Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Health Economics 24 (2005) 775–793

JOURNAL OF HEALTH ECONOMICS

www.elsevier.com/locate/econbase

# Information asymmetry, insurance, and the decision to hospitalize

Åke Blomqvist[a], Pierre Thomas Léger[b,c,*]

[a] *Department of Economics, National University of Singapore, Singapore*
[b] *HEC Montréal, 3000 Côte-Ste-Catherine, Montréal, Qué., Canada H3T 2A7*
[c] *CIRANO and CIRPÉE, Montréal, Qué., Canada H3T 2A7*

Received 1 June 2003; received in revised form 1 August 2004; accepted 11 December 2004
Available online 10 February 2005

## Abstract

We analyze the problem of second-best optimal health insurance in the context of a model in which patients and doctors must decide not only on an aggregate quantity of health services to use in treating various kinds of illness, but also have a choice between different kinds of providers (in particular, outpatient services rendered by primary-care physicians or inpatient services provided by hospital-based specialists). We consider well-informed patients' choices of provider when they have conventional insurance so they only pay part of the cost of their health services, as well as the equilibrium strategies of doctors and patients when there is patient-provider asymmetry; in the latter case we also analyze a managed-care insurance setup under which doctors are paid by capitation. We find that under certain plausible conditions, second-best optimal managed-care plans with supply-side incentives dominate second-best optimal conventional plans that rely on cost control through demand-side cost sharing.

*JEL classification:* I1

*Keywords:* Primary care; Specialist care; Hospitalization; Insurance; Capitation

[*] Corresponding author.
*E-mail address:* pierre-thomas.leger@hec.ca (P.T. Léger).

0167-6296/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2004.12.001

## 1. Introduction

Much of the health economics literature has focussed on the effects of different payment mechanisms and insurance schemes on the utilization of medical services. In the presence of conventional service benefit insurance, individuals will want to use medical care beyond efficient levels (the traditional moral hazard problem). Furthermore, physicians that are paid for each service they provide (in a fee-for-service system) may not only be willing to supply inefficiently large volumes of care, but may also have incentives to encourage utilization (the problem of supplier-induced demand). In order to reduce the problems associated with moral hazard and supplier-induced demand, insurers have used demand-side incentives (such as patient cost-sharing through co-insurance and deductibles), as well as supply-side incentives aimed at providers (such as paying physicians through salary or capitation, or hospitals through episode-based prospective reimbursement).

Formal models dealing with these issues have generally been specified so as to involve only one type of medical care (see, for example, Blomqvist (1991), Ma and McGuire (1997), Ellis and McGuire (1986)). That is, they have abstracted from the fact that the health services sector produces many types of care, using a variety of different kinds of inputs. For example, certain kinds of disease may be treated through a combination of physician services and pharmaceuticals. In other cases, there may be substitutability between outpatient services provided by primary-care physicians and services provided by hospital-based specialists.[1] Although the latter may be necessary for individuals that suffer from particularly complex and severe forms of illness, excessive use of specialist and hospital care may be inefficient and inappropriate. First, for certain kinds of illness, primary-care physicians may be able to provide treatment at lower cost. Furthermore, specialist in-hospital care is more likely to be invasive and risky, and thus should only be used when medically warranted (Frank and Clancy, 1997). Providing incentives and information to ensure that patients and doctors use the appropriate type of care is thus important both from a health perspective, and for economic reasons. In this paper, we analyze the effects of various kinds of demand- and supply-side incentives in the context of a model in which patients and doctors must consider not only the aggregate quantity of health services to use in treating various kinds of illness, but also have a choice between different kinds of providers, in particular, outpatient services rendered by primary-care physicians or inpatient services provided by hospital-based specialists.

Although theoretical work on the economics of referrals to hospitals and specialists is limited and quite recent, there is a growing empirical literature that has examined physician referral patterns.[2] Overall, 4.5% of visits to primary-care physicians in the US result in a referral (Frank and Clancy, 1997).[3] Furthermore, although hospital admissions are rela-

[1] For example, certain forms of progressive heart disease and rheumatoid arthritis may be treated through non-surgical and surgical means.

[2] Theoretical work on referrals include Shortell (1972), Bradford and Martin (1996), Glazer and McGuire (1992) as well as related work by Pauly (1979) and Wolinsky (1993). An interesting recent contribution is Mariñoso and Jelovac (2003); their focus is on the importance of accurate diagnosis by GPs in determining the appropriateness of the referral decision.

[3] Based on American survey data from the National Ambulatory Medical Care Survey (NAMCS) for the years 1985–1992.

tively rare (approximately 10% of individuals in the Rand Study experienced one or more admissions in a year), hospitalization episodes are very costly so that the cost of hospital care accounts for a large portion of health care costs.[4] On average, each referral results in US$3,000 in hospital charges and professional fees (Glenn et al., 1987).

Potentially important factors that may influence the use of specialist and hospital care include whether or not patients are allowed to seek such care on their own (that is, without a referral from a primary-care provider). Although many health-care systems and managed-care plans prohibit patient self-referrals to specialist care and in-hospital care, others do not, and it has been estimated that in the US, 30–50% of all specialist consultations take place as a result of self-referrals (Forrest and Reid, 1997). In American managed-care plans, a common device for affecting the use of hospital services is to require a 'second opinion' before approval is given for hospitalization. This may be one reason that HMO patients are less likely to be hospitalized compared to their non-HMO counterparts (Newhouse, 1993).

Empirical work provides some evidence that the rate of hospitalization is influenced by incentives both on the supply side and on the demand side. With respect to supply-side factors, there is evidence to suggest that primary-care physicians who are paid on the basis of fee-for-service are less likely to refer patients than are physicians paid through capitation (Grembowski et al., 1998). Furthermore, in cases where primary-care physicians have a role as gatekeepers (that is, a referral from a primary-care doctor is required for a patient to receive treatment by a specialist or in hospital), it has been found that gatekeepers who face financial risks when they refer (that is, who have to pay some of the cost of specialist and hospital care used by their patients) are less likely to refer to specialists (Martin et al., 1989; Hurley et al., 1991; Gravelle et al., 2002). Patients also appear to respond to demand-side incentives when making decisions with respect to specialist care. Shortell and Vahovich (1975) find that patients with higher third-party coverage are more likely to use specialist care. Furthermore, among persons that belong to a government plan, those who have supplemental insurance are more likely to use specialist care than those who do not (Shea et al., 1999). The Rand data also suggest, although weakly, that patient cost-sharing reduces total hospital expenditures (Newhouse, 1993).

In this paper, we extend the study of second-best demand- and supply-side incentives to a model in which we explicitly consider the interaction between insurance and the choice between primary care and in-hospital care. We find that such a model yields certain new insights for both types of incentives.

With respect to conventional insurance plans in which utilization is influenced by demand-side incentives (patient cost-sharing), we find that the moral-hazard problem associated with overutilization of services from a given provider may be significantly exacerbated because patient cost-sharing will also influence the patient's choice of provider (i.e., their decision to be hospitalized). This effect may be an important one in searching for the optimum degree of patient cost-sharing, and is likely to be particularly significant in assessing the effect of plans in which there is a lower degree of cost-sharing for hospital care because it

[4] For example, in the Rand 'Free Plan' (no co-payments or deductibles), the likelihood of any use of medical care was 86.8%, while the likelihood of one or more hospital admissions was 10.3%. Furthermore, the average total expenditure (per person per year) was $982 (1991 dollars) with $536 dollars of that in in-patient expenditures (Table 3.2, page 40; Newhouse, 1993).

tends to be used in cases of serious illness. An interesting finding is that managed-care plans that use patient cost-sharing as the principal cost-control mechanism, but control hospital utilization through a second-opinion requirement, may yield a substantially more efficient pattern of care than plans that rely on patient cost sharing alone.[5]

In some models that explore the effects of different insurance arrangements in an environment of information asymmetry between providers and patients, it has been shown that paying primary-care physicians through capitation may be efficient in the sense that it reduces excessive health services utilization (Hillman et al., 1989; Stearns et al., 1992; Léger, 2000). However, this result may not hold in a model such as ours when primary-care physicians advise patients not only regarding the use of their own services, but also regarding the advisability of the services of other providers such as in-hospital care provided by a specialist. Indeed, primary-care physicians paid via capitation have an incentive to over-refer to hospital, since this may reduce the physician's workload without affecting his or her income. On the other hand, primary-care physicians paid via fee-for-service may under-refer to hospital (in comparison with an efficient rate) since services provided by hospital-based specialists do not generate additional income. In our model, we analyze consequences of both kinds of incentives, and possible mechanisms for overcoming them, in designing second-best optimal insurance arrangements.

The rest of the paper is organized as follows. In the second section, we specify the basic model and consider the problem of second-best optimal insurance when both patients and doctors are fully informed in the sense that the patient has the same information as the doctor with respect to the patient's illness and the effectiveness of medical treatment, and insurance is of the conventional type with providers being paid through fee for service. In Section 3, we then consider the case where there is asymmetric information in the sense that doctors know both the patient's illness severity and the effectiveness of different types of treatment, but patients do not. We analyze this case both with conventional insurance, and with insurance through managed-care plans in which doctors are paid through capitation or salary. Conclusions are drawn in Section 4.

## 2. The basic model with fully informed patients

To simplify the analysis, we assume that individuals have at most one illness episode in each time period and that all sick consumers suffer from the same kind of illness, although the degree of severity may differ from patient to patient. Depending on the degree of severity, the illness may either be treated by a primary-care physician (henceforth referred to as a GP) or by an in-hospital specialist. Hospital treatment is assumed to involve more advanced technology than GP treatment, and also to require more costly equipment and a larger number of highly trained personnel. For example, hospital treatment may take the form of an operation performed by one or more specialists (surgeons, anesthesiologists) assisted by a team of nurses in an operating theater.

Intuitively, one can think of individuals as having an initial endowment of health, and illness as constituting an exogenous loss of part of this endowment. To partially offset

[5] Cost-control mechanisms such as these used in the managed-care industry are discussed in Glied (2000).

this loss, individuals utilize health services to produce health. One may think of GP and in-hospital specialist services as alternative inputs that can be used in producing health. When the amount to be produced is relatively small (that is, when the exogenous loss of health is relatively small), treatment by a GP may be sufficient and less costly. However, one can think of the production of health via GP services as being subject to diminishing returns (and therefore, to rising average cost per unit) at relatively small quantities. For those whose illness shocks are large and who therefore want to produce a large amount of health, treatment by specialists in a hospital (i.e., an operation) may be a more efficient choice. That is, the average cost of producing a large amount of health for a given individual may be lower if the patient is hospitalized than if he tries to produce the same amount of health through a large amount of GP services. For some kinds of illness, diminishing returns may be so strong that there is an upper limit on the amount of health that can be produced using GP services (i.e., beyond this quantity, the marginal and average cost becomes infinite). In such cases, hospitalization is the only alternative.

While GP services may entail a higher average cost per unit of health when large amounts are to be produced, hospital services are likely to have higher average cost per unit when only a small number of units are to be produced (i.e., when the patient's illness is less severe). As noted in the introduction, the more advanced treatment methods used in hospital may sometimes be more risky and invasive than the care that GPs provide. The risk of complications associated with invasive treatment can be regarded as a fixed cost of hospital treatment which raises the average cost per unit of small quantities.

The nature of the production process in a hospital (e.g., if the patient undergoes an operation) can also be thought of as involving an indivisibility. For example, it is not possible for an individual to undergo half an operation or half a diagnostic procedure, such as an MRI scan (or for two individuals to share the benefits of a single operation or procedure). Formally, one can represent this indivisibility by the restriction that each hospitalization episode entails production of at least a fixed minimum number of units of health. The cost of this minimum number can be thought of as an *episode-specific fixed cost*, with each unit below this fixed minimum having a marginal cost of zero.[6]

Using subscript *G* to denote GP services and *S* to denote in-hospital specialist services, we represent the above considerations by defining two cost functions $C_G(q)$ and $C_S(q)$, where $q$ is the quantity of health produced during a given illness episode. Using a lower-case letter for marginal cost (that is, defining $C'_J(q) = c_J(q)J = G, S)$, we assume $c_G(q) > c_S(q), \forall q$. The shapes of the marginal cost functions in Fig. 1 reflect the diminishing returns to health services in producing health, giving rise to increasing marginal costs. Consistent with the indivisibility of hospital services production, the marginal cost of hospital services is zero up until the minimum quantity $q^0$ but positive thereafter. Moreover, we assume that there exists some $\widehat{q}$ such that $C_S(q) < C_G(q)$ for $q > \widehat{q}$; i.e., $\widehat{q}$ is that (relatively large) break-even value of $q$ where the episode-specific fixed cost of hospitalization is just offset by the lower marginal cost for each unit produced in a hospital. Diagrammatically, this is the point where the difference between the areas under the marginal cost curves $c_G(q)$ and $c_S(q)$ is equal to the episode-specific fixed costs for hospital services. We denote this fixed cost by $F_S$.

[6] For simplicity, we assume that there is no episode-specific fixed cost associated with GP services. If there is, there will be some range of illness severity for which the patient's optimum choice will be 'no care'.

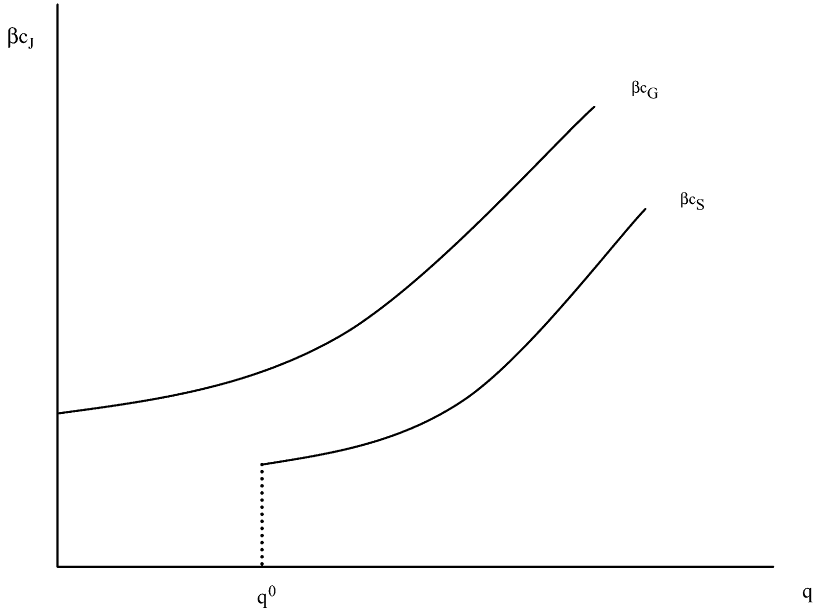

Fig. 1. Provider-specific cost functions.

Initially, we assume that there is perfect competition in the provision of health services, so that provider charges exactly cover total costs for each quantity provided.[7]

In the first version of the model, we assume that patients have the same information about their illness and the effectiveness of different kinds of medical treatment as their doctors and choose both the quantity of care and the provider, at given prices (i.e., from a 'price list' per episode and quantity of health produced by each provider). We also assume that patients can purchase actuarially fair insurance prior to the revelation of illness severity.

The insurer can only observe patients' illness severity imperfectly, so state-contingent contracts are not possible. Instead, cost control is in the form of some degree of patient cost sharing (i.e., demand-side incentives). Given their insurance coverage, and once illness severity is revealed, patients will choose whether to receive care from a primary-care physician (GP), or to enter a hospital to be treated by a hospital-based specialist. They will also choose what quantity of health to produce (by choosing the quantity of services to utilize).

Formally, we specify a model in which the representative consumer *i*'s utility depends on consumption $X$ and health $H$. Health, in turn depends on the value of a state variable $\theta$ which we interpret as an illness severity variable or an 'illness shock' (with large values of $\theta$ corresponding to more severe illnesses). Given the patient's use of health services, $H$ is then defined as $H = q - \theta$. Ex ante (when buying insurance), the patient does not know

[7] Even though the cost curves in Fig. 1 display rising marginal cost *per unit of health*, the marginal cost per unit of health services (physician visits, hospital days) may be constant. In the case of hospital services, however, it may be more realistic to think of a two-part pricing scheme: One charge for major one-time procedures (such as an operation), then a per-diem charge that depends on the length of the patient's stay.

what $\theta$ is going to be, though it is assumed that its cumulative distribution function $F(\theta)$ is known. *Ex post*, however (when choosing what provider to use and what services to buy), the patients knows $\theta$. We assume that the patient has a conventional insurance contract with co-payment rate $\beta$ and premium $\alpha$. In each state, the patient receives a (state-independent) income $I$. Since $\theta$ is known *ex post*, the patient can maximize utility in each state given $\theta$; the patient maximizes utility by choosing a provider $J$ where $J$ may be either $G$ or $S$, and a quantity of health production $q$. That is, in each state, the patient solves the problem

$$\max_{J,q} U(X, H), \; J = G \text{ or } S \tag{1}$$

subject to

$$X = I - \beta C_J(q) - \alpha \tag{2}$$

and

$$H = q - \theta. \tag{3}$$

To solve this problem, the patient finds the two quantities $q_J$ that are optimal when $J$ is $G$ and $S$, respectively, and compares the levels of maximized utility for each provider type. For future reference, denote these maximized utilities by $V_J(\theta, \beta)$, $J = G, S$. For either provider type, the first-order condition corresponding to the optimal choice of $q$ is

$$U_x(-\beta c_J(q)) + U_H = 0 \tag{4}$$

where

$$U_i \equiv \frac{\partial U}{\partial i}, \; i = X, H. \tag{5}$$

For a given $\theta$ and $J$, (4) defines a demand curve for health as a function of the marginal price of health produced by the $J$-th provider, $\beta c_J(q)$. A sufficient condition for this demand curve to be downward-sloping is that $U_{XH} = U_{HX} > 0$ (see Appendix A). For simplicity, we henceforth impose the assumption that this condition holds.

We now show:

**Proposition 1.** *If conditions are such that care provided by G will be chosen for some $\theta$ while S will be chosen for other $\theta$, then S will be chosen if and only if $\theta > \theta^C$ for some critical value $\theta^C$*. That is,

$$V_S(\theta, \beta) \geq V_G(\theta, \beta) \quad \text{if} \quad \theta \geq \theta^C. \tag{6}$$

To establish Proposition 1, we can first easily show the following lemma.

**Lemma 1.** *The quantity of care demanded for a particular type of care* (*either GP or in-hospital specialist care*) *is increasing in illness severity for $J = G$ or $S$, i.e.*

$$\frac{\partial_q J(\theta)}{\partial \theta} > 0 \tag{7}$$

**Proof.** See Appendix B.

To prove Proposition 1, consider first (in Fig. 2) the quantity $q_G(\theta)$ which denotes the optimal quantity the consumer chooses if provision were from a GP. Then define the compensated demand curve $D(\theta, G)$ by finding the quantities the consumer would choose

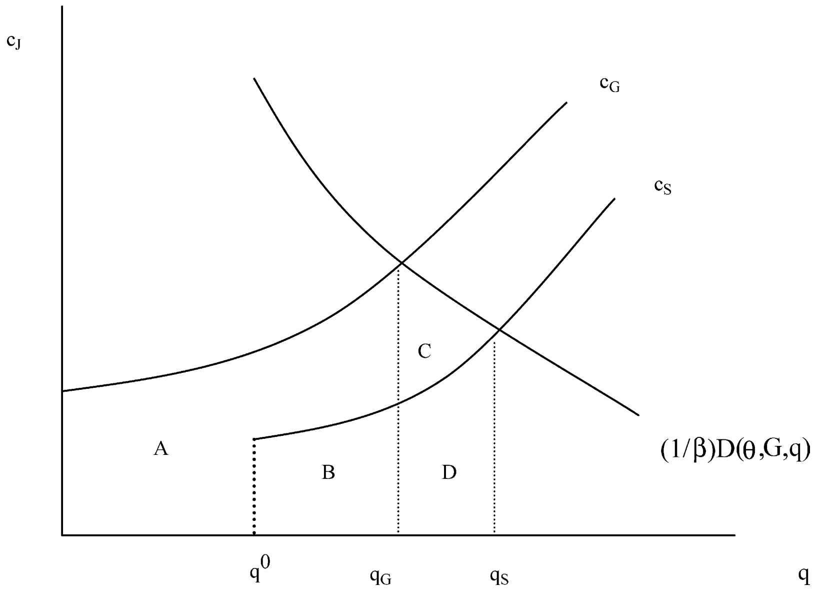

Fig. 2. Illness severity and choice of provider.

at different marginal costs if his net income were continuously varied so as to keep his utility at $V_G(\theta, \beta)$. By definition $V_G(\theta, \beta) = U(I - \alpha - \beta C_G(q_G), q_G - \theta)$. Therefore, at the point $q_S$ on this compensated demand curve we can write the consumer's utility as $U(I - \alpha - \beta C_G(q_G) - Z, q_S - \theta)$, where $Z$ is the compensating variation necessary to keep utility constant as the consumer utilizes $q_S$ units of health (rather than $q_G$). Thus, $Z$ is a measure of the incremental value of the additional health services $q_S - q_G$. But now note that the *actual* cost of utilizing $q_S$ is given by $C_S(q_S)$. Thus, if

$$\beta(C_S(q_S) - C_G(q_G)) < Z \tag{8}$$

then choosing $q_G$ cannot be optimal, and provision should instead be from $S$. □

Condition (8) can be illustrated diagrammatically as in Fig. 2. To interpret the figure, it is helpful to observe that (8) can be rewritten as (9):

$$F_S - \int_0^{q_G} (c_G(q) - c_S(q))\mathrm{d}q + \int_{q_G}^{q_S} c_s(q_S)\mathrm{d}q - \frac{1}{\beta}\int_{q_G}^{q_S} D(\theta, G, q) < 0 \tag{9}$$

where we have used the definitions of marginal cost $c_J(q)$, $J = G, S$ (note that $c_S(q) = 0$ for $0 < q < q^0$), and the fact that the compensating variation $Z$ is just the area under the compensated demand curve $D(\theta, G, q)$ between $q_G$ and $q_S$. In Fig. 2, we illustrate (9) diagrammatically. The last term in (9) corresponds to area $(C) + (D)$ while the third term is area $(D)$ and the second term is area $(A)$. Diagrammatically, therefore, condition (8) is equivalent to the condition that $F_S$ < area $(A) + (C)$. The proof of Proposition 1 is completed by observing that an increase in $\theta$ increases $Z$ (diagrammatically, it shifts the compensated demand curve to the right, increasing $(A)+(C)$). Thus, there is only one value of $\theta$ for which (8) will hold with equality.

The following corollary is also immediate.

**Corollary.** *In the neighbourhood of the critical value of* $\theta^C$, (*where* $\beta(C_S(q_S) - C_G(q_G)) = Z$) *total expenditures if hospital care is chosen is larger than total expenditure if GP care is chosen.*

### *2.1. Choice of provider and efficient insurance*

We now examine the role of insurance in the above model; more specifically, we consider how a change in insurance coverage will alter both the mix and quantities of health services purchased.

Although a decrease in the co-insurance rate (for a given provider type) will lead to greater health-services consumption (the well-know moral hazard problem), a change in the insurance parameter $\beta$ will also lead to a change in the critical value $\theta^C$. That is, it will change the illness severity at which the patient will switch to hospital care.

**Proposition 2.** *For a given insurance premium* $\alpha$, *a decrease in* $\beta$ (*the co-insurance rate*) *will decrease the critical value of* $\theta^C$, *i.e., the critical point where the patient switches from* '*GP care*' *to* '*in-hospital specialist care*' *will occur at lower severity of illnesses.*

**Proof.** Observe that at the original $\beta$, the left-hand side of (9) must equal zero at $\theta = \theta^C$. A reduction in $\beta$ will be equivalent to an upward shift in the curve $(1/\beta)D(\theta, G, q)$ that determines the limits of integration and the value of area $(A)+(C)$. But an upward shift of this curve will increase area $(A)+(C)$ above the fixed cost $F_S$ so at a lower value of $\beta$ the left-hand side of (9) will be less than zero. Consequently, it will be in the patient's best interest to switch to hospital care at a $\theta < \theta^C$. □

In the health economics literature, consideration has also been given to what is an efficient value of $\beta$, i.e., an efficient degree of patient cost sharing. Below we show that the solution to this problem will, in general, depend on both the tendency for a decrease in $\beta$ to affect the optimal quantity of care given by a particular provider, and on its influence on the critical value of $\theta$ determining the choice of provider.

An analysis of the problem of second-best cost sharing requires consideration of the effect of patients' behaviour on insurance premiums. Assuming competitive insurance markets so that premiums are actuarially fair, the $\alpha$ in this model can be written as:

$$\alpha = (1-\beta)\int_{\theta}(C_J(q(\theta)))\mathrm{d}F(\theta) \tag{10}$$

where $F(\theta)$ is the cumulative distribution of $\theta$, and $J = G$ for $\theta < \theta^C$ and $J = S$ for $\theta \geq \theta^C$. The solution to the optimal insurance problem consists in finding that $\beta$ which maximizes the consumer's expected utility subject to (10), where $q(\theta)$ and $\theta^C$ are determined as analyzed above.

Consumers' expected utility is given by:

$$EU(\beta, \theta^C, \alpha(\beta, \theta^C)) = \int_\theta U(I - \alpha - \beta C_J(q(\theta)), q(\theta) - \theta)\mathrm{d}F\theta. \tag{11}$$

In the next proposition, we show that the nature of the moral-hazard related inefficiency in this model consists not only in over-utilization for a given provider, but also in a tendency to switch providers at an inefficiently low level of illness severity $\theta$ (one can interpret this as a second form of *ex post* moral hazard). We demonstrate this by showing that the consumer's expected utility under an actuarially fair insurance contract would increase if he could be induced to use a larger critical value $\theta^C$ at which to switch from GP care to in-hospital specialist care than would be individually optimal without such a restriction.

**Proposition 3.** *An insured consumer will choose an inefficiently small critical value $\theta^C$ at which to be hospitalized.*

**Proof.** Observe that at the critical value $\theta^C$, $V_G(\theta^C, \beta) = V_S(\theta^C, \beta)$. Consider now the effect of a change in the critical value $\theta^C$ on expected utility, i.e. $\mathrm{d}E/\mathrm{d}\theta^C$.

$$\frac{\mathrm{d}E}{\mathrm{d}\theta^C} = V_S(\theta^C, \beta) - V_G(\theta^C, \beta) + \frac{\mathrm{d}E}{\mathrm{d}\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}\theta^C} = 0 + \frac{\partial E}{\partial \alpha}[C_G(q) - C_S(q)](1-\beta). \tag{12}$$

Given that $\partial E/\partial\alpha < 0$ and $C_G(q) - C_S(q) < 0$, $\mathrm{d}E/\mathrm{d}\theta^C > 0$. □

Proposition 3 has implications for the efficient degree of cost sharing, or equivalently, the design of a second-best insurance plan. To see this, note that by treating $\theta^C$ as an endogenous variable, we can differentiate consumer's expected utility with respect to $\beta$, and obtain:

$$\frac{\mathrm{d}EU}{\mathrm{d}\beta} = \left(\frac{\partial EU}{\partial \beta} + \frac{\partial EU}{\partial \alpha}\frac{\partial \alpha}{\partial \beta}\right)\Bigg|_{\theta^C} + \frac{\mathrm{d}\theta^C}{\mathrm{d}\beta}\left(\frac{\partial EU}{\partial \theta^C} + \frac{\partial EU}{\partial \alpha}\frac{\partial \alpha}{\partial \theta^C}\right) \tag{13}$$

The first term within round brackets reflects the standard trade-off between the incremental loss from less complete insurance and the reduction in the conventional moral-hazard effect as the degree of cost sharing is increased, holding $\theta^C$ constant (if $\theta^C$ were given, this term would have to be zero in a second-best optimal plan). However, if (13) is evaluated at the critical value that the consumer would choose for a given value of $\beta$ and $\alpha$, the first term inside the second set of brackets would be zero. Moreover, since $C_G(q) < C_S(q)$, the insurance premium $\alpha$ is decreasing in $\theta^C$. Therefore, if the critical value $\theta^C$ is chosen by the consumer, Eq. (13) would be positive at the value of $\beta$ where the first term in round brackets would be zero (since $\theta^C$ increases with $\beta$ and $\partial EU/\partial\alpha < 0$).

Taking this effect into account, it is clear that the optimum degree of cost sharing is higher when the effect through the choice of critical value $\theta^C$ is taken into account, than it would be for a fixed $\theta^C$. Moreover, suppose it were possible for the insurer to verify the value of $\theta$. If this could be done at no cost, an insurance policy that specified optimally chosen values of both $\beta$, and $\theta^C$ would involve a $\theta^C$ higher than what consumers themselves would choose at any given $\beta$, but would give a higher expected utility than a policy specifying an optimally chosen cost-sharing parameter $\beta$ alone (i.e., it would 'delay' hospitalization but yield a higher expected utility). Managed-care plans requiring a second opinion before covering hospitalization, but in other ways similar to conventional insurance, can be regarded as a real-world example consistent with this finding.

In the preceding analysis, we assumed that both doctors and patients perfectly and costlessly observed the value of $\theta$.[8] In reality, of course, the degree of seriousness of a person's illness can generally be established only after the doctor's time and other resources have been used to establish a diagnosis. Even after a diagnosis has been made, some residual uncertainty may remain.[9] In some cases, the nature and cost of the diagnostic procedure used by GPs may be different from, and cheaper than, those made in hospitals, but at the same time a diagnosis performed in hospital (perhaps using more advanced equipment) may yield more precise information than one performed by a GP.

The analysis in the preceding pages may be extended to the case with diagnostic uncertainty on the part of both doctors and patients. A key issue that arises in such an extension is whether or not the uncertainty is symmetric (i.e., whether one can continue to assume that at any given stage, doctors and patients face the same degree of imperfect information).

Introducing uncertainty and costly diagnosis would give rise to a number of complications even in the case of symmetric information. For example, it would require consideration of cases in which a patient would seek a diagnosis from one type of provider, but would subsequently decide to receive treatment from another.

Although we do not formally analyze this case, we believe that the basic intuition in the preceding cases would continue to be valid. That is, for a given diagnosis, a more fully insured patient would choose treatment in hospital at a lower degree of illness severity than a patient with a higher co-payment rate. Similarly, more fully insured patients would be more likely to undergo costly (hospital-based) diagnostic procedures than those with less complete insurance.

## 3. Asymmetric patient–doctor information

Many health economists would describe the assumption of no information asymmetry between doctors and patients as unrealistic, especially when taken in conjuction with the presumed information asymmetry between providers and insurers. Indeed, if both doctors and patients had the same information about $\theta$ in each illness episode it would seem that

[8] However, we implicitly assumed that $\theta$ could not be observed by the insurer, so that state-contingent insurance was not possible.

[9] The paper by Mariñoso and Jelovac (2003) consider a case where there are only two degrees of illness severity, but where the probability of the correct diagnosis being made is a function of the effort the doctor puts forth in making it. They study the design of payments mechanisms intended to elicit the optimal degree of diagnostic effort by doctors, given that the hospital is the more appropriate place to treat severe cases, while less severe cases should be treated outside hospital, by GPs. They also find conditions under which it is efficient to impose a gate-keeping rule under which patients must have a referral from a GP before being admitted to hospital. Implicitly, Mariñoso and Jelovac also assume that there is information asymmetry between doctors and patients, as patients comply with doctors' referral recommendation whatever the doctor's diagnosis. Indeed, patient information plays no role in their model, as patients are assumed to follow mechanical rules of thumb in deciding whether to consult a GP or a specialist when they are ill.

contracts contingent on this information (state-contingent contracts) would be possible, especially if the information also were available to a third party. In this section, therefore, we extend the analysis to the case where there is information asymmetry between doctors and patients and patients can only imperfectly observe the value of $\theta$.

More precisely, assume that the distribution $F(\theta)$ from which illness severity is drawn is bounded by $\theta^0, \theta^L$ and is subdivided into $L$ intervals $[\theta^{l-1}, \theta^l]$, $l = 1, \dots, L$. Although the patient does not observe the exact value of $\theta$, we assume that he or she can distinguish between these intervals (classes of illness); that is, the patient knows in which interval his or her true $\theta$ is located. However, there is information asymmetry: A physician can costlessly observe each patient's precise $\theta$ (can costlessly diagnose the patient's illness severity). Initially, we continue to assume that doctors, both GPs and hospital-based specialists, are paid on the basis of fee-for-service. We also assume that doctors know the boundaries of the intervals that define the patient's information.[10]

With information asymmetry, the patient has to rely on the advice of the doctor in order to decide on the amount of treatment. If there is perfect competition in the market for health services (an assumption that we implicitly made in the previous section), providers are indifferent as to how many units they sell to individual patients, and so, have no reason to exploit their information advantage in order to sell additional units. However, most analysts believe that the markets for physician and hospital services are better characterized as monopolistically competitive. In a market with monopolistic competition, sellers who can increase the amount demanded by individual buyers through the advice they give, have an incentive to do so.[11]

In the present context, doctors (GPs and hospital-based specialists) typically have an incentive to tell patients that the value of their illness parameter is at the upper end of the relevant interval. The exception is for patients whose illness severity lies in the interval which contains $\theta^C$ (the critical value at which a well-informed patient would switch from GP services to in-hospital specialist services). In that interval, a GP's incentive is to report a value just below $\theta^C$, while a specialist would report a value at the upper end of the interval.

[10] We assume implicitly that neither patients, nor the insurance provider, can infer ex post whether or not the treatment was appropriate (within each illness intervals). The assumption that providers can perfectly observe $\theta$ is maintained in order to prevent the analysis from becoming too complicated. For papers that attempt to model explicitly the case with imperfect but asymmetric information on the two sides of the market see footnote 11.

[11] Formal models of asymmetric information in medical care include Dranove (1998) and Rochaix (1989). Both authors specify probability distributions that link patients' beliefs about the way they should be treated, to the 'true' underlying illness conditions, and employ models in which the patient's problem is whether to accept or reject a doctor's treatment recommendation based on their beliefs about their illness condition and their beliefs about the doctor's information and strategy. The solution depends in part on either the cost of not being treated (Dranove) or of obtaining a recommendation from another doctor (Rochaix). Our approach simplifies the problem both by the way we specify patient beliefs and because we model the quantity of treatment as being decided by the patient; asymmetric information remains important, however, because it influences the way the patient treats information conveyed by doctors in making the quantity decision. In some models featuring information asymmetry between doctors and patients, an attempt is made to allow formally for the influence of professional ethics in explaining the recommendations that doctors make to patients, by specifying that the doctor's utility depends on her own welfare and that of the patient's. Although we do not doubt that professional ethics play a significant role in the decisions of many individuals in the real world, we do not follow this approach, in part because we believe that the influence of such ethical constraints may not be independent of economic incentives.

Assuming that patients correctly perceive their physicians' incentives, they realize that in reality, their illness severity is unlikely to always be at the upper end of the relevant intervals. However, they have no way of finding out what the true value of illness is. As a result, they must decide on the quantity of treatment to receive (and, in the interval which includes the critical value $\theta^C$ (which we denote by $r$), from what provider), knowing only which interval they are in. Thus, the quantity chosen will depend only on the interval, not on the value of $\theta$ within the interval.

Assuming patients know that distribution function $F(\theta)$, for a given interval $l$, insurance premium $\alpha$, and co-insurance rate $\beta$, the patient maximizes expected utility for the interval by choice of a single value $q_J^l$. The first order conditions for each interval are given by:

$$\int_{\theta^{l-1}}^{\theta^l} [U_X(q_J^l, \theta)(-\beta c_J) + U_H(q_J^l, \theta)]\mathrm{d}F(\theta) = 0 \tag{14}$$

where, $J = G$ for intervals $l = 1, \ldots, r-1$ and $J = S$ for intervals $l = r+1, \ldots, L$. For $l = r$, $J$ may be $G$ or $S$ depending on which choice yields the higher level of utility at the quantity that maximizes expected utility.

As before, an actuarial fairness constraint of type (10) but with a constant quantity $q_J^l$ in each interval, will hold in equilibrium. It can be written as

$$\alpha = (1-\beta)\sum_{l=1}^{L} C_J(q_J^l)P(l) \tag{15}$$

where $P(l) = \int_{\theta^{l-1}}^{\theta^l} \mathrm{d}F(\theta)$.

We can once again consider the problem of finding the insurance contract $\{\beta, \alpha(\beta)\}$ that is second-best optimal in the sense of balancing appropriately the moral-hazard loss associated with overutilization of health services against the gains from more complete insurance. In solving this problem, however, one must take account of the possibility that the function $\alpha(\beta)$ may have a discontinuity if there is a value of $\beta^C$ such that the consumer switches from $G$ to $S$ in the $r$-th interval. Using reasoning similar to that employed in establishing Proposition 1, it can be shown that if there is such a $\beta^C$, it will be the case that the consumer chooses $J = G$ for $\beta > \beta^C$ and $J = S$ for $\beta < \beta^C$. For a given $\alpha$, it can also be shown that around $\beta^C$, the actuarial-fair premium will increase by the finite amount $(1-\beta^C)P(r)[C_S(q_S^r) - C_G(q_G^r)]$. Other things equal, this discontinuity makes it more likely that the second-best optimal value of $\beta$ would be just slightly above $\beta^C$, as we consider the optimal degree of cost sharing for a variety of cost and demand conditions. Intuitively, this once again suggest that incentives and rules affecting the decision of whether or not to hospitalize patients in marginal cases are important in designing real world insurance plans.

### *3.1. Managed care*

In the previous section, we assumed that patients were covered by a conventional insurance plan in which they themselves decided what quantity of services to utilize, given their information about illness severity. Furthermore, physicians were assumed to be paid on the basis of fee for service, and their role was limited to supplying the quantity the patients decided to utilize, given their insurance contract; the insurer's role was that of a passive payer of bills. In this section, we consider insurance plans in which the insurer takes a more active role in influencing the services their patients utilize, i.e., managed-care plans.

One way we distinguish managed-care plans from conventional insurance is by assuming that in managed-care plans physicians are paid by a method such as capitation or salary. With these modes of payments, physicians have no incentive to exploit their information advantage for the purpose of inducing additional demands for their services. The other fundamental difference between our representation of managed care and conventional insurance is that for the former, we assume that the insurance contract specifies (through rules imposed on the physicians) what quantity of care the patient will receive in different circumstances, and from what provider.

Given this new specification, the question arises as to how managed-care contracts can be enforced when the insurer cannot observe the patient's illness severity. The answer to this question is that we don't assume contracts to be specified in such a way as to provide for a different quantity of services for each value of $\theta$. Instead, we assume contracts that specify a single quantity of services for each of the intervals referred to above (i.e., the intervals that represent the degree to which the patient knows his or her illness severity). Consequently, even though a doctor paid by capitation or salary has an incentive to downplay the patient's illness severity, his ability to effectively do so is limited by the patient's knowledge of the lower bound of the interval in which the true illness severity lies.

Note also, that the joint incentive for patient and doctor to over-state illness severity to the insurer under fee-for-service is not present under managed care since the physician's incentive under, for example, capitation is to understate, not overstate, illness severity. Even though the patient's and physicians' opposing incentives do not induce the doctor to reveal his knowledge of the true value of $\theta$, it at least enables the insurer to effectively enforce a contract that specifies a different value of services for each interval (even though the insurer cannot directly observe the illness severity or even the interval in which it falls). Note also that under managed-care contracts of this type, costs can be contained without relying on patient cost sharing.[12] Thus, we assume initially that the managed-care contracts have zero cost sharing.

The preceding paragraphs refer to those intervals in which only a single provider would have been chosen under full information. Consider now the critical interval containing $\theta^C$ (where the patient first seeks in-hospital care). Recall that we denoted this interval as $r$-th interval. Although a general practitioner paid by capitation would have an incentive to refer a patient to hospital anywhere in the $r$-th interval, the true value $\theta$ of the patient's illness severity can also (by assumption) be observed by other doctors. Thus, a managed-care plan can specify that a patient will not be treated in hospital unless the primary-care physician's referral is validated by an independent diagnosis (a 'second opinion'). In managed-care plans where specialists also are paid through salary or capitation, this requirement may be essentially self-enforcing, since hospital doctors are also assumed to observe the patient's true illness severity. As a consequence, they can therefore refuse to accept patients with a

[12] Baumgardner (1991) is an early paper that characterizes managed-care plans as insurance that uses specified quantities of care, rather than patient cost sharing, as a way of limiting costs.

$\theta$ below a contractually specified level.[13,14] For this reason, a managed-care contract can credibly specify a critical $\theta^C$ in this interval such that the patient will be treated in hospital if and only if the $\theta$ observed by the doctors is above that level, as well as separate quantities to be supplied depending on where the patient is treated. Formally, therefore a managed-care contract of this type will take on the following form:

$$\{q_G^l(l=1,\ldots,r-1), q_G^r, \theta^C, q_S^r, q_S^l(l=r+1,\ldots,L);\alpha\} \tag{16}$$

where $\alpha$ is the actuarially fair premium of the form given by (10). That is, the contract will specify a given quantity $q_G$ of GP services in each of the lower intervals, given quantities $q_G$ and $q_S$ in the lower and upper parts of the $r$-th interval, a critical value $\theta^C$ dividing this interval, and quantities of in-hospital specialist services $q_S$ in each of the upper intervals.

If the insurance market is competitive, the equilibrium contract is the one that maximizes the representative consumer's expected utility by choice of the $L+1$ quantities referred to above, and $\theta^C$, subject to the actuarial fairness constraint. The necessary first order conditions are:

$$\int_{\theta^{l-1}}^{\theta l} U_H(q_J^l,\theta)\mathrm{d}F(\theta) - \lambda(C_J(q))P(l) = 0 \tag{17}$$

for $l=1,\ldots,r-1,r+1,\ldots,L$ and $P(l)=\int_{\theta^{l-1}}^{\theta^l}\mathrm{d}F(\theta)$ and where $J=G$ for $r+1,\ldots,r-1$ and $J=S$ for $r+1,\ldots,L$

$$\int_{\theta^{r-1}}^{\theta^C} U_H(q_G^l,\theta)\mathrm{d}F(\theta) - \lambda(C_G(q^r))P(r,G) = 0 \tag{18}$$

$$\int_{\theta^C}^{\theta^r} U_H(q_S^l,\theta)\mathrm{d}F(\theta) - \lambda(C_S(q^r))P(r,S) = 0 \tag{19}$$

where $P(r,G)$ is the proportion of patients that fall in that $r$-th interval who use GP care and $P(r,S)$ is the proportion of patients that fall in that $r$-th interval and use in-hospital specialist care. An optimal choice of $\theta^C$ requires,

$$U(q_S^r(\theta^C)) - U(q_G^r(\theta^C)) - \lambda(C_S(q_S^r(\theta^C)) - C_S(q_G^r(\theta^C))) = 0. \tag{20}$$

The actuarial fairness constraint defining ($\alpha$) is of the form given by (15) and $\lambda$ is the Lagrange multiplier associated with the actuarial fairness constraint.

Clearly, a contract of this form will yield an expected utility that is lower than a full-information state-contingent contract. However, a more interesting question is whether it will yield a higher expected utility than the second-best optimal conventional contract under information asymmetry.

At first glance, one might expect that optimally chosen quantities under a managed-care contract would necessarily yield a higher expected utility than under patient cost sharing,

[13] Note that hospital-based specialists are also paid by capitation, they have no incentive to treat patients unnecessarily.

[14] In their model of referrals (one characterized by quasi-altruistic fundholding physicians), Gravelle et al. (2002) assume that specialists are able to perfectly assess the patients benefits of specialty care and can therefore refuse all patients who are referred to them but do not meet an exogenously given benefit threshold.

since managed-care contracts with no patient cost sharing provides for the same level of consumption of non-health goods and services in each state. However, if the marginal utility of consumption depends on health status, it is theoretically possible that the implicit redistribution of consumption across states with conventional cost sharing *raises* expected utility. If such is the case, the lack of any patient cost sharing under managed care may, paradoxically, yield a lower expected utility than under a second-best conventional contract.

By the same token, there is no reason why some degree of patient cost sharing could not be part of a second-best optimal managed-care contract. If interval specific cost-sharing parameters $\beta_l$ are permitted, it can be shown that a second-best managed-care contract will dominate a second-best conventional insurance contract.[15] Thus, we have Proposition 4.

**Proposition 4.** *Under imperfect information, the optimal managed-care contact of the form* (16) *with appropriately chosen interval-specific cost-sharing parameters $\beta_l$ yields higher expected utility than the optimal conventional contract of the form* $\{\alpha, \beta(\alpha)\}$.

**Proof.** Under a conventional contract with imperfect information, the values of health services utilization and consumption are both constant in each interval $l$, but are chosen so as to satisfy restrictions of the form (17)–(20). With a managed-care contract with interval-specific cost-sharing parameters, the constant levels of health services utilization and consumption in each interval can be optimally chosen without restrictions. □

Note that Proposition 4 would still be true if consumers in conventional plans always knew whether or not their illness severity parameter $\theta$ were above or below the optimum critical value of $\theta^C$ and could choose appropriately among providers in the $r$-th interval. In practice, however, a substantial part of the efficiency gains achievable through a second-best optimal managed-care plan of the form (16) may be due to the fact that consumers in conventional plans do not know where in the $r$-th interval they are and, as a result, can only choose one type of provider in that interval. If they consistently choose $S$ (that is, choose in-hospital specialist care), total costs are likely to be considerably higher than they would be if those below $\theta^C$ would choose $G$. Indeed, studies of the reason why HMOs in the US are able to provide care at costs below those of conventional plans have pointed to less utilization of hospital services as an important part of the explanation.

## 4. Conclusion

In this paper we have extended the analysis of the interaction between insurance and health services utilization to the case where there is a choice for consumers with different illness severity not only with respect to the quantities of services to utilize, but also among types of providers with different cost conditions; our main example has been the choice between outpatient primary-care physicians and treatment in hospital.

Our analysis shows that consideration of the patient's incentive to choose between outpatient and hospital care is important for finding the efficient degree of patient cost sharing in

[15] If we allow for interval specific cost-sharing, the actuarial fairness constraint and first-order-conditions would have to be modified in an obvious way.

models of second-best optimal conventional insurance. Patients with lower degrees of cost sharing have too small an incentive to choose the lowest-cost provider. The efficiency loss associated with this effect is in addition to that associated with the tendency of consumers with lower cost sharing to overutilize services from given providers.

Generally this result holds as well when it is assumed that there is information asymmetry between patients and providers, even though in this case outpatient providers paid via fee-for-service may have an incentive to *understate* patients' illness severity in certain circumstances, in order to discourage them from seeking hospital care.

We also consider the case where insurance takes the form of prepayment plans in which the quantity of care in different states is not chosen by the patient but is specified in the insurance contract. If it is assumed that the patient's illness state is costlessly observable by patients and insurers as well as by doctors, it would be possible to design a prepayment plan of this form that is first-best optimal both in the sense of making patients utilize the efficient volume of services given the choice of provider, *and* to choose efficiently between the two kinds of provider in given illness states.

If there is asymmetric information in the sense that illness severity cannot be perfectly observed by patients and insurers, first-best prepayment contracts cannot be credibly enforced. However, second-best prepayment plans can be designed through managed-care contracts under which providers are subject to supply-side incentives to control service utilization (for example, by being paid through salary or capitation), and the quantity of care promised under the plan is contingent on the consumers' (imperfect) information regarding their illness severity. Moreover, through restrictions such as requiring a second opinion before a patient is hospitalized, such plans can induce a more efficient pattern of hospital vs outpatient treatment than in conventional plans. Although any plan under imperfect information clearly must yield lower expected utility than a first-best prepayment plan would, we find that a second-best optimal managed-care plan dominates a second-best optimal conventional plan with cost control through demand-side cost sharing, at least if it allows for some degree of interval-specific cost sharing.

Although we believe that these results are of considerable interest, their significance of course is tempered by the restrictiveness of the assumptions built into the models from which they are derived. In particular, the assumption that all consumers face the same probability distribution for the illness severity parameter rules out consideration of problems with cream skimming and adverse selection. Another important assumption is that even in the cases where patients and insurers cannot observe precisely the patient's illness severity parameter, they can observe the quantities of services that providers render. If these quantities are imperfectly observable as well, the superiority of managed-care plans over conventional insurance is no longer guaranteed. Moreover, the assumption that providers can costlessly observe illness severity rules out consideration of the possible separation between diagnosis and treatment, with diagnosis being sought from one provider and treatment from another.[16]

[16] A referee has also noted that our specification rules out consideration of cases where treatment of a given illness involves the input of *both* physician *and* hospital services. Although we have not formally analyzed this case, we believe that our conclusions would generally be unchanged in the case where the choice of provider was limited to 'GP only' or 'hospital + GP services', if we continued to assume a major episode-specific fixed cost associated with hospital treatment.

## Acknowledgements

We thank seminar participants at McGill, National University of Singapore, HEC Montréal, CIRANO, the CEA annual meeting (Montréal), the SCSE annual meeting (Québec), the European Conference on Health Economics (Paris), the European Economics Association Annual Meeting (Venice), the 5th Biennial Conference on the Industrial Organization of Health Care (sponsored by the VA Management Group and the ISP, Department of Economics, Boston University), especially our discussant Lise Rochaix and two referees. Léger thanks HEC Montréal (Fond Mercure) and SSHRC for financial support. The usual caveats apply.

## Appendix A

**Proof.** ($\partial q/\partial \beta < 0$).

Let $F(q, \beta, \theta) = U_X(-\beta c_J(q)) + U_H$

We know by the implicit function theorem that $\partial q/\partial \theta = -(F_\beta/F_q)$.

We know that $F_q > 0$ if $U_{XX}(\cdot,\cdot) < 0$, $U_{HH}(\cdot,\cdot) < 0$ and $U_{XH}(\cdot,\cdot) = U_{HX}(\cdot,\cdot) \geq 0$ (sufficient but not necessary)

and,

$$-F_\beta = c_J U_X(\cdot,\cdot) - \beta c_J U_{XX}(C_J(q)) + U_{HX}(\cdot,\cdot)(C_J(q)) > 0 \tag{21}$$

if $U_{HX}(\cdot,\cdot) \geq 0$ and $C_J(q) \geq \alpha'(\beta)$ (sufficient but not necessary).

Thus, $\partial q/\partial \beta < 0$. □

## Appendix B

**Proof.** ($\partial q/\partial \theta > 0$).

We know that by the implicit function theorem that $\partial q/\partial \theta = -(F_\theta/F_q)$.

Where,

$$F_\theta = \beta c_J U_{XH}(\cdot,\cdot) - U_{HH}(\cdot,\cdot) > 0 \tag{22}$$

if $U_{XH}(\cdot,\cdot) \geq 0$ and $U_{HH}(\cdot,\cdot) < 0$ (sufficient but not necessary)

and where,

$$-F_q = -(\beta c_J)^2 U_{XX}(\cdot,\cdot) + \beta c_J(U_{XH}(\cdot,\cdot) + U_{HX}(\cdot,\cdot)) - U_{HH}(\cdot,\cdot) > 0 \tag{23}$$

if $U_{XX}(\cdot,\cdot) < 0$, $U_{HH}(\cdot,\cdot) < 0$, and $U_{XH}(\cdot,\cdot) = U_{HX}(\cdot,\cdot) \geq 0$ (sufficient but not necessary).

Thus, $\partial q/\partial \theta > 0$. □

## References

Baumgardner, J., 1991. The interaction between forms of insurance contracts and types of technical change in medical care. Rand Journal of Economics 22, 36–53.

Blomqvist, Å., 1991. The doctor as a double agent: information asymmetry, health insurance and medical care. Journal of Health Economics 10, 411–432.

Bradford, W., Martin, R., 1996. An Economic Theory of Referrals: Applications to the Medical Profession. University of New Hampshire, Department of Economics, Working paper.

Dranove, D., 1998. Demand inducement and the physician/patient relationship. Economic Inquiry 26, 281–298.

Ellis, R.P., McGuire, T., 1986. Provider behavior under prospective reimbursement. Journal of Health Economics 5, 121–151.

Forrest, C.B., Reid, R.J., 1997. Passing the Baton: HMOs' influence on referrals to specialty care. Health Affairs 16, 151–162.

Frank, P., Clancy, C.M., 1997. Referrals of adult patients from primary care: demographic disparities and their relationship to HMO insurance. The Journal of Family Practice 45, 47–53.

Glazer, J., McGuire, T.G., 1992. The Economics of Referrals. Boston University Industry Studies Program Discussion Paper Series 20.

Glenn, J.K., Lawler, F.H., Hoerl, M.S., 1987. Physician referrals in a competitive environment: an estimate of the economic impact of a referral. Journal of the American Medical Association 258, 1920–1923.

Glied, S., 2000. Managed Care. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics, Vol. 1A. Elsevier Science, North Holland, pp. 707–753.

Gravelle, H., Dusheiko, M., Sutton, M., 2002. The demand for elective surgery in a public system: time and money prices in the UK National Health Service. Journal of Health Economics 21, 423–449.

Grembowski, D.E., Cook, K., Patrick, D.L., Roussel, A.E., 1998. Managed care and physician referral. Medical Care Research and Review 55, 3–31.

Hillman, A.L., Pauly, M.V., Kerstein, J.J., 1989. How financial incentives affect physicians' clinical decisions and the financial performance of health maintenance organizations? New England Journal of Medicine 321, 86–92.

Hurley, R.E., Freund, D.A., Gage, B.J., 1991. Gatekeeper effects on patterns of physician use. The Journal of Family Practice 32, 167–174.

Léger, P.T., 2000. Quality control mechanisms under capitation payment for medical services. Canadian Journal of Economics 33, 564–586.

Ma, A., McGuire, T., 1997. Optimal health insurance and provider payment. American Economic Review 87, 685–704.

Mariñoso, B.G., Jelovac, I., 2003. GPs payment contracts and their referral practice. Journal of Health Economics 22, 617–635.

Martin, R.P., Diehr, P., Price, K.F., Richardson, W.C., 1989. Effect of a gatekeeper plan on health services use and charges: a randomized trial. American Journal of Public Health 79, 1628–1632.

Newhouse, J.P., Insurance Experiment Group, 1993. Free For All? Lessons from the RAND Health Insurance Experiment. Harvard University Press, Cambridge.

Pauly, M.V., 1979. The ethics and economics of kickbacks and fee splitting. Bell Journal of Economics 10, 344–352.

Rochaix, L., 1989. Information asymmetry and search in the market for physicians' services. Journal of Health Economics 8, 53–84.

Shea, D., Stuart, B., Vasey, J., Nag, S., 1999. Medical physician referral patterns. Health Services Research 34, 332–348.

Shortell, S., 1972. A Model of Physician Referral Behavior: a Test of Exchange Theory in Medical Practice. Center for Health Administration Studies, University of Chicago Research Series 31.

Shortell, S., Vahovich, S., 1975. Patient referral differences among specialties. Health Services Research 10, 146–161.

Stearns, S.C., Wolfe, B.L., Kindig, D.A., 1992. Physician response to fee-for-service and capitation payment. Inquiry 29, 416–425.

Wolinsky, A., 1993. Competition in a market for informed experts' services. RAND Journal of Economics 24, 380–398.