the heterogeneity of patients' medical needs. In particular, there is good reason to believe that many of the egregious features of the current health care system in Canada are a direct consequence of poorly designed payment schemes. The fact that doctors who are paid under a fee-for-service scheme schedule ten-minute consultations for all patients, regardless of the complexity of their health problems, is a consequence of the fact that existing fee-for-service schedules do not provide physicians with appropriate incentives to offer longer visits to patients in poorer health. Doctors would in fact choose to provide longer consultations if the fee-for-service schedule were adjusted appropriately. Similarly, fee-for-service physicians will engage in preventive care if they are compensated for their efforts.

The second message of the analysis is that since individual physicians are less

able to bear risk than is the public sector, then whenever there is uncertainty about the costs of meeting the needs of a particular patient it will be less costly to meet patients' needs when physicians are paid under a fee-for-service scheme than when they are paid on a capitation basis. Importantly, this means that if preventive services are socially desirable, then it is cheaper to deliver these services on a fee-for-service basis than under capitation. Moreover, the analysis provides insight into the incentives that a physician paid under a capitation scheme faces when considering the provision of preventive care. The physician's decision depends entirely on whether the benefits to the physician of a reduction in the expected number of return visits outweigh the physician's certain increase in costs due to this extra effort. In particular, a highly risk-averse physician may over-provide preventive services.

Some of the readers may object to one of the key assumptions underlying this analysis, namely that the demand for physician services is exogenously determined given the literature on physician-induced demand. Whereas such an assumption may be overly heroic in countries with a large stock of physicians, it seems very reasonable in countries such as Canada where there are serious shortages of primary care providers, and retiring physicians have difficulties finding anyone to take over their practice. In such an environment there is no reason to believe that doctors have any incentive to encourage return visits that are not actually medically necessary. Moreover, with looming retirements of primary care providers and limited numbers of new family physicians entering the labour force in many countries, this

story may soon become the rule rather than the exception.

Another feature of this analysis is that the risk-type of the patient, as well as the services provided (complex versus straightforward assessment) are both assumed to be costlessly observable. Although these are strong assumptions, they are nonetheless defensible. In particular, rules concerning liability mean that physicians are required to keep careful charts. Chart audits enable a third party to verify whether or not the care provided for a given patient was appropriate in view of key descriptors of the patient's state of health. Additionally, the development of electronic medical records means that this information will be more readily available, including outside a given physician's clinic.

An obvious wrinkle that has been neglected in this analysis is that of the importance of continuity of care: patients requiring a second visit are indifferent between seeing the same physician twice, or two different doctors. The medical literature has shown that continuity of care is an important determinant of patient satisfaction, and of overall quality of care. It would be worthwhile incorporating considerations of continuity of care into this model. It can be expected that at a Pareto efficient outcome it would no longer be the case that the case load should be shared equally across all physicians of the same type: some physicians - whose patients had a higher-than-average demand for follow-up visits - should see more patients than others. Whether or not a fee-for-service payment system can be designed to procure the desired medical services more cost-effectively than a capitation-based system remains an open question.

REFERENCES

Allen, R. and P. Gertler (1991) Regulation and the provision of quality to heterogeneous consumers: the case of prospective pricing of medical service, *Journal of Regulatory Economics*, Vol.3 pp.361-375.

Barham, V., Wooders, M. H. (1998) "First and Second Welfare Theorems for Economies with Collective Goods" In *Topics in Public Finance* by D. Pines, E. Sadka, and I. Zilcha, Eds., Cambridge University Press, pp. 57-88.

Chalkley, M., and J. M. Malcomson (1998) Contracting for health services when patient demand does not reflect quality, *Journal of Health Economics,* Vol.17 pp. 1-19.

Eggleston, K. (2000) Risk selection and optimal health insurance-provider payment systems, *Journal of Risk and Insurance,* Vol. 67(2) pp. 173-196.

Eggleston, K. (2005) Multitasking and mixed systems for provider payment, *Journal of Health Economics,* Vol. 24 pp. 211-223.

Ellis, R. P. (1998) Creaming, skimping and dumping: provider competition on the intensive and extensive margins, *Journal of Health Economics,* Vol. 17(5) pp. 537-555.

Ellis, R. P., and T. G. McGuire (1986) Provider behavior under prospective reimbursement, *Journal of Health Economics,* Vol. 5 pp. 129-151.

Ellis, R.P., and T. G. McGuire (1990) Optimal payment systems for health services, *Journal of Health Economics,* Vol. 9 pp. 375-396.

Evans, R. G. (1974) "Supplier-induced Demand: Empirical Evidence and Implications" In *The economics of health and medical care by* M. Perlman, Ed., Amsterdam: North Holland, pp. 162-173.

Léger, P.T. (2008) "Physician Payment Mechanisms" In *Financing Health Care: New Ideas for a Changing Society* by M. Lu and E. Jonsson, Eds., Wiley-VCH Press, pp. 149-176.

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

McGuire, T. G. (2000) "Physician Agency" In *Handbook of Health Economics* by A. J. Culyer, and J. P. Newhouse, Eds., Vol.1, Amsterdam: Elsevier, pp. 462-535.

Newhouse, J. P. (1996) Reimbursing health plans and health providers: efficiency in production versus selection, *Journal of Economic Literature,* Vol. 34 pp. 1236-1263.

Pauly, M. V. (1980) *Doctors and their workshops: economic models of physician behaviour*, Chicago: University of Chicago Press.

Pauly, M. V. (1984) Is cream-skimming a problem for the competitive medical market? *Journal of Health Economics*, Vol. 3 pp. 87-95.

Scotchmer, S., (1994) "Public Goods and the Invisible Hand" In *Modern Public Finance* by J. M. Quigley and E. Smolensky, Eds., Cambridge, Massachusetts, and London, England: Harvard University Press, pp. 93-119.

Shimmura, K. (1988) Effects of different remuneration methods on general medical practice: a comparison of capitation and fee-for-service payment, *International Journal of Health Planning and Management*, Vol. 3 pp. 245-258.

Stearns, S., B. L. Wolfe, and D. A. Kindig (1992) Physician responses for fee-for-service and capitation payment, *Inquiry*, Vol. 29(4) pp. 416-425.

Wooders, M. (1978) Equilibria, the core, and jurisdictions structures in economies with a local public good, *Journal of Economic Theory,* Vol. 18 pp. 328-348.

CHAPTER 4

Comparative Efficiency Assessment of Primary Care Models Using Data Envelopment Analysis

4.1 Introduction

This paper[1] undertakes an efficiency comparison of four distinct models of primary health care service delivery in Ontario using the methodology of Data Envelopment Analysis (DEA): fee-for-service practices including Family Health Groups (FFS/FHGs), health service organizations (HSOs), family health networks (FHNs), and community health centres (CHCs). The analysis draws on data collected between 2005 and 2006 as part of a multidisciplinary project funded by the Ontario Ministry of Health and Long-Term Care entitled "Comparison of Models of Primary Health Care in Ontario", further, CoM. A comprehensive study of the performance of these models has been lacking (Muldoon *et al.*, 2006). This paper is the first attempt to compare the efficiency outcomes of four distinct models of primary care in Ontario.

Previous studies of efficiency in the provision of primary care have employed the number of visits and tests as an intermediate output measure (Huang and McLaughlin 1989, Andes *et al.* 2002, Linna *et al.* 2003, Kirigia *et al.* 2004, and Rosenman and Friesner 2004). Relatively few papers (Salinas-Jimenez and

[1] A shorter version of this essay has been prepared for publication and is co-authored by Rose Anne Devlin, Vicky Barham, William Hogg, Simone Dahrouge and Grant Russell. Olga V. Milliken is the leading author of this publication.

Smith 1996, and Wagner *et al.* 2003) have incorporated measures of quality of care into the measurement of efficiency. In contrast, this paper integrates several performance indicators measuring health service delivery and quality of care, and evaluates their cumulative influence on efficiency of primary care practices.

In Ontario, the longest standing model with the largest number of participating physicians is the fee-for-service model in which physicians are paid on a per-service basis with relatively flat fee schedule across services. Since physician income increases with the quantity of services provided under this payment model, and because physicians are better able to evaluate the patient's health care needs than are patients themselves, some claim that FFS physicians will be volume-driven (Evans 1974, Pauly 1980, Arrow 1986, and McGuire 2000). The Cochrane Review (Gosden *et al.* 2000) of the empirical literature concludes that fee-for-service, as compared to capitation, results in more primary care visit contacts, and more diagnostic and curative services, but fewer hospital referrals and fewer repeat prescriptions. Family Health Groups (FHGs) are a new model of service delivery in Ontario that is similar to pure fee-for-service. In the CoM study traditional fee-for-service physicians and FHG physicians are lumped together and referred to as FFS. Currently there are about 7,439 pure FFS physicians in Ontario servicing some 9.2 million patients and about 2,536 FHG physicians servicing 3.7 million patients (Coulson 2005, Muldoon *et al.* 2006).

In an attempt to control health care costs and to address concerns with quality of primary care service delivery, the government has adopted other remuneration

and delivery schemes. It has long been argued (Ellis and McGuire 1986, Ma 1994, and Newhouse 1996) that when patients are rostered, and physicians paid a flat fee per patient in their practice, doctors have powerful incentives to provide services in a cost-effective manner. Equally importantly, capitation-based payment schemes provide cost predictability for public health authorities. This view underlay the introduction of health service organizations in Ontario in 1975 (Gillett *et al.* 2001). HSO physicians roster (i.e., register) their patients and are paid a monthly capitation fee, differing with age and gender. The fee is partially clawed back if a rostered patient is provided with primary care service outside the HSO. One idea behind the rostering approach was that it would foster a closer patient/physician relationship, with the capitation payment mechanism promoting preventive services and better quality of care, in general, compared to a volume-driven FFS scheme. Multidisciplinary teams, it was claimed, would be able to deliver appropriate and more cost-effective care in comparison to the FFS regime. In this light, the introduction of nurse practitioners and other health professionals, who could unburden physicians of certain tasks, was encouraged. Some nurse practitioner positions and other programs have been financed through government grants in HSOs. According to Muldoon *et al.* (2006) there are 49 HSOs in Ontario with 160 physicians serving 255,000 patients. A single HSO can have multiple geographically distinct sites of operation.

A weakness with capitation-based physician remuneration is that this encourages physicians to engage in cream-skimming, that is, to roster only those patients

whom they expect to be relatively healthy, or to offload costs to the non-capitated sector, e.g., by referring patients to specialists (Newhouse 1984, Pauly 1984, Dranove 1987, Allen and Gertler 1991, Ellis 1998, and Mulligan 2002). Mixed payment mechanisms – combining some element of fee-for-service and some element of capitation-based payment – have therefore been proposed as a means of providing physicians with balanced incentives to provide quality care to patients with a variety of risk profiles, while still providing reasonable control of the overall cost of delivery of primary health care (Ellis and McGuire 1986, 1993, Ma 1994, Ellis 1998, Chalkey and Malcomson 1998, Jelovac 2001, and Jack 2005). In Ontario, Family Health Networks (FHNs) were introduced in 2001 to provide comprehensive care for their patients 24 hours a day, seven days a week. Physicians working in these networks are paid under a blended scheme, combining a capitation fee for rostered patients with specific bonuses or fee-for-service payments that encourage doctors to provide preventive services and services additional to those specified in the roster agreement. FHN physicians receive a bonus for each new patient rostered and fee-for-service payments at a rate of 10% of the provincial schedule for most services.

Whereas each of the previous three models positively link physician income to some measure of service provision, the fourth model pays physicians a straight salary. Community Health Centres (CHCs) are community-oriented, and with a primary mandate to surmount barriers to health care by underserved populations and to address disadvantaged populations' needs. Introduced widely across On-

tario in the 1980s, these interdisciplinary teams are one-stop health care shops for patients including physicians, nurse practitioners, nurses, physiotherapists, chiropodists, social workers and other health and community health professionals. Economic theory suggests that salaried physicians can be expected to see a low volume of patients and will therefore provide higher-quality care either in the environment where the volume of services and quality are substitutes (Laffont and Tirole 1993), or under the circumstances where an incentive pay (i.e. fee-for-service rates) directs the allocation of physician's time and effort towards some tasks which are readily observable and verifiable (e.g. volume of services) and away from others which are difficult to monitor and quantify, e.g. quality care (Holmstrom and Milgrom 1991). Empirical studies included in the Cochrane Review (Gosden *et al.* 2000), as well Sorensen and Grytten (2003) (focusing on Norway), and Devlin and Sarma (2007) (focusing on Canada) show that salary payments are associated with a lower number of patient visits and tests compared to fee-for-service contracts, and that switching from a salary contract to a fee-for-service contract would increase the number of visits and tests. These authors do not find that an increased number of visits/procedures is associated with better patient care. Yalnizyan and Macdonald (2005) report on several studies supporting the view that CHCs provide higher quality care than FFS. Proser (2005) cites several U.S. studies, which found evidence that community health centres improve management of chronic conditions and birth outcomes.

Physicians in Canada continue to be encouraged to move into new organiza-

tional structures with non-FFS payment mechanisms and new forms of service delivery, such as multidisciplinary teams with more comprehensive arrays of services, and group practices with patient and after-hours service sharing.[2] The trend of moving towards capitation and incentive payment schemes is also prevalent in many countries with public health care systems. In this view, a comprehensive analysis of different organizational forms of physician practices and accompanying payment mechanisms seem to be of particular significance for meaningful policy-making.

4.2 Methodology

DEA is a nonparametric linear programming technique (Farrell 1957, and Charnes *et al.* 1978). It measures relative efficiencies, rather than absolute efficiencies, as it judges performance relative to others in the sample, and not against a theoretically constructed absolute measure of efficiency. The DEA approach permits the evaluation of practice sites' performance by a single efficiency score which takes account of the variety of different outputs produced, the quality of those outputs, and the inputs used. The definition of efficiency underlying DEA is illustrated in figure1, for a production technology consisting of two outputs and one input. The area south-west of GCDE in figure1 indicates all possible combinations of output 1 and 2 per unit of the input. Empirically constructed on a

[2] See Canadian Institute for Health Information (CIHI) on the status of alternative payment programs for physicians in Canada 2003-2004. Also, the Government of Ontario recently introduced a new primary care model – Family Health Teams (FHTs), which consist of physicians, nurse practitioners, nurses, dieticians, pharmacists, social workers and other health providers, who will provide comprehensive care seven days per week. By the end of 2007-08, 150 FHTs are planned to be fully operational across province of Ontario (Government of Ontario, 2007).

particular sample of practice sites, the efficiency frontier GCDE represents points with maximal output combinations given one unit of input. Practice sites C and D are located on the efficiency frontier GCDE, hence they receive a score of 1. Practice site A does not lie on the frontier and, thus, is inefficient relative to C and D. The inefficiency of site A is measured by the ratio OA/OB, and is less than 1.

DEA allows several input and output variables, with different units of measurement, to be combined. Thus, distinct output measures, such as performance indicators (for quality of care and health service delivery), service volume and intensity, can be incorporated into the analysis. Additionally, the DEA method does not require weights to be assigned *a priori* to each input and output, instead weights are assigned by the DEA program to present every practice site in the best possible light against the others. Moreover, DEA does not require the specification of a production or a cost function. Finally, unlike parametric estimation procedures, the technique is only moderately vulnerable to the sample size. The DEA method has been extensively used for estimating efficiencies (see Hollingsworth, 2003, for review), and to the best of the author's knowledge, there are no theoretical grounds to prefer any one parametric or non-parametric method of efficiency measurement over another in the health care sector (Giuffrida and Gravelle 2001).

However, DEA is not without drawbacks. Because the efficiency frontier is constructed from the sample data, it is vulnerable to data inconsistencies, outliers, or possible errors. Moreover, in the absence of time-series data for the same sample, the non-parametric nature of the method does not permit a disentangling of real

efficiency from random fluctuations. In addition, the choice of inputs and outputs is mostly *ad hoc*; researchers have to rely largely on their understanding of the production processes of the industry in the study.

The output from the DEA is an array of efficiency scores which may be used to rank the practices within each model and across models; for the most part, the analysis of this paper concerns itself with this latter ranking. Importantly, these scores depend critically upon the variables used to measure inputs into production and the outputs. In the context of this paper, efficiency scores may vary because of variation across practices in the mix of patients treated. Specifically, practices which specialize in high-need patients may operate differently than otherwise. If so, then simply measuring inputs and outputs without regard to the patients being treated will yield "efficiency" scores that are not very meaningful from a policy perspective. To address this problem, a two-stage procedure is used. First, efficiency scores employing the DEA method are calculated and then, these scores are used to compare the performance across models. Second, a variety of patient characteristics are utilized to examine whether they can explain the variation in the inter-model efficiency scoring.

4.3 Data and the choice of input and output variables

The cross-sectional data used in this paper were collected under the CoM project over the period 2005-2006. 137 practice sites in Ontario were chosen randomly, stratified by model type. Information on the recruitment and representativeness of the practices is presented in Hogg et al. (2008). This paper is based

on an analysis of 109 practice sites from the four models under study: 19 CHCs, 27 FHN sites, 32 FFS sites and 31 HSO sites. 28 practice sites were excluded from the total study sample due to the lack of expenditure and/or patient data or inconsistencies in the reported variables.

As the DEA approach requires that primary care practice sites be as homogeneous as possible, only the clinical primary care component of the services provided is considered. For the purpose of this study, clinical primary care is defined as one-on-one encounters with a physician, a nurse practitioner, a registered practical nurse, a nurse or a nursing assistant for the purpose of clinical medical care. While clinical primary care services are virtually the only services provided in HSOs, FHNs and FFS, CHCs provide additional services. The most challenging task was to disentangle the cost of the clinical services component from the additional services provided by CHCs, such as group activities, outreach community services, counselling and education. Each CHC was requested to provide the personnel data and annual operations expenditure data attributed directly to clinical primary care services. Capital costs were distributed to the clinical primary care component proportionally to the office space occupied by the clinical primary care service unit; while overhead costs, including administration expenditures, were distributed proportionately to the clinical primary care budget.

The primary care practice sites in this study consist mostly of family physicians with few specialists present. Both types of physicians are addressed as physicians in this paper. The term 'provider' refers to both physicians and nurse practitioners

(NPs). To the best of author's knowledge no consensus is found in the health care literature with regard to the substitutability of doctors and nurse practitioners (Laurant *et al.* 2005). The DEA results presented here are calculated under the assumption that 50 per cent of NP's work is equivalent to that of a physician, thus, in the calculations one NP equals one-half of a physician. Although the ranking of the different practice sites change slightly under alternative assumptions regarding this substitutability ratio, the collective ranking of the models does not.

The number of patients in the practice site relies on numbers that are reported by the practice site as having been seen at least once within the past year. Practice site annual costs are comprised of physician incomes, the salaries of medical and administrative personnel, operating costs, and maintenance and capital costs (including rent and depreciation of capital assets). Output measures include the average number of visits per patient in the practice site, and performance indicators measuring technical quality of care and health service delivery. These indicators are based on the best practices reported in the health care literature (Shi *et al.* 2001) and on established guidelines; they have been calculated by researchers involved in the CoM project using the data extracted from patient's charts and patient surveys. The calculation procedure is available in Hogg *et al.* (2008). Table 1 provides the definition of the variables used in the analyses.

Table 2 provides some descriptive statistics of the output and the input variables. Notice that the figures in table 2 are not adjusted for patient characteristics which can potentially influence outcomes in primary care service delivery; these are

adjusted at the second stage of the analysis. CHCs have the highest means for four out of the seven output variables when compared to the other three models (table 2). The Wilcoxon-Mann-Whitney test confirms that the CHCs are statistically different from the other models in health promotion, chronic disease management, comprehensiveness and visits per patient.[3] In addition, notice that the means for the CHCs are not the lowest across models except for the continuity of care indicator. Moreover, the prevention scores for CHCs are not statistically different from those of FHN sites[4], which have the highest mean, or from FFS sites[5], which have the second highest mean. Overall, a perusal of the outcome means would suggest that the CHC model performs well in terms of quality of care and performance measures; and that on average CHC providers see their patients as often as do providers in FFS sites[6]. The HSO model is statistically different from every other model in both continuity and access to primary care services scores and has the highest average for these variables[7]. FHNs and CHCs lead in delivery of preventive services.

---

[3] The Wilcoxon-Mann-Whitney test, a non-parametric version of the independent samples t-test does not assume that the dependent variable is normally distributed. The results are: for health promotion, $z = -2.90$, p-value $= 0.0037$; for chronic disease management, $z = -2.53$, p-value $= 0.011$; for comprehensiveness, $z = -3.47$, p-value $= 0.0005$; for visit per patient, $z = -3.22$, p-value $= 0.0013$.

[4] According to the Wilcoxon-Mann-Whitney test $z = -0.223$, p-value $= 0.823$.

[5] According to the Wilcoxon-Mann-Whitney test $z = 1.131$, p-value $= 0.258$.

[6] The result of the Wilcoxon-Mann-Whitney test indicates no statistical difference between the CHC and the FFS model in average visits per patient ($z = 0.789$, p-value $= 0.430$).

[7] For continuity of care performance a pair-wise Wilcoxon-Mann-Whitney test concludes that HSO and FFS sites are different ($z = -2.449$, p-value $= 0.0143$), HSO and FHN sites are different ($z = -2.207$, p-value $= 0.027$), and HSO and CHCs are different ($z = -3.641$, p-value $= 0.0003$). For access to primary care service a pair-wise Wilcoxon-Mann-Whitney test reveals that HSO and FFS are different ($z = -5.735$, p-value $= 0.000$), HSO and FHN sites are different ($z = -4.242$, p-value $= 0.000$), and HSO and CHCs are different ($z = -4.599$, p-value $= 0.000$).

Total practice site costs per provider, practice site costs (excluding physicians' income) per provider, total practice site costs per patient, and provider-patient ratio are input variables used in the DEA (table 2). In general, the mean values for these variables are the lowest in the FFS model and the highest for the CHC model. These differences are statistically significant.

4.4 DEA Scenarios and Results

The practice output comprises of seven output variables described in table 2. Three different input scenarios are estimated, reflecting each input variable. Also, the sensitivity of the model's ranking to the exclusion or inclusion of physician income in the cost data is examined.

Table 3 synthesizes the efficiency scores by presenting the number of practice sites found at decile intervals from 0 to 1, where 1 is the most efficient practice site. There is clearly significant variation across models and within a model; some differences are discernible for the input scenarios.

4.4.1 Input Scenario 1: Total Cost per Provider

In scenario 1, with the total cost per provider as the input measure, the mean of the resulting efficiency scores is 0.6 (standard deviation: 0.17). The distribution of these efficiency scores by quartiles is shown in figure 2a. The first quartile shows the highest efficiency scores and the fourth quartile the lowest; each bar indicates the percentage of the total number of practice sites for each model in a respective quartile. Observe that the FFS model has the highest representation

in the first and the second quartile compared to the other models. 40 percent of CHC practice sites are located in the lowest quartile, while another 37 per cent of its sites are in the third quartile. The efficiency scores of both the HSO and FHN sites are more evenly distributed across all quartiles in comparison to FFS and CHCs. However, the FHN model appears to perform better in terms of efficiency scoring than the HSO model: FHN sites have higher representation in the first and the second quartile, whereas HSO sites – in the third and the forth quartile. The explanation for the models' ranking is that the correlation between the total cost per provider variable and the resulting efficiency score variable is very high - 87 per cent. This means that the input variable is driving the results: CHCs have, on average, the highest total cost per provider and FFS sites have the lowest total cost per provider.

In this scenario the influence of physicians' incomes on efficiency score ranking is investigated. As only about one half of physicians in each practice site were approached to self-report their annual before-taxes, there was some concern regarding the accuracy of the physician income data for the FFS, FHN and the HSO models. Given the potential biases, physicians' income were excluded from the cost data. This worsens the ranking of the CHC practices (see figure 2b), suggesting that the low CHC ranking are driven by costs other than physician salaries.

Another potentially useful way of presenting the results from the DEA analysis is to examine the characteristics of the 10 per cent highest-ranked and 10 per cent lowest-ranked practice sites. Table 4 presents this information for the total-cost-

per-provider scenario. The FFS model is represented most in the top decile, while no CHCs are found in this group. CHC and HSO sites are represented most in the bottom decile. Comparing the top and bottom deciles, high efficiency practice sites have more visits per patient per year, undertake more preventive measures, are better in chronic disease management, promote health more actively, and provide more comprehensive care. Efficient practice sites have fewer nurses and fewer nurse practitioners than the least efficient group, and are more likely to be in an urban area than in a rural one. Top performers also employ fewer administrative personnel. Interestingly, however, is that there are almost no differences with respect to continuity of care, access, and the number of physicians in the practice.

4.4.2 Input Scenario 2: Total Cost per Patient

The total cost per patient is used as the input measure in scenario 2. The resulting DEA efficiency scores have a mean of 0.44 (standard deviation: 0.23). The top performer in this scenario is the FFS model with 40 per cent of its sites in the first quartile and over 70 per cent in the first two quartiles, followed by the HSO model with 58 per cent of all sites in the first two quartiles (see figure 3). The ranking of CHC practice sites is worsened: almost 80 per cent of sites located in the lowest efficiency score quartile are CHCs. This is explained by the fact that the total cost per patient in CHCs is significantly higher, on average, than in other models and CHC providers have on average smaller patient loads. Furthermore, the correlation between total costs and total patient roster sizes is the lowest for the CHC model: in other words, it is not the number of patients that is driving

up the costs of CHCs.[8]

4.4.3 Input Scenario 3: Provider per Patient

Using the provider-per-patient ratio as the input measure, the average efficiency score for the models is 0.41 with a standard deviation of 0.21. The best performers are HSO and FFS models, which are well represented in the first and the second quartile (see figure 4). The worst performer is again the CHC model, which is highly represented in the last quartile and is least represented in the first two quartiles.

The efficiency score ranking of this scenario reinforces the findings just described: CHCs have a relatively high provider/patient average compared to that of other models. The FFS and HSO models have relatively low means in this category. Again, the efficiency ranking is largely determined by the input variable: the correlation coefficient between the provider/patient variable for the whole sample and the resulting efficiency score is high at 79 per cent.

The comparison of the described above scenarios indicates that the FFS model is the most efficient performer, dominating the first and the second quartiles in scenarios 1 and 2 and sharing this top position with the HSO model in scenario 3. The HSO model is the next best performer in scenarios 2 and 3, but not in scenario 1 where the FHN model shines. The worst performer is the CHC model in all three scenarios.

[8] For the CHC model, the correlation coefficient is 21.8 per cent; it is 73.8 per cent for the FFS model, 81.7 per cent for the FHN model, and 80 per cent for the HSO model.

4.5 Controlling for Organizational and Patient Characteristics

Variation in the efficiency scores can be caused by characteristics of the practice sites themselves as well as the characteristics of the patients served by these practice sites. Thus, the second stage of the analysis involves a regression technique where the practice sites' efficiency scores (the dependent variables) are explained by patient and other environment characteristics. The choice of regression technique varies according to the problem at hand. Linna *et al.* (2003) for instance, employed a Tobit procedure to examine the influences of a number of factors on efficiency scores in oral health care provision. Tobit is an ideal procedure whenever the data are censored at one or both ends of the distribution. In the problem considered in this paper, very few observations were found at the high end – only five observations out of 109 in scenario 1 – and no observations were found at the low end of the range. As a result, an ordinary least squares (OLS) technique is employed in the analysis that follows (using a Tobit procedure, however, makes very little difference to the results).

The explanatory variables used in the analysis are defined in table 5; their descriptive statistics are presented in table 6. The organizational structure of a practice is captured by dummy variables representing a particular model. The reference case is the FFS model. The patient profile variables are defined on the level of a practice site and are obtained from exit questionnaires collected from patients. Patient characteristics comprise of age, gender, various socio-economic and health-related characteristics. Rurality index and the practice site's experience

with the given model are added to control for practice environment. The regression results are presented in table 7 for each of the three input scenarios. The OLS model for scenario 1 has the highest explanatory power (adjusted R^2 is equal to 0.274), and it is also the scenario in which we have the most confidence as it uses data on the number of providers per practice rather than on the number of patients. The impact of several patient profile variables on efficiency changes dramatically by scenario. Because of the poor quality of the information on the number of patients in each practice upon which the last two scenarios are based, scenario 1 is the most reliable and hence the discussion below focuses mainly on this case.

Table 7 reveals that after adjusting for patient characteristics CHCs are less efficient than FFS practices in scenarios 1 and 2. However, when provider per patient is used as the input, the difference between CHCs and FFSs is negative but statistically weak (20 per cent level of significance). As before, whenever total costs are used as part of the input measure, CHCs fare very badly. HSOs are less efficient than FFS when costs are used, especially in scenario 1, whereas FHNs are less efficient than FFS in scenario 1, but not otherwise.

Regression coefficients for model dummy variables inform on the percentage difference in efficiency scores across the different models, holding constant all other influences. In scenario 1 the estimated coefficient on the CHC dummy variable is -0.194. This means that, in comparison with the FFS model, once the influences described in Table 7 are taken into account, the average CHC efficiency score is 28 per cent lower (i.e., 0.194 divided by the average FFS efficiency score of 0.691).

The average HSO score is 17 per cent lower than the average FFS practice, while the score of the FHN is 9 per cent lower.

Certain patient characteristics help to explain the efficiency scores. In scenario 1, ceteris paribus, the proportion of patients over age 65 contributes negatively to the efficiency score of the practice site, an influence which is reversed in scenarios 2 and 3. The estimated coefficient on MALE is positive (at the 12 per cent level of significance) in scenario 1, and negative and statistically significant for the other two scenarios. Notice that HSOs improve their efficiency ranking in scenarios 2 and 3 compared to scenario 1, and they are also characterized by the largest share of patients over 65 years old and of male patients compared to the other models. In contrast, the CHC model has the lowest share in these patient characteristics.

For the most part, having more immigrants among a patient population does not seem to affect efficient scores, corroborating the findings of Sarma *et al.* (2007), and consistent with the 'healthy-immigrant' effect of McDonald and Kennedy (2004) and Deri (2006). Surprisingly, the socio-economic status of patients is not found to influence the efficiency score. Having a higher percentage of patients who are unemployed does not affect very much the efficiency of practices, nor does the income and education of patients. Multicollinearity across the model dummy variables and various patient characteristics may be thwarting our attempt to tease out the effect of patient socio-economic variables in the regression analysis. The auxiliary regressions of each dummy variable on all patient characteristics, rurality index and practice year variable demonstrate that the CHC dummy is strongly as-

sociated with patients of low socio-economic background (particularly, unemployed and low income), and with female patients (all three are significant at 2 per cent level or lower, adjusted R^2 is equal to 0.42). The HSO dummy is associated with male patients, low share of immigrants, and practice year variables (all three are significant at the 4.5 per cent level or lower, with an adjusted R2 is equal to 0.22). The FHN dummy is positively associated with the number of chronic conditions per patient and practice year variable (both significant at 4 per cent level or lower, adjusted R^2 is equal to 0.39). Unfortunately, the exclusion of the variables which are correlated with the model dummies does not improve the significance level of other patient characteristics in explaining the variation in efficiency scores.

The results reported in table 7 show that the fact that the practice serves patients with particular needs does affect efficiency. The larger the proportion of patients who perceive their health as being good, the more efficient the practice site becomes (especially in scenario 1). Similarly, the larger the proportion of patients with conditions lasting more than one year, the lower the efficiency score (again, in scenario 1). Patients with several chronic conditions appear to have a positive impact on efficiency in scenario 1; however this variable is highly negatively correlated with the good health variable, so its estimated coefficient may be unstable. Eliminating both the long-term conditions and chronic conditions variables from the analysis increases the adjusted R^2 for the regression; the impact of 'Good Health' on efficiency persists.

Geographic location and other environmental characteristics captured by the

rurality index have a positive influence on efficiency scores in scenario 1. Practices in rural areas seem to be more efficient, ceteris paribus. Once again, the results for the other two scenarios are different, likely because HSOs - which rank higher in scenario 2 and 3 - are located mostly in one urban area. The age of the practice site does not seem to have a statistically significant impact on efficiency. The fact that the number of HSO practices was frozen for many years, and that the FHN model is relatively new, may mean that there is not enough variation in the practice year variable to exert a significant effect on efficiency.

4.6 Conclusion, limitations, and policy implications

This paper compares and contrasts four different models of primary care delivery in Ontario with a view to identifying, if possible, which model is the most efficient and under what circumstances. It is clear that how output and inputs are measured matters. One of the innovative elements of the data set used in the paper is its focus on both qualitative and quantitative indicators of output. And one of the clear advantages of the DEA procedure is its ability to incorporate a variety of inputs and outputs and then weigh them in order to present each practice in the best possible light.

If one looks only at the qualitative indicators of output alone, CHC practices perform reasonably well. On average, they achieve scores higher than the other three models for three of the eight performance indicators: health promotion, chronic disease management, and comprehensiveness and fare relatively well on prevention and access. The HSO model has the highest average for the continuity

of care and access to primary care services variables. FHNs are the best in terms of preventive services.

However, once costs are added to the mix in the DEA analysis, CHCs are the least efficient practice sites virtually across the board, whereas the FFS model performs the best. The efficiency scores of both the HSO and FHN sites are more evenly distributed across all quartiles in comparison to FFS and CHCs. The data show that these efficiency score rankings are driven by the costs of running the practice.

A number of reasons explain the poor efficiency scores of the CHCs. The link between performance indicators and costs may be non-linear, and it therefore may be relatively inexpensive to achieve a low level of performance, but very costly to push these indicators beyond any given threshold[9]. Practices with higher-than-average performance on quality of care and service delivery indicators may require many more resources than practices providing a lower quality of care. Another reason is related to the limitations of the study which focused on clinical primary care; thus the collected expenditure/cost data do not take account of the costs associated with shifting health care from primary care providers onto hospitals, emergency rooms, specialists and outsourced diagnostic services, as well as the costs of prescribed drugs. The costs absorbed by patients, insurance payers, provincial governments and society as a whole are likely to be significant. If the better quality care

[9] Increasing immunization from the 60 per cent level to the 80 per cent level may require additional resources beyond those that are necessary to achieve an increase from 40 per cent to 60 per cent. Home visits, additional phone calls and letters are examples of extra effort that require more resources than a single reminder from a doctor's office.

provided by CHCs were to reduce these expenses it could diminish or eliminate the cost differences found between the CHC model and the others. Also, it may well be that CHCs are too small relative to the large fixed costs necessary to operate a multidisciplinary health centre. These high up-front costs must be spread over a large patient population before the CHCs are able to operate at an average cost that is at least comparable to other practice types. Finally, it is worth pointing out that formidable challenges were encountered when trying to disentangle the primary care component from the array of other health care, community and social services provided by CHCs. It seems that CHC providers perform clinical primary care services which are not provided in other models (e.g. diagnostics, chiropody). Moreover, it is simply impossible to perfectly separate medical and administrative full-time equivalents and overhead costs associated with the clinical primary care component from other services provided within an individual CHC.

It is also interesting to ask why fee-for-service practices fare so well in the DEA analysis. Part of the story, undoubtedly, is that the broader costs of the FFS approach are borne outside of the practices themselves – costs such as the reputed over-use of specialists. Nevertheless, FFS physicians clearly face incentives to see as many patients as possible given that their remuneration depends upon the number of visits conducted per period of time, while, at the same time, they would want to minimize the costs of running their individual practices because physicians themselves are the residual claimants to the proceeds of the practice.

Overall, the efficiency scores of the FHN and HSO models lie somewhere in-

between FFS and CHC practices. In terms of quality of care, FHNs and HSOs are on average at least as good as the FFSs or better, particular, in the area of prevention – not surprisingly since physicians in these models receive financial bonuses, which vary with the achieved level of preventive services.

In terms of the costs of remunerating physicians, however, FHNs and HSOs are overall more expensive than FFS practices. This may be explained by the fact that physicians who are paid prospectively on a capitation basis have to be compensated for bearing the financial risks associated with running their practices, while CHC physicians who are also paid prospectively (on a salary base) do not bear such risk, as they do not own the centres.

The findings of this paper show clearly that practice type matters. How practices are organized and how physicians are remunerated affect the costs associated with providing patient care. However, one cannot say unequivocally that one type of primary care model dominates. In particular, further research, which better tracks the relationship between the primary care model and the use of other health system resources, is necessary in order to better understand which approach makes the best use of public resources.

Table 1: Input and output measures

Variable name	Definition
I. Output Variables	
1) Intensity of Service Utilization	
Visits/patient	Average number of visits per patient
2) Technical Quality of Care Indicators	
Prevent	Prevention score [0,1]
Chronic_Mngt	Chronic disease management score [0,1]
Health Promo	Health promotion score [0,1]
3) Service Delivery Indicators	
Access	Access to primary care services score [0,1]
Contin	Continuity of care score [0,1]
Compreh	Comprehensiveness of care score [0,1]
II. Input Variables	
Total cost per provider	Full practice site's expenditure per provider's FTE[a]
Cost per provider	Practice site's expenditure excluding physicians' income per provider's FTE
Total cost per patient	Full practice site's expenditure per patient
Provider/patient	Provider's FTEs per 1000 patients

[a] One provider's FTE is equivalent to one physician's FTE or two nurse practitioner's FTE.

Table 2: Descriptive Statistics[a]

Model	CHC		FFS		FHN		HSO	
Variable	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Output Variables								
Visits/patient[b]	5.2	*0.94*	5	*1.55*	4.06	*1.23*	3.68	*0.88*
Prevent	59	*16.6*	53.8	*14*	60.5	*14.3*	48.5	*12.6*
Chronic_Mngt	71.7	*16.3*	57.9	*20*	59.3	*19.3*	62.2	*16.2*
Health Promo	49.6	*10.8*	40.6	*10.4*	42.5	*9.7*	40.1	*12.7*
Access	3.0	*0.19*	2.94	*0.217*	3.0	*0.28*	3.3	*0.16*
Contin	3.4	*0.21*	3.5	*0.20*	3.5	*0.16*	3.6	*0.16*
Compreh	76.5	*13.1*	61.9	*16.8*	61.3	*17.4*	60.6	*15.1*
Input Variables								
Total cost per provider	390789	*67893*	250962	*61496*	292584	*76073*	313110	*74680*
Cost per provider[c]	247008	*67030*	93126	*38497*	116432	*55793*	129228	*62768*
Total cost per patient	480.6	*242.5*	172.8	*76.5*	214.5	*73.8*	199.4	*83.2*
Provider/patient	1.23	*0.58*	0.72	*0.37*	0.77	*0.3*	0.67	*0.34*

[a] The statistics are calculated on the practice site's level and then averaged across models

[b] In efficiency score calculations output measure *average visits per patient in one year* may create an unfavourable bias against sites in which physicians take longer time with a patient during a single visit rather than booking several visits. However, the comparison of efficiency scores with the "visit" variable and without it indicates no significant difference in model ranking.

[c] Practice cost excludes physicians' income

Table 3: DEA efficiency scores by input scenario and model

RANGE	SCENARIO 1					SCENARIO 2					SCENARIO 3				
	All N=109	CHC n=19	FFS n=32	FHN n=27	HSO n=31	All N=109	CHC n=19	FFS n=32	FHN n=27	HSO n=31	All N=109	CHC n=19	FFS n=32	FHN n=27	HSO n=31
1.0[a]	5	0	3	2	0	6	0	4	1	1	2	0	1	0	1
0.9-0.999	1	0	1	0	0	3	0	0	2	1	1	0	0	0	1
0.80-0.899	9	0	5	1	3	1	0	0	0	1	5	1	2	1	1
0.70-0.799	14	1	5	5	3	4	1	1	1	1	4	1	1	2	0
0.6-0.699	23	3	10	6	4	6	1	2	1	2	8	0	3	1	4
0.5-0.599	24	4	2	8	10	12	0	8	1	3	9	0	3	3	3
0.4-0.499	22	8	4	3	7	20	1	6	4	9	18	1	6	4	7
0.3-0.399	9	2	1	1	5	21	0	5	10	6	21	5	4	4	8
0.2-0.299	2	1	0	1	0	24	8	4	6	6	27	5	9	9	4
0.1-0.199	0	0	0	0	0	10	7	1	1	1	13	6	1	3	3
mean	0.604	0.504	0.691	0.628	0.559	0.438	0.255	0.520	0.439	0.467	0.410	0.310	0.439	0.387	0.460
max	1	0.779	1	1	0.857	1	0.709	1	1	1	1	0.831	1	0.848	1
min	0.289	0.293	0.393	0.289	0.359	0.081	0.081	0.196	0.198	0.125	0.115	0.118	0.118	0.167	0.115
sd	0.169	0.122	0.172	0.172	0.144	0.226	0.165	0.227	0.233	0.198	0.211	0.192	0.224	0.186	0.215

[a] A score of one indicates the most efficient practice sites.

Table 4: 10 per cent highest-ranked and 10 per cent lowest-ranked practice sites

STATISTICS	Score (Scenario 1)	Model	Visits per patient	Prevent	Chronic_Mngt	Health Promo	Contin	Access	Compreh	Physician's FTEs	NP's FTEs	Nurse's FTEs	Admin FTEs	Patients self-reported	RI[c]
TOP 10% (N=11)															
mean	0.928		5.4	63.4	55	43.1	3.48	2.98	60.7	2.45	0	0.32	1.79	4431	5.30
sd	0.072		2.0	17.5	22	13.0	0.24	0.30	15.0	0.96	0	0.64	1.12	6268	3.44
min	0.838	CHC (0%)[a]	2.7	32	19	22.2	2.85	2.42	27.0	1.00	0	0.00	0.50	1000	3.3
max	1	FFS (25%)[b]	10.2	94	81	60.9	3.71	3.42	80.0	4.00	0	2.00	4.5	23000	12.3
BOTTOM 10% (N=11)															
mean	0.362		4.21	53.6	68.5	37.9	3.56	3.05	57.0	2.55	0.59	1.43	4.19	5552	8.31
sd	0.041		1.08	12.0	16.9	6.6	0.13	0.27	11.2	2.20	1.07	1.11	3.61	5882	5.92
min	0.289	FFS (3%)	2.27	32.0	36	28.0	3.27	2.51	40.0	1	0	0	1	1500	0
max	0.398	CHC and HSO (16%)	5.7	73.0	88	47.1	3.75	3.39	73.0	7.5	3	4	12	20000	21.5

[a]min of variable *Model* denotes the model which is represented the least in the sample; the number in the brackets denotes the percentage of the model's sites among the top 10%.

[b]max of variable *Model* denotes the model which is represented the most in the sample; the number in brackets denotes the percentage of the model's sites among the bottom 10%.

[c]RI denotes rurality index reflecting population density, travel distances to secondary and tertiary hospital care, supply of family doctors and specialists, relative availability of imaging services and ambulance services, weather conditions and social indicators (Ontario Medical Association). Higher index is associated with relatively remote areas with lower level of services available.

Table 5: Explanatory variables for regression analysis

Variable name	Definition
	I. Organizational Structure
FFS	Reference case (=1 if the practice site is a CHC site, 0 otherwise)
CHC	CHC dummy (=1 if the practice site is a CHC site, 0 otherwise)
FHN	FHN dummy (=1 if the practice site is a FHN site, 0 otherwise)
HSO	HSO dummy (=1 if the practice site is a HSO site, 0 otherwise)
Practyr	The number of years in operation in this model
	II. Patient Profile (a proportion of patients with a certain characteristic to the total number of patients who filled out exit questionnaires)
	1. Age/gender profile
Age65	Proportion of patients age 65 or over
Male	Proportion of male patients
	2. Socio-economic status
Immig2	Proportion of patients-immigrants who have been to Canada for 2 or fewer years[a]
Unempl	Proportion of unemployed, excluding housewives and househusbands and those who study
LowIncome	Proportion of patients living in a household with the household income lower than $20,000 per an equivalent household member
LowEdu	Proportion of patients with education less than high school
	3. Health status of patients
GoodHealth	Proportion of patients who perceived their health being excellent or very good[b]
Cond	Proportion of patients with physical, mental or emotional condition that have lasted or are likely to last longer than one year
Chronic	The number of chronic conditions per patient (Average per practice)
	III. Environmental variable
RI	Rurality index. Higher index is associated with relatively remote areas with lower level of services available.

[a]A threshold of 5 years yields equal significance in the analysis.
[b]A variable representing the proportion of patients who perceived their health being fair or poor has not been identified as a significant predictor.

Table 6: Summary statistics for explanatory variables in regression analysis (mean, min, max, standard deviation)

Model	Age65	Male	Immig2	Unempl	Low Income	LowEdu	Good Health	Cond	Chronic	Practyr	RI
CHC											
mean	14.5	26.1	18.7	12.8	0.33	23	41.6	47	1.9	17.5	13.5
min	0	0	0	0	0.13	5.9	23.1	19	0.4	5.5	3.3
max	41.4	52.6	83.3	25.9	0.56	42.9	56.5	77.8	2.9	30	52
sd	11.9	13.2	22.4	8.2	0.15	10.5	8.5	14.8	0.5	7.3	17.1
FFS											
mean	20.2	34.6	10.4	5	0.15	14	46	42.9	1.75	16.3	13.1
min	0	7.4	0	0	0	0	26	27.6	1.2	0	0
max	43.8	100	37.1	17.4	0.37	35.6	65.7	70.4	2.98	40	76.1
sd	9.7	17.7	12.1	4.7	0.09	10	9.9	10.8	0.47	10.2	18.5
FHN											
mean	22.8	34.3	3.3	3.6	0.15	14.1	47.6	40.7	1.8	2.34	18.3
min	6.1	12.2	0	0	0	0	29.2	27	1	0.25	0.81
max	40.9	55.3	12	9.7	0.33	31.1	79.6	53.3	2.76	6	65.4
sd	9.9	13	3.7	3.1	0.09	9.5	12.6	6.8	0.45	1.17	20.8
HSO											
mean	24.8	39.8	2.91	3.43	0.15	16.4	50.7	37.7	1.68	15.3	8
min	0	25	0	0	0.02	0.03	35.3	21.2	1.11	2	0.81
max	53.1	56.8	20	12.9	0.32	37	76.5	61.5	2.52	36	42.1
sd	11.1	9.69	4.21	3.33	0.08	8.51	10.1	10	0.35	6.7	9.23
Total											
mean	21.2	34.6	7.89	5.54	0.18	16.3	47	41.6	1.78	12.7	12.9
min	0	0	0	0	0	0	23.1	19	0.41	0	0
max	53.1	100	83.3	25.9	0.56	42.9	79.6	77.8	2.98	40	76.1
sd	11	14.3	13	5.87	0.12	9.98	10.8	10.9	0.45	9.39	16.9

Table 7: The influence of patient and practice characteristics on efficiency

Explanatory variable	Scenario 1		Scenario 2		Scenario 3	
	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Constant	0.518	0.000	0.649	0.001	0.547	0.004
CHC	-0.194	0.000	-0.244	0.002	-0.091	0.197
HSO	-0.120	0.005	-0.090	0.131	-0.017	0.753
FHN	-0.063	0.197	-0.074	0.291	-0.058	0.366
Age65	-0.005	0.004	0.006	0.026	0.007	0.002
Male	0.002	0.118	-0.003	0.062	-0.003	0.038
Immig2	0.000	0.800	-0.000	0.890	0.001	0.661
Unempl	0.003	0.388	-0.004	0.425	-0.009	0.086
LowIncome	-0.011	0.954	0.152	0.584	0.118	0.647
LowEdu	-0.003	0.158	0.003	0.359	0.004	0.151
GoodHealth	0.004	0.033	0.001	0.707	0.000	0.923
Cond	-0.003	0.137	0.001	0.772	0.001	0.548
Chronic	0.083	0.119	-0.143	0.059	-0.113	0.105
RI	0.002	0.013	-0.002	0.253	-0.003	0.023
Practyr	0.000	0.884	0.001	0.689	-0.000	0.943
Adj. R^2	0.2740		0.181		0.193	

Figure 1: Efficient Frontier

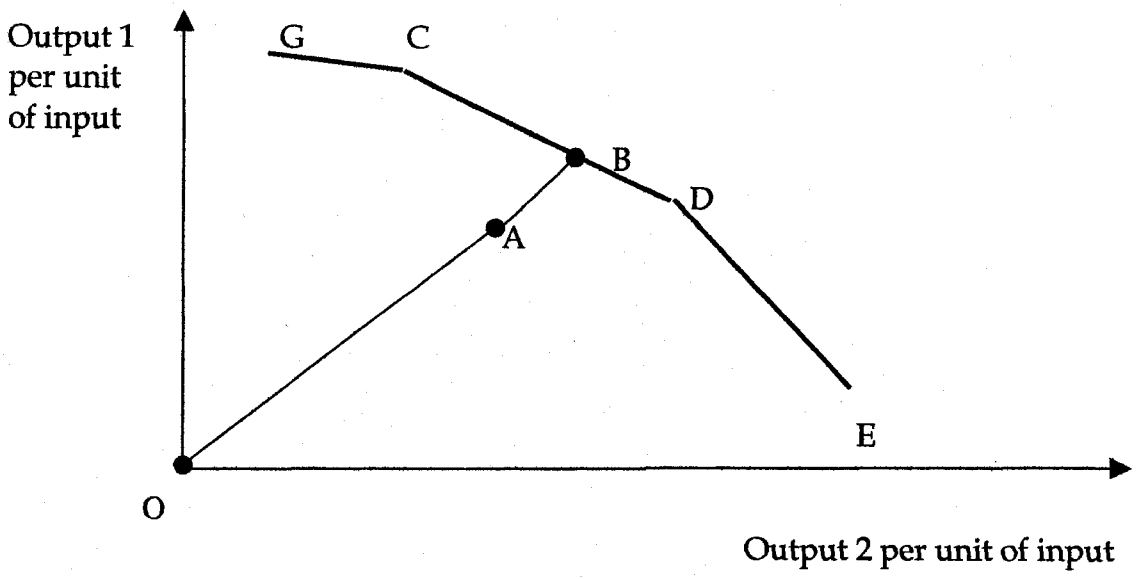


Figure 2a

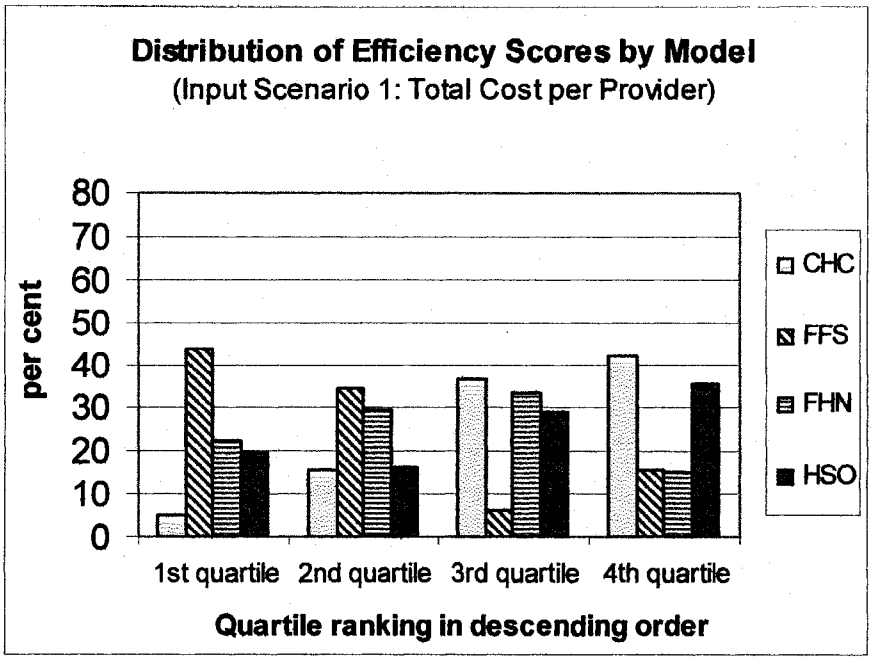


Figure 2b

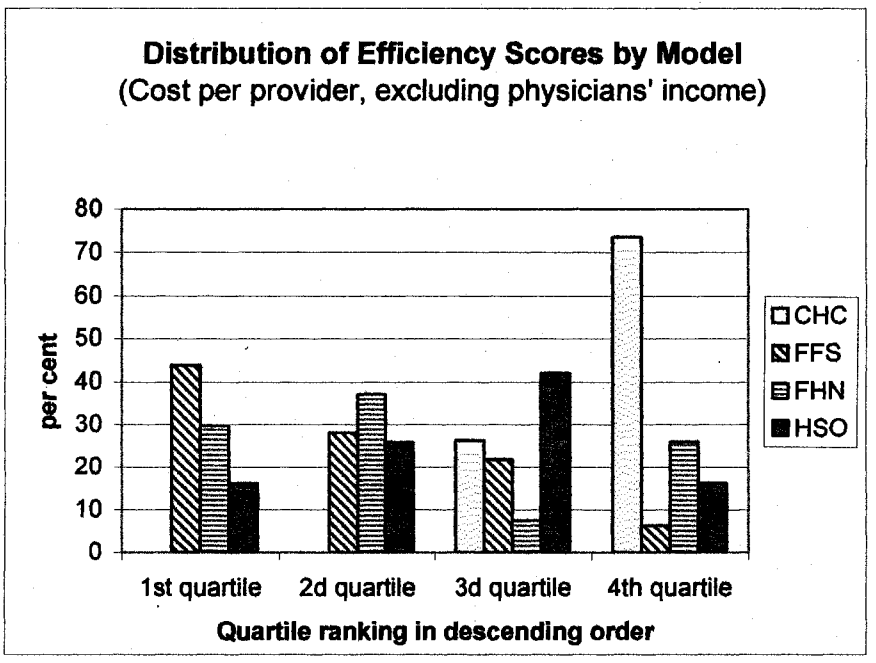


Figure 3

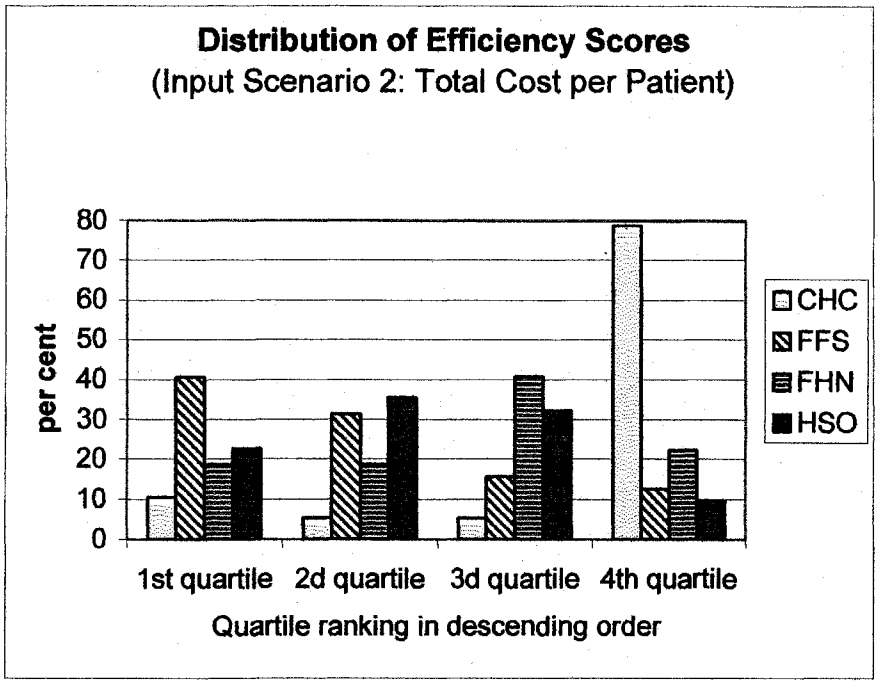
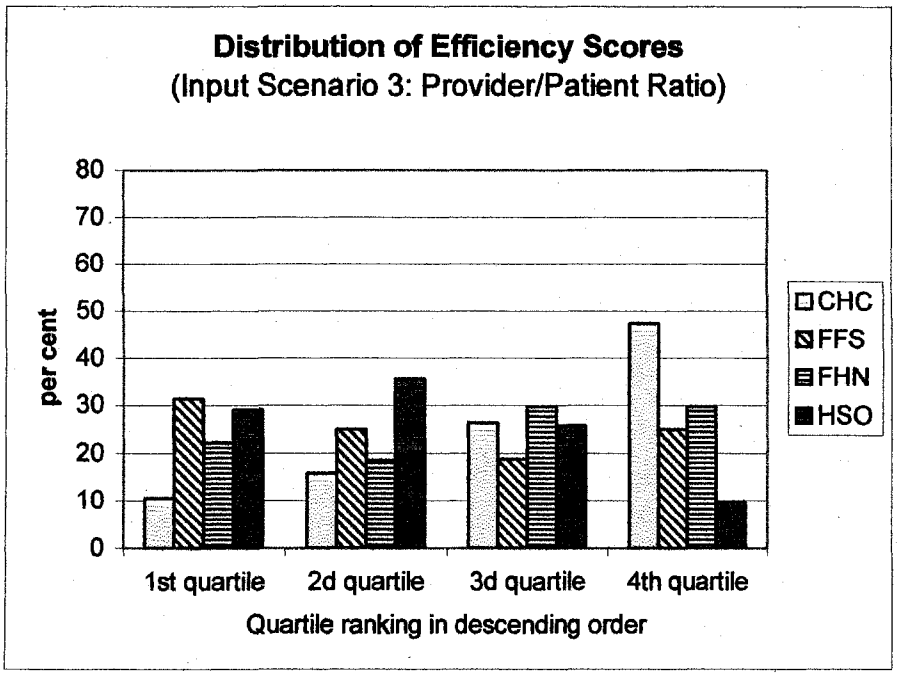


Figure 4



REFERENCES

Andes, S., L. M. Metzger, J. Kralewski, and D. Gans (2002) Measuring efficiency of physician practices using data envelopment analysis. *Managed Care*, Vol. 11(11) pp. 48-54.

Allen, R. and P. Gertler (1991) Regulation and the provision of quality to heterogeneous consumers: the case of prospective pricing of medical services, *Journal of Regulatory Economics*, Vol. 3 pp. 361-375.

Arrow, K. J. (1986) "Agency and the Market" In *Handbook of Mathematical Economics* by K. J. Arrow, and M. D. Intriligator, Eds., North Holland: Elsevier Science Publishers, Vol.3 pp.1183-1195.

Chalkley, M. and J. M. Malcomson (1998) Contracting for health services when patient demand does not reflect quality, *Journal of Health Economics*, Vol. 17 pp. 1-19.

Charnes, A., W. W. Cooper, and E. Rhodes (1978) Measuring the efficiency of decision-making units, *European Journal of Operational Research*, Vol.2 pp. 429-444.

Coulson, D. (2005) Primary health care models in Ontario. Toronto, Ontario, Coulson & Associates.

Deri, C. (2006) What lessons can we learn from the body mass of immigrants? University of Ottawa.

Devlin, R., and S. Sarma (2007) Do physician remuneration schemes matter? The Case of Canadian Family Physicians, Canadian Health Economics Study Group.

Dranove, D. (1987) Rate-setting by diagnosis related groups and hospital specialization, *RAND Journal of Economics*, Vol.18 pp. 417-427.

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

Ellis, R. P. and T. G. McGuire (1986) Provider Behaviour under prospective reimbursement: cost sharing and supply, *Journal of Health Economics*, Vol.6 pp. 129-151.

Ellis, R. P. and T. G. McGuire (1993) Supply-side and demand-side cost sharing in health care, *Journal of Economic Perspectives,* Vol.7(4) pp. 135-151.

Evans, R. G. (1974) "Supplier-Induced Demand: Empirical Evidence and Implications" In *The Economics of Health and Medical Care* by M. Perlman, Ed., Amsterdam: North Holland, pp. 162-173.

Farrell, M. J. (1957) The measurement of productive efficiency, *Journal of the Royal Statistical Society*, Series A, Vol.120(3) pp. 252-281.

Gillet, J., B. Hutchison and S. Birch (2001) Capitation and primary care in Canada: Financial incentives and the evolution of health service organizations, *International Journal of Health Services*, Vol.31(3) pp. 583-603.

Giuffrida, A., and H. Gravelle (2001) Measuring performance in primary care: Econometric analysis and DEA, *Applied Economics*, Vol.33 pp. 163-175.

Gosden, T., F. Forland, I. S. Kristiansen, M. Sutton, B. Leese, A. Giuffrida, M. Sergison, and L. Pedersen (2000) Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians, The Cochrane Database Of Systematic Reviews, (3) CD002215.

Hogg, W., S. Dahrouge, G. Russell, R. Geneau, E. Kristjansson, L. Muldoon, S. Johnston. (2008) The comparison of models of primary care in Ontario study (COMP-PC): Methodology of a multifaceted cross-sectional practice-based study, University of Ottawa.

Hollingsworth, B. (2003) Non-parametric and parametric applications measuring efficiency in health care, *Health Care Management Science*, Vol. 6(4) pp. 203-218.

Holmstrom, B. and P. Milgrom (1991) Multitask principal-agent analyses: incentive contracts, asset ownership and job-design, *Journal of Law, Economics and Organization*, Vol.7 pp. 24-52.

Huang, Y. G., and C. P. McLaughlin (1989) Relative efficiency in rural primary health care: an application of data envelopment analysis, *Health Services Research*, Vol.24(2) pp. 143-158.

Jack, W. (2005) Purchasing health care services from providers with unknown altruism, *Journal of Health Economics*, Vol. 24 pp. 73-93.

Jelovac, I. (2001) Physicians' payment contracts, treatment decisions and diagnosis accuracy, *Health Economics*, Vol.10 pp. 9-25.

Kirigia, J. M., A. Emrouznejad, L. G. Sambo, N. Munguti, and W. Liambila (2004) Using data envelopment analysis to measure the technical efficiency of public health centers in Kenya, *Journal Of Medical Systems*, Vol.28(2) pp. 155-166.

Laffont, J.-J. and J. Tirole (1993) *A Theory of Incentives in Procurement and Regulation*, Cambridge, MA: MIT Press.

Laurant, M., D. Reeves, R. Hermens, J. Braspenning, R. Grol, and B. Sibbald (2005) Substitution of doctors by nurses in primary care. The Cochrane Database Of Systematic Reviews (2), CD001271.

Linna, M., A. Nordblad, and M. Koivu (2003) Technical and cost efficiency of oral health care provision in Finnish health centres, *Social Science and Medicine*, Vol.56(2) pp. 343-353.

Ma, C. A. (1994) Health care payment systems: cost and quality incentives, *Journal of Economics & Management Strategy*, Vol. 3(1) pp. 93-112.

McDonald, J. T., and S. Kennedy (2004) Insights into the 'healthy immigrant effect': health status and health service use of immigrants to Canada, *Social Science and Medicine*, Vol.59(8) pp. 1613-1627.

McGuire, T. (2000) "Physician Agency" In *Handbook of Health Economics* by A. J. Culyer, and J. P. Newhouse, Eds., Amsterdam: Elsevier.

Muldoon, L., M. Rowan, R. Geneau, W. Hogg, and D. Coulson (2006) Models of primary care service delivery in Ontario: Why such diversity? *Healthcare Management Forum*, Winter pp. 18-23.

Mulligan, P. K. (2002) Capitation: the wrong direction for primary care reform, *Canadian Family Physician*, February.

Newhouse, J. P., (1984) Cream skimming, asymmetric information, and a competitive insurance market, *Journal of Health Economics*, Vol.3 pp. 97-100.

Newhouse, J. P. (1996) Reimbursing health plans and health providers: efficiency in production versus selection, *Journal of Economic Literature*, XXXIV, September pp. 1236-1263.

Pauly, M. V. (1980) *Doctors and Their Workshops: Economic Models of Physician Behaviour*, Chicago: University of Chicago Press.

Pauly, M. (1984) Is cream-skimming a problem for the competitive medical market? *Journal of Health Economics*, Vol.3 pp. 87-95.

Proser, M. (2005) Deserving the spotlight. Health centers provide high-quality and cost-effective care, *Journal of Ambulatory Care Management*, Vol.28(4) pp. 321-330.

Rosenman, R., and D. Friesner (2004) Scope and scale inefficiencies in physician practices, *Health Economics*, Vol.13(11) pp. 1091-1116.

Salinas-Jimenez, J., and P. Smith (1996) Data envelopment analysis applied to quality in primary health care, *Annals of Operational Research*, Vol.67 pp. 141-161.

Sarma, S., R. Devlin, and W. Hogg (2006) Physician production of primary care in Ontario: Evidence from the 2004 National Physician Survey, Canada. Canadian Health Economics Study Group.

Shi, L., B. Starfield, and J. Xu (2001) Validating the Adult Primary Care Assessment Tool, *Journal of Family Practice*, Vol.50(2), E1.

Sorensen, R. J., and J. Grytten (2003) Service production and contract choice in primary physician services, *Health Policy*, Vol.66(1) pp. 73-93.

Wagner, J., D. Shimshak, and M. Novak (2003) Advances in physician profiling: the use of DEA, *Socio-Economic Planning Sciences*, Vol.37 pp. 141-163.

Yalnizyan, A., and D. Macdonald (2005) CHC cost effectiveness: a review of the literature. Association of Ontario Health Centres, June.