A Shirking Theory of Referrals

Damien S. Eldridge

Discussion Paper No. A07.05
ISBN 1 92137 7259
ISSN 1441 3213 November 2007

A shirking theory of referrals

Damien S. Eldridge*
Department of Economics and Finance
La Trobe University
Bundoora, Vic, 3086
Australia
E-mail: d.eldridge@latrobe.edu.au

November 2007

Abstract

Many service industries, including the medical and legal professions in some countries, display a gated structure. Rather than approaching a final producer directly, a consumer will first seek a referral from an intermediary. In this paper, we provide one possible explanation for such an industry structure. If the outcome of a transaction depends on producer effort, which is unobservable and unverifiable, then the market may fail to generate a Pareto optimal outcome. This is the standard moral hazard problem. If consumers had a long-run relationship with producers, this type of market failure might be avoided. However, in some industries, consumers will only have a short-run relationship with producers. A gatekeeping intermediary may provide an opportunity for reputation effects to apply in such a setting. By aggregating many potential consumers, gatekeeping intermediaries can create an artificial long-run relationship between a consumer and a producer. This long-run relationship reduces the incidence of shirking on the part of the producer.

1 Introduction

The potential for moral hazard problems to result in market failure is well understood.[1] Indeed, there is a large literature on the design of contracts to alleviate moral hazard problems.[2] This literature focuses on a static setting in which the principal and the agent interact only once. Except in very restrictive circumstances, the market failure can at best be only partially mitigated. For

*This paper is a modified version of Chapter 2 of the author's PhD dissertation (Eldridge [18], [19]). The author would like to thank Peter Bardsley, Suren Basov, Stephen Donald, Ken Hendricks, Preston McAfee, David Sibley, Max Stinchcombe, Andrew Whinston and Tom Wiseman for helpful advice and comments.

[1] See, for example, Arrow ([3], [4]) and Pauly ([48], [49]).

[2] Useful surveys of this literature are provided by Hirshleifer and Riley ([29]), Laffont and Martimort ([39]), Macho-Stadler and Perez-Castrillo ([40]) and Mas-Colell et al ([44]). The central references include Arrow ([3]), Grossman and Hart ([26]), Hermalin and Katz ([28]), Holmstrom ([31]), Jewitt ([34]), Mirlees ([45]), Rogerson ([57]), Ross ([58]), Shapiro and Stiglitz ([62]) and Shavell ([63], [64]).

this reason, moral hazard and other problems involving asymmetric information are often used to justify a variety of consumer protection related regulations.[3] But such regulations are potentially costly and sometimes ineffective.

In many settings involving moral hazard, the transacting parties interact more than once. Repeated interaction potentially allows for greater alleviation of moral hazard than is possible in a static setting, through the use of dynamic punishment strategies and reputation effects.[4] Unfortunately, there are also many occasions in which parties to a transaction do not repeatedly interact. In the absence of repeated interaction, reputation cannot be relied upon to deter moral hazard. We might expect moral hazard problems to be rampant in markets characterised by few or infrequent interactions between the trading parties. Indeed, Mooney and Ryan ([46], p. 134) raise exactly this concern in relation to health care markets:

> "Whilst it is possible that repeated interactions constrain doctors' behaviour, since doctors will not want patients to lose faith in them, such repeated interaction will not take place in all sectors of the health care market. Whilst a dynamic model may be applicable to the GP-patient interaction (since there will be repeated interactions), it is less clear how applicable such a model is to the specialist-patient interaction."

Is consumer protection regulation the only safeguard available in such settings or can institutions be devised that might capture the benefits of a long-run relationship? This paper explores one possible solution. It involves the creation of intermediaries that generate an artificial long-run relationship between the transacting parties by aggregating many short-run relationships. In effect, the intermediaries act as a surrogate long-term partner, leveraging their own repeated relationship with the two transacting parties. This allows the short-run agents to build up a reputation for quality and the short-run principals vicarious access to that reputation.

2 Motivating examples

Variations of this industry structure can be found in many professional service industries, including the medical and legal professions. In many countries, these professions are organised around a gatekeeper. Access to the ultimate producer frequently requires a referral from an intermediary. In this paper, we will explicitly model the organisation of health care markets. However, the model readily translates into the organisation of some other service industries, including the legal profession.

In the medical industries of some Commonwealth countries, it is unusual for patients to visit a specialist without first obtaining a referral from a general

[3] See, for example, the discussions in Damania and Round ([14]), Hadfield et al ([27]) and Smith ([65]).

[4] See, for example, Abreu et al ([1], [2]), Atkeson and Lucas ([5]), Fudenberg et al ([23]), Radner ([53], [54]), Radner et al ([55]), Rogerson ([56]), Rubenstein and Yaari ([60]), Spear and Srivastava ([66]), Stigler ([67], p. 179), Thomas and Worrall ([69]) and Townsend ([72]).

practitioner (GP).[5] The GP is essentially the family doctor. He typically sees a patient many times throughout the patient's life, treating a variety of minor illnesses and referring the patient to an appropriate specialist for more serious complaints. As such, the patient and the GP interact repeatedly over a long period of time. Furthermore, because the GP has a pool of patients, he will typically encounter particular diseases many times. As such, the GP has the opportunity to develop a long-run relationship with particular specialists. While some severe or chronic complaints might require repeated interaction between a patient and a specialist as well, many patient-specialist relationships are inherently short-run. In such circumstances, the GP can potentially leverage his long-run relationships with both patients and specialists to induce an artificial long-run relationship between a patient and a specialist.

While the medical industry in the United States of America is not formally organised in the same way as it is in some Commonwealth countries, some of the key institutions in the US health sector enforce similar arrangements. Health maintenance organizations (HMOs) and preferred provider organizations (PPOs) combine health insurance and the provision of medical care.[6] HMOs typically require patients to see one of their gatekeeping medical practitioners before being referred to an approved specialist. By controlling which specialists receive patients and monitoring which patients are treated by a particular specialist, the HMO can effectively play a role similar to that of the GP in many Commonwealth countries. PPOs are essentially just a less restrictive form of HMO. They have no requirement for the patient to visit a gatekeeper before seeing a specialist. They do, however, provide financial incentives for patients to visit specialists on their list of preferred providers. By controlling which specialists are on this list, they can effectively punish specialists who are suspected of shirking. Just like the GP in our earlier example, HMOs and PPOs have a long-run relationship with both specialists and patients. They can leverage the financial clout provided by their relatively large customer base to punish specialists suspected of shirking.[7]

While we focus on the medical industry example in this paper, there are other industries with a gated structure that might in part be explained by this theory of intermediation, including the legal industry. The structure of the legal industry in some Commonwealth countries appears to be similar to that of the medical industry in those countries. A client needing legal services first visits a solicitor. If the service required is relatively minor, the solicitor may be able to take care of it himself. But if the client's case is going to trial, the solicitor might choose to brief a barrister, who will then represent the client at court. A client might have a repeated relationship with a solicitor because solicitors

[5] Commonwealth countries in which many patients obtain a referral from a general practitioner before seeking the services of a specialist include Australia ([50], p. 421; [12], part 2, p. 3), New Zealand ([59], section 2, p. 14) and the United Kingdom ([10]). This arrangement does not necessarily apply to all medical specialties within these countries. For example, a patient would probably seek a referral before visiting an opthamologist in Australia, but would be unlikely to do so before visiting an optometrist.

[6] Many undergraduate textbooks on health economics contain a discussion of HMOs and PPOs. A particularly good source is Folland et al ([22]).

[7] Clearly, HMOs and PPOs can punish specialists for a number of other undesirable behaviours too. Constraining costs by punishing suspected over-servicing may be one such concern. We are not suggesting that HMOs and PPOs exist solely to punish shirking by medical specialists. We are, however, suggesting that this is one of many roles that they can play.

can handle estate planning, conveyancing and many other legal matters that do not require representation at trial. Even if a particular client does not have a repeated relationship with a solicitor, the client might have sought out the solicitor's services on the recommendation of a friend or family member who has previously employed him. A sequence of such recommendations can give rise to a sequence of clients that might be thought of as a single long-run client. A solicitor typically has many clients, so that repeated relationships between a solicitor and a barrister might also occur. In a similar fashion to a GP in the medical example, the solicitor may be able to leverage his long-run relationship with a particular barrister to ameliorate any moral hazard problem that might normally arise because of the short-run relationship between his client and the barrister.

As with the respective medical industries, the legal industry in the United States of America is not formally organised like it is in some Commonwealth countries. Once again, however, organisations have evolved that, among other things, play a similar role to solicitors in those Commonwealth countries. While many lawyers and some legal practices might specialise in a particular area of the law, multi-purpose law firms also exist. These firms are able to help their clients in a number of disparate and unrelated legal matters. With the exception of possible savings due to economies of scope arising from sharing fixed overhead costs, it is unclear why clients would not prefer to seek out different legal practices for different legal problems. The generation of an artificial long run relationship provides one further argument in favour of the use of multi-product firms. By increasing the extent of interaction between particular clients and the legal practice, multi-purpose law firms provide the client with greater opportunity to punish poor performance. The amount of business the firm loses if the client drops them is greater than it would be if they were a single purpose practice.

3 A competitive model of health care markets

Consider an economy with three groups of agents who live forever. These groups are patients, general practitioners (GPs) and medical specialists. Let patients be indexed by $i \in \{1, 2, \cdots, I\}$, GPs by $j \in \{1, 2, \cdots, J\}$ and specialists by $k \in \{1, 2, \cdots, K\}$. We will assume that there are an infinite number of patients ($I \to \infty$) and specialists ($K \to \infty$), but only a finite number of GPs ($J < \infty$).[8]

In each period, a patient is randomly allocated a disease state ($d \in \{0, 1\}$). Patients may be either well ($d = 0$) or sick ($d = 1$). Following this, each sick patient can choose whether or not to seek treatment. Treatment can sometimes result in a cure, improving the patient's health status for that period. The probability that a disease is cured by treatment increases with the amount of

[8] The reason we assume that there are a countably infinite number of patients, a countably infinite number of specialists and only a finite number of GPs is that it allows us to use the standard version of the strong law of large numbers to make inferences about the number of sick patients in a GPs patient pool in each period. An alternative to this set of assumptions would be to assume that the number of patients is uncountably infinite, the number of specialists is countably infinite and the number of GPs is countably infinite. In order to analyse this alternative version of the model presented in this paper, we would need to use the techniques outlined in Judd ([35]).

effort the specialist devotes to the treatment.[9] Patients can seek a referral to the specialist from a GP if they believe this will increase the probability that high effort treatment is provided. Both referrals and treatments come at a price. For budget constrained patients, the benefits of an increased probability of good health need to be weighed against the foregone consumption of other goods that expenditure on health care entails. We will assume that patients visit neither a GP nor a specialist when they are healthy.[10]

All agents in this economy are price takers who behave as though the existing prices are exogenously specified. We will focus on stationary equilibria for this economy, so that prices don't change over time. The price per referral from any GP is w, while the price per treatment from any specialist is r.

For payoff purposes, time is assumed to be discrete in this economy. Time periods are indexed by $t \in \{0, 1, 2, \cdots\}$, with payoffs occurring at the end of each period. In each period, the market opens and the agents interact within the market. Note that not all agents move at once in the market. The market process involves sequential moves by various agents. Thus the timing of the moves in the market process is important. We will maintain the assumption that time is discrete and index time within a period by $s \in \{0, 1, 2, \cdots\}$. In this fashion, each point in time can be given a unique time stamp of the form $(t, s) \in \{0, 1, 2, \cdots\}^2 = \mathbb{Z}_+^2$.

3.1 The timing of the market in each period

At the beginning of each period ($s = 0$), Nature randomly chooses a disease state for each patient, $d_i \in \{0, 1\}$. Each patient's disease state is chosen as an independent draw from some common distribution, $\Pi : \{0, 1\} \rightarrow [0, 1]$. The probability that any given patient is sick in any given period is π, while the probability that any given patient is well in any given period is $(1 - \pi)$. While each patient's disease state is private information, the distribution from which disease states are drawn is common knowledge.

At $s = 1$, having observed their disease state for the current period, patients choose whether or not to seek treatment. If patients choose to seek treatment, then they also choose whether or not to seek a referral from a GP. If they seek a referral, they choose which GP to visit at $s = 2$. If not, they choose which of the specialists that treats their disease type to visit. Recall that patients who are healthy are assumed to seek neither treatment nor referral.

We will assume that GPs follow up on the outcomes from treatment of any of the patients they refer. In this fashion, the GP knows the entire history of outcomes for each of his previous referrals at the start of each period. At $s = 3$, GPs choose the specialists to which they will refer their patients. Patients who seek a referral are assumed to follow the GPs advice and seek treatment from the specialist to which they are referred. We will assume that each GP refers all of his sick patients in a given period to the same specialist. This assumption is not essential. However, it does simplify the analysis when GPs have finite

[9] While this paper focuses solely on a potential moral hazard problem in the treatment market, a potential adverse selection problem in the treatment market is considered in a companion paper (Eldridge [20]).

[10] If the equilibrium prices for referrals and treatment are positive, this assumption is not needed. Even if these prices are zero, we could avoid making this assumption by introducing an opportunity cost of time (perhaps in the form of foregone leisure) into the model.

patient pools. The reason for this is that it allows specialists to estimate the size of a GP's patient pool from the number of patients that GP refers to him. The assumption is relatively innocuous when GPs have infinite patient pools, which is the case that we will focus on in this paper.

Following this, at $s = 4$, specialists choose how much effort to devote to treating each patient. For each patient, they can independently choose either high effort ($e = 1$) or low effort ($e = 0$). The effort choice affects the probability of the patient being cured.

Finally, at $s = 5$, Nature chooses whether or not each patient is cured. If a patient is cured, he will have good health in that period ($h = 1$), while if the patient is not cured, he will have bad health ($h = 0$). We will assume that high effort on the part of the treating specialist always results in the patient being cured, while low effort results in a cure with probability $\theta \in (0, 1)$. If the patient chose not to seek treatment, he will not be cured.

3.2 Player objectives

Every agent in this game is assumed to maximise the discounted present value of a sequence of per-period von Neuman Morgernstern expected utility functions. Furthermore, they all have a common rate of time preference, represented by the stationary discount factor $\delta \in [0, 1)$. Thus differences in the preferences of the three groups of agents arise from differences in their per-period preferences. These are outlined below.

3.2.1 Patients

Patients all have identical per-period preferences defined over their expenditure on health care (p) and their health state (h). These preferences may be represented by a quasi-linear per-period Bernoulli utility function of the form

$$u(h_t, p_t) = B(h_t) - p_t,$$

where $B(0)$ is normalised to zero and $B(1) = B > 0$.

The health state in each period is a random variable and may vary across patients. It depends on whether or not the patient is sick, whether or not treatment is sought and, if so, the effort devoted to treating the patient. Since each patient knows their disease status before having to make any decisions about treatment, the probability of good health in period t is given by

$$\gamma_t = \begin{cases} 1 & \text{if either } d = 0 \text{ or high effort treatment is received when } d \neq 0; \\ \theta & \text{if } d \neq 0 \text{ and low effort treatment is received;} \\ 0 & \text{if } d \neq 0 \text{ and no treatment is received.} \end{cases}$$

Note that the probability of good health depends on the amount of effort exerted by a specialist when treating a patient. This is not observed by patients, so that they do not know the exact probability of good health in period t after making their treatment and referral decisions. Let $\lambda_{i,t}$ denote a patient's belief that he will receive high effort treatment if he seeks a referral. In this paper, we will restrict our attention to symmetric equilibria in which each patient holds the same beliefs. Thus we can set $\lambda_{i,t} = \lambda_t$ for all $i \in \{1, 2, \cdots, I\}$. With these beliefs, each patient's subjective estimate of the probability of good health in period t is given by

$$\mu_t = \begin{cases} 1 & \text{if } d = 0; \\ \lambda_t + \theta(1 - \lambda_t) & \text{if } d \neq 0 \text{ and treatment is sought;} \\ 0 & \text{if } d \neq 0 \text{ and no treatment is sought.} \end{cases}$$

Expenditure on health care in any given period may also vary across patients. It will depend on whether or not the patient seeks treatment and, if so, whether or not the patient also seeks a referral. We will assume that patients do not seek a referral if they do not also desire treatment. Thus a patient's expenditure on health care is given by

$$p_t = \begin{cases} 0 & \text{if neither treatment nor referral is sought;} \\ r & \text{if treatment is sought without a referral;} \\ w + r & \text{if both treatment and referral are sought.} \end{cases}$$

Thus a patient's per-period expected utility is

$$Eu(h_t, p_t) = \mu_t B - p_t.$$

Patients do not know their future disease states. However, they do not have to make any decisions in any given period prior to observing their disease state in that period. As such, a patient's remaining lifetime expected utility after he has observed his disease state in any given period can be conveniently represented as:

$$U(p_t; d_t, H_{i,t}) = \mu_t B - p_t + \delta M,$$

where $H_{i,t}$ is the entire history that is observed by the patient prior to period t and M is the expected continuation payoff to the patient. The expected continuation payoff is the next period value of the total utility the patient expects to receive from all subsequent periods following the completion of the current period's stage game. Clearly the expected continuation payoff will be a function of the distribution of disease states within the population (Π), the patient's future referral and treatment decisions and the effort that specialists devote to treating the patient.

3.2.2 General practitioners

GPs are assumed to be risk-neutral. The Bernoulli utility function that represents their per-period preferences is simply their per-period profit. Assuming that they have a constant marginal cost of k per referral and no fixed costs, their per-period profits are

$$v(w, n_{j,1,t}) = n_{j,1,t}(w - k) = \sum_{i=1}^{I}(w - k)1_{i,j,t}.$$

The number of referrals a particular GP makes in period t is equal to the number of sick patients he has in that period ($n_{j,1,t}$). Note that the indicator variable ($1_{i,j,t}$) takes on the value one if patient i obtains a referral from GP j in period t and the value zero otherwise.

There are two sources of uncertainty that affect a GP's payoffs. First, there is the fact that $n_{j,1,t}$ is a random variable. We show later in this paper that this source of uncertainty disappears if the GP has an infinite patient pool, since the number of his patients who are sick in any particular period is almost surely infinite. The second source of uncertainty relates to the number of patients that will visit him in future periods. This may, in part, depend on the GP's success in motivating specialists to exert high effort when treating patients that are referred by him. If a GP makes his referral decision before observing the number of patients that visit him during period t, then his expected lifetime utility can be conveniently written as

$$V(\cdot) = V^B(\cdot) = E(n_{j,1,t} \,|n_{j,t})(w - k) + \delta Q = \pi n_{j,t} + \delta Q,$$

where Q is the GP's expected continuation payoff. However, if the GP does not make his referral decisions until after he has observed the number of patients that are seeking his referral services in the current period, then his lifetime expected utility becomes

$$V^A(\cdot) = n_{j,1,t}(w - k) + \delta Q.$$

As we show later in this paper, if a GP has an infinite patient pool, then $n_{j,1,t}$ converges almost surely to $\pi n_{j,t}$. As such, $V^A(\cdot)$ converges almost surely to $V^B(\cdot)$. Since we are interested in the case in which GPs have infinite patient pools, we will focus on $V^B(\cdot)$.

3.2.3 Medical Specialists

We will assume that all medical specialists have the same per-period preferences. These preferences are defined over the price they receive for providing treatment (r) and the effort they devote to that treatment (e). We will assume that these preferences may be represented by a quasi-linear Bernoulli utility function that is additively separable across patients. Each specialist's per-period per-patient preferences are given by

$$\widehat{z}(r, e) = r - C(e),$$

where treatment effort can either be high ($e = 1$) or low ($e = 0$). The cost of low effort, $C(0)$, will be normalised to zero, while the cost of high effort is $C(1) = C > 0$. Note that an implication of additive separability across patients is that the marginal disutility of effort is constant. It does not vary with the total amount of effort being exerted on all of the patients treated by a specialist in any given period. Let $n_{k,t}$ denote the number of patients a particular specialist has in period t and the variable $1_{i,k,t}$ indicate whether or not patient i was treated by this specialist k in period t. Since specialist per-period preferences are additively separable across patients, they may be represented by a per-period Bernoulli utility function of the form

$$z(r, e) = \sum_{i=1}^{I} \{r - C(e_{i,k,t})\} 1_{i,k,t} = n_{k,t} r - \sum_{i=1}^{I} C(e_{i,k,t}) 1_{i,k,t}.$$

As was the case with the GPs, the only uncertainty that affects medical specialists relates to the number of patients that will seek their treatment services in future periods. This may depend on a number of factors, including the specialist's current effort choices, the actual health outcomes following low effort treatment and the incidence of the disease in future periods. The specialist's expected lifetime utility can be conveniently written as

$$Z(\cdot) = n_{k,t}r - \sum_{i=1}^{I} C\left(e_{i,k,t}\right) 1_{i,k,t} + \delta Y,$$

where Y denotes the specialist's expected continuation payoff. This payoff will clearly depend on the number of patients that visit him in the future. While this is in part random, varying with Nature's decisions about disease incidence, it will also depend on the future decisions of patients and GPs.

3.3 The nature of equilibria

In this paper, we impose some sequential rationality restrictions on the set of acceptable equilibria for the supergame. We do this by solving the game by backwards induction. Since the market process in each period involves incomplete information that is not always revealed following its completion, there will be many non-singleton information sets in the supergame. As such, we would expect the set of equilibrium strategy profiles to depend on players' beliefs about the prior history of the game at each of their information sets. In an infinitely repeated game, these beliefs can be rather complicated.

In order to avoid the complicated beliefs that can arise in infinitely repeated games, we will make use of the competitive nature of our model of health care markets. In particular, we will solve the model in three stages. First, we will consider an infinitely repeated game between a representative patient and a representative specialist in the absence of GPs. This will provide a benchmark for the outcome if a patient chooses to self-refer. We will then consider an infinitely repeated game between a representative GP and a representative specialist. We will assume that the GP has a constant patient pool of infinite size. This will provide a benchmark for the outcome if a patient seeks a referral from a GP. Finally, we will consider a representative patient's choice between self-referral and seeking a referral from a GP.

4 The patient-specialist supergame

Our explanation for the structure of gated industries focuses on the role of intermediaries in ameliorating the market failure resulting from a static moral hazard problem. In order to pursue this line of reasoning, we need to understand the outcomes that result in such industries when intermediaries are not present. These outcomes are analysed in this section.

Suppose that there are no GPs. In these circumstances, the only decisions that a patient makes are whether to seek treatment and, if so, which specialist to visit. The only decisions that a specialist needs to make involve the amount of effort to devote to treating each patient that visits him.

A desire to impose some credibility restrictions on the use of punishment threats is implicit in our decision to solve the entire supergame by backwards induction. We will maintain this approach within the patient-specialist supergame. In order to impose some degree of sequential rationality on the set of acceptable equilibria for the patient-specialist supergame, we will restrict our attention to perfect public equilibria.[11] Perfect public equilibria have two desirable properties, both of which simplify the process of finding sequentially rational Nash equilibria for a supergame with imperfect monitoring. The first property is belief independence. It is known that beliefs exist which will support a perfect public equilibrium as a perfect Bayesian equilibrium.[12]. As such, we know that any perfect public equilibrium is sequentially rational without needing to calculate the actual beliefs that make it so. The other desirable property of perfect public equilibria is that they are recursive, in the sense that, from any point in time, they will induce a perfect public equilibrium in every subsequent continuation game.

However, we would like to extend this backwards induction reasoning to the stage game itself. Recall that each specialist gets to make all of his effort choices after he observes which patients are seeking treatment from him in that period, as well as any continuation payoffs that the patients can credibly promise. As such, we will solve each specialist's problem first, conditional on the patients' strategy choices. We will then solve each patient's problem under the assumption that the specialists will respond accordingly.

Specialists' strategy choices in the stage game simply amount to a choice of treatment effort for each patient that seeks treatment from them. We assumed earlier that the specialists' payoffs were additively separable across patients. Furthermore, patients cannot directly communicate treatment outcomes with each other in this model. As such, there is no direct gain for a specialist from linking his effort choices across patients. We will assume throughout this section that each specialist chooses the effort he will devote to treating each of his patients independently of the effort devoted to treating his other patients. When choosing the amount of effort to devote to treating a particular patient, the specialist will simply way up the expected lifetime utility of exerting high effort against that of exerting low effort. In each case, the expected lifetime utility will clearly depend on the expected continuation payoffs promised by the patient.

Patients' strategy choices consist of three components in this model. These components are a treatment decision, the choice of specialist in the event that treatment is chosen and a credible statement about their future treatment and specialist choices if they happen to get sick again. These future strategy choices can be represented by the choice of a continuation payoff for that specialist. This continuation payoff can vary with treatment outcomes.

4.1 The treatment choices of specialists

Since we wish to solve the specialist's problem first, suppose that a representative patient (i) who is sick has decided to seek treatment from some specialist (k). The patient will be able to motivate high effort from this specialist if and only if he can credibly promise continuation payoffs that will ensure that both the specialist's high effort incentive compatibility constraint and participation constraint are satisfied. Recall that the patient only observes the outcome of

[11] The seminal papers on the perfect public equilibrium concept are Abreu et al ([2]) and Fudenberg et al ([23]). Useful discussions of the concept can be found in Fudenberg and Tirole ([24], chapter 5, sections 5 and 6) and Mailath and Samuelson ([42], chapter 7).

[12] See Fudenberg et al ([23], pp. 8-9). More recent work on belief-free equilibria can be found in Ely et al ([21]).

the treatment and not the effort devoted to treatment by the specialist. Furthermore, the patient only observes his own health outcomes and not those of other patients treated by the specialist. Similarly, the specialist only observes the patient's health outcomes when he treats the patient and not when the patient is treated by another specialist. Since we are restricting our attention to public strategies, the continuation payoffs can only be conditioned on the patient's history of health outcomes following treatment by this specialist. With prices fixed, the only punishment available to a patient is to dump the treating specialist.

This dumping strategy could be employed temporarily, with the patient refusing to visit that particular specialist at any time in the next T periods or the next T times he is sick. Alternatively, it could be employed permanently, with the patient refusing to ever seek treatment from that specialist again. The patient could choose to trigger the punishment only after a series of bad outcomes, or if the proportion of bad outcomes exceeds some threshold. Alternatively, the patient could trigger the punishment after only a single bad outcome. The most extreme punishment that could be chosen involves the patient permanently dumping the specialist if there is ever a bad outcome.[13] The extreme punishment strategy requires the specification of two continuation payoffs, one for histories in which the patient always has good outcomes following treatment ($V(1)$) and one for histories in which there is at least one bad outcome ($V(0)$). We will focus on this strategy in the analysis below.

Proposition 1 *The specialist will prefer to provide high effort treatment to a patient rather than low effort treatment if and only if the treatment price matches or exceeds some threshold price.*

Proof. Under the extreme punishment strategy employed by the patient, the specialist's high effort incentive compatibility constraint is:

$$r - C + \delta V(1) \geq r + \delta\left[\theta V(1) + (1-\theta)V(0)\right],$$

which simplifies to:

$$V(1) \geq V(0) + \frac{C}{\delta(1-\theta)}.$$

Since punishment involves dumping the specialist forever, we can set $V(0) = 0$. Furthermore, since the price of treatment is exogenous, the highest continuation payoff for a history of only good outcomes that a patient could credibly offer is a constant stream of the static payoff to high effort. This is:

$$V(1) = \sum_{t=0}^{\infty} \delta^t \pi(r - C) = \frac{\pi(r-C)}{(1-\delta)}.$$

Substituting these continuation payoffs into the high effort constraint yields:

$$\frac{\pi(r-C)}{(1-\delta)} \geq \frac{C}{\delta(1-\theta)}.$$

[13] Recall that in this model, bad health outcomes following treatment can only occur if the specialist devotes low effort to that treatment.

After some rearranging, this becomes:

$$r \geq \left[\frac{(1-\delta) + \pi\delta(1-\theta)}{\pi\delta(1-\theta)} \right] C,$$

or alternatively:

$$r \geq \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)} \right] C.$$

We will call this inequality the threshold price inequality. ■

The threshold price referred to in proposition 1 is the lowest price at which a specialist will be willing to provide high effort treatment rather than low effort treatment. Specifically, in the absence of GPs, the threshold price is given by:

$$\widehat{r} = \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)} \right] C.$$

Recall that $\pi \in (0,1)$ and $\theta \in (0,1)$. Furthermore, specialists are neither perfectly patient nor perfectly impatient, so that $\delta \in (0,1)$. As such, the threshold price inequality in Proposition 1 implies the following result.

Proposition 2 *If high effort treatment is to be provided, then the price of such treatment must exceed the marginal cost of such treatment.*[14]

This result is somewhat unusual for a competitive economy. It is generated by the asymmetric information that is present in the market for treatment services. The gap between the threshold treatment price and the disutility incurred by a specialist that provides high effort treatment is an information rent that must be paid in order to induce specialists to provide high effort treatment.

While the threshold price inequality in Proposition 1 guarantees that any specialist that provides treatment will prefer providing high effort treatment to low effort treatment, we still need to establish the circumstances under which a specialist would want to provide treatment of either variety. To simplify matters, we will assume that the specialist's reservation utility has been normalised to zero.

First, let us establish conditions under which the specialist will prefer to provide high effort treatment than provide no treatment whatsoever. If the specialist refuses to treat a patient, he will receive no surplus from that transaction. It is possible that the patient could punish such behaviour in a similar way to the punishment used for a bad health outcome from treatment. However, no such punishment is necessary to induce treatment in those cases where the patient can motivate high effort treatment from the specialist.

Proposition 3 *If the high effort incentive compatibility condition holds, then the specialist will prefer to provide high effort treatment to the patient rather than not treating the patient at all.*

Proof. We know from Proposition 2 that if the high effort incentive compatibility constraint is satisfied, then $r > C$. This is sufficient to ensure that the

[14] This result is similar to that obtained by Klein and Leffler ([36]).

specialist would receive positive surplus if he provides high effort treatment to the patient. Since the specialist receives no surplus if he refuses to treat the patient, he will prefer to provide high effort treatment to the patient rather than not treat the patient at all. ■

Note that the result in Proposition 3 holds, even if no dynamic punishment for non-treatment is used by the patient. If patients are able to motivate high effort from the specialist, they can automatically ensure participation.

Suppose instead that the high effort incentive compatibility constraint does not hold. In this case, the specialist will only provide low effort treatment, if any treatment is provided at all.

Proposition 4 *If the high effort incentive compatibility constraint does not hold, then the specialist will (weakly) prefer to provide low effort treatment to the patient over not treating the patient at all if the prevailing treatment price is non-negative.*

Proof. If the specialist provides low effort treatment, then he only incurs the disutility associated with low effort treatment. Thus his cost of treatment is $C(0) = 0$. As such, any non-negative price for treatment will be sufficient to induce the specialist to offer treatment, even if the patient does not employ any dynamic punishments for non-treatment. ■

Note that if the treatment price is positive, then the specialist will earn a positive surplus from the transaction. Even in the absence of dynamic punishments for non-treatment, this still exceeds the surplus from non-treatment, which is zero. If the price is zero, then in the absence of dynamic punishments for non-treatment, the specialist will be indifferent between providing low effort treatment to the patient and not treating the patient at all. We will adopt the standard convention and assume that when the specialist is indifferent between providing low effort treatment and no treatment, the specialist chooses to provide low effort treatment.

The results in Proposition 3 and Proposition 4 ensure that motivating specialists to provide treatment is not a problem in this economy. The only question is whether they will provide high effort treatment or low effort treatment. The conditions under which high effort treatment will be provided are given by the threshold price inequality in Proposition 1. If this high effort incentive compatibility condition does not hold, then low effort treatment will be provided.

4.2 The treatment choices of patients

Patient's preferences depend on both their health state and their expenditure on health care. A sick patient will only seek treatment if the expected benefits in terms of a higher probability of good health exceed the cost of the treatment. A patient who is not sick will not seek treatment, since doing so will involve a cost but yield no benefit. As such, we will focus on the treatment choices of a sick patients. Since there are no GPs present in this hypothetical economy, the patient cannot seek a referral. As such, if the patient chooses to seek treatment, the only expenditure incurred will be the treatment price, so that $p = r$. Prior to seeking treatment in any given period, a sick patient does not know the amount of effort that will be exerted by the treating specialist. As such, his expected utility from treatment is:

$$U(h,p) = \mu B - r + \delta M,$$

where M is the patients expected continuation payoff if he seeks treatment in the current period and $\mu = \lambda + \theta(1-\lambda)$ is the patient's belief that he will be cured following treatment in the current period.

In principle, we could allow the continuation payoff to vary with the decision to seek treatment for disease k and the health state following any such treatment in every period up until and including the current period. The reason for this is that these will be observed by both the representative patient and the representative specialist. However, given the competitive nature of this model, we will assume that specialists do not condition their future strategy choices on the history of treatment choices or the health outcomes of their patients in the current period. As such, from a patient's point of view, the continuation payoff does not vary with the public history of either the treatment choices for disease k or the public history of health outcomes following any such treatment. Hence we can set all of the patient's continuation payoffs in the current period equal to M. Given this, the patient's expected utility if he does not seek treatment is:

$$U^0 = \delta M.$$

Proposition 5 *If the treatment price is not too high, a sick patient will seek treatment.*

Proof. A sick patient will seek treatment if and only if the following individual rationality constraint is satisfied:

$$\mu B - r + \delta M \geq \delta M.$$

This constraint simplifies to the following restriction on the treatment price:

$$r \leq \mu B.$$

Thus, so long as the treatment price does not exceed the expected benefit to the patient from treatment, he will seek treatment. ■

A patient who is using the extreme dumping strategy outlined previously will know whether or not the treatment price is at least as large as the high effort threshold price. As such, the patient will know whether or not he will receive high effort treatment. This allows us to be more specific about the patient's decision to seek treatment.

Proposition 6 *If the treatment price matches or exceeds the threshold treatment price, a sick patient will seek treatment whenever $r \leq B$. If the treatment price is less than the threshold treatment price, a sick patient will seek treatment whenever $r \leq \theta B$.*

Proof. If the treatment price matches or exceeds the threshold price, then the patient knows that he will receive high effort treatment. As such, $\lambda = 1$ and hence $\mu = 1$. Thus the maximum treatment price that the patient will be willing to pay in this case is $r = B$. If the treatment price is less than the threshold price, then the patient knows that he will receive low effort treatment. As such,

$\lambda = 0$ and hence $\mu = \theta$. Thus the maximum treatment price that the patient will be willing to pay in this case is $r = \theta B$. ■

Clearly, the extreme dumping strategy is designed to induce the specialist to provide high effort treatment. Assuming that a patient uses the extreme dumping strategy, we have characterised the range of prices for he will seek and receive high effort treatment. We have also characterised the range of prices for which he will seek and receive low effort treatment. However, we have not yet established that a sick patient would prefer high effort treatment to low effort treatment.

Proposition 7 *A sick patient will always prefer to receive high effort treatment rather than low effort treatment for any given treatment price.*

Proof. The payoff to a sick patient who receives high effort treatment is $B-r+\delta M$. The payoff to a sick patient who receives low effort treatment is $\theta B-r+\delta M$. The patient will prefer high effort treatment over low effort treatment if and only if:

$$B - r + \delta M \geq \theta B - r + \delta M.$$

This expression simplifies to:

$$(1-\theta)\, B \geq 0.$$

Since $\theta \in (0,1)$ and $B > 0$, this inequality is always satisfied. As such, a sick patient will prefer high effort treatment to low effort treatment for any given treatment price. ■

We now know that a patient will prefer high effort treatment to low effort treatment and that he can motivate a specialist to provide high effort treatment if the treatment price is sufficiently high.

4.3 Long-run and short-run relationships

If a patient is able to motivate high effort treatment from the specialist, then he is said to have a long-run relationship with that specialist. If a patient is not able to motivate high effort treatment from a specialist, then he is said to have a short-run relationship with that specialist. We have already found conditions on the prevailing treatment price that will allow us to characterise the relationship between a patient and a specialist as either long-run or short-run. However, it is perhaps more intuitive to define a short-run relationship between a patient and a specialist in terms of the probability that a patient will need the services of a specialist in any given period. After all, if that probability is sufficiently low, a patient that is being treated in the current period will be unlikely to require treatment for the foreseeable future. Given that the specialist is not perfectly patient ($\delta < 1$), he is likely to ignore any impact on this patient's future demand for his services when choosing his current effort level. Recall that threshold treatment price for ensuring high effort treatment was a function of the probability that the patient will get sick in any given period. As such, we can rearrange the high effort incentive compatibility condition to provide a restriction on the probability that a patient gets sick in any given period.

Proposition 8 *A patient has a short-run relationship with a specialist if and only if at least one of the following three conditions hold: (a) $(r-C) \leq 0$, (b) $\pi < \widehat{\pi}$ and (c) $\widehat{\pi} > 1$.*

Proof. Recall that a patient can motivate high effort treatment from a specialist if and only if:

$$r \geq \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C.$$

Since this requires that $(r - C) > 0$, we know that the patient will receive low effort treatment if $(r - C) \leq 0$. We can rearrange the threshold price inequality to obtain:

$$\pi \geq \widehat{\pi} = \frac{(1-\delta)C}{\delta(1-\theta)(r-C)}.$$

Thus a patient will have a short-run relationship with a specialist if $\pi < \widehat{\pi}$. Finally, note that $\pi \in [0, 1]$ since it is a probability. As such, if $\widehat{\pi} > 1$, then the patient will have a short-run relationship with the specialist. ■

If none of the conditions in Proposition 8 hold, then the patient will have a long-run relationship with the specialist.

Proposition 9 *The patient will have a long-run relationship with the specialist if and only if all of the following conditions hold: (a) $(r - C) > 0$, (b) $\pi \geq \widehat{\pi}$ and (c) $\widehat{\pi} \leq 1$.*

5 The GP-specialist supergame

Suppose that patients only have short-run relationships with specialists. They might be willing to seek a referral from a GP if they thought that this would result in high effort treatment and obtaining the referral was not too costly. In this section, we examine the circumstances under which a GP will be able to motivate a specialist to provide high effort treatment to all of the patients that are referred to the specialist by him. In order to incorporate the idea that each GP has a large patient pool that is stable in size, we will ultimately assume that each GP has an infinite patient pool. This assumption is required to ensure that it will be rational for specialists to hold static expectations with respect to the size of GP patient pools. A specialist has static expectations about a GP's patient pool if, in any given period, he believes that the number of patients utilising the GP's services will remain at its level in that period forever. We will begin the analysis by assuming that the representative GP has a finite patient pool of size n. The infinite patient pool assumption will be implemented by taking limits as $n \to \infty$.

5.1 Optimal deviation by a specialist

Consider a representative GP, j, who currently has a patient pool of size n. Suppose that $n_{j,1}$ of these patients are sick in the current period and that the GP j chooses specialist k to treat all of these patients. Can the GP motivate the specialist to exert high effort whenever the specialist is treating patients referred by him? Before considering this question, we will need to make some simplifying assumptions about the nature of competition. First, we continue to assume that all agents are price takers and that prices are set exogenously. In addition to this, we will assume that specialists have static expectations with respect to the size of GP patient pools. In any given period, they believe that the number of patients utilising a GP's services will remain at its current level forever. Justifications for this assumption are provided later in this paper, when the price formation process is considered.

Suppose that a specialist decides to deviate and shirk in his treatment of at least one of GP j's patients. What is the specialist's optimal deviation? This amounts to determining how many of the $n_{j,1}$ patients should receive low effort treatment.

Proposition 10 *If a specialist chooses to shirk when treating any patient referred to him by a particular GP in a particular period, then he will shirk when treating every patient referred to him by that GP in that period.*

Proof. If $n_{j,1} = 1$, this question is easy to answer. The only possible deviation from high effort treatment for all of GP j's patients is to shirk for that lone patient. When $n_{j,1} > 1$, the specialist could choose to provide low effort treatment for all of these patients or just for some subset of them. From the specialist's point of view, all of the patients referred by a particular GP in any given period are identical. Thus we need only consider the number of these patients that receive low effort treatment and not their individual identities. Let $m_{j,1}$ denote the number of patients referred to the specialist by GP j in the current period that receive low effort treatment. If the GP employs an all or nothing punishment strategy, then only two continuation payoffs need to be specified. These are the payoff to only good health outcomes, $V(1)$, and the payoff if there are any bad outcomes, $V(0)$.

We will assume that the GP follows up on the treatment outcomes for all of the patients he refers at the end of each period. Thus the GP can condition the specialist's continuation payoffs on whether or not a bad outcome occurs for any of the patients he referred to the specialist in the current period or in any past period. Given this, the payoff to the specialist from providing low effort treatment to $m_{j,1}$ of the $n_{j,1}$ patients referred by GP j is

$$\widehat{U}_s(m_{j,1}; n_{j,1}) = n_{j,1}r - (n_{j,1} - m_{j,1})\, C + \delta \left[\theta^{m(j,1)} V(1) + \left(1 - \theta^{m(j,1)}\right) V(0)\right].$$

If the GP's punishment for a bad outcome is to sack the specialist, so that $V(0) = 0$, this becomes

$$\widehat{U}_s(m_{j,1}; n_{j,1}) = n_{j,1}r - (n_{j,1} - m_{j,1})\, C + \delta\theta^{m(j,1)} V(1).$$

Differentiating this with respect to the number of patients for which the specialist shirks ($m_{j,1}$), we obtain

$$\frac{\partial \widehat{U}_s(m_{j,1}; n_{j,1})}{\partial m_{j,1}} = C + \delta\theta^{m(j,1)} \ln(\theta) V(1).$$

Since $\theta \in (0, 1)$, so that $\ln(\theta) < 0$, the sign of this derivative is ambiguous. We could assume that the derivative is always positive, but that would place strong restrictions on the size of the disutility of effort. Notice that, since $\ln(\theta) > -\infty$ because $\theta > 0$, we have

$$\lim_{m(j,1) \to \infty} \delta\theta^{m(j,1)} \ln(\theta) V(1) = \delta(0) \ln(\theta) V(1) = 0.$$

Hence, for sufficiently large $m_{j,1}$, the derivative will be positive. Indeed, the derivative is monotonically increasing in $m_{j,1}$, since

$$\frac{\partial^2 \widehat{U}_s(m_{j,1};n_{j,1})}{\partial m_{j,1}^2} = \delta\theta^{m(j,1)} \left[\ln(\theta)\right]^2 V(1) > 0 \text{ for all } m_{j,1}.$$

If the first derivative of $\widehat{U}_s(m_{j,1};n_{j,1})$ with respect to $m_{j,1}$ is positive for all $m_{j,1}$, then the unique optimal deviation is for the specialist to shirk for all of GP j's patients. If the derivative is negative for sufficiently low $m_{j,1}$, then there are two possibilities for the optimal deviation. One possibility is that the specialist will not want to shirk at all, so that $m_{j,1} = 0$, in which case there is no deviation. However, if the specialist is going to shirk for any of GP j's patients, he will choose to shirk for all of them. The optimal deviation is thus $m_{j,1} = n_{j,1}$. ■

The intuition behind this result is clear. The marginal benefit of shirking is simply the avoided cost of effort for the additional patient (C), which is constant. It does not change as the number of patients referred by GP j that receive low effort treatment increases. However, the additional probability of being detected (and hence the expected marginal cost of shirking) falls as the number of patients referred by GP j that receive low effort treatment increases. As such, if the specialist chooses to shirk when treating any of GP j's patients in any given period, he will shirk when treating all of GP j's patients in that period.

We are now in a position to derive the maximum payoff that a specialist will receive if he shirks for any of GP j's patients when GP j is employing the extreme punishment strategy outlined earlier.

Proposition 11 *If a GP employs the extreme punishment strategy, the maximum payoff to a specialist who fails to provide high effort treatment for all of the patients referred by that GP is*

$$\widehat{U}_s(deviate) = n_{j,1}r + \delta\theta^{n(j,1)}V(1).$$

Proof. We have already shown that the optimal deviation for a specialist involves setting $m_{j,1} = n_{j,1}$. Substituting this into the expression for the specialists payoff yields

$$\widehat{U}_s(deviate) = \widehat{U}_s(n_{j,1};n_{j,1}) = n_{j,1}r + \delta\theta^{n(j,1)}V(1).$$

■

A GP will be able to motivate high effort treatment from a specialist for all of his patients if and only if the payoff to specialist from providing only high effort treatment exceeds both the payoff to his optimal deviation and the payoff to refusing

to treat the patients. The conditions under which a specialist would prefer to provide high effort treatment to all of the patients referred from GP j are derived in the next subsection of this paper. The conditions under which the specialist will choose to treat all of GP j's patients are considered in the following subsection of this paper.

5.2 Incentive compatibility constraints for specialists

A specialist will provide high effort treatment to all of the patients that are referred to him by a particular GP only if the payoff to doing so exceeds the payoff that he would receive from shirking for at least some of these patients. This requires that the treatment price be sufficiently high.

Proposition 12 *A specialist will weakly prefer to provide high effort treatment to all of the patients that are referred to him by a particular GP in a given period over the provision of low effort treatment to one or more of these patients if the treatment price exceed some threshold price. The threshold price is given by*

$$\widehat{r}(n_{j,1}, n_j) = \left[1 + \frac{n_{j,1}(1-\delta)}{\pi_j n_j \delta \left(1 - \theta^{n(j,1)}\right)}\right] C.$$

Proof. The payoff to the specialist from providing high effort treatment for all of GP j's patients is

$$\widehat{U}_s(no\ deviation) = n_{j,1}(r - C) + \delta V(1).$$

Thus the specialist will choose to provide high effort treatment to every patient referred by GP j only if the payoff from doing so matches or exceeds the largest possible payoff from not doing so. This requires that

$$n_{j,1}(r - C) + \delta V(1) \geq n_{j,1} r + \delta \theta^{n(j,1)} V(1),$$

which can be rearranged to obtain

$$V(1) \geq \frac{n_{j,1} C}{\delta \left(1 - \theta^{n(j,1)}\right)}.$$

■

In determining the equilibrium continuation payoff, $V(1)$, we need to remember that the specialists have static beliefs about the size of each GP's patient pool, n_j.[15] Given this, the highest continuation payoff for a history of only good outcomes that a GP can credibly offer is a constant stream of the expected static payoff to high effort. This is

$$V(1) = \sum_{t=0}^{\infty} \delta^t \pi n_j (r - C) = \frac{\pi n_j (r - C)}{(1 - \delta)}.$$

Note that $(r - C)$ is the net payoff per patient when the specialist exerts high effort, while πn_j is the expected number of GP j's patients that will be sick in any given period. Substituting this into the high effort ICC we obtain

$$\frac{\pi n_j (r - C)}{(1 - \delta)} \geq \frac{n_{j,1} C}{\delta \left(1 - \theta^{n(j,1)}\right)},$$

[15] Strictly speaking, this only makes sense if GPs have infinite patient pools. All that specialists observe is the number of patients that are referred to them by a particular GP in any given period ($n_{j,1}$). As such, they need to infer the size of the GP's patient pool (n_j) on the basis of this information. Since $n_{j,1} \sim bin(\pi, n_j)$, the specialist will view n_j as a non-degenerate random variable if the GP has a finite patient pool. However, if the GP has an infinite patient pool, then $n_{j,1}$ is almost surely infinite. This greatly simplifies the statistical inference problem facing the specialist. If a specialist receives an infinite number of referrals from a particular GP, then he knows that the GP has an infinite patient pool.

which can be rearranged to yield

$$r \geq \left[1 + \frac{n_{j,1}\left(1-\delta\right)}{\pi n_j \delta \left(1 - \theta^{n(j,1)}\right)} \right] C.$$

If the prevailing treatment price does not fall below this threshold, then the GP will be able to assure his patients that they will provided with high effort treatment by this specialist in the current period.

We will denote the threshold price, below which a GP cannot ensure high effort treatment for all of his patients, by $\widehat{r}(n_j, n_{j,1})$. Note that when GPs have finite patient pools, this threshold price is a random variable. The reason for this is that the threshold price is a function of the number members of a GP's patient pool who are sick in a particular period. This means that a patient cannot be sure that any particular GP will be able to motivate high effort treatment on the part of a specialist, even if he knows both the prevailing treatment price and the size of each GPs patient pool. This situation can be avoided if GPs have infinite patient pools.

Proposition 13 *If each GP has an infinite patient pool, then the threshold treatment price is almost surely $\frac{C}{\delta}$.*

Proof. Consider a GP who has a patient pool of size n_j. Suppose that $n_{j,1}$ of these patients are sick in a particular period. The threshold treatment price for such a GP will be

$$\widehat{r}(n_j, n_{j,1}) = \left[1 + \frac{n_{j,1}\left(1-\delta\right)}{\pi n_j \delta \left(1 - \theta^{n(j,1)}\right)} \right] C = \left[1 + \frac{\alpha_j \left(1-\delta\right)}{\pi \delta \left(1 - \beta_j\right)} \right] C,$$

where $\alpha_j = \frac{n_{j,1}}{n_j}$ is the proportion of the GPs patients who are sick in that period and $\beta_j = \theta^{n(j,1)}$ is the probability that a specialist who shirks when treating all of these patients does not get caught. Determining what happens to the threshold price as the size of a GP's patient pool approaches infinity requires us to determine what happens to $n_{j,1}$ as $n_j \to \infty$. This is not straightforward, as the relationship between $n_{j,1}$ and n_j is stochastic. Indeed, $n_{j,1}$ can be viewed as the number of negative outcomes in a random sample of n_j Bernoulli trials, where the probability of a negative outcome on any given trial is π. As such, $n_{j,1}$ is a binomially distributed random variable, with parameters n_j and π. The specific number of patients that are referred to a specialist by a GP in any given period is simply a particular realisation of this underlying random variable. Unfortunately, it is this actual realisation that enters a specialist's high effort incentive compatibility constraint and hence the threshold price. In finite samples, any particular realisation of $n_{j,1}$ could occur with positive probability. However, in an infinite sample, we can use limiting arguments to show that the relative proportion of negative outcomes (α_j) is almost surely equal to the probability of a negative outcome in a single trial (π). This in turn allows us to show that each GP almost surely has an infinite number of sick patients in each period. Furthermore, the probability that any specialist who shirks when treating all of these patients is not caught is almost surely equal to zero. The combination of these limiting results allows us to show that the threshold price for any GP who has an infinite patient pool is almost surely a constant.

First, we need to show that α_j almost surely converges to π as n_j approaches infinity. Recall that

$$\alpha_j = \frac{n_{j,1}}{n_j} = \frac{1}{n_j} \sum_{i(j)=1}^{n(j)} 1_{i(j),1},$$

where $1_{i(j),1}$ is an indicator variable that takes on the value one if a particular member of GP j's patient pool, patient $i(j)$, has the disease in the current period and zero otherwise. The summation is over GP j's entire patient pool for the current period. Not that each of these indicator variables is a Bernoulli random variable that takes on the value one with probability π and zero otherwise. As such, $\left\{1_{i(j),1}\right\}_{i(j)=1}^{n(j)}$ is a sequence of independent and identically distributed Bernoulli random variables. Furthermore, note that

$$E\left(1_{i(j),1}\right) = \pi(1) + (1-\pi)(0) = \pi \text{ for all } i_j \in \{1, 2, \cdots, n_j\}.$$

Thus, from the strong law of large numbers[16], we know that

$$\Pr\left(\lim_{n(j)\to\infty} \alpha_j = \pi\right) = 1.$$

Hence we can conclude that the relative proportion of sick patients in any given period for a particular GP (α_j) is almost surely equal to the probability that any individual patient is sick in any given period (π) if the GP has an infinite patient pool. Now we want to show that the probability that any specialist who shirks when treating all of a GPs patients in any given period is not caught is almost surely equal to zero when the GP has an infinite patient pool. Since $\beta_j = \theta^{n(j,1)}$ and $\theta \in (0,1)$, this will be clearly be the case if $n_{j,1}$ approaches infinity as n_j approaches infinity. Note that

$$n_{j,1} = \left(\frac{n_{j,1}}{n_j}\right) n_j = \alpha_j n_j.$$

Furthermore,

$$\lim_{n(j)\to\infty} n_j = \infty.$$

Thus we can conclude that

$$n_{j,1} \xrightarrow{a.s.} (\pi)(\infty) = \infty.$$

[16] See Billingsley ([8], pp. 85-86) for a discussion of the strong law of large numbers. Note that when GP patient pools are finite, the number of members in a GP's patient pool is an integer. As such, when we take the limit as this number approaches infinity, we are restricting our attention to the set of natural numbers. In effect, there is a one-to-one correspondence between each member of a GP's patient pool and each element of the set of natural numbers when that GP has an infinite patient pool. As such, each GP has a countable number of patients. This ensures that the standard version of the strong law of large numbers applies in the model considered in this paper. If we had assumed that each GP had a continuum of patients instead of a countably infinite number of patients, we would have needed to use the techniques mentioned in Judd ([35]). The reason for this is that each GP would have had an uncountable number of patients in that case.

Hence we know that the probability that any specialist who shirks when treating all of a GPs patients in any given period is not caught is almost surely equal to zero when the GP has an infinite patient pool.

We are now in a position to look at what happens to the threshold price, $\widehat{r}(n_j, n_{j,1})$, as the size of GP's patient pools get very large. Note that $\widehat{r}(n_j, n_{j,1})$ is a continuous function of α_j and β_j as long as $\beta_j \neq 1$. Furthermore, since $\theta \in (0, 1)$ ensures that $\beta_j \in [0, 1)$, we do not need to worry about the potential discontinuity at $\beta_j = 1$. This means that

$$\widehat{r}(n_j, n_{j,1}) \xrightarrow{a.s.} \left[1 + \frac{\pi(1-\delta)}{\pi\delta(1-0)}\right] C = \left[\frac{\delta + 1 - \delta}{\delta}\right] C = \frac{C}{\delta}.$$

Thus we have established that the threshold treatment price for any GP with an infinite patient pool is almost surely $\frac{C}{\delta}$. ■

5.3 Participation constraints for specialists

While we have established the conditions under which a specialist will prefer providing high effort treatment to low effort treatment, we still need to establish that the specialist would prefer providing high effort treatment for all of the patients referred by a particular GP to not providing some or all of them with any treatment. If the specialist refuses to treat any of a GP's referrals, he will receive no surplus from that transaction. It is possible that the GP could punish such behaviour by refusing to refer any future patients to that specialist. However, as in the case without GPs, no such punishment is necessary to induce treatment in those cases where the GP can motivate high effort treatment from the specialist.

Proposition 14 *If the high effort incentive compatibility constraint is satisfied for a specialist with respect to a particular GP, then the specialist will prefer to provide high effort treatment to any patient referred by that GP in a given period, rather than not treat the patient at all.*

Proof. Since specialists are not perfectly patient ($\delta \in (0, 1)$), the threshold price must exceed the cost of high effort treatment. Thus we must have $r \geq \widehat{r}(n_j, n_{j,1}) > C$. This is sufficient to ensure that the specialist would receive positive surplus if he provides high effort treatment to the patient. Since the specialist will receive zero surplus from any patient he refuses to treat, he will prefer to provide high effort treatment rather than no treatment whatsoever. ■

Thus the specialist will prefer to provide high effort treatment to the patient than not treat the patient at all, even if no dynamic punishment for non-treatment is used by the referring GP. If a GP is able to motivate high effort from the specialist, he can automatically ensure participation.

Suppose instead that the high effort incentive compatibility constraint does not hold. In this case, the specialist will only provide low effort treatment to any patient referred by the GP in that period, if any treatment is provided at all.

Proposition 15 *If the high effort incentive compatibility constraint is not satisfied for a specialist with respect to a particular GP, then the specialist will weakly prefer to provide low effort treatment to any patient that is referred to him by that GP in a given period, rather than not treat the patient at all.*

Proof. If the specialist only provides low effort treatment for each of the patients referred by the GP, then he only incurs the disutility associated with low effort treatment, $C(0) = 0$, for each of those patients. As such, any non-negative price for treatment will be sufficient to induce the specialist to offer at least low effort treatment, even if the GP does not employ any dynamic punishments for non-treatment. ■

5.4 Participation constraints for GPs

GP's will be willing to provide referral services if and only if the discounted present value of their expected revenues exceeds that of their expected costs. Like all of the other players in this economy, GPs are price takers. As such, the only way their future revenue can be affected is by patients choosing not to utilise their referral services. Since prices and costs are exogenously fixed and constant across time in this economy, patients cannot induce specialists to provide treatment at a price below cost now in return for their future custom at above cost prices. Thus, GPs will provide their referral services if and only if the price per referral is at least as high as the cost per referral ($w \geq k$).

Proposition 16 *GPs will offer referral services if and only if the referral price exceeds the marginal cot of a referral.*

Proof. Recall that there are no fixed costs associated with providing referral services in this economy. Furthermore, the variable costs are constant. As such, the marginal cost of a referral equals the average cost per referral. Given this, the proposition follows from the above arguments. ■

6 Industry structure with exogenous prices

The structure of the health care industry will be jointly determined by the decisions of patients, general practitioners and specialists. We have characterised the conditions under which patients can motivate high effort treatment from specialists by themselves and the conditions under which GPs can motivate high effort treatment from specialists for all of their patients. We have also analysed the conditions under which GPs will be willing to offer their referral services and specialists will be willing to offer their treatment services. Finally, we have analysed the conditions under which a patient will demand treatment services alone. All that remains is for us to determine the circumstances under which a patient will prefer to seek both treatment and a referral over both treatment alone and no treatment whatsoever. We will then be in a position to describe how the structure of the health care industry will vary with both the treatment price and the referral price.

6.1 The referral choices of patients

In order to determine the circumstances under which a patient will seek a referral, we need to compare the payoff that a patient gets from obtaining a referral and treatment with both the payoff that the patient would get if he sought treatment alone and the payoff he would get without treatment. We can ignore the possibility that a patient will seek a referral alone because it would not improve his expected health status but it would use resources that could otherwise be spent on consumption. The treatment outcomes facing a patient who chooses not to seek a referral will be the same as those in the absence of GPs. Recall that these treatment outcomes varied with the prevailing treatment price as follows:

$$\textit{treatment outcome} = \begin{cases} \text{high effort treatment,} & \text{if } r \in [\widehat{r}_1, B]\,; \\ \text{low effort treatment,} & \text{if } r \in [0, \min\{\widehat{r}_1, \theta B\})\,; \\ \text{no treatment,} & \text{otherwise;} \end{cases}$$

where

$$\widehat{r}_1 = \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C.$$

The patient's continuation payoff is not affected by the current period outcome. As such, we can focus on the current period payoffs facing the patient. In the absence of a referral, these are

$$EU(h, r) = \begin{cases} B - r & \text{if } r \in [\widehat{r}_1, B]\,; \\ \theta B - r & \text{if } r \in [0, \min\{\widehat{r}_1, \theta B\})\,; \\ 0 & \text{otherwise.} \end{cases}$$

Now suppose that a patient seeks a referral. The treatment outcomes for a patient who has a referral are

$$\textit{treatment outcome} = \begin{cases} \text{high effort treatment,} & \text{if } r \in \left[\frac{C}{\delta}, B\right]\,; \\ \text{low effort treatment,} & \text{if } r \in \left[0, \min\left\{\frac{C}{\delta}, \theta B\right\}\right)\,; \\ \text{no treatment,} & \text{otherwise.} \end{cases}$$

The patient's payoffs if he seeks a referral are

$$EU(h, r + w) = \begin{cases} B - r - w & \text{if } r \in \left[\frac{C}{\delta}, B\right]\,; \\ \theta B - r - w & \text{if } r \in \left[0, \min\left\{\frac{C}{\delta}, \theta B\right\}\right)\,; \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 17 *A necessary condition for a patient to seek a referral is that* $r \in \left[\frac{C}{\delta}, \widehat{r}_1\right)$.

Proof. Clearly, if $r \in [\widehat{r}_1, B]$, then the patient will choose not to seek a referral if $w > 0$. When $w = o$, the patient will be indifferent between seeking a referral and self-referring. In these circumstances, we will assume that the patient self-refers. As such, whenever, $r \in [\widehat{r}_1, B]$, patients will not seek referrals. Furthermore, if $r \in \left[0, \frac{C}{\delta}\right)$, then patients who seek treatment will receive low effort treatment regardless of whether or not they have a referral. As such, these patients will not seek a referral either. Thus, a necessary condition for patients to seek a referral is that $r \in \left[\frac{C}{\delta}, \widehat{r}_1\right)$. ■

While this is a necessary condition for a patient to seek a referral, it is not a sufficient condition. When treatment prices satisfy $r \in \left[\frac{C}{\delta}, \widehat{r}_1\right)$, a patient will receive high effort treatment if he obtains a referral and low effort treatment if he self-refers. As such, his expected health benefits will be higher if he obtains a referral. However, his treatment costs will also be higher unless treatment is free. As such, a patient will seek a referral only if the additional expected benefits from receiving high effort treatment match or exceed the cost of a referral.

Proposition 18 *A patient will seek a referral if and only if both* $r \in \left[\frac{C}{\delta}, \widehat{r}_1\right)$ *and* $w \leq (1-\theta) B$.

Proof. we have already established that a patient will not seek a referral unless $r \in \left[\frac{C}{\delta}, \widehat{r}_1\right)$. Even if this condition is satisfied, the payoff to obtaining a referral must be at least as high as the payoff to self-referring if the patient is to seek a referral. This requires that

$$B - r - w \geq \theta B - r,$$

which can be rearranged to yield

$$w \leq (1-\theta) B.$$

■

6.2 The equilibrium industry structure

We have established the circumstances under which patients will seek a referral and treatment, seek treatment alone and seek neither treatment nor referral. We have also established the conditions under which GPs will offer their referral services and specialists will offer their treatment services. The market outcome will vary with the prevailing treatment and referral prices. The relationship between market outcomes and prices is summarised in Table 1.

Table 1: Market outcomes

Circumstances	**Market Outcomes**
$r \in \left[\frac{C}{\delta}, \widehat{r}_1\right)$, $w \leq (1-\theta) B$, $r + w \leq B$	High effort treatment, referral
$r \in [\widehat{r}_1, B]$	High effort treatment, no referral
$r \in \left[0, \min\left\{\frac{C}{\delta}, \theta B\right\}\right)$	Low effort treatment, no referral
$r \in \left[\frac{C}{\delta}, \min\left\{\theta B, \widehat{r}_1\right\}\right)$, $w < k$	Low effort treatment, no referral
Otherwise	No treatment, no referral

These market outcomes can be illustrated in (r, w)-space. A variety of possible outcomes are illustrated in Figures 1 to 3. Note that the presence of GPs allows for the existence of a region in (r, w)-space in which patients will choose to seek a referral. This will result in them getting high effort treatment where, in most cases, they would not do so otherwise. This provides the foundation for a demand driven explanation for the existence of GPs. When treatment and referral prices fall in this region, patients will prefer to have the option of seeking a referral from a GP. The reason for this is that GPs are able to motivate high effort treatment from specialists when prices fall in this region, while patients cannot do so. Furthermore, the additional health benefits that patients expect to receive from high effort treatment exceed the additional cost of seeking a referral when prices fall in this region.

In general, specialists do not like the presence of GPs. The reason for this is that they need to provide high effort treatment, which involves a higher disutility of effort for them, but they do not receive any additional remuneration. However, there are some circumstances in which both patients and specialists prefer to have GPs present. These situations involve treatment prices that satisfy $r \in \left(\frac{C}{\delta}, \theta B\right]$, where this interval is non-empty. In the absence of GPs, patients would not seek treatment and specialists would earn no profits. If GPs are present, then patients will seek both a referral and treatment. Specialist will provide high effort treatment and earn positive profits. If $k \leq w \leq \min\{(1-\theta)B, B-r\}$, then patients, GPs and specialists will all weakly prefer the presence of GPs to their absence in such circumstances. As such, there are some cases in which a gated industry structure weakly Pareto dominates an ungated industry structure when prices are exogenous. Circumstances such as these occur for some treatment prices in Figure 2.

6.3 The impact of chronic illnesses

Despite the fact that the model employed in this paper explicitly incorporates only a single disease, it is sufficiently flexible to allow a comparison between the industry structures that might prevail for diseases.with different characteristics Suppose that there are a number of different types of disease and that a different group of specialists treats each disease type. All of the preceeding analysis carries over to each of these disease types. As such, we can compare the industry outcomes for each disease type by examining the impact of a change in the probability that a patient gets sick on the structure of the health care industry. For example, a patient with a chronic disease is much more likely to require treatment in any given period that a patient who does not have a chronic disease. As such, we can examine the impact of a chronic disease by considering what happens to health industry outcomnes when the probability opf illness (π) increases.

Proposition 19 *The threshold price above which an individual patient can motivate high effort treatment from a specialist is a decreasing function of the probability of the patient contracting a particular illness in any given period.*

Proof. Recall that the threshold price above which an individual patient can motivate high effort treatment from a specialist is given by

$$\widehat{r}_1 = \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C.$$

Figure 1: Market outcomes when $\theta B < \frac{C}{\delta} < \widehat{r}_1 < B$.

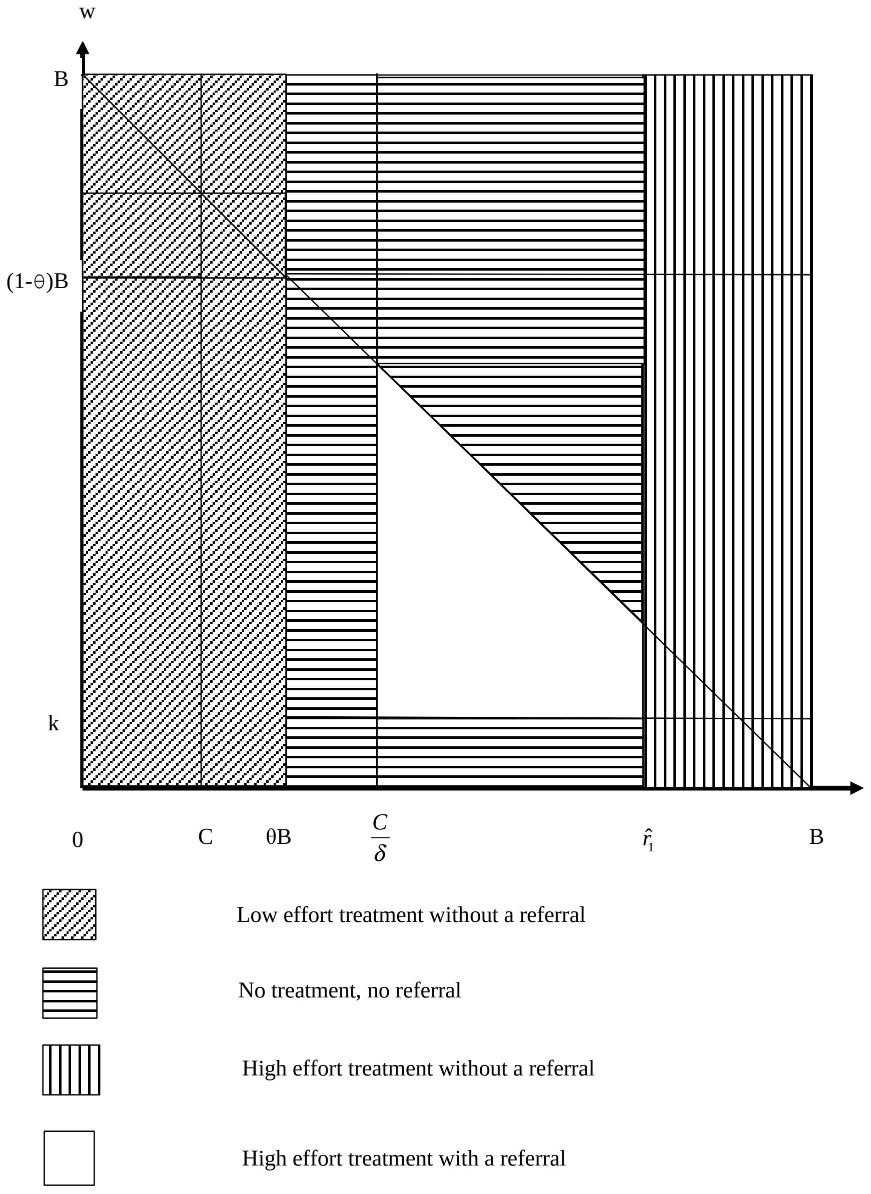

w
B
(1-θ)B
k
r
0
C
θB
$\frac{C}{\delta}$
$\hat{r}_1$
B
Low effort treatment without a referral
No treatment, no referral
High effort treatment without a referral
High effort treatment with a referral

Figure 2: Market outcomes when $\frac{C}{\delta} < \theta B < \widehat{r}_1 < B$.

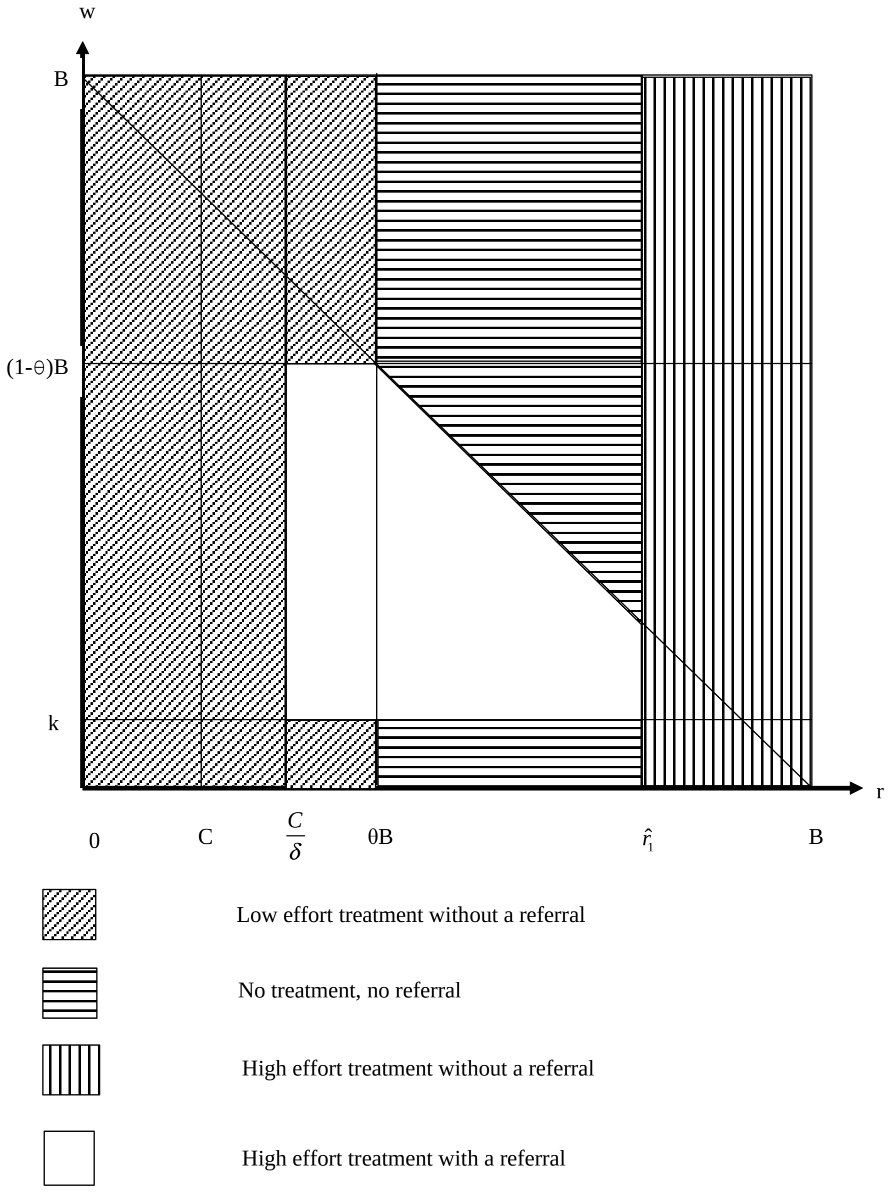
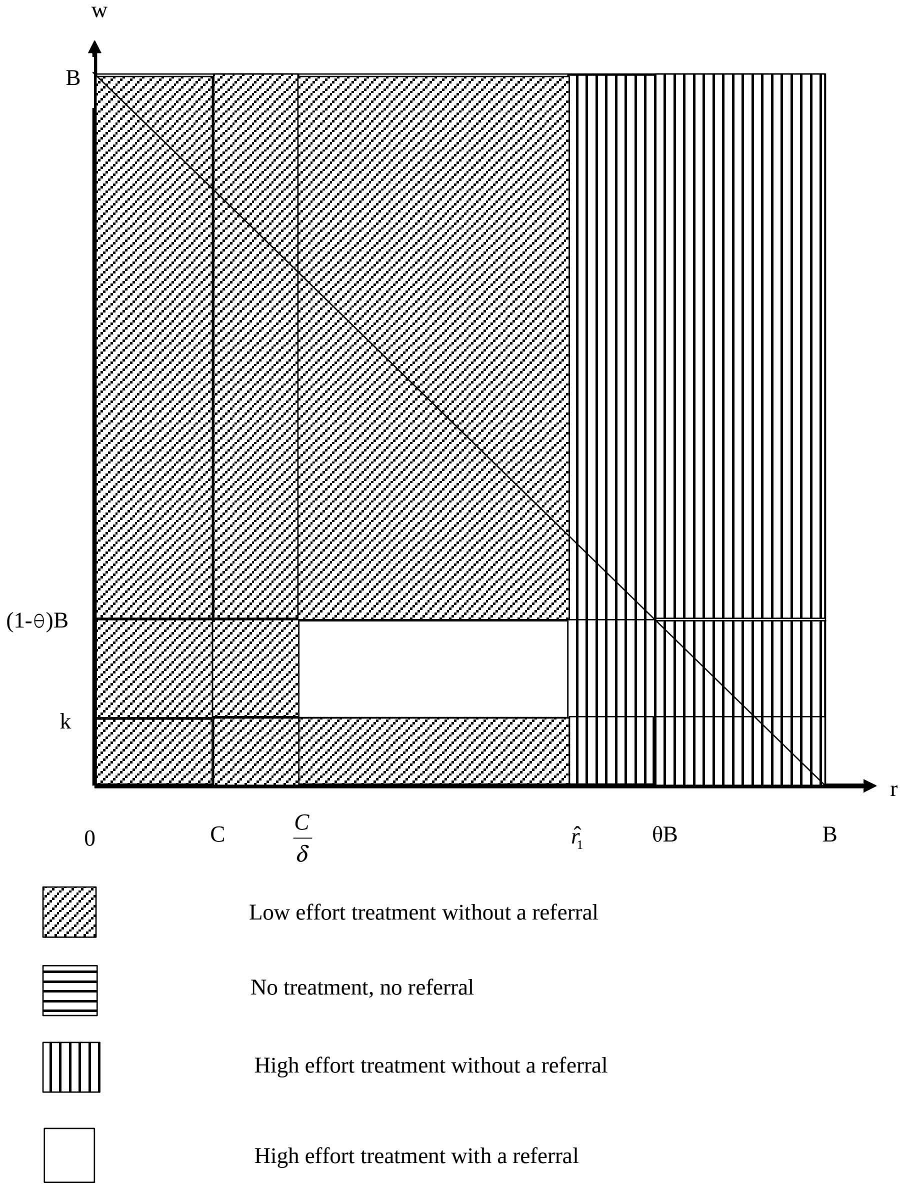

w
B
(1-θ)B
k
r
0
C
C/δ
θB
r̂1
B
Low effort treatment without a referral
No treatment, no referral
High effort treatment without a referral
High effort treatment with a referral

Figure 3: Market outcomes when $\frac{C}{\delta} < \widehat{r}_1 < \theta B$.


w
B
(1-θ)B
k
r
0
C
$\frac{C}{\delta}$
$\hat{r}_1$
θB
B
Low effort treatment without a referral
No treatment, no referral
High effort treatment without a referral
High effort treatment with a referral

Partially differentiating this expression with respect to π yields

$$\frac{\partial \widehat{r}_1}{\partial \pi} = \frac{-(1-\delta)C}{\pi^2 \delta(1-\theta)} < 0,$$

since $\delta \in (0,1)$, $\theta \in (0,1)$, $\pi \in (0,1]$ and $C > 0$. ■

Proposition 20 *The threshold price above which a GP with an infinite patient pool can motivate high effort treatment from a specialist does not vary the probability of a patient contracting a particular illness in any given period.*

Proof. Recall that the threshold price above which A GP with an infinite patient pool can motivate high effort treatment from a specialist is almost surely given by

$$\widehat{r}_\infty = \frac{C}{\delta}.$$

Partially differentiating this with respect to π yields

$$\frac{\partial \widehat{r}_\infty}{\partial \pi} = 0.$$

As such, the threshold price almost surely facing a GP with an infinite patient pool will be the same for all diseases that occur with positive probability. Thus the threshold price will be the same for both chronic and rare diseases. ■

We can use these two reults to examine the impact of disease incidence on the structure of the health care industry. A chronic illness will have a high probability of occurence in each period. This increases the likelihood that a patient will be able to motivate high effort treatment without the need for a referral. As such, it reduces the likelihood that a patient will seek a referral. This can been seen in each of Figures 1 to 3. As the probability of a disease occuring in any given period increases, $\widehat{r}_1$ falls while $\widehat{r}_\infty$ stays the same. Hence if nothing else changes, the unshaded area that represents the referral price and treatment price combinations for which a patient will seek a referral will shrink as π increases.[17]

7 Conclusion

We have used reputation effects to explain the organisation of many professional service industries, including the medical and legal professions. The main focus has been on explaining the existence of gatekeeping intermediaries who refer consumers to one of many ultimate producers. Examples of such intermediaries include general practitioners in the health care industry and solicitors in the legal industry. The explanation for the existence of such intermediaries that is provided in this paper focuses on their role as a reputation monitor. The GPs keep track of the treatment outcomes for each patient they refer to a particular specialist. GPs have large patient pools because they provide referral services for many different types of disease. As such, they will observe many more treatment outcomes with a particular specialist than any individual patient will observe. Furthermore, the fact that a GP has a large patient pool also means

[17] Recall, however, that π is a probability. As such, π canot exceed one.

that if he discovers evidence of shirking on the part of the specialist, he can punish the specialist much more effectively than could any individual patient. The potential loss of future business from a GP with a large patient pool is much more significant than the potential loss of future business from a single patient.

There is a related literature that uses reputation to explain the existence of institutions, primarily firms[18], unions[19] and retailers[20]. This literature builds on earlier work examining the extent to which reputation effects and market forces provide an incentive for parties to exert effort when such effort is costly and unobservable.[21] In this section, we compare the model presented in this paper with other reputation based theories of the existence of institutions. In particular, we compare it with Kreps' theory of firms ([38]), Hogan's theory of unions ([30]) and Biglaiser and Friedman's theory of retailers ([7]).

Kreps ([38]) provides a reputation-based explanation for the existence of firms. His starting point is a static moral hazard problem similar to the one considered in this paper. The outcome of a transaction between a consumer and a producer depends on some action taken by the producer which is unobservable to the consumer at the time of the transaction. For example, a consumer's satisfaction with a product may depend on its quality, which might not be observed until the product is actually used, well after the time of purchase. Furthermore, this quality level may be unverifiable to third parties. If low quality products are cheaper to produce, then the producer may have an incentive to pretend that a low quality product is really a high quality product. However, if the producer has a long-run relationship with the consumer, he runs the risk of losing that consumers future custom if he misleads the consumer about product quality. As such, repetition may overcome the static moral hazard problem. Unfortunately, consumers will not always have a long-run relationship with the producer. Kreps shows that if the outcomes of previous transactions can be communicated to future customers, then the fact that any individual customer only has a short-run relationship with the producer is irrelevant. What matters is that the producer can be punished in the future for any current transgressions.

The analysis presented in this paper strengthens the foundations of Kreps model in two ways. Kreps' model assumes that the outcomes of current transactions can be accurately and costlessly communicated to future consumers. It also assumes that the terms of trade between consumers and producers are fixed because of the existence of competition for trading partners. However, Kreps does not explicitly examine either the communication process or the price formation process. In this paper, we have provided a natural means of communicating past outcomes in the form of gatekeeping intermediaries that monitor the outcomes of transactions that result from their referrals. Furthermore, we have explicitly modelled the process of price formation. This has allowed us to provide foundations for the fixed terms of trade assumption and to determine the equilibrium terms of trade.

Hogan ([30]) provides a reputation-based explanation for the existence of

[18] See Kreps ([38]) and Tadelis ([68]) in particular.

[19] See Hogan ([30]), MacLeod and Malcolmson ([41]) and Malcolmson ([43]) in particular.

[20] See Biglaiser ([6]) and Biglaiser and Friedman ([7]) in particular.

[21] The important earlier papers include Cooper and Ross ([13]), Diamond ([15]), Holmstrom ([32]), Klein and Leffler ([36]) and Shapiro ([61]). More recent work along these lines can be found in Horner ([33]).

unions. He considers a moral hazard problem between a worker and a firm in which output depends on employee effort which is costly for the employee to provide. The effort provided by any given employee is assumed to be observable to both the employee and the employer, but is both unobservable and unverifiable to third parties. As a result, the employer must use an implicit contract to motivate high effort on the part of employees. If the employer and an employee have only a short-run relationship, the employer will have an incentive to renege on any promised high effort payments for employees, even if they provide high effort. This problem can be at least partially overcome if the employer and the employees have a long-run relationship. However, if the firm's production technology exhibits diminishing marginal returns to labour, implicit contracts will be insufficient to achieve first-best employment levels. The presence of a union in this setting increases employment and thereby allows efficiency losses to be reduced. The reason for this is that the union is able to monitor the behaviour of the employer and inform its members if the employer has reneged on a contract with any them. Note that the individual employees cannot undertake this monitoring role themselves because they do not observe the effort choices of other employees. The union is assumed to possess a technology that enables it to observe the effort choices of its members. This technology is too expensive for individual employees to utilise in the absence of the union. The monitoring costs incurred by the union are recovered through union membership fees.

Unlike the model considered in this paper, the employer and the employees have a long-run relationship in Hogan's model. As such, reputation can play a role in reducing the occurrence of opportunistic behaviour by the firm, even if the union is not present. The presence of the union simply enhances the effectiveness of these reputation effects. A gatekeeping intermediary in Hogan's model would look more like a temporary recruitment agency. The very nature of temporary employment would ensure that temporary employees have only a short-run relationship with the employer. As such, in the absence of a temporary recruitment agency, the employer would have a strong incentive to renege on any payments that were promised in return for the provision of high effort by the employee. This would in turn provide an incentive for the employee to only provide low effort. However, if the employer obtains his temporary employees through a temporary recruitment agency, then that agency will have a long-run relationship with the firm. Furthermore, if that agency refers its workers to many different firms, then it will have a long-run relationship with the temporary employees that use its referral services. As such, the temporary recruitment agency may be able to leverage its long-run relationships with employers and temporary employees to create an artificial long-run relationship between the employers and the temporary employees. As such, the use of temporary recruitment agencies may allow for equilibria in which the employees provide high effort and the employers do not renege on their promise to pay extra for the provision of high effort.

Biglaiser and Friedman ([7]) provide a reputation-based theory of the existence of retailers. They consider a moral hazard problem in which consumers do not observe the quality of a good until after they have purchased it. As such, producers will have an incentive to mislead consumers about the quality of the goods they produce. This incentive is reduced if the producer sells his products through retailers that also stock the products of other producers rather than directly to the public. The reason for this is that the retailers will lose future sales on their other products if they do not punish a producer for misleading consumers about the quality of his products.

In many respects, Biglaiser and Friedman's model is the closest in spirit to the one employed in this paper. In both models, an intermediary has a long-run relationship with producers because he sells their products to many different consumers. Similarly, in both models an intermediary has a long-run relationship with consumers because he sells many different products that they might wish to purchase. However, there are also a number of differences. One relatively minor difference relates to the party that chooses to use the services of an intermediary. In Biglaiser and Friedman's model, the producers choose to use an intermediary to market their goods to consumers. In the model employed in this paper, consumers choose to use an intermediary to access the services of a producer.

There are also more significant differences between the two models. Biglaiser and Friedman allow quality to vary continuously, while specialist effort can only take on one of two values in the model considered in this paper. In Biglaiser and Friedman's model, consumers learn the quality of the products they purchase following the transaction. If they purchased the product from an intermediary, the intermediary also learns the quality of the product after the transaction has been completed. As such, there is never any uncertainty about whether or not they have been deceived by the producer. However, in the model employed in this paper, patients and general practitioners cannot always infer the effort choices of specialists. In order to allow for the possibility that producers may mislead consumers about the quality of their products, Biglaiser and Friedman incorporate a signaling component into their model. No signaling components are incorporated into the model considered in this paper. One final difference in the structure of the two models relates to the length of the relationship between consumers and producers. In Biglaiser and Friedman's model, in the absence of retailers, demand in every period will be the same if the producer never defects. The proportional decrease in demand is identical to the proportion of customers that were deceived in the previous period. This is consistent with consumers having a long-run relationship with the producer. The main focus of this paper, on the other hand, is on situations in which patients have only a short-run relationship with specialists.

One of the most significant differences between Biglaiser and Friedman's analysis and the analysis presented in this paper relates to the type of equilibria that are considered. Even when retailers are absent, Biglaiser and Friedman focus on equilibria in which producers do not mislead consumers about the quality of their products. The introduction of retailers reduces the cost to a producer of signalling his chosen quality level and reduces the price that a consumer must pay to receive a product of that quality level. While a similar result was obtained in the model employed in this paper for the case in which both patients and GPs could motivate high effort treatment from specialists, this case was not the main focus of this paper. The main focus of this paper was on situations in which patients could not motivate high effort treatment from specialists, but GPs could do so. We showed that there existed equilibria in which this was the case that were preferred by patients to the outcomes when GPs were not present. Patients preferred this equilibrium because the additional expected benefit they received from high effort treatment exceeded the additional cost of such treatment. In some cases, specialists also preferred the presence of GPs. In these cases, the additional revenue that specialists received from providing high effort treatment exceeded the additional cost of providing high effort treatment.

8 Appendix: Industry structure with endogenous prices

Throughout this paper, we have assumed that prices are set exogenously. This assumption can be viewed as a black-box for any price formation process that generates a uniform price. It does not really matter whether health care markets are perfectly competitive, imperfectly competitive or even monopolised, as long as price discrimination is not present. However, in the previous section on industry structure when prices are exogenous, we imposed a slightly stronger assumption. This assumption involved the equilibrium treatment price being the same when GPs are present as it is when they are absent. In this section, we relax this assumption and examine market outcomes in the context of an explicit price formation process. Equilibrium prices are assumed to be the outcome of Bertrand competition. We will allow specialists to offer two different prices, one for high effort treatment and one for low effort treatment. However, since effort is not observable and treatment outcomes are not verifiable, the high effort treatment price will need to satisfy the high effort incentive compatibility constraint if it is to be credible. The difference between the high effort incentive compatibility constraint for patient-specialist interactions and the corresponding constraint for GP-specialist interactions suggests that the equilibrium treatment price may vary with the presence of GPs in some cases under this price formation process.

8.1 Price formation without GPs

In the absence of GPs, standard Bertrand competition arguments suggest that the equilibrium high effort treatment price will be

$$r_{11}^{*} = \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C.$$

This is the lowest price at which patients will believe that specialists will provide high effort treatment. Similarly, standard Bertrand arguments suggest that the equilibrium low effort treatment price will be

$$r_{10}^{*} = 0.$$

Patients will demand high effort treatment if and only if

$$B - \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C \geq \theta B,$$

which can be rearranged to yield

$$C \leq \left[\frac{\pi\delta\,(1-\theta)^{2}}{\pi\delta\,(1-\theta) + (1-\delta)}\right] B.$$

Thus, if the cost of providing high effort treatment is not too high, then the prevailing treatment price will be r^*_{11}. However, if the cost of high effort is too high, then the prevailing treatment price will be zero.

8.2 Price formation with GPs

When GPs are present, we need to determine both the equilibrium treatment price and the equilibrium referral price. In this section, we will allow specialists to offer three types of treatment service. They can offer high effort treatment to patients with a referral, high effort treatment to patients without a referral and low effort treatment. Once again, each of these outcomes needs to be self-enforcing. We have already described the candidate treatment prices in the absence of a referral. As such, we only need consider the case in which a patient seeks both a referral and treatment. If GPs have infinite patient pools, standard Bertrand arguments suggest that the equilibrium treatment price will be

$$r^*_{\infty 1} = \frac{C}{\delta}.$$

Furthermore, assuming there are an infinite number of potential GPs who stand ready to enter at zero cost, Bertrand competition among GPs will result in an equilibrium referral price of $w^* = k$. As such, patients will seek both a referral and high effort treatment if and only if

$$B - \frac{C}{\delta} - k \geq \max\left\{ B - \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C, \theta B \right\}.$$

If a patient would prefer high effort treatment without a referral to no treatment whatsoever, then this becomes

$$B - \frac{C}{\delta} - k \geq B - \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C,$$

which can be rearranged to obtain

$$k \leq \left\{ \frac{(1-\delta)\left[1-\pi\left(1-\theta\right)\right]}{\pi\delta\left(1-\theta\right)} \right\} C.$$

This is equivalent to

$$C \geq \left\{ \frac{\pi\delta\left(1-\theta\right)}{(1-\delta)\left[1-\pi\left(1-\theta\right)\right]} \right\} k.$$

On the other hand, if a patient would prefer low effort treatment to high effort treatment without a referral, then the patient will seek high effort treatment with a referral if and only if

$$B - \frac{C}{\delta} - k \geq \theta B,$$

which can be rearranged to obtain

$$k \leq (1-\theta)B - \frac{C}{\delta}.$$

This is equivalent to

$$C \leq \delta \left[(1-\theta) B - k \right].$$

Thus we know that the prevailing treatment price will be $r^*_{\infty 1}$ if either

$$\left\{ \frac{\pi\delta(1-\theta)}{(1-\delta)\left[1-\pi(1-\theta)\right]} \right\} k \leq C \leq \left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B,$$

or

$$\left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B \leq C \leq \delta \left[(1-\theta) B - k \right].$$

8.3 Market outcomes with endogenous prices

There are three possible outcomes in this market. The first outcome involves all patients obtaining both a referral and high effort treatment. The second outcome involves all patients obtaining high effort treatment without a referral. The third case involves all patients obtaining low effort treatment without a referral.

All patients will obtain both a referral and high effort treatment if either

$$\left\{ \frac{\pi\delta(1-\theta)}{(1-\delta)\left[1-\pi(1-\theta)\right]} \right\} k \leq C \leq \left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B$$

or

$$\left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B \leq C \leq \delta \left[(1-\theta) B - k \right].$$

In these cases, the equilibrium treatment price will be

$$r^*_{\infty 1} = \frac{C}{\delta},$$

while the equilibrium referral price will be

$$w^* = k.$$

Suppose that

$$\Omega = \left\{ \varpi : \left\{ \frac{\pi\delta(1-\theta)}{(1-\delta)\left[1-\pi(1-\theta)\right]} \right\} k \leq \varpi \leq \left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B \right\}$$

and

$$\Psi = \left\{ \psi : \left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B \leq \psi \leq \delta \left[(1-\theta) B - k \right] \right\}.$$

Note that it is possible that Ω might be an empty set. Similarly, it is possible that Ψ might be an empty set. In order for patients not to obtain a referral, we need both $C \notin \Omega$ and $C \notin \Psi$.

All patients will obtain high effort treatment without a referral if both $C \notin \Omega \cup \Psi$ and

$$C \leq \left[\frac{\pi\delta(1-\theta)^2}{\pi\delta(1-\theta)+(1-\delta)} \right] B.$$

In this case, the equilibrium treatment price will be

$$r_{11}^{*} = \left[1 + \frac{(1-\delta)}{\pi\delta(1-\theta)}\right] C,$$

while the referral market will not exist.

All patients will obtain low effort treatment without a referral if both $C \notin \Omega \cup \Psi$ and

$$C > \left[\frac{\pi\delta\left(1-\theta\right)^{2}}{\pi\delta\left(1-\theta\right)+\left(1-\delta\right)}\right] B.$$

In this case, the equilibrium treatment price will be

$$r_{10}^{*} = 0,$$

while the referral market will not exist.

References

[1] Abreu, D, D Pearce and E Stacchetti (1987), "Optimal cartel equilibria with imperfect monitoring", *Journal of Economic Theory 39(1)*, pp. 251-269.

[2] Abreu, D, D Pearce and E Stacchetti (1990), "Toward a theory of discounted repeated games with imperfect monitoring", *Econometrica 58(5)*, pp. 1041-1063.

[3] Arrow, KJ (1963), "Uncertainty and the welfare economics of medical care", *American Economic Review 53(5)*, pp. 941-973.

[4] Arrow, KJ (1968), "The economics of moral hazard: Further comment", *American Economic Review 58(3)(Part 1)*, June, pp. 537-539.

[5] Atkeson, A and R Lucas (1992), "On efficient distribution with private information", *Review of Economic Studies 59(200)*, pp. 427-453.

[6] Biglaiser, G (1993), "Middlemen as experts", *Rand Journal of Economics 24(2)*, pp. 212-223.

[7] Biglaiser, G and JW Friedman (1994), "Middlemen as guarantors of quality", *International Journal of Industrial Organization 12(4)*, pp. 509-531.

[8] Billingsley, P (1995), *Probability and measure (third edition)*, John Wiley and Sons, USA.

[9] Bolton, P and M Dewatripont (2005), *Contract theory*, Massachusetts Institute of Technology Press, USA.

[10] British Medical Association (2007), "NHS hospital treatment and information about specialists (updated June 2007)", Information available online at (http://www.bma.org.uk/ap.nsf/Content/infospecialists). Downloaded on 6 November 2007.

[11] Commonwealth of Australia (2005), *General practice in Australia: 2004*, GP Communications and Business Improvement Unit, Budget and Performance Branch, Primary Care Division, Department of Health and Ageing, Canberra, May (ISBN: 0-642-82-676-5).

[12] Commonwealth of Australia (Undated), "Medicare welcome kit (English version)", available online as a portable document format file at (http://www.medicareaustralia.gov.au/resources/welcome_kits/english/ ma_welcome_kit_english_medicare.pdf). Downloaded on 6 November 2007.

[13] Cooper, R and TW Ross (1984), "Prices, product qualities and asymmetric information: The competitive case", *Review of Economic Studies 51(2)*, pp. 197-207.

[14] Damania, R and D Round (2000), "The economics of consumer protection: Introduction", *Australian Economic Papers 39(4)*, pp. 403-407.

[15] Diamond, DW (1989), "Reputation acquisition in debt markets", *Journal of Political Economy 97(4)*, pp. 828-862.

[16] Dixit, A (1987), "Trade and insurance with moral hazard", *Journal of International Economics 23(3 and 4)*, November, pp. 201-220.

[17] Ehrlich, I and GS Becker (1972), "Market insurance, self-insurance and self-protection", *Journal of Political Economy 80(4)*, July-August, pp. 623-648.

[18] Eldridge, DS (2007a), *Essays in microeconomic theory*, Doctor of Philosophy Dissertation, The University of Texas at Austin.

[19] Eldridge, DS (2007b), "A shirking theory of referrals", Chapter 2 (pp. 4-60) in Eldridge, DS (2007a).

[20] Eldridge, DS (2007c), "A learning theory of referrals", Chapter 3 (pp. 61-98) in Eldridge, DS (2007a).

[21] Ely, JC, J Horner and W Olszewski (2005), "Belief-free equilibria in repeated games", *Econometrica 73(2)*, pp. 377-415.

[22] Folland, S, AC Goodman and M Stano (1993), *The economics of health and health care (second edition)*, Prentice-Hall, USA.

[23] Fudenberg, D, D Levine and E Maskin (1994), "The folk theorem with imperfect public information", *Econometrica 62(5)*, pp. 997-1039.

[24] Fudenberg, D and J Tirole (1995), *Game theory*, Massachusetts Institute of Technology Press, USA.

[25] Green, EJ and RH Porter (1984), "Noncooperative collusion under imperfect price information", *Econometrica 52(1)*, pp. 87-100.

[26] Grossman, S and O Hart (1983), "An analysis of the principal-agent problem", *Econometrica 51(1)*, pp. 7-45.

[27] Hadfield, GK, R Howse and MJ Trebilcock (1998), "Information based principles for rethinking consumer protection policy", *Journal of Consumer Policy 21(2)*, pp. 131-169.

[28] Hermalin, BE and ML Katz (1991), "Moral hazard and verifiability: The effects of renegotiating in agency", *Econometrica 59(6)*, pp. 1735-1754.

[29] Hirshleifer, J and J Riley (1992), *The analytics of uncertainty and information*, Cambridge Surveys of Economic Literature, Cambridge University Press, Great Britain.

[30] Hogan, C (2001), "Enforcement of implicit employment contracts through unionization", *Journal of Labor Economics 19(1)*, pp. pp. 171-195.

[31] Holmstrom, B (1979), "Moral hazard and observability", *Bell Journal of Economics 10(1)*, pp. pp. 74-91.

[32] Holmstrom, B (1999), "Managerial incentive problems: A dynamic perspective", *Review of Economic Studies 66(1)*, pp. 169-182.

[33] Horner, J (2002), "Reputation and competition", *American Economic Review 92(3)*, pp. 644-663.

[34] Jewitt, I (1988), "Justifying the first-order approach to principal-agent problems", *Econometrica 56(5)*, pp. 1177-1190.

[35] Judd, KL (1985), "The law of large numbers with a continuum of iid random variables", *Journal of Economic Theory 35(1)*, February, pp. 19-25.

[36] Klein, B and KB Leffler (1981), "The role of market forces in assuring contractual performance", *Journal of Political Economy 89(4)*, August, pp. 615-641.

[37] Kocherlakota, NR (1996), "Implications of risk sharing without commitment", *Review of Economic Studies 63(4)*, pp. 595-609.

[38] Kreps, DM (1990), "Corporate culture and economic theory", Chapter 4 (pp. 90-143) in Alt, JE and KA Shepsle (Editors) (1990), *Perspectives on positive political economy*, Cambridge University Press, USA.

[39] Laffont, JJ and D Martimort (2002), *The theory of incentives: The principal-agent model*, Princeton University Press, USA.

[40] Macho-Stadler, I and JD Perez-Castrillo (2001), *An introductrion to the economics of information: Incentives and contracts (second edition)*, Oxford University Press, Great Britain.

[41] MacLeod, B and JM Malcolmson (1989), "Implicit contracts, incentive compatibility and involuntary unemployment", *Econometrica 57(2)*, pp. 447-480.

[42] Mailath, GJ and L Samuelson (2006), *Repeated games and reputations: Long-run relationships*, Oxford University Press, USA.

[43] Malcolmson, JM (1983), "Trade unions and economic efficiency", *Economic Journal 93 (Supplement: Conference Papers)*, pp. 51-65.

[44] Mas-Colell, A, MD Whinston and JR Green (1995), *Microeconomic theory*, Oxford University Press, USA.

[45] Mirlees, JA (1999), "The theory of moral hazard and unobservable behaviour: Part 1", *Review of Economic Studies 66(1)*, pp. 3-21.

[46] Mooney, G and M Ryan (1993), "Agency in health care: Getting beyond first principles", *Journal of Health Economics 12(2)*, pp. 125-135.

[47] Myerson, R (1979), "Incentive compatibility and the bargaining problem", *Econometrica 47(1)*, pp. 61-73.

[48] Pauly, MV (1968), "The economics of moral hazard: Comment", *American Economic Review 58(3)(Part 1)*, June, pp. 531-537.

[49] Pauly, MV (1974), "Overinsurance and public provision of insurance: The roles of moral hazard and adverse selection", *Quarterly Journal of Economics 88(1)*, February, pp. 44-62.

[50] Powell-Davies, G and D Fry (2005), "General practice in the health system", Chapter 10 (pp. 420-464) of Commonwealth of Australia (2005), *General practice in Australia: 2004*, GP Communications and Business Improvement Unit, Budget and Performance Branch, Primary Care Division, Department of Health and Ageing, Canberra, May (ISBN: 0-642-82-676-5).

[51] Prescott, EC and RM Townsend (1984), "Pareto optima and competitive equilibria with adverse selection and moral hazard", *Econometrica 52(1)*, January, pp. 21-46.

[52] Prescott, ES and RM Townsend (2002), "Collective organizations versus relative performance contracts: Inequality, risk sharing and moral hazard", *Journal of Economic Theory 103(2)*, pp. 282-310.

[53] Radner, R (1982), "Monitoring cooperative agreements in a repeated principal-agent relationship", *Econometrica 49(5)*, pp. 1127-1148.

[54] Radner, R (1985), "Repeated principal-agent games with discounting", *Econometrica 53(5)*, pp. 1173-1198.

[55] Radner, R, R Myerson and E Maskin (1986), "An example of a partnership game with discounting and uniformly inefficient equilibria", *Review of Economic Studies 53(1)*, pp. 59-69.

[56] Rogerson, W (1985a), "Repeated moral hazard", *Econometrica 53(1)*, pp. 69-76.

[57] Rogerson, W (1985b), "The first-order approach to principal-agent problems", *Econometrica 53(6)*, pp. 1357-1368.

[58] Ross, SA (1973), "The economic theory of agency: The principal's problem", *American Economic Review: Papers and Proceedings 63(2)*, pp. 134-139.

[59] Royal New Zealand College of General Practitioners (2005), *Your guide to New Zealand general practice*, Version 3, Royal New Zealand College of General Practitioners, New Zealand, July (ISBN: 0-9582429-1-7).

[60] Rubinstein, A and ME Yaari (1983), "Repeated insurance contracts and moral hazard", *Journal of Economic Theory 30(1)*, pp. 74-97.

[61] Shapiro, C (1983), "Premiums for high quality products as returns to reputations", *Quarterly Journal of Economics 98(4)*, pp. 659-679.

[62] Shapiro, C and JE Stiglitz (1984), "Equilibrium unemployment as a worker discipline device", *American Economic Review 74(3)*, pp. 433-444.

[63] Shavell, S (1979a), "Risk sharing and incentives in the principal and agent relationship", *Bell Journal of Economics 10(1)*, pp. 55-73.

[64] Shavell, S (1979b), "On moral hazard and insurance", *Quarterly Journal of Economics 93(4)*, pp. 541-592.

[65] Smith, RL (2000), "When competition is not enough: Consumer protection", *Australian Economic Papers 39(4)*, pp. 408-425.

[66] Spear, SE and S Srivastava (1987), "On repeated moral hazard with discounting", *Review of Economic Studies 54(4)*, pp. 599-617.

[67] Stigler, GJ (1971), "Can regulatory agencies protect the consumer?", Chapter 11 (pp. 178-188) in Stigler, GJ (1975), *The citizen and the state: Essays on regulation*, University of Chicago Press, USA.

[68] Tadelis, S (1999), "What's in a name? Reputation as a tradable asset", *American Economic Review 89(3)*, pp. 548-563.

[69] Thomas, J and T Worrall (1990), "Income fluctuation and asymmetric information: An example of a repeated principal-agent problem", *Journal of Economic Theory 51(2)*, pp. 367-390.

[70] Tirole, J (1988), *The theory of industrial organization*, Massachusetts Institute of Technology Press, USA.

[71] Tirole, J (1996), "A theory of collective reputations (with applications to the persistence of corruption and to firm quality)", *Review of Economic Studies 63(1)*, pp. 1-22.

[72] Townsend, R (1982), "Optimal multiperiod contracts and the gain from enduring relationships under private information", *Journal of Political Economy 90(6)*, pp. 1166-1186.

Recent Discussion Papers

School of Business Discussion Papers are available from the Research Officer, School of Business, La Trobe University VIC 3086, Australia.

06.01 Shawn Chen-Yu Leu – A New Keynesian Perspective of Monetary Policy in Australia.

06.02 David Prentice – A re-examination of the origins of American industrial success.

06.03 Elisabetta Magnani and David Prentice – Outsourcing and Unionization: A tale of misallocated (resistance) resources.

06.04 Buly A. Cardak and Chris Ryan – Why are high ability individuals from poor backgrounds under-represented at university?

06.05 Rosaria Burchielli, Donna M. Buttigieg and Annie Delaney – Why are high ability individuals from poor backgrounds under-represented at university?

06.06 László Kónya and Jai Pal Singh – Exports, Imports and Economic Growth in India.

07.01 Rosaria Burchielli and Timothy Bartram – What makes organising work? A model of the stages and facilitators of organising.

07.02 László Kónya and Jai Pal Singh - Causality between Indian Exports, Imports, and Agricultural, Manufacturing GDP.

07.03 Buly A. Cardak and Chris Ryan – Participation in Higher Education: Equity and Access – Are Equity-based Scholarships an Answer?

07.04 Samantha Farmakis-Gamboni and Divid Prentice – Does Reducing Union Bargaining Power Increase Productivity