Provider Competition in a Dynamic Setting[*]

Marie Allard[†] Pierre Thomas Léger[‡] Lise Rochaix[§]

March 24, 2006

Abstract

In this paper, we examine provider and patient behaviour where effort is non-contractible and competition between providers is modeled in an explicit way. More specifically, we construct a model where physicians repeatedly compete for patients and where patients' outside options are solved for in equilibrium. In our model, physicians are characterized by an individual-specific ethical constraint which allows for unobserved heterogeneity in the physicians market. Allowing for unobserved heterogeneity in the physicians market introduces uncertainty in the patient's expected treatment if he were in fact to leave his current physician to seek care elsewhere. We also introduce switching costs associated with moving from one physician to another and, uncertainty in the treatment-outcome relationships. Our model can generate equilibria which are consistent with real-world observations. That is, our model can generate treatment heterogeneity, unstable physician-patient relationships and, over-treatment (a form of defensive medicine). Our model also suggests several avenues which may lead to more efficient provision of care.

JEL classification: I10, I18, J24, C30

Keywords: Physician Payment Mechanisms, Physician heterogeneity, Competition, Information Asymmetry, Insurance.

[*] We thank seminar participants at the 4th Health Economics Workshop (Oslo), DELTA/CREST-LEI (Paris), 3e Journées d'économie publique LAGV (IDEP-Marseille), CHESG (Toronto) and HEC Montréal, and Albert Ma, Jean Charles Rochet, Rob Clark and Chris Auld. Allard thanks the CEPREMAP, where part of this work was undertaken, for its hospitality. The usual caveats apply.

[†] HEC Montréal

[‡] HEC Montréal, CIRANO and CIRPÉE; corresponding author: pierre-thomas.leger@hec.ca

[§] IDEP-GREQAM-Université de la Méditerranée (Aix-Marseille II)

1 Introduction

The provision of medical services includes different forms of care. Some forms, such as hospitalizations, testing and pharmaceuticals are observed by patients, physicians and insurers. Other forms, such as physician time and effort, are unobservable by third parties and thus are non-contractible. A number of different mechanisms including physician monitoring and/or payment schemes, which seek to encourage the efficient provision of unobservable forms of care, have been put forth. In this paper, we propose that competition between health-care providers can serve as an alternative way of dealing with this issue. More specifically, the fact that unsatisfied patients can leave their current physician for a competing one may provide an important mechanism for encouraging the desired provision of such non-contractible care.

According to Gaynor and Vogt (2000), physician competition has been somewhat ignored in the literature partly because of the lack of concentration in the physicians market. The authors argue that as a consequence of the lack of concentration, the market is unlikely to exhibit anti-competitive behaviour. However, the presence of information asymmetry between patients and physicians, the proliferation of prospective payments which may encourage sub-optimal care, and the discretionary powers held by physicians, points to a role for competition and/or monitoring in the physicians market. Although monitoring (either directly or through medical malpractice litigation) may be a way to address these market imperfections, there is still a need to study other mechanisms such as competition in order to determine how to achieve the efficient provision of care.[1]

We build a model which is related to several papers (Ellis and McGuire, 1986; Ma, 1994; Ma and McGuire, 1997; Ellis, 1998; Gal-Or, 1999) while exploiting competition between similar providers (for example, between GPs) in a specific way. More specifically, we build a model in which physicians provide observable medical care (q) and unobservable effort (ϵ). Patients observe their post-treatment health can must decide on whether to stay with their current provider or,

[1] For a discussion of monitoring see Léger (2000). For a discussion of medical malpractice see Danzon (2000).

pay a switching cost and seek care elsewhere. Each physician is characterized by an individual-specific ethical constraint which specifies the minimal amount of effort she is willing to provide. These ethical constraints allow for unobserved heterogeneity in the physicians market. Unobserved heterogeneity introduces uncertainty in the patient's expected treatment if he were in fact to leave his current physician to seek care elsewhere.

We initially examine a situation in which switching costs are absent and the treatment-outcome relationship is certain. We find that competition creates an important incentive for physicians and leads to the desired provision of care and stable patient-physician relationships. However, if switching costs are present the effect of competition is dampened and leads some physicians to provide more care than others (i.e., heterogeneity in effort levels exist in equilibrium). We also introduce uncertainty in the treatment-outcome relationship and find several results which are consistent with real-world observations.

First, our model predicts over-provision of care. Although the over-provision of care is generally associated with the fear of medical-malpractice litigation (i.e., defensive medicine (Danzon, 2000)), we find an alternative (albeit related) reason for the over-provision of care: the fear of losing a patient to a competing physician. Second, uncertainty between treatment and outcome also leads to some unstable physician-patient relationships in equilibrium. That is, a proportion of patients switch physicians in each period. This result is consistent with empirical evidence which shows that four to eleven per cent of patients switch physicians in each year.[2] Our model suggests interesting avenues which may lead to the more efficient provision of care. For example, our model suggests that reducing switching costs and improving the treatment-outcome relationship will lead to more efficient provision of care, reduced heterogeneity in treatment from different types of physicians and lower patient turnover.

Although we are unaware of any other model which has all of the important features noted above and which can generate all of the predictions which are consistent with empirical observations, our

[2] See Sobero (2001) for a complete review of the literature on patients switching physicians.

work is related to several papers on competition in the physicians market. In Rochaix (1989), the patient's ability to consult a competing physician imposes an implicit constraint on his physician's discretionary power. More specifically, a physician risks losing her patient if the former's diagnosis differs greatly from the latter's prior expectations about illness severity. The threat of losing patients leads physicians to recommend a treatment intensity that is closer to the full information solution (a result which holds in the presence of only a small number of informed patients). Rochaix, however, does not deal with the issue of non-observable (and thus, non-contractible) effort. In Allard *et al.* (2001), the authors study compensation of health-care providers in a principal-agent framework where information asymmetry exists between providers and the regulatory agent. In their model, physicians are differentiated by their productivity. Patients, who are assumed to be identical, choose the physician who offers them the greatest net benefit. In equilibrium, competition in the physicians market equalizes net benefits among patients, i.e., the 'market constraint' leads physicians to exert non-contractible effort in order to attract patients. Our paper differs from theirs in several respects, most notably, by introducing patient heterogeneity. Furthermore, unlike Allard *et al.,* our model can generate both treatment heterogeneity and patient turnover in equilibrium.

Finally, our paper is related to Ma and McGuire (1997) who derive optimal health insurance and physician payment plans in a setting where medical services include both an observable component (q) and an unobservable (to third parties) effort (ϵ). In their model, physician effort is observed by the patient prior to the latter's quantity decision. Without this assumption, they argue, payment mechanisms cannot provide incentives to exert this costly effort. Ma and McGuire also examine the role of competition. However, in their model physicians compete with an exogenously given outside option (i.e., where the patient can obtain a given utility if he decided to leave), and by introducing patient heterogeneity with respect to their out-of-pocket cost for using different physicians. Our model also includes two types of care (q and ϵ), but we adopt a dynamic framework. Doing so allows us to: (i) relax the assumption that effort is observable to the patient prior to his quantity decisions, and (ii) endogenize the outside option if the patient were to leave his current physician.

Endogenizing this outside option is an important feature of our model.

The remainder of the paper is organized as follows. In section 2, we describe the model. In section 3, we solve the model in a static setting. We resolve the model in a repeated-game setting in section 4. Conclusions are drawn in section 5.

2 The Model

In this section we introduce a dynamic model characterizing the relationship between physicians, patients and insurance providers. As in Ma and McGuire (1997), treatment following an illness requires two forms of medical input: (i) observable medical care denoted by q, and (ii) unobservable physician effort denoted by ϵ. Medical care (q) is defined as any form of observable and contractible medical treatment. On the other hand, effort (ϵ) may be thought of as all valued forms of care which are not observable to third parties and thus non-contractible. These forms of care may include the physician's time and effort spent in researching and providing the appropriate treatment, monitoring the patient's progress and communicating with the patient (see Wedig *et al.*, 1989). We further assume a mixed physician payment scheme which consists of both a per-unit-of-q reimbursement and a prospective payment. This prospective component will ultimately serve to compensate physicians for the effort they exert, given that this form of care cannot be reimbursed on a per-unit basis.

Before competition (for patients) begins, a population of measure one of patients is assumed to be equally allocated to a population of measure one of physicians. Competition is introduced in our model by adopting a multi-period setting in which patients can move from one physician to another. Because we adopt such a framework, our model is best suited to potentially long-term relationships between patients and providers (for example, between patients and their family practitioners or, in the case of a chronic illness, between patients and their specialists).

The timing of the game is as follows:

Stage 1:

The physician-payment and insurance parameters are contracted upon. It is at this stage that the patient purchases an actuarially-fair insurance policy at a premium α.

Stage 2:

With probability π, the patient becomes ill and requires medical treatment. If the patient is ill, he draws θ from a known distribution of illness $F(\theta)$.[3] We assume that the patient perfectly observes his illness severity, but the third-party payer does not. If the patient is not ill, the 'period' ends (i.e., the patient does not seek medical treatment, remains healthy for one full period and returns, in the repeated-game setting, to stage 1 in the next period).

Stage 3:

A patient with illness severity θ seeks medical treatment. In our model it is assumed that ϵ and q are chosen simultaneously by the physician and the patient, respectively, i.e., neither patient nor physician can base his or her decision on the other's choice.[4,5] We assume, however, that the quantity q is purchased (on behalf of the patient) by the physician at a cost of ω per unit.

Stage 4:

Once medical care and effort have been provided, the patient's ex post health, denoted by H, is revealed. We assume that ex post health is perfectly observable to the patient yet unobservable to the third party. We also assume initially that the health production function ($h(\theta, q, \epsilon)$) is deterministic . We relax this somewhat restrictive assumption by allowing for uncertainty in the link between treatment and outcome in section 4.2.3. Once the physician has treated the patient, the latter pays γpq where γ denotes the co-payment rate and p denotes the price per unit of q, and the physician receives a net payment $(p - \omega)$ for each unit of q provided and a prospective payment (denoted by δ) which serves to compensate for effort.[6] For simplicity, we assume that the net payment per q is zero: $p = \omega$.

[3] In this setup, we can think of θ as representing a single illness with a severity distribution or a composite measure which maps different types of illnesses and their severity into a single dimension.

[4] We differ from Ma and McGuire (1997) in this respect, i.e., we relax their somewhat restrictive assumption that the patient observes the effort provided by his physician before choosing the quantity of medical care.

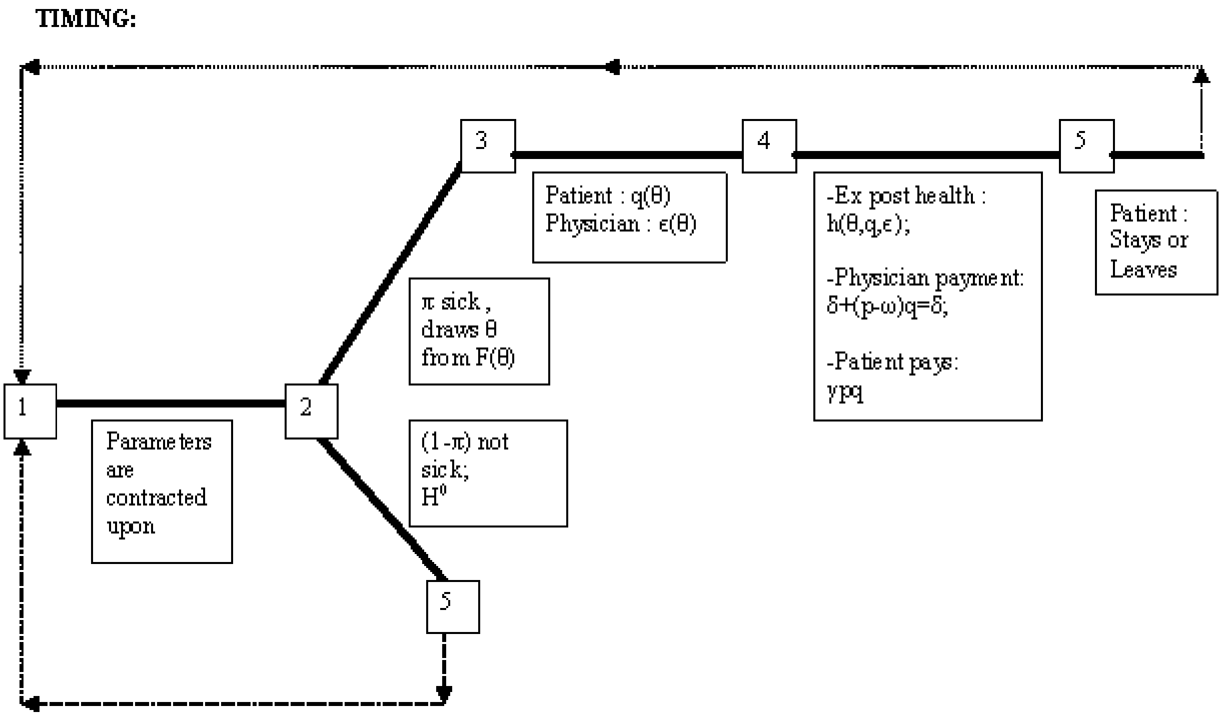
[5] Allowing the patient to choose the quantity of medical service (q) is equivalent to the physician proposing a schedule of treatments and prices. Because greater levels of q are associated with greater costs (i.e., a higher co-payment), the patient will choose the quantity which maximizes his expected utility.

[6] Thus, in this framework, physicians only receive payment if the patient is ill and seeks medical care.

Stage 5:

Because the patient observes his illness severity (θ), chooses the quantity of medical care (q) provided and observes his health outcome (H), he can perfectly infer his physician's effort (ϵ). Based on this information, the patient may choose to leave his current physician. For simplicity, we assume that each period is characterized by a new draw from the illness distribution, i.e., we do not consider the 'dynamic' aspect of health.[7]

Figure 1:


TIMING:
1
Parameters are contracted upon
2
π sick, draws θ from F(θ)
(1-π) not sick; H⁰
5
3
Patient : q(θ)
Physician : ε(θ)
4
-Ex post health : h(θ,q,ε);
-Physician payment: δ+(p-ω)q=δ;
-Patient pays: γpq
5
Patient : Stays or Leaves

We next describe each player in greater detail.

[7] Because the patient draws from the illness severity distribution independently in each period, cream-skimming issues are not dealt with here. That is, because all patients are identical before each period begins, physicians will not be able to select less or more costly patients.

The Patient:

The patient's per-period expected utility is given by:

$$EU = (1-\pi)U(C, H^0) + \pi \int_\theta U(C, h(\theta, q, \epsilon))dF(\theta), \tag{1}$$

where

$$C = I - \alpha - \gamma pq. \tag{2}$$

We assume a separable utility function for $U(C, H)$:

$$U(C, H) = u(c) + h(\theta, q, \epsilon),$$

where $u' > 0$ and $u'' < 0$. In (2) C denotes the patient's consumption while I denotes the state-independent income. We define $H^0 \equiv h(0, 0, 0)$ to be the patient's health in the absence of illness. We also assume that the patient is risk-neutral with respect to his health.

The Physician:

In our model, we introduce unobserved heterogeneity in the physicians market in a simple way. Each physician is characterized by a λ parameter where $\lambda \in [0, 1]$. If for a given illness severity θ, the patient were to choose an effort $\widetilde{\epsilon}$ (henceforth referred to as the patient's desired level of effort), a physician λ would never be willing to provide less than $\lambda\widetilde{\epsilon}(\theta)$. One can think of this illness-specific desired effort ($\widetilde{\epsilon}(\theta)$) as the effort that the patient would himself choose under full information.[8] As a result, a physician with $\lambda = 1$ would never be willing to provide less than the patient's desired level of effort ($\epsilon = \widetilde{\epsilon}(\theta)$). However, a physician with $\lambda = 0$ could provide the minimal amount of effort possible ($\epsilon = 0$).[9] Thus, each physician will be characterized by an ethical constraint which gives the minimum proportion of the desired effort level to be provided. We also assume that physician types are distributed according to a known distribution $\Gamma(\lambda)$.

[8] One could also consider $\widetilde{\epsilon}(\theta)$ as the illness-specific appropriate level of effort set by medical protocol or norms. Although a more complicated model could allow for $\widetilde{\epsilon}(\theta, \gamma)$, that is for an illness-co-payment specific level of effort, the results would not change qualitatively in the model without switching costs. However, in the model with switching costs, heterogeneity in treatments could lead to heterogeneity in the optimal co-payment. Thus, we would have to consider the possibility of patients acting strategically through their choice of γ (for example, the possibility that patients increase their co-payment to extract more effort). We leave this potential extension to future research.

[9] One can think of effort $\epsilon = 0$ as the minimal amount of effort below which the physician's effort would be observably insufficient.

Each physician is assumed to have a per-patient per-period utility denoted by V which is increasing in income (M) and decreasing in effort (ϵ). Thus, the physician's per-patient per-period expected utility is given by:

$$EV = (1-\pi)V(0,0) + \pi V(M,\epsilon), \tag{3}$$

where $M = \delta + (p-\omega)q$ when the patient seeks medical treatment. We assume a separable utility function for $V(M,\epsilon)$:

$$V(M,\epsilon) = M - c(\epsilon),$$

where $c(e)$ denotes the cost of effort and where $c' > 0$ and $c'' > 0$.

The Insurer:

We assume that the market for insurance is perfectly competitive. The actuarially-fair health-insurance premium for physician services is thus given by:

$$\alpha = \pi \int_\theta ((1-\gamma)pq(\theta) + \delta(\epsilon(\theta)))dF(\theta), \tag{4}$$

where $q(\theta)$ and $\epsilon(\theta)$ denote the equilibrium quantity of medical services and effort, respectively.

3 The Static Framework

In this section, we examine the static setting by shutting down Stage 5 in the game described above. Examining our model without its competitive feature will serve as a benchmark.

It is well known in the literature that the first-best health insurance policy would provide state-contingent treatments (in our case, illness contingent levels of q and ϵ). In our case, optimal illness-contingent levels of q and ϵ can be obtained by solving the patient's ex ante problem. That is, optimal levels of q and ϵ can be obtained by maximizing the patient's expected utility (1) subject to his budget constraint (2), the physician-participation constraint (that will be satisfied if the physician's expected utility (3) is greater than some exogenously given value $\overline{V}$), and an actuarially-fair health-insurance premium (4). However, a state-contingent contract of this type is infeasible given that illness severity, effort levels and post-treatment health are not verifiable and thus non-contractible (Arrow, 1963).

As noted above, the patient observes his illness severity and his ex post health but does not observe his physician's type. Also recall that the physician chooses effort level ϵ while the patient simultaneously chooses medical care q. It is obvious that in a static setting the physician will never wish to provide effort beyond the minimum amount determined by her ethical constraint, i.e., for a given illness severity θ, the physician λ will provide $\lambda\widetilde{\epsilon}(\theta)$ irrespective of the prospective payment. This is simply because increasing the effort beyond the minimum amount, which is utility decreasing for the physician, does not yield a larger prospective payment for the physician.

For a given co-payment γ and a specific realization of θ, the patient's expectation with respect to his physician's effort is given by $E_{\lambda}(\lambda\widetilde{\epsilon}(\theta)) = \int_0^1 \lambda d\Gamma(\lambda)\widetilde{\epsilon}(\theta)$. Thus a patient with illness θ solves:

$$\max_q U(I - \alpha - \gamma pq, h(\theta, q, E_{\lambda}(\lambda\widetilde{\epsilon}(\theta)))).$$

For a given co-payment γ and a specific illness severity θ, the equilibrium will be characterized by homogeneity in quantities ($q^*(\theta)$) chosen by the patients and yet also by heterogeneity in efforts ($\epsilon^*(\theta)$) provided by the physicians (where the equilibrium efforts will be distributed between 0 and $\widetilde{\epsilon}(\theta)$). As a result, how much effort the patient receives is simply a function of his illness severity and the physician type he has been assigned to. We assume that physicians cannot turn patients away based on their illness severity.

To ensure the participation of all physicians (i.e., irrespective of type), the prospective payment must (at least), in expectation, compensate effort provided by the physician of type $\lambda = 1$. If θ were observable, an illness-specific prospective payment ($\delta(\theta)$) would have to be paid to all physicians irrespective of their type. However, given that the illness severity is not observable by the insurer, the equilibrium prospective payment (δ^*), which is paid to the physician prior to the realization of θ, must be illness independent and based on its expectation, i.e.,

$$\delta^* = \int_{\theta} \delta^*(\theta) dF(\theta) = \int_{\theta} c(\widetilde{\epsilon}(\theta)) dF(\theta).$$

Next, the actuarially-fair insurance premium α is given by:

$$\alpha(\gamma) = \pi \int_{\theta} ((1-\gamma)pq^*(\theta))dF(\theta) + \pi\delta^*. \tag{5}$$

Given our assumption of perfect competition in the insurance market, insurers will be indifferent between all co-payment levels (i.e., each co-payment level is associated with an insurance premium that yields zero expected profits). As a result, the equilibrium co-payment (γ^*) will maximize the patient's expected utility. This equilibrium co-payment balances the expected utility gains of more complete insurance with the utility loss of a higher insurance premium. Thus, the equilibrium actuarially-fair insurance premium (α^*) is simply given by (5) evaluated at γ^*.

Given the results provided above, we can characterize both the patient's and physician's ex post utility. The patient's ex post utility is given by:

$$U(I - \alpha^* - \gamma^* pq^*(\theta), h(\theta, q^*(\theta), \lambda\widetilde{\epsilon}(\theta))),$$

where we recall that the quantity q^* is chosen based on the realization of θ and the *expected* effort level to be provided by his physician. Ex post health, however, is a function of the realization of illness severity θ, q^* and the *true* effort provided by the physician. Thus, if the patient's physician is of a type greater than the expected type ($\lambda > E_{\lambda}(\lambda\widetilde{\epsilon}(\theta))$), then the patient will be treated with more effort than expected. In such a case, the patient will have chosen a quantity q^* which is too large (small) if q and ϵ are substitutes (complements).

In this setting all physicians receive the same compensation (i.e., irrespective of their type): $M = \delta^* + (p - \omega)q^*(\theta) = \delta^*$ because $p = \omega$. However, physician ex post utility is type dependent, i.e.,

$$V(\delta^* + (p-\omega)q^*, \lambda\widetilde{\epsilon}(\theta)) = \delta^* - c(\lambda\widetilde{\epsilon}(\theta)).$$

Thus, in equilibrium, all but the physician with $\lambda = 1$ will receive a prospective payment which over-compensates for effort provided (in expected terms).

The above result, where all physicians provide their respective minimum effort, is consistent with Ma and McGuire's statement that: "the alternative assumptions - that physician effort decision is made either simultaneously with, or after the patient's quantity decision- are unpalatable: in both cases, neither the patient's quantity choice nor the payment contract can provide any incentive for the physician to undertake costly actions."(p. 690). In the next section we show that this is not necessarily the case when competition is introduced in a dynamic setting. That is, we show that physicians may undertake costly effort even if physician effort is chosen simultaneously with the patient's quantity decision when physicians repeatedly compete for patients.

4 The Dynamic Framework

In this section, we turn our attention to a richer model where competition between providers plays a central role. In a repeated game setting, the patient's ability to move from one physician to another may serve to encourage physicians to provide treatment levels beyond those determined by their ethical constraints. In the following sections, we require that strategies be 'credible' in the sense that if a physician provides less than some level of effort, the patient will seek care elsewhere. This eliminates the possibilities of empty threats.

4.1 The Patient's and Physician's Strategies

In this section, we define the patient's and physician's strategies in a repeated-game framework.

The Patient's Strategy:

If the patient leaves his current physician, he receives at the end of the first period:

$$U(I - \alpha - \gamma p q^* - \kappa, h(\theta, q^*, \epsilon)), \tag{6}$$

and expect to receive in the future (at least):

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Leave} = \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^*, h(\theta, q_t^*, \epsilon_t^{\exp}(\theta))) dF(\theta), \tag{7}$$

where κ is included to represent financial and/or psychic costs associated with moving from one physician to another, where ρ denotes the patient's discount factor, and where $\epsilon_t^{\exp}(\theta)$, $t = 2, 3, ...$, denotes the patient's expectation about the future stream of efforts he would receive (from the outside physician(s)) if he were to leave his current physician using an optimal exit strategy.[10] In order to calculate this stream of expected efforts ($\epsilon_t^{\exp}(\theta)$, $t = 2, 3, ...$), the patient must consider all possible exit strategies. That is, he must consider all exit strategies of the form: leave the current physician if the effort provided by this physician is below a given level (i.e., below a reservation effort) assuming: (i) that he would randomly draw from the pool of physicians if he were to leave, and (ii) that each drawn physician would provide her minimum effort. The patient must then calculate the discounted expected utility associated with each of these exit strategies. The stream of expected efforts ($\epsilon_t^{\exp}(\theta)$, $t = 2, 3, ...$) in (7) represents the stream of expected efforts associated with the exit strategy that yields the patient the greatest discounted expected utility. We denote the reservation effort level associated with the optimal exit strategy as $\epsilon^{\exp}(\theta)$.

We assume a **random matching technology**. That is, a patient who decides to leave his current physician is randomly matched to a new physician. One can imagine that this random matching is done through a centralized system similar to those offered by many medical association and health agencies. We further assume that physicians are not initially capacity constrained. That is, given the initial population of physicians of measure one who have been allocated an equal share of the population of patients measure one, each physician can accept several new patients. This does not imply, however, that physicians could not at some point become capacity constrained. We return to the issue of capacity constraints later on.

It is important to note that q^* in (6) is based both on the current period illness severity and the expected effort provided by the patient's current physician. However, q_t^* in (7) is based both on the illness severity and on the effort expected to be provided by the outside physician(s) if the patient were to leave.

In order to characterize the expected present value of not leaving, we must define how patients form their expectations (i.e., their beliefs) regarding future effort levels to be provided by their

[10] Outside physician(s) is pluralized as patients are not limited as to how many times they may leave their current physician for a randomly drawn outside one.

current physician. Recall that the patient observes θ, chooses q, observes ex post health and thus can perfectly infer the effort provided to him by his physician in the current period. Although a patient cannot perfectly infer his physician's type, he can infer what type his physician is not. That is, a patient who draws θ can always infer an upper bound for his physician's type. More specifically, a physician who provides ϵ (given θ) must be characterized by a $\lambda \in [0, \epsilon/\widetilde{\epsilon}]$ where $\widetilde{\epsilon}$ is the desired effort level for the particular value of θ. We denote $\lambda^{\max} = \epsilon/\widetilde{\epsilon}$. In the following sections we assume that patients will base their expectations regarding their current physician's future behaviour on this $\lambda^{\max}$.[11] While basing future behaviour on $\lambda^{\max}$, rather than any other value in the interval $[0, \lambda^{\max}]$, may appear to be somewhat limiting and arbitrary, we show in the appendix why these are the only beliefs which survive in equilibrium.

Given the above discussion, the patient who remains with his current physician would receive in the current period:

$$U(I - \alpha - \gamma p q^*, h(\theta, q^*, \epsilon)), \tag{8}$$

and expect to receive in the future (at least):

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Stay} = \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q^*, h(\theta, q^*, \lambda^{\max}\widetilde{\epsilon}(\theta)))dF(\theta). \tag{9}$$

We now write the patient's strategy based on (6), (7), (8), and (9). That is, the patient will be willing to leave his current physician if and only if:

$$\begin{aligned} & U(I - \alpha - \gamma p q^* - \kappa, h(\theta, q^*, \epsilon)) + \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^*, h(\theta, q_t^*, \epsilon_t^{\exp}(\theta)))dF(\theta) \\ > \; & U(I - \alpha - \gamma p q^*, h(\theta, q^*, \epsilon)) + \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q^*, h(\theta, q^*, \lambda^{\max}\widetilde{\epsilon}(\theta)))dF(\theta). \end{aligned} \tag{10}$$

If we assume, for the time being, that transaction costs are absent (i.e., $\kappa = 0$), we can rewrite (10)

[11] Although it is possible for a physician for whom $\lambda^{\max}\widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$ to provide efforts greater than $\epsilon^{\exp}(\theta)$ in the future, those for whom $\lambda\widetilde{\epsilon}(\theta) > \epsilon^{\exp}(\theta)$ have no choice but to do so. Therefore, it is reasonable to believe that, ceteris paribus, physicians with $\lambda\widetilde{\epsilon}(\theta) > \epsilon^{\exp}(\theta)$ will provide greater effort in the future than physicians with $\lambda^{\max}\widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$.

as:

$$\begin{aligned} & \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^*, h(\theta, q_t^*, \epsilon_t^{\exp}(\theta))) dF(\theta) \\ > & \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q^*, h(\theta, q^*, \lambda^{\max} \widetilde{\epsilon}(\theta))) dF(\theta). \end{aligned} \tag{11}$$

Given that $\sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^*, h(\theta, q_t^*, \epsilon_t^{\exp}(\theta))) dF(\theta)$ represents the discounted expected utility associated with the optimal exit strategy with corresponding effort $\epsilon^{\exp}(\theta)$, we can rewrite the patient's strategy as: leave (stay with) the current physician if $\epsilon^{\exp}(\theta) > (\leq)\ \lambda^{\max} \widetilde{\epsilon}(\theta)$.

The Physician's Strategy:

We now turn our attention to the physician's strategy. A physician for whom $\lambda \widetilde{\epsilon}(\theta) \geq \epsilon^{\exp}(\theta)$ will provide effort according to her ethical constraint (i.e., $\lambda \widetilde{\epsilon}(\theta)$). By doing so, the physician will minimize her effort costs without risking the loss of her patient. However, a physician for whom $\lambda \widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$ (i.e., for whom the effort determined by her ethical constraint is less than the effort required to maintain her patient into the next period) will provide $\epsilon^{\exp}(\theta)$ if providing such effort yields greater discounted expected utility than providing her minimum effort and losing her patient i.e., if and only if:

$$V(\delta, \epsilon^{\exp}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} \int_{\theta} V(\delta, \epsilon^{\exp}(\theta)) dF(\theta) \geq V(\delta, \lambda \widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} V_t^{DEV}, \tag{12}$$

where β denotes the physician's discount rate. In (12), $\sum_{t=2}^{\infty} \beta^{t-1} V_t^{DEV}$ represents the future discounted utility associated with losing one's patient. It is important to note that this discounted utility is endogenous to the model and will be solved for in equilibrium.

It is important to recall that the problem described by (12) pertains to one physician's decision about whether or not to keep one of her patients. In a setting where the physician is not capacity constrained, the physician's likelihood of being randomly assigned a new patient is independent of whether or not she retains her current ones. We discuss this issue in detail later on.

4.2 Equilibrium Analysis

In this section we solve for the equilibrium given the patient's and the physician's strategies described above. Although the equilibrium is achieved instantaneously, we adopt a sequential reasoning when solving for the equilibrium for presentation sake only. In section 4.2.1. we solve for the equilibrium assuming that (i) switching costs are absent (i.e., $\kappa = 0$) and (ii) the treatment outcome relationship is certain (i.e., the health production function is deterministic). We relax each of these assumptions in sections 4.2.2. and 4.2.3., respectively.

4.2.1 The case with no switching costs and a deterministic health production function

In this section, we solve for a pooling equilibrium assuming no switching costs. More specifically, we show that under certain conditions, a pooling equilibrium can be achieved where all physicians provide their patients with desired levels of effort, irrespective of their ethical constraint. We also show that this equilibrium is characterized by stable patient-physician relationships.

Recall that before competition begins, patients are assumed to be equally allocated across physician types who are not initially capacity constrained. Recall from (11) that a patient who is currently with a physician for whom $\lambda^{\max}\widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$ would wish to leave. For such a patient, the threat of leaving is credible. Define $\widehat{\lambda}$ to be the physician type that would provide exactly $\epsilon^{\exp}(\theta)$ if she were to strictly follow her ethical constraint (i.e., $\widehat{\lambda}\widetilde{\epsilon}(\theta) = \epsilon^{\exp}(\theta)$ or $\widehat{\lambda} = \frac{\epsilon^{\exp}(\theta)}{\widetilde{\epsilon}(\theta)}$). Thus, a patient would be willing to leave his current physician if his current physician's $\lambda^{\max} < \widehat{\lambda}$. Consequently, all physicians of type $\lambda < \widehat{\lambda}$ will wish to provide the effort required to maintain their patient into the next period, i.e., $\epsilon^{\exp}(\theta) = \widehat{\lambda}\widetilde{\epsilon}(\theta)$ (assuming that condition (12) holds at $\widehat{\lambda}\widetilde{\epsilon}(\theta)$). On the other hand, a patient who is currently with a physician for whom $\lambda^{\max}\widetilde{\epsilon}(\theta) \geq \epsilon^{\exp}(\theta)$ would not be willing to leave. Consequently, all physicians of type $\lambda \geq \widehat{\lambda}$ will provide the effort determined by their ethical constraint (i.e., $\lambda\widetilde{\epsilon}(\theta)$) without risk of losing their patients.

Given the partial results provided above, we can see that effort levels should no longer be distributed between $[0, \widetilde{\epsilon}(\theta)]$ but rather between $[\epsilon^{\exp}(\theta), \widetilde{\epsilon}(\theta)]$=$[\widehat{\lambda}\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$, with a mass of the physicians providing precisely $\widehat{\lambda}\widetilde{\epsilon}(\theta)$. This is, however, not the full story. Patients must now calculate a new $\epsilon^{\exp}(\theta)$, say $\epsilon_1^{\exp}(\theta)$, by considering all potential exit strategies, assuming this time: (i) that they would randomly draw from the pool of physicians if they did leave, and, (ii) that each physician with a $\lambda \leq \widehat{\lambda}$ would provide $\widehat{\lambda}\widetilde{\epsilon}(\theta)$ while those with a $\lambda > \widehat{\lambda}$ would provide their minimum effort. Consequently, all physicians for whom $\lambda\widetilde{\epsilon}(\theta)$ is less than this new $\epsilon_1^{\exp}(\theta)$ should provide effort exactly equal to $\epsilon_1^{\exp}(\theta)$ while the rest should provide the effort determined by their ethical constraint. Defining $\widehat{\lambda}_1 = \frac{\epsilon_1^{\exp}(\theta)}{\widetilde{\epsilon}(\theta)}$, all physicians of type $\lambda < \widehat{\lambda}_1$ will wish to provide $\epsilon_1^{\exp}(\theta)$ in order to maintain their patient into the next period. Hence efforts should no longer be distributed between $[\epsilon^{\exp}(\theta), \widetilde{\epsilon}(\theta)]$=$[\widehat{\lambda}\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$ but rather between $[\epsilon_1^{\exp}(\theta), \widetilde{\epsilon}(\theta)]$=$[\widehat{\lambda}_1\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$ where it is obvious that $\epsilon_1^{\exp}(\theta) > \epsilon^{\exp}(\theta)$. Using the same rationale, it can easily be shown that the only level of effort which survives in equilibrium is the desired effort ($\widetilde{\epsilon}(\theta)$) i.e., the equilibrium is characterized by a degenerate distribution of efforts where $\epsilon^*(\theta) = \epsilon^{\exp}(\theta) = \widetilde{\epsilon}(\theta)$. Thus, heterogeneity in physician types (i.e., with different ethical constraints) nonetheless leads to homogeneity in effort levels.

Obviously, given that patients will always be provided with the desired effort ($\widetilde{\epsilon}(\theta)$), they will always choose quantity q^* accordingly.[12] Thus, this equilibrium will be characterized by homogeneity in treatment and stable patient-physician relationships (i.e., patients will not move from one physician to another in equilibrium).

The above equilibrium, however, requires patient switching to be zero and that no physician has any incentive to deviate and provide a level of effort below $\widetilde{\epsilon}(\theta)$, i.e. $\forall\lambda$,

$$V(\delta^*, \widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} \int_{\theta} V(\delta^*, \widetilde{\epsilon}(\theta)) dF(\theta) \geq V(\delta^*, \lambda\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} V_t^{DEV} \tag{13}$$

which is simply condition (12) where $\epsilon^{\exp}(\theta) = \widetilde{\epsilon}(\theta)$. In the above discussion, we did not, however, elaborate on the physician's discounted utility if she deviated from the patient's desired effort ($\widetilde{\epsilon}(\theta)$) and subsequently lost her patient, i.e., $\sum_{t=2}^{\infty} \beta^{t-1} V_t^{DEV}$. Notice that, if the physician provided insufficient effort and the patient did leave, the patient would be reassigned to the same physician

[12] It is important to note that the quantity q^* chosen in equilibrium will not correspond to the first-best level because of the presence of insurance which will lead to the well known problem of moral hazard.

with probability 0 in the following period, given the continuum of physicians and the fact that no physician is initially capacity constrained. Thus, the discounted value of losing one's patient is given by: $\sum_{t=2}^{\infty} \beta^{t-1} V_t(0,0)$.

As a result, in order to ensure a pooling equilibrium (i.e., where all physicians provide the desired level of effort) it must be the case that for all physicians and for every θ:

$$V(\delta^*, \widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} \int_{\theta} V(\delta^*, \widetilde{\epsilon}(\theta)) dF(\theta) \geq V(\delta^*, \lambda \widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} V_t(0,0). \tag{14}$$

Consequently, a pooling equilibrium will be guaranteed if the prospective payment (δ^*) is large enough to ensure that the least ethical physician (the physician with a $\lambda = 0$) will provide the desired level of effort ($\widetilde{\epsilon}(\theta)$) (and keep her patient) rather than provide the minimum effort ($\epsilon = 0$) (and lose her patient).

This δ^* is such that for the physician with $\lambda = 0$,

$$V(\delta^*, \widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} \int_{\theta} V(\delta^*, \widetilde{\epsilon}(\theta)) dF(\theta) \geq V(\delta^*, 0) + \sum_{t=2}^{\infty} \beta^{t-1} V_t(0,0) \tag{15}$$

is satisfied for every θ.

Given that the prospective payment is paid prior to the realization of illness severity θ, condition (15) must hold for all illness severities θ. Thus, the prospective payment will need to be relatively high to ensure that the least ethical physician would provide the most ill patient with his desired level of effort. Obviously, such a prospective payment would yield rents to all physicians.[13] We assume henceforth that (15) holds with equality for the least ethical physician at the highest value of θ.

Because quantity $q^*(\theta)$ and effort $\widetilde{\epsilon}(\theta)$ will always be chosen in equilibrium, the actuarially-fair

[13] Even though physicians yield positive rents (because of the high prospective payments), one could easily imagine a mechanism where physicians would have to pay an upfront payment to participate in the health-care market. Such a mechanism could be used to transfer some of these rents back to the patients.

insurance premium is given by[14]:

$$\alpha^* = \pi \int_\theta ((1-\gamma^*)pq^*(\theta))dF(\theta) + \pi\delta^*. \tag{16}$$

Finally, the patient's ex post utility is given by:

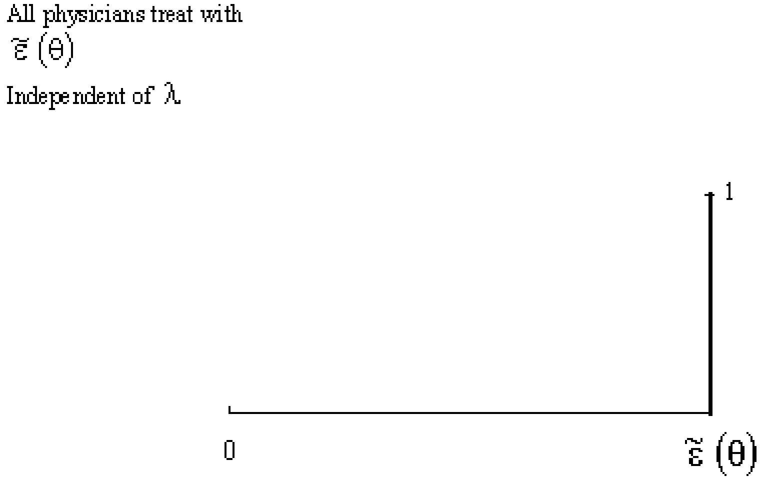
$$U(I - \alpha^* - \gamma^* pq^*(\theta), h(\theta, q^*(\theta), \widetilde{\epsilon}(\theta))).$$

Proposition 1

In the absence of switching costs the equilibrium strategies are such that: (i) physicians will provide their patients with desired levels of effort ($\widetilde{\epsilon}(\theta)$) if the prospective payment is sufficiently large (i.e., if condition (15) is satisfied for the physician with a $\lambda = 0$ at the highest illness severity), and provide effort according to their ethical constraint ($\lambda\widetilde{\epsilon}(\theta)$) otherwise; (ii) patients will stay with their current physician if they receive effort greater than or equal to the desired level of effort ($\widetilde{\epsilon}(\theta)$) and leave otherwise. Thus, a pooling equilibrium can be achieved with everyone receiving their desired level of effort $\widetilde{\epsilon}(\theta)$, where quantities of medical care are chosen optimally $q^*(\theta)$, and where patient-physician relationships are stable.

[14] In (16), γ^* is the equilibrium co-payment, i.e., the one which balances the patient's expected utility gains of fuller insurance with the loss of a higher insurance premium.

Figure 2:



In the following section, we examine the case where switching costs are present.

4.2.2 The case with switching costs

In this section, we introduce patient switching costs and show that, under certain conditions, the equilibrium will be characterized by some heterogeneity in physician effort where a proportion of physicians will provide effort beyond that determined by the ethical constraint. Under such a scenario, physician-patient relationships will remain stable.

Recall that, in the above section we began by showing that if a patient were currently with a physician identified by a $\lambda^{\max} < \widehat{\lambda}$, then he would be willing to leave for another physician if and only if:

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Leave} - \sum_{t=2}^{\infty} \rho^{t-1} U^{Stay} > U(I - \alpha - \gamma p q^*, h(\theta, q^*, \epsilon)) - U(I - \alpha - \gamma p q^* - \kappa, h(\theta, q^*, \epsilon)). \quad (17)$$

Suppose now that the switching costs κ are such that condition (17) exactly binds for a particular patient. That is, for this patient the present utility loss of switching from his current physician ($\lambda^{\max} < \widehat{\lambda}$) is just compensated by the expected future discounted utility gains of receiving $\widehat{\lambda}\widetilde{\epsilon}(\theta)$ (i.e., the effort level associated with the optimal exit strategy). Denote this particular patient's physician's $\lambda^{\max}$ as $\lambda^c(\kappa)$. All physicians with a $\lambda < \lambda^c(\kappa)$ should then behave like $\lambda^c(\kappa)$ in order to keep their patients. Consequently, a proportion n of physicians (i.e., those with $\lambda < \lambda^c(\kappa)$) should provide effort such that their patients infer $\lambda^{\max} = \lambda^c(\kappa)$, while the rest should provide effort according to their own ethical constraint (i.e., $\lambda\widetilde{\epsilon}(\theta)$).[15] As a result, effort levels should be distributed between $[\lambda^c(\kappa)\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$ with a proportion n of physicians treating precisely at $\lambda^c(\kappa)\widetilde{\epsilon}(\theta)$. Patients must now, however, calculate a new $\epsilon^{\exp}(\theta)$, say $\epsilon_1^{\exp}(\theta)$, by considering all potential exit strategies, assuming this time: (i) that they would draw randomly from the pool of physicians if they did leave, and (ii) that each physician with a $\lambda < \lambda^c(\kappa)$ would provide $\lambda^c(\kappa)\widetilde{\epsilon}(\theta)$ while those with a a $\lambda \geq \lambda^c(\kappa)$ would provide their minimal effort. We can now identify a new critical effort level ($\lambda^{c\prime}(\kappa)\widetilde{\epsilon}(\theta)$) which would leave patients just indifferent between receiving $\lambda^{c\prime}(\kappa)\widetilde{\epsilon}(\theta)$, and receiving $\epsilon_1^{\exp}(\theta)$ but having to pay the fixed cost κ. Therefore, physicians with a $\lambda < \lambda^{c\prime}(\kappa)$ will wish to behave like $\lambda^{c\prime}(\kappa)$ (i.e., provide $\lambda^{c\prime}(\kappa)\widetilde{\epsilon}(\theta)$) while those with a $\lambda \geq \lambda^{c\prime}(\kappa)$ will wish to provide their minimal effort. As a result, efforts will now be distributed between $[\lambda^{c'}(\kappa)\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$. Using the same rationale, we can identify the equilibrium critical effort (and its corresponding $\lambda^*(\kappa)$) which leaves a proportion of patients just indifferent between: (i) paying κ, leaving and expecting to receive the stream of efforts $\epsilon_t^{*\exp}(\theta)$, $t = 2, 3, ...$, and (ii) staying with their current physician

[15] As before, although it is possible for a physician characterized by a $\lambda^{\max} < \lambda^c(\kappa)$ to provide effort greater than the expected amount in the future (i.e., greater than $\lambda^c(\kappa)\widetilde{\epsilon}(\theta)$), those with a $\lambda > \lambda^c(\kappa)$ have no choice but to do so. Therefore, it is reasonable for the patient to base his expectations about his current physician's future treatments on his current physician's $\lambda^{\max}$).

and receiving the stream of constant efforts $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$. That is, for these patients:

$$\begin{aligned} & U(I-\alpha-\gamma p q^*-\kappa, h(\theta, q^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta))) + \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta} U(I-\alpha-\gamma p q_t^*, h(\theta, q_t^*, \epsilon_t^{*\,\exp}(\theta)))dF(\theta) \\ = \; & U(I-\alpha-\gamma p q^*, h(\theta, q^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta))) + \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta} U(I-\alpha-\gamma p q^*, h(\theta, q^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta)))dF(\theta) \end{aligned} \tag{18}$$

where the stream of efforts $\epsilon_t^{*\,\exp}(\theta)$ is associated with the equilibrium exit $\epsilon^{*\,\exp}(\theta)$.[16] Thus, in equilibrium, a proportion n^* of physicians (i.e., those characterized by a $\lambda < \lambda^*(\kappa)$) will provide $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$ while the rest (i.e., those characterized by $\lambda \geq \lambda^*(\kappa)$) will treat according to their ethical constraint $\lambda\widetilde{\epsilon}(\theta)$. This equilibrium is dependent on physicians not wanting to deviate and lose their patient, i.e., for all physicians with $\lambda < \lambda^*(\kappa)$ and for all illness severities (θ) :[17]

$$V(\delta^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty}\beta^{t-1}\int_{\theta} V(\delta^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta))dF(\theta) \geq V(\delta^*, \lambda\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty}\beta^{t-1}V_t(0,0). \tag{19}$$

Consequently, an equilibrium will be guaranteed if the prospective payment (δ^*) is large enough to ensure that the least ethical physician will provide $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$ (and keep her patient) rather than provide the minimum effort $(\epsilon = 0)$ (and lose her patient).

This δ^* is such that for the physician with $\lambda = 0$,

$$V(\delta^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty}\beta^{t-1}\int_{\theta} V(\delta^*, \lambda^*(\kappa)\widetilde{\epsilon}(\theta))dF(\theta) \geq V(\delta^*, 0) + \sum_{t=2}^{\infty}\beta^{t-1}V_t(0,0) \tag{20}$$

is satisfied for every θ.

Given that the prospective payment is paid prior to the realization of illness severity θ, condition (20) must hold for all illness severities θ. Thus, the prospective payment will need to be relatively high to ensure that the least ethical physician will provide the most ill patient with $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$. Obviously, such a prospective payment would yield rents to all physicians. We assume henceforth that (20) holds with equality for the least ethical physician at the highest value of θ.

[16] Note that everyone who is provided effort greater than $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$ would not be willing to pay the fixed costs to leave i.e., their threat of leaving is non-credible. However, their physicians are unable to provide them with less effort as they are providing their patients according to their ethical constraint.

[17] Notice again that as long as physicians are not initially capacity constrained the value of loosing one's patient is $\sum_{t=2}^{\infty}\beta^{t-1}V_t(0,0)$ as the probability that the patient will return is 0.

Recall that, before competition begins, patients are equally distributed across physician types. Therefore, in the first period, patients have no information regarding their physician's type. However, they know that in equilibrium efforts will be distributed between $[\lambda^*(\kappa)\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$ with the expected effort equal to:

$$n^*\lambda^*(\kappa)\widetilde{\epsilon}(\theta) + \int_{\lambda^*(\kappa)}^{1} \lambda\widetilde{\epsilon}(\theta) d\Gamma(\lambda). \tag{21}$$

Thus, given a particular illness severity θ, the patient will choose the quantity of medical services q^* based on this expected effort.

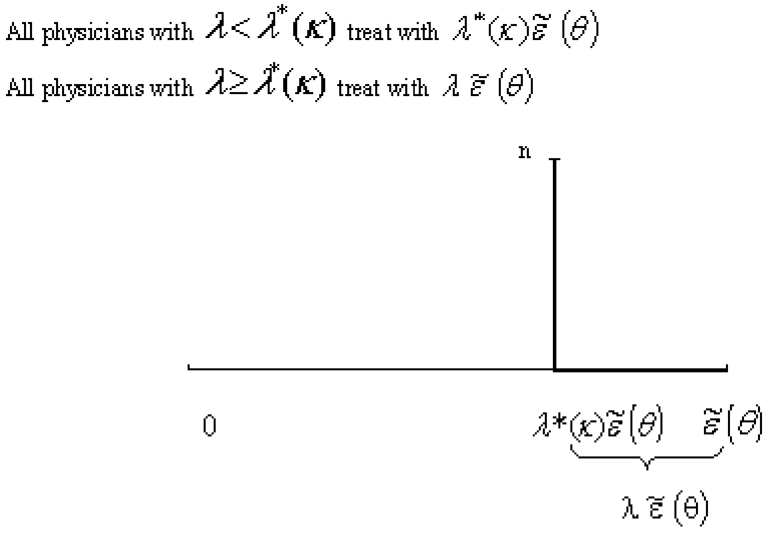
After one period, the patient's physician's $\lambda^{\max}$ is revealed. Given that the patient will remain with the same physician for all periods and that this physician will provide effort equal to $\lambda^{\max}\widetilde{\epsilon}(\theta)$, the patient will choose q^* based on $\lambda^{\max}\widetilde{\epsilon}(\theta)$ rather than (21). The long-run actuarially fair insurance premium (α^*) will be based on the equilibrium prospective payment and the expected medical expenditures.[18]

Proposition 2:

In the presence of switching costs, the equilibrium strategies are such that: (i) physicians with $\lambda \leq \lambda^*(\kappa)$ *will provide their patients with* $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$ *if the prospective payment is sufficiently large (i.e., if condition (20) is satisfied for the physician with a* $\lambda = 0$ *at the highest illness severity), and provide effort according to their ethical constraint (*$\lambda\widetilde{\epsilon}(\theta)$*) otherwise, (ii) physicians with* $\lambda > \lambda^*(\kappa)$ *will provide their patients with their minimum effort (*$\lambda\widetilde{\epsilon}(\theta)$*); (iii) patients will stay with their current physician if they receive effort equal to or greater than a either* $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$*, and leave otherwise. Thus, an equilibrium can be achieved where everyone receiving at least* $\lambda^*(\kappa)\widetilde{\epsilon}(\theta)$*, where quantities of medical care are chosen optimally* $q^*(\theta)$*, and where patient-physician relationships are stable.*

[18] Again recall that the equilibrium co-payment γ^* is the one which balances the patient's expected utility gains of fuller insurance with the loss of a higher insurance premium.

Figure 3:



Although the effect of competition is dampened with the introduction of switching costs, competition nonetheless ensures a lower-bound on the effort provided $(\lambda^*(\kappa)\tilde{\epsilon}(\theta))$. Thus, the above equilibrium is characterized by heterogeneity in effort ϵ and quantity of care q for a given illness severity θ. Furthermore, as switching costs tend to zero, the proportion of physicians treating their patients with desired effort will tend to one. This may have important implications from a policy perspective. In fact, according to our model, any mechanism which reduces the costs (both psychic and financial) of moving from one physician to another will lead physicians to provide their patients with their desired levels of treatment. That is, reducing switching costs reduces the negative effect (in terms of treatment) of being randomly assigned to a less ethical physician.[19]

[19] It is important to recall that, in the above, we assumed that patients are risk-neutral with respect to their health. We make this assumption uniquely to keep things as simple as possible. It can be shown, however, that introducing risk-aversion in health is quite simple and leads to results which are qualitatively identical to those presented in this section (i.e., qualitatively identical to those found when introducing non-trivial switching costs).

4.2.3 Uncertainty in the health production function

In this section, we relax the assumption that the relationship between health care provision and post-treatment health is deterministic, while assuming no switching costs ($\kappa = 0$). Although adding uncertainty in the treatment-outcome relationship could take many forms, we adopt a simple form to highlight the possibility that unexpected (i.e., positive or negative) outcomes may occur which are not directly the result of the treatment received. We show that, under certain conditions, an equilibrium will be characterized by all physicians will treat their patients with a unique, illness dependent, effort which is greater than the desired effort. In such an equilibrium, physicians thus practice defensive medicine. Furthermore, some patient-physician relationships will be unstable.

We assume henceforth that the relationship between effort provided by the physician and its effect on health is subject to a random component. More specifically, we now define the health production function as:

$$h(\theta, q, \epsilon + \mu) = h(\theta, q, \epsilon^R),$$

where the realized effort $\epsilon^R = \epsilon + \mu$, and where μ represents an i.i.d. random component. For simplicity, we assume that $\mu \in [\underline{\mu}, \overline{\mu}]$ where $\Phi(\mu)$ is symmetric and single-peeked with $E(\mu) = 0$.[20] We also assume that although the realization of μ is unobservable, its distribution, $\Phi(\mu)$, is common knowledge.

Because of the random component μ, the patient can no longer perfectly infer the physician's effort. For example, if the patient observes that his ex post health is unsatisfactory it may be because: (i) the physician provided insufficient effort, or (ii) the physician provided sufficient effort but the random component μ was negative.

We now turn to solving for the equilibrium by examining the patient's expected outside option if he were to leave his current physician. Again, we assume that patients are initially equally distributed across physician types who are not initially capacity constrained.

[20] We assume that the support of μ (i.e., $[\underline{\mu}, \overline{\mu}]$) is small relative to the support of potential effort ϵ (i.e., $[0, \widetilde{\epsilon}]$).

As before, a patient who leaves his current physician, under the optimal exit strategy, can expect to receive a stream of efforts $\epsilon_t^{\exp}(\theta)$, $t = 2, 3, ...$ (which is the same stream of efforts he could expect to receive in the previous sections where we assumed a deterministic health production function, given that $E(\mu) = 0$).

As a result, a patient who leaves his current physician would receive in the current period:

$$U(I - \alpha - \gamma p q^*, h(\theta, q^*, \epsilon + \mu)), \tag{22}$$

and expect to receive in the future (at least):

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Leave} = \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^*, h(\theta, q_t^*, \epsilon_t^{\exp}(\theta))) dF(\theta). \tag{23}$$

In (22), q^* is based on the realization of illness severity and the patient's expectation about his current physician's effort. Also notice that (23) is identical to the patient's outside option (7) in the model with a deterministic health production function.

If the patient stays, he would receive in the current period:

$$U(I - \alpha - \gamma p q^*, h(\theta, q^*, \epsilon + \mu)).$$

Because of the uncertain relationship between treatment and post-treatment health, the patient will be unable to perfectly infer the effort provided to him by his current physician. This makes determining his expectation about the future efforts to be provided to him by his current physician that much more difficult. Take for example the case where the patient has drawn illness severity θ, has chosen quantity q and observes post-treatment health to be H. From this, the patient can infer ϵ^R. Notice, given the assumptions made about μ, that ϵ^R is an unbiased estimate of the real effort provided by the physician i.e., $E(\epsilon^R) = \epsilon$. By using the unbiased estimate ϵ^R rather than ϵ, the patient can estimate his current physician's $\lambda^{\max}$ for the case where the relationship between treatment and outcome is uncertain (which we denote as $\lambda_{unc}^{\max}$, where $\lambda_{unc}^{\max} = \frac{\epsilon^R}{\widetilde{\epsilon}}$). Further notice that $\lambda_{unc}^{\max}$ is an unbiased estimator of $\lambda^{\max}$. Consequently, if a patient has drawn a $\mu < 0$ ($\mu > 0$), then $\epsilon^R < \epsilon$ ($\epsilon^R > \epsilon$) and $\lambda_{unc}^{\max} < \lambda^{\max}$ ($\lambda_{unc}^{\max} > \lambda^{\max}$).

Using $\lambda^{\max}_{unc}$ (rather than $\lambda^{\max}$) to form expectations about the current physician's future effort provision, the patient can expect to receive in the future (at least):

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Stay} = \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q^{*}, h(\theta, q^{*}, \lambda^{\max}_{unc} \widetilde{\epsilon}(\theta))) dF(\theta)$$

if he stays with his current physician.

The patient will thus be willing to leave his current physician for a randomly drawn outside physician if and only if:

$$\begin{aligned} & \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^{*}, h(\theta, q_t^{*}, \epsilon_t^{\exp}(\theta))) dF(\theta) \\ > \; & \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q^{*}, h(\theta, q^{*}, \lambda^{\max}_{unc} \widetilde{\epsilon}(\theta))) dF(\theta). \end{aligned}$$

Given that $\sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q_t^{*}, h(\theta, q_t^{*}, \epsilon_t^{\exp}(\theta))) dF(\theta)$ represents the discounted expected utility associated with the optimal exit strategy with corresponding reservation effort $\epsilon^{\exp}(\theta)$, the patient will leave his current physician if and only if $\epsilon^{\exp}(\theta) > \lambda^{\max}_{unc} \widetilde{\epsilon}(\theta)$.

If, as before, we denote the physician type who would provide exactly $\epsilon^{\exp}(\theta)$ if she were to strictly follow her ethical constraint as $\widehat{\lambda}$, then the patient would be willing to leave his current physician if and only if his current physician's $\lambda^{\max}_{unc} < \widehat{\lambda}$.

We now turn our attention to the physicians. All physicians characterized by a λ where $\lambda\widetilde{\epsilon}(\theta) + \underline{\mu} < \widehat{\lambda}\widetilde{\epsilon}(\theta)$ (i.e., $\lambda < \widehat{\lambda} - \frac{\mu}{\widetilde{\epsilon}(\theta)} = \widehat{\lambda} + \frac{\overline{\mu}}{\widetilde{\epsilon}(\theta)}$), are at some risk of losing their patient. Although all such physicians could provide $\widehat{\lambda}\widetilde{\epsilon}(\theta) + \overline{\mu}$ to perfectly insure themselves against the loss of a patient, doing so may not be optimal. In fact, all physicians with a $\lambda < \widehat{\lambda} - \frac{\mu}{\widetilde{\epsilon}(\theta)}$ will wish to maximize their expected utility by choosing an optimal level of effort $\epsilon^{\dagger}(\theta)$, where each effort level is associated with a probability of keeping one's patient. In order to determine the optimal effort $\epsilon^{\dagger}(\theta)$, physicians characterized by $\lambda < \widehat{\lambda} - \frac{\mu}{\widetilde{\epsilon}(\theta)}$ must solve the following dynamic programming problem:

$$\begin{aligned} W(Patient &= 1) = \max_{\epsilon(\theta)} \{ V(\delta, \epsilon(\theta)) \\ + \beta[\Pr(\epsilon^{\exp}(\theta), \epsilon(\theta), \mu) W(Patient' &= 1) + (1 - \Pr(\epsilon^{\exp}(\theta), \epsilon(\theta), \mu)) W(Patient' = 0)] \}, \quad (24) \end{aligned}$$

where $W(Patient = 1)$ denotes the discounted expected utility of having a patient and where $W(Patient = 0)$ denotes the discounted utility of losing a patient (which is given by $\sum_{t=1}^{\infty}\beta^{t-1}V_t^{DEV}(0,0)$). We denote $\Pr(\epsilon^{\exp}(\theta), \epsilon(\theta), \mu)$ as the probability of retaining the patient into the next period (i.e., the probability that given the realization of θ, $\epsilon(\theta)$, the random draw from $\Phi(\mu)$, and the reservation effort associated with the patient's optimal exit strategy $\epsilon^{\exp}(\theta)$, that the physician's $\lambda_{unc}^{\max} > \widehat{\lambda}$). It is important to underline the fact that the probability of losing one's patient (i.e., $1 - \Pr(\epsilon^{\exp}(\theta), \epsilon(\theta), \mu)$) is the probability that $\epsilon(\theta) + \mu < \widehat{\lambda}\widetilde{\epsilon}(\theta)$.[21]

Notice that:

(i) $\epsilon^{\dagger}(\theta) \geq \widehat{\lambda}\widetilde{\epsilon}(\theta) - \overline{\mu}$ (if not, $\lambda_{unc}^{\max}$ will always be less than $\widehat{\lambda}$ which would always result in the physician losing her patient);

(ii) $\epsilon^{\dagger}(\theta) \leq \widehat{\lambda}\widetilde{\epsilon}(\theta) - \underline{\mu}$ (if not, $\lambda_{unc}^{\max}$ will always be greater than $\widehat{\lambda}$ and the patient will stay for sure -but the same could also be achieved at $\epsilon^{\dagger}(\theta) = \widehat{\lambda}\widetilde{\epsilon}(\theta) - \underline{\mu}$ which would require less effort);

From the above solution to (24), the physician will provide effort equal to: (i) $\lambda\widetilde{\epsilon}(\theta)$ if $\epsilon^{\dagger}(\theta) < \lambda\widetilde{\epsilon}(\theta)$ or (ii) $\epsilon^{\dagger}(\theta)$ if $\epsilon^{\dagger}(\theta) > \lambda\widetilde{\epsilon}(\theta)$.

From the above partial solution, the effort levels should now be distributed between $[\epsilon^{\dagger}(\theta), \widetilde{\epsilon}(\theta)]$ with a mass of physicians providing exactly $\epsilon^{\dagger}(\theta)$ (and with realized efforts $(\epsilon^{R}(\theta))$ distributed between $[\epsilon^{\dagger}(\theta) + \underline{\mu}, \widetilde{\epsilon}(\theta) + \overline{\mu}]$). However, this is not full the story. Patients can now calculate a new $\epsilon^{\exp}(\theta)$, say $\epsilon_1^{\exp}(\theta)$, by considering all potential exit strategies, assuming this time: (i) that they would randomly draw from the pool of physicians if they left their current physician, and (ii) that each physician for whom $\epsilon^{\dagger}(\theta) > \lambda\widetilde{\epsilon}(\theta)$ would provide $\epsilon^{\dagger}(\theta)$, while physicians for whom $\epsilon^{\dagger}(\theta) < \lambda\widetilde{\epsilon}(\theta)$ would provide $\lambda\widetilde{\epsilon}(\theta)$ (i.e., their minimal effort). Thus, patients should now be willing to leave their current physician if $\epsilon^{R}(\theta) < \epsilon_1^{\exp}(\theta)$. If, as before, we denote $\widehat{\lambda}_1$ to be the physician type that would provide exactly $\epsilon_1^{\exp}(\theta)$ if she were to strictly follow her ethical constraint, patients

[21] The present value of losing a patient is given by $\sum_{t=1}^{\infty}\beta^{t-1}V_t(0,0)$. Again, this is because a patient who leaves his current physician will be reassigned to the same physician with probability zero. Although a share of patients will leave their physician each period, these patients will be reassigned to physicians independently of whether or not these same physicians lost a patient in the same period. Thus, the probability of being assigned a new patient in each period is independent of the physician's action and thus does not figure into the physician's dynamic programming problem. As will be shown later on, no physician will be capacity constrained in equilibrium.

will leave their current physician if $\lambda_{unc}^{\max} < \widehat{\lambda}_1$.

Given this new credible threat, physicians must solve a new dynamic programming problem, one similar to that defined in (24). That is, the physician must resolve (24) but where $\Pr(\epsilon_1^{\exp}(\theta), \epsilon(\theta), \mu)$ (i.e., the probability that the patient remains) reflects the patient's new strategy (i.e., the patient's new expected outside option).

Notice that the new $\epsilon^\dagger(\theta)$ will be such that:

(i) $\epsilon^\dagger(\theta) > \widehat{\lambda}_1\widetilde{\epsilon}(\theta) - \overline{\mu}$ (if not, $\lambda_{unc}^{\max}$ will always be less than $\widehat{\lambda}_1$ which would result in the physician always losing her patient);

(ii) $\epsilon^\dagger(\theta) < \widehat{\lambda}_1\widetilde{\epsilon}(\theta)$ $-\underline{\mu}$ (if not, $\lambda_{unc}^{\max}$ will always be greater than $\widehat{\lambda}$ and the patient will stay for sure- but the same could be achieved at $\epsilon^\dagger(\theta) = \widehat{\lambda}\widetilde{\epsilon}(\theta) - \underline{\mu}$ which would require less effort). Consequently, physicians will provide effort: (i) $\lambda\widetilde{\epsilon}(\theta)$ if $\epsilon^\dagger(\theta) < \lambda\widetilde{\epsilon}(\theta)$ or (ii) $\epsilon^\dagger(\theta)$ if $\epsilon^\dagger(\theta) > \lambda\widetilde{\epsilon}(\theta)$.

We continue with the above logic until all physicians treat with the same effort $\widetilde{\epsilon}(\theta)$. Patients can thus expect to receive effort $\widetilde{\epsilon}(\theta)$ if they leave. Thus, if $\epsilon^R < \widetilde{\epsilon}(\theta)$, or equivalently, if $\lambda_{unc}^{\max} < 1$, the patient would be willing to leave. Given the symmetric distribution of μ, half of patients would leave if their physician provided exactly $\widetilde{\epsilon}(\theta)$. Providing exactly $\widetilde{\epsilon}(\theta)$ and losing their patient with probability $\frac{1}{2}$ may not be optimal. In order to determine the optimal effort $\epsilon^\dagger(\theta)$ under the patient's new strategy, physicians must solve a revised version of the dynamic programming model (i.e., assuming that patients will leave if $\epsilon^R(\theta) < \widetilde{\epsilon}(\theta)$); leading to a new distribution of observed effort $\epsilon^R(\theta) \in [\epsilon^\dagger(\theta) + \underline{\mu}, \epsilon^\dagger(\theta) + \overline{\mu}]$. At this point, however, patients need not change their threat point. No patient would be willing to leave a physician who has provided at least $\widetilde{\epsilon}(\theta)$, given that this is the desired effort (greater efforts would lead to greater prospective payments which would not justified from a patient's utility maximizing standpoint). Similarly, no physician would be willing to provide less than $\epsilon^\dagger(\theta)$ given their patient's strategy. In equilibrium, this optimal level of effort ($\epsilon^\dagger(\theta)$) is denoted as $\epsilon^*(\theta)$.[22]

Two important predictions of the model with uncertainty should be highlighted. First, in

[22] Assuming that a condition analogous to (15) is satisfied.

equilibrium, physicians over-treat their patients in order to partially insure themselves against negative shocks which could lead to the loss of a patient. This type of over-treatment can be considered an additional form of defensive medicine In the literature, the most commonly cited form of defensive medicine results from partial insurance against potential medical-malpractice litigation (Danzon, 2000). Second, our model predicts that a proportion of patients (i.e., those who receive an effort level less than $\widetilde{\epsilon}(\theta)$) will leave their current physician for another in each period. This prediction is consistent with the data which show that four to eleven per cent of patients switch physicians annually.

As before, the discounted utility of losing a patient is given by $\sum_{t=2}^{\infty}\beta^{t-1}V_t(0,0)$. This is simply because a patient who leaves his current physician will be reassigned to the same physician with probability zero. Furthermore, the probability of being assigned a new patient (resulting from a patient leaving a competing physician) is also independent of whether or not the physician, herself, has lost a patient in the current period. This would not, however, be the case if physicians reached their capacity constraint. In equilibrium, no physician will ever become capacity constrained (as long as the capacity constraint is not too low). This is because, in each period, the probability of loosing a patient and gaining a new one are equal.[23] Thus, to reach her capacity constraint, a physician would have to systematically maintain patients (i.e., systematically draw relatively high μ) while also systematically being assigned patients who have left other physicians. Thus, these probabilities do not figure into the physician's dynamic programming program.

The prospective payment must be such that for the physician with $\lambda = 0$,

$$V(\delta^*, \epsilon^*(\theta)) + \sum_{t=2}^{\infty}\beta^{t-1}\int_{\theta}V(\delta^*, \epsilon^*(\theta))dF(\theta) \geq V(\delta^*, 0) + \sum_{t=2}^{\infty}\beta^{t-1}V_t(0,0) \tag{25}$$

is satisfied for every θ.

[23] For example, take the case where there are 1000 physicians and 100000 patients. Each physician initially has 100 patients. If the probability of losing a patient is 4 per cent, each physician can expect to lose 4 patients each period. However, each physician can also expect to gain 4 patients each period. Thus, each physician can expect to maintain the same amount of patients each year. Thus, in equilibrium, physicians maintain, on average, 100 patients and do not reach their capacity constraint.

Given that the prospective payment is paid prior to the realization of illness severity θ, condition (25) must hold for all illness severities θ. Thus, the prospective payment will need to be relatively high to ensure that the least ethical physician would provide the most ill patient with his desired level of effort. Obviously, such a prospective payment would yield rents to all physicians. We assume henceforth that (25) holds with equality for the least ethical physician at the highest value of θ.

Because quantity $q^*(\theta)$ and effort $\epsilon^*(\theta)$ will always be chosen in equilibrium, the actuarially-fair insurance premium is given by:

$$\alpha^* = \pi \int_\theta ((1-\gamma^*)pq^*(\theta))dF(\theta) + \pi\delta^*.$$

Finally, the patient's ex post utility is given by:

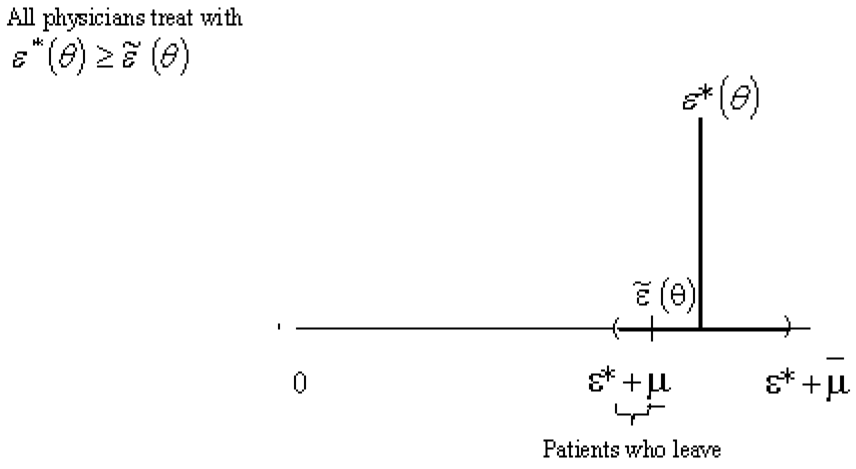
$$U(I - \alpha^* - \gamma^* pq^*(\theta), h(\theta, q^*(\theta), \epsilon^R(\theta))).$$

That is, although all patients will choose the same amount of observable care q^* (for a given illness severity θ), pay the same insurance premium α^*, and pay the same out of pocket expenses $\gamma^* q^*$, their ex post utility will depend on $\epsilon^R(\theta) = \epsilon^*(\theta) + \mu$. Thus, in equilibrium, ex post health will not only depend on the random draw from the illness severity distribution $F(\theta)$, but also the random draw from the distribution of $\Phi(\mu)$.

Proposition 3

In the absence of switching and in the presence of uncertainty the equilibrium strategies are such that: (i) all physicians (irrespective of their λ) will treat their patients with a unique, illness dependent, effort $\epsilon^*(\theta)$ which is greater than the desired effort ($\widetilde{\epsilon}(\theta)$) if the prospective payment is sufficiently large (i.e., if condition (25) is satisfied for the physician with $\lambda = 0$ facing the highest illness severity), and provide effort according to their ethical constraint $\lambda\widetilde{\epsilon}(\theta)$ otherwise; (ii) patients will stay with their current physician if they receive effort greater than or equal to $\epsilon^*(\theta)$, and leave otherwise. Thus, a pooling equilibrium can be achieved with everyone receiving the same level of effort for each illness severity, where quantities of medical care are chosen optimally $q^*(\theta)$, and where some patient-physician relationships are unstable.

Figure 4:


All physicians treat with
$\varepsilon^*(\theta) \geq \widetilde{\varepsilon}(\theta)$
$\varepsilon^*(\theta)$
$\widetilde{\varepsilon}(\theta)$
0
$\varepsilon^* + \mu$
$\varepsilon^* + \bar{\mu}$
Patients who leave

Results from this section suggest that any mechanism which reduces uncertainty in the treatment-outcome relationship will lead to a more efficient provision of care (i.e., reduce over-treatment) and will lead to more stable physician-patient relationships.

5 Conclusion

In this paper we examine the role of competition in the physicians market as a means of encouraging physicians to provide desired levels of care in a setting characterized by information asymmetry. In order to examine this role, we adopt a repeated game setting and solve for equilibria supported by credible threats. Our framework is distinguished, most notably, from the previous literature by this dynamic element as well as the introduction of unobserved heterogeneity in the physicians market, which allows us to endogenize patients' outside options.

In the static framework, we show that all physicians will provide their minimum amount of unobservable effort, i.e., the amount determined by their ethical constraint. Consequently, the equilibrium is characterized by heterogeneity in effort (conditional on a given illness severity). In the dynamic framework, however, we show that competition may serve as an important mechanism to induce the desired provision of unobserved elements of medical care. More specifically, we show that, under certain conditions, competition may provide enough incentives for all physicians to provide their patients with their desired level of effort irrespective of physicians' ethical constraints. We also show that the introduction of switching costs may dampen the effect of competition yielding some heterogeneity in treatments. Nonetheless, competition provides a lower bound on the provision of effort in the presence of such switching costs. Finally, we show that in the presence of an uncertain treatment-outcome relationship, physicians will wish to over-provide care and a proportion of patients will leave their physician in each period. That is, both defensive medicine and unstable patient/physician relationships will occur. It is worth noting that our results do not depend on the patient observing his physician's effort prior to treatment decisions (as suggested by Ma and McGuire (1997)) nor does it require the patient's knowledge of their physician's type.

Our model suggests that even in the presence of switching costs and uncertainty, competition may create important incentives in markets where certain valued inputs are unobservable to both consumers and third parties. Our results also suggest that public policies which seek to reduce switching costs and/or uncertainty in the treatment-outcome relationship are likely to contribute to a more efficient provision of care.

References

[1] Allard, M., Cremer, H., and M. Marchand (2001) ‘Incentive Contracts and the Compensation of Health Care Providers,’ *Economie Publique* 3, 37-54.

[2] Arrow, K (1963): ‘Uncertainty and the Welfare Economics of Medical Care,’ *American Economics Review* 53, 941-69.

[3] Blomqvist, Å (1991) ‘The doctor as double agent: Information asymmetry, health insurance, and medical care,’ *Journal of Health Economics* 10, 411-422.

[4] Danzon, P. (2000) ‘Liability for Medical Malpractice,’ in A.J. Culyer and J.P. Newhouse eds. Handbook of Health Economics (Amsterdam: Elsevier Science)

[5] Dranove, D. (1988) ‘Demand inducement and the physician-patient relationship,’ *Economic Inquiry* 26, 28-98.

[6] Ellis, Randall (1998) ’Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins,’ *Journal of Health Economics* 17, 537-55.

[7] Ellis, Randall and Thomas McGuire (1986) ‘Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply’ *Journal of Health Economics* 5, 129-51.

[8] Gal-Or, Esther (1999) ‘Optimal Reimbursement and Malpractice Sharing Rules in Health Care Markets,’ *Journal of Regulatory Economics* 16, 237-65.

[9] Gaynor, M. and W. Vogt (2000) ‘Antitrust and Competition in Health Care Markets’ in: A.J. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics,* (Amsterdam: Elsevier Science North-Holland), Chapter 27.

[10] Léger, P.T. (2000) ‘Quality Control Mechanisms under Capitation Payment for Medical Services,’ *Canadian Journal of Economics* 33, 564-88.

[11] Ma, Ching-to Albert (1994) ‘Health Care Payment Systems: Cost and Quality Incentives,’ *Journal of Economics and Management Strategy* 3, 93-112.

[12] Ma, Ching-to Albert and Thomas G. McGuire (1997) ‘Optimal Health Insurance and Provider Payment,’ *American Economic Review* 87, 685-704.

[13] Rochaix, L., (1989) ‘Information Asymmetry and Search in the Market for Physicians’ Services,’ *Journal of Health Economics* 8, 53-84.

[14] Sobsero, M.E.W. (2001) ‘Do capitated primary physicians “encourage” their high utilization patients to leave their practice?,’ Dissertation, The University of Rochester, Rochester, New York.

[15] Wedig, Gerald, Mitchell, Janet B. and Jerry Cromwell (1989) ‘Can Optimal Physician Behavior Be Obtained Using Price Controls?,’ *Journal of Health Politics, Policy and Law* 14, 601-20.

Appendix

Suppose that when forming expectations about his current physician’s future effort, the patient does not use $\lambda^{\max}$ but rather uses the conditional expectation of λ given $\lambda^{\max}$. That is, by inferring his physician’s $\lambda^{\max}$, the patient knows that his physician’s actual $\lambda \in [0, \lambda^{\max}]$ and therefore takes the expected value of his current physician’s λ based on this interval, i.e., $\lambda^1 = \int_0^{\lambda^{\max}} \lambda d\Gamma(\lambda)$. Thus, equation (10) can be rewritten by replacing $\lambda^{\max}$ by λ^1. That is, the patient’s strategy is simply to leave (stay with) his current physician if $\epsilon^{\exp}(\theta) > (\leq)\ \lambda^1\widetilde{\epsilon}(\theta)$. Notice, however, that no matter how much effort a physician provides, her patient will always leave. Take, as an extreme case, a physician who is characterized by a $\lambda = 1$. This physician will always provide her patient with the desired level of effort ($\widetilde{\epsilon}(\theta)$). However, the patient would infer a $\lambda^{\max} = 1$ and thus a $\lambda^1 = \int_0^1 \lambda d\Gamma(\lambda) = \frac{1}{2}$ (under a symmetric distribution). If $\epsilon^{\exp}(\theta) > \frac{1}{2}\widetilde{\epsilon}(\theta)$, then the patient will always leave his physician, no matter how much effort he receives. Given that the patient will always leave, the physician should always provide her minimal effort. If every physician provides her minimal effort, however, the patient's strategy of leaving his current physician if $\epsilon^{\exp}(\theta) > \lambda^1\widetilde{\epsilon}(\theta)$ is irrational. The patient should leave his current physician if and only if $\epsilon^{\exp}(\theta) > \lambda^{\max}\widetilde{\epsilon}(\theta) = \lambda\widetilde{\epsilon}(\theta)$. It is also important to note that the above rationale holds for any belief $\lambda^1 \in [0, \lambda^{\max}]$.