ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

HEALTH SERVICE GATEKEEPERS

James M. Malcomson

Number 169

September 2003

Manor Road Building, Oxford OX1 3UQ

Health Service Gatekeepers*

James M. Malcomson
University of Oxford, U.K.
Email: james.malcomson@economics.oxford.ac.uk

First version: April 20, 2001
This version: September 4, 2003

*I would like to thank Philippe Choné, Raymond Deneckere, Ian Jewitt, two anonymous referees and participants in the Fifth Biennial Conference on the Industrial Organization of Health Care, Meredith, New Hampshire, September 23-25, 2001 for helpful discussion and comments. The support of the Economic and Social Research Council (ESRC) is gratefully acknowledged. The research was funded by ESRC award number R000236723.

Abstract

Incentive contracts for gatekeepers who control patient access to specialist medical services provide too weak incentives to investigate cost further when expected cost of treatment is greater than benefit. Making gatekeepers residual claimants with a fixed fee from which treatment costs must be met (as with full insurers who are themselves gatekeepers) provides too strong incentives when expected cost is less than benefit. Giving patients the choice between a gatekeeper with an incentive contract and one without is unstable. With one scenario, patients always prefer the latter. With another, patients have incentives to acquire information that makes incentive contracts ineffective.

Keywords: Gatekeepers, Patient referrals, General practitioners, Fundholding, Medical insurance, Incentive contracts

JEL classifications: I11, I18

1 Introduction

An important role in many health services is that of a *gatekeeper* who controls patients' access to specialist medical services. This paper is concerned with the contractual basis under which gatekeepers operate. The need for a gatekeeper arises when patients pay below market prices for services at the time they use them, either because they have private insurance or because the services are publicly funded. Without a gatekeeper, insured patients will make use of specialist services up to the point at which the marginal benefit to them equals the marginal cost to them. If the marginal cost is below the market price, specialist services are overused. But will a gatekeeper actually ensure that specialist medical services are used when appropriate? And how does that depend on the contractual basis under which gatekeepers operate?

Private medical insurance in the US has not traditionally used gatekeepers—patients could themselves choose when and where to make use of specialist services. But many managed care plans use gatekeeper arrangements, either by requiring a referral from a specified primary care physician before consulting a specialist (Glied (2000)) or by fully insuring provision only if it is supplied, or authorised, by the responsible health maintenance organisation (HMO). In the publicly-funded British National Health Service (NHS), the gatekeeper role is filled by a general practitioner (GP), see Scott (2000). To receive non-emergency specialist care, a patient has to be referred by a GP. Many private medical insurers in Britain also require this. A gatekeeping physician may provide valuable information to patients about specialist services but it is the control of access to services that makes the physician a gatekeeper.

An HMO that itself both carries out the gatekeeper role and provides full insurance pays for the cost of specialist treatment and thus retains cost savings from not referring a patient, or from referring the patient for a less expensive specialist service. This contractual arrangement is referred to here as a *paying gatekeeper*. There has been concern in the US that managed care plans have been too much concerned with saving money and too little with patient welfare. See Cutler and Zeckhauser (2000, pp. 629-631) and Glied (2000, pp. 739-740) for discussions of the empirical evidence. In the British NHS, there has been some experimentation with contractual arrangements for gatekeepers. In the traditional arrangement a GP receives a capitation fee for each patient on her list but does not pay for specialist services—the costs of these are met by the health authority. Concern that GPs did not take sufficient account of these costs led to the introduction in 1991 of arrangements under which GPs could become *fundholders*. Fundholding GPs were allocated a budget out of which they were expected to meet the costs of many of the specialist services for which they referred patients. For those services, they were thus paying gatekeepers. Concern about the working of the fundholding system led to its abolition in April 1999. See Glennerster, Matsaganis, Owens and Hancock (1993) for more details on fundholding and an evaluation from a social policy perspective.

Having gatekeepers bear the full cost of specialist medical treatment is not the only way to make them more cost conscious. Any incentive contract with reward decreasing in the cost of specialist treatment gives a gatekeeper a reason to find out more about cost. There would

seem to be two potential ways in which this may improve services: (1) of those patients who would otherwise have been referred, some may be referred at a lower cost while others are not referred because the cost is found to exceed the benefit, and (2) of those patients for whom referral would not otherwise have been thought cost effective, some may be referred because the cost is discovered to be lower. Set against any such gains must be the time, effort and money gatekeepers incur to acquire additional information. The issues of concern in this paper are what feasible contractual arrangements are most effective at inducing gatekeepers to put in that time, effort and money where it is most worthwhile. Can one improve on the paying gatekeeper arrangement by which the gatekeeper is made, in effect, residual claimant, as with HMOs that themselves act as both gatekeepers and insurers and many of the referrals by GP fundholders in Britain? And where patients have the choice between types of gatekeeper (as they did in the fundholding system in Britain), will those types for whom that time, effort and money is most worthwhile actually choose a gatekeeper with an incentive contract in preference to one without?

The paper uses the following framework to analyse these issues. Patients differ in the cost of, and benefit from, specialist treatment resulting from referral. With a low level of effort, a gatekeeper observes the expected benefit and receives an initial signal of cost. On the basis of that signal, the gatekeeper can decide whether to refer the patient. With more effort, the gatekeeper may find out more information about cost and so make a better-informed judgement about whether referral is worthwhile. A gatekeeper who retains some of the cost savings has more incentive to incur that effort.

A number of messages come across from the analysis. It is important to distinguish between medical conditions for which the gatekeeper observes a cost signal *before* deciding how much to investigate from those for which the gatekeeper observes a cost signal only *after* deciding how much to investigate. In the former case, giving a gatekeeper a reward that decreases with cost *never* realises gains of the type identified in (2) above. As for the potential gains identified in (1) above, a paying gatekeeper has too strong incentives to carry out further investigations, in the sense that such investigations are carried out for more patient types than is socially efficient. As a result, fewer patient types are referred than would be the case with efficient incentives. This contrasts with other agency problems in which making the agent residual claimant for costs ensures efficient decisions. The reason for the difference is that here referral decisions affect benefit as well as cost and the gatekeeper is not residual claimant for benefit. As long as the cost of specialist treatment can be monitored effectively even if benefit cannot, one can improve on the contractual arrangement by which the gatekeeper pays the cost, though it is not always possible to induce fully efficient decisions. This result has implications for the contractual arrangements for gatekeepers. It also has implications for the organisation of private health insurance. More efficient than having an HMO itself act as both gatekeeper and insurer is for insurance to be provided by a third party insurer who employs the gatekeeper on a contractual basis that does not make the gatekeeper residual claimant. There is also a further message for the organisation of health services. Patients who have a choice between

a gatekeeper with an incentive contract and one without when they seek a referral will never prefer the former. Introducing patient choice between a fundholding and a non-fundholding GP, as in the British NHS, is not a sensible organisational structure.

The messages that come across are somewhat different in the case of medical conditions for which the gatekeeper observes a cost signal only *after* deciding how much investigation to do. In contrast to the previous case, there may then be gains of the type identified in (2) above from giving a gatekeeper incentives. Also in contrast to the previous case, patients may prefer to consult a gatekeeper with an incentive contract to one without. They are more likely to be referred by a gatekeeper without an incentive contract if the cost signal indicates an expected cost less than benefit and by a gatekeeper with an incentive contract if that signal indicates an expected cost greater than benefit. There is thus a premium to patients who find out more about the costs likely to arise from referral in their case before choosing between different gatekeeper arrangements (that is, *before* choosing their type of managed care plan or GP). But, if patients become too well informed, adverse selection becomes so serious that even gatekeepers with incentive contracts lose all incentive to incur additional effort and cease to behave differently from those without incentive contracts. In this case too, therefore, the model predicts that coexistence of the two arrangements is unstable when patients can become informed. For testing that prediction empirically, it is unfortunate that fundholding in the British NHS was abolished too soon to find out whether the prediction is correct.

The model used here is a form of principal-agent model with both selection among types unknown to the principal (because patient costs and benefits are not observed by the insurer or health authority) and moral hazard (arising from unverifiable gatekeeper effort). It differs from the standard procurement model with unknown types and unverifiable effort (see, for example, Laffont and Tirole (1993, Ch. 1)) in that there the principal is trading off lower information rent to the agent against sub-optimal agent effort to reduce cost when the agent knows the type before the contract is set and cost is verifiable *ex post*. If there is selection between types, procurement occurs for an interval of lowest cost types. In contrast, here information rents are not an issue because the agent does not know a patient's type before the contract is set. Instead the principal wishes the agent (the gatekeeper) to select between types on the difference between benefit and cost, with cost but not benefit verifiable *ex post*. In the literature on contracts specifically for health care, much of the concern has been with ensuring that *all* types of patient are treated, with an optimal trade-off between quality and cost, so there is no selection. See Chalkley and Malcomson (2000) for a survey. An exception is Ma and Riordan (2002, Section 6), in which physicians select on the basis of benefit. There, however, the cost of treatment is known and the same for all patient types, so there is no role for physician effort to find out about cost and no moral hazard. The physician's only role is to influence the cutoff benefit above which a patient is treated, which greatly simplifies the problem of inducing efficient selection. Moreover, the only incentive contracts considered for the physician are those with a fixed charge for agreeing to the patient receiving treatment.

The paper is organised as follows. The next section sets out the model used for the analysis.

Section 3 considers the case in which a gatekeeper receives the cost signal before deciding whether to investigate further and can, therefore, condition that decision on the signal, Section 4 the case in which a gatekeeper has to decide how much to find out about the cost before receiving a cost signal. Section 5 discusses empirical evidence relating to, and some practical implications of, the model. Section 6 contains concluding remarks.

2 The model

Patients seeking specialist treatment must visit a gatekeeper physician who decides whether to refer them. Patients are of different types, each characterised by a benefit from referral b known by all parties to lie in the interval $\left[\underline{b}, \bar{b}\right]$, with $0 < \underline{b} \leq \bar{b}$, and a cost of treatment resulting from that referral c known by all parties to lie in the finite interval $\left[\underline{c}, \bar{c}\right]$, with $0 < \underline{c} < \bar{b} < \bar{c}$. Benefit is measured net of any disutility of treatment experienced by the patient. Patient types are distributed in the population according to the commonly known joint distribution $H(c, b)$, with associated density function $h(c, b)$. A patient's type is unknown to the gatekeeper before agreeing to take on the patient. To avoid cream skimming (declining to take on more costly patients) of the kind discussed by Newhouse (1989) for the US and Matsaganis and Glennerster (1994) for the UK, payments would need to be conditioned on characteristics observed by the gatekeeper at that stage. Cream skimming is a potentially important issue but not the concern of the present paper.

Patients do not pay directly for the services of either the gatekeeper or the specialist. They are fully insured by either a private insurer or a publicly-funded health authority. To allow for the deadweight loss of raising public funds by taxation, it is convenient to adopt the convention that the cost c includes a premium $\alpha \geq 0$ to account for that loss. The actual monetary payment for the specialist services is thus $c/(1 + \alpha)$.

With loss of reservation utility $\underline{U}$, a gatekeeper visited by a patient observes the patient's benefit of referral b and receives a signal $s \in S = \left[\underline{s}, \bar{s}\right]$ that provides information about the patient's cost of treatment resulting from referral.[1] Uncertainty about cost given this signal may arise from uncertainty either about the cost of a particular medical intervention (because, for example, of the possibility of complications) or about which medical intervention the specialist will select. One interpretation of s is as the set of "presenting symptoms" readily apparent to any physician who examines the patient. An example might be severe abdominal pain. If the patient is referred on the basis of this information alone, the treatment cost resulting from the referral will depend on what the specialist determines to be the cause. This is captured

[1] The benefit b relevant to the present analysis is the gatekeeper's assessment of it after eliciting information from the patient about, for example, how much pain the medical condition causes. How to elicit such information is a concern of any gatekeeper, whether or not having an incentive contract, and so is not analysed here. The results that follow continue to hold when the gatekeeper receives only a *signal* of expected benefit if, conditional on that signal and the signal s, actual benefit is uncorrelated with actual cost. (This does not, of course, imply that benefit is uncorrelated with cost *ex ante*—only that knowing cost provides no information about benefit *additional to* that provided by the signal.) Allowing for this complicates the exposition without adding further insights. Some amendments are required in other cases.

by the following specification. For a given patient type (c, b), the signal s has commonly known probability density function $g(s; c, b)$. More convenient than using the conventional distribution function associated with $g(s; c, b)$ is to define

$$G(X; c, b) = \int_{s \in X} g(s; c, b)\, ds, \text{ for } X \subseteq S. \tag{1}$$

This function specifies the probability that s lies in the set X for given (c, b). It is the same as the conventional distribution function for $s = s'$ when $X = \left[\underline{s}, s'\right]$. The gatekeeper's probability assessment $f(c; s, b)$ that a patient has true cost c conditional on information (s, b) is given by Bayes' Rule as

$$f(c; s, b) = \frac{g(s; c, b)\, h(c, b)}{\int_{\underline{c}}^{\bar{c}} g(s; c, b)\, h(c, b)\, dc}. \tag{2}$$

Let $F(c; s, b)$ denote the associated distribution function and $c(s, b) = E\{c \mid s, b\}$ the mean cost conditional on (s, b). This paper is concerned with medical conditions for which the probability distributions have the property that the upper support of $f(c; s, b)$ exceeds b given s. Thus, given any signal s, there is always positive probability that the actual cost resulting from referral exceeds the benefit. For any (s, b) for which this is not the case, referral is always efficient.

By incurring utility loss $e > 0$, the gatekeeper may learn more about cost before deciding whether to refer the patient. Specifically, the gatekeeper learns the true cost c with probability π $(0 < \pi < 1)$ but nothing more than s with probability $1 - \pi$.[2] The gatekeeper is risk neutral, so e can be interpreted as either monetary cost or disutility of another kind. It might, for example, correspond to the time taken to give the patient a thorough examination or to the cost of supplies used to carry out tests in the office. Here it is referred to simply as *effort*. The gatekeeper receives no direct utility or disutility from referring the patient given the information available but, if operating under an incentive contract, the decision may have financial consequences. Within this framework, two different scenarios are analysed.

Assumption 1 *The gatekeeper observes the cost signal s and benefit b before deciding whether to incur effort e.*

Assumption 2 *The gatekeeper must decide whether to incur effort e without first observing the cost signal s and benefit b. The gatekeeper learns b and either s or c.*

Assumption 1 is appropriate when, as with many medical services, investigations are sequential. Assumption 2 is appropriate when there is a choice between two investigations that are either mutually exclusive or such that it is never worthwhile to carry out both. It is also appropriate when "effort" corresponds to buying equipment or setting up procedures before the patient visits the gatekeeper.

The information structure in the model is as follows:

[2] An obvious extension would have $e \in \left[\underline{e}, \bar{e}\right]$ continuously variable and the probability $\pi(e)$ of discovering the true cost an increasing function of e. Some implications of this extension are mentioned below.

1. whether a patient has visited the gatekeeper and, if referred, the cost of treatment c are verifiable at no cost;

2. the benefit b and the signal s are observed by both the gatekeeper and the patient but are not verifiable;

3. whether the additional effort e is incurred and the cost c, if learnt in advance of the referral decision, are observed only by the gatekeeper but the gatekeeper can choose to credibly reveal c if learnt.

In the present context, it is natural that the patient is better informed about the benefit b and the signal s than about whether the gatekeeper has discovered the actual cost c. The patient will, for example, have some idea what it would feel like to be made well again and will be aware of the symptoms that induced a visit to the gatekeeper in the first place. In contrast, a patient presenting with, to use the previous example, severe abdominal pain is unlikely to know whether the physician has located a tumour unless the physician reveals this. Locating a tumour will reveal information about the cost of treatment c, for example, the need for a specific surgical procedure. Moreover, once located, it is likely to be relatively straightforward for the gatekeeper to demonstrate the presence of the tumour to another physician, which motivates the specification that information discovered as a result of the effort e can be credibly revealed if the gatekeeper wishes. The specification that the patient observes b and s perfectly, but not whether the gatekeeper learns c, serves to bring out this difference between b and s on the one hand and c on the other.[3]

A patient can limit adverse gatekeeper decisions by requesting a second opinion from another physician in the way standard in medical practice. That physician, also with utility loss $\underline{U}$, observes the same benefit b and signal s as the gatekeeper, and can verify c if the gatekeeper has learnt this before making the referral decision and chooses to reveal it. If paid a fixed fee of $\underline{U}$, she has no incentive to report other than truthfully, given the information available to her, on whether referral is warranted.[4]

Proposition 1 *Suppose a second opinion is at the payer's expense. Suppose also the gatekeeper's contract is such that a second opinion confirming the gatekeeper's referral decision leaves the gatekeeper's payoff unaffected but one overturning the gatekeeper's referral decision*

[3] If instead the patient observed only noisy signals of b and s, the second opinion mechanism to be discussed shortly as a way for the patient to limit adverse gatekeeper decisions would result in second opinions sometimes actually being sought.

[4] In principle, the payer could require a second opinion in every case, use this to verify b and s, and thus make payment to the gatekeeper conditional on b and s. It is assumed here that $\underline{U}$ is too large to make this worthwhile and that the maximum penalty that can be imposed on the gatekeeper is too small to make random checks cost effective. It is also assumed that it is too costly to be worthwhile setting up any mechanism that both induces the patient and the gatekeeper to reveal their common observations directly and does not undermine the purpose of providing the patient with health insurance in the first place. For example, permitting the patient to bribe the gatekeeper to agree to referral (which the patient would be willing to do up to the amount b) and requiring the gatekeeper to pay all the costs resulting from referral would ensure that the gatekeeper receives the full social benefit of her decision. It would, however, leave the patient in the position of being effectively uninsured.

results in a payoff to the gatekeeper strictly lower than if the gatekeeper had made the opposite decision. Then no second opinion is sought and:

1. *under Assumption 1: if $b \geq c(s,b)$, the gatekeeper refers the patient unless she has learnt c and $c > b$; if $b < c(s,b)$, the gatekeeper makes referral decisions unconstrained by the possibility of a second opinion;*

2. *under Assumption 2: if s is learnt and $b \geq c(s,b)$ or if c is learnt and $b \geq c$, the gatekeeper refers the patient; if s is learnt and $b < c(s,b)$ or if c is learnt and $b < c$, the gatekeeper makes referral decisions unconstrained by the possibility of a second opinion.*

Proof. Assumption 1. Suppose $b \geq c(s,b)$. Since the patient observes b and s, he will request a second opinion (which will overturn the gatekeeper's decision) unless referred or unless the gatekeeper reveals c and $c > b$. Having the decision overturned reduces the gatekeeper's payoff, so the gatekeeper will refer the patient unless she has learnt c and $c > b$. In neither case can the patient gain by requesting a second opinion. Suppose now $b < c(s,b)$. A second opinion will result in the patient not being referred unless the gatekeeper has learnt c and chooses to reveal it. The gatekeeper thus receives no penalty from refusing referral. If the gatekeeper refers the patient, the patient will not ask for a second opinion because this will result in non-referral and $\underline{b} > 0$. So the gatekeeper is unconstrained by the possibility of a second opinion. Again, the patient does not gain by requesting one.

Assumption 2. Under Assumption 2, either s or c is learnt but not both. If s is learnt, the same argument applies as for Assumption 1 except that there is no possibility that c is also learnt. If c is learnt, the patient will request a second opinion unless referred or unless the gatekeeper reveals c and $c > b$. Since in this case the gatekeeper has not learnt s, she will have to reveal c to justify non-referral. If $b \geq c$, the second opinion will overturn a decision not to refer, reducing the gatekeeper's payoff, so the gatekeeper will refer the patient and the patient will not request a second opinion. If $b < c$, the same argument applies as for Assumption 1 when $b < c(s,b)$. ■

In the absence of a second opinion, the only verifiable information is whether a patient visits the gatekeeper and, if the patient is referred, the cost of the treatment resulting from that referral. A contract for payment to the gatekeeper can therefore be denoted by a payment P_0 for a patient who visits but is not referred and a payment $P(c) \leq P_0$ for each $c \in \left[\underline{c}, \bar{c}\right]$ for a patient referred with the resulting treatment cost c. The restriction to $P(c) \leq P_0$ for all $c \in \left[\underline{c}, \bar{c}\right]$ ensures that, by Proposition 1, the gatekeeper can be induced not to refer a patient for whom $b < c(s,b)$, or $b < c$ if c is learnt before the referral decision, because it is straightforward to ensure that the gatekeeper's contract satisfies the other contract conditions specified there. A gatekeeper has an *incentive contract* if $P(c) < P_0$ for some $c \in \left[\underline{c}, \bar{c}\right]$. Because a second opinion is never sought, it imposes no costs on the payer. Thus giving the patient the right to seek a second opinion at the payer's expense is a costless mechanism for ensuring that referrals are made if and only if the benefit exceeds the expected social cost given the information available to the

gatekeeper. For a public sector payer it is efficient. For a private insurer, competition in the insurance market will ensure that it, or some equally costless and effective mechanism, is part of the insurance contract with the patient.[5]

One incentive contract is a fixed fee paid to the gatekeeper whether or not the patient is referred, with the gatekeeper paying the cost of specialist treatment. This is the paying gatekeeper arrangement discussed in the Introduction. In that case, the net reward to the gatekeeper for a patient referred with cost c is $P(c)$ given by

$$P(c) = P_0 - c/(1+\alpha)\,, \text{ for } c \in \left[\underline{c}, \bar{c}\right], \tag{3}$$

where P_0 is the fixed fee. (Recall that c is the social cost of treatment resulting from referral and that the gatekeeper incurs only the private cost $c/(1+\alpha)$.) For an HMO that both provides full insurance and acts as gatekeeper, the net reward is also given by (3) with P_0 the insurance premium. (In this case, there is no deadweight loss from taxation, so $\alpha = 0$). To achieve a net reward function other than (3) with full private insurance, the gatekeeper needs to be independent of the insurer. Thus, the choice of reward function has implications for the organisation of private health insurance.

A gatekeeper without an incentive contract has $P(c) = P_0$ for all $c \in \left[\underline{c}, \bar{c}\right]$ and thus receives payment P_0 whether or not a patient is referred. Such a gatekeeper has no incentive to incur effort e under either Assumption 1 or Assumption 2 and makes referral decisions on the basis of the information (s, b). In view of Proposition 1, she therefore refers patients for whom the benefit of treatment b is no less than the expected cost $c\,(s, b)$ given the information (s, b), that is patients with $s \in S(b)$ defined by

$$S(b) = \{s \mid c\,(s, b) \leq b\}\,. \tag{4}$$

She does not refer patients for whom the benefit is less than the expected cost, those with $s \notin S(b)$. These referral rules are efficient conditional on effort e not being incurred. For such a gatekeeper to be prepared to assess patients, she must receive payment of at least the lost reservation utility $\underline{U}$ but there is no need for her to be paid more. Thus the payer sets $P_0 = \underline{U}$.

A natural benchmark for a publicly-funded health service is the social welfare arising from the gatekeeper's choices. That is also a natural benchmark for a private insurer in a perfect insurance market when committing to a gatekeeper arrangement before individuals choose their

[5] Ellis and McGuire (1990) and Ma and Riordan (2002) assume that health care decisions maximize a weighted sum of patient and physician utility as the result of an (unmodelled) bargaining process. (A similar outcome occurs when patient care directly enters physician preferences as discussed by Newhouse (1970) and Ellis and McGuire (1986)). This alternative mechanism for limiting physician decisions adverse to the patient would, however, result in a gatekeeper with no incentive contract always referring a patient who sought referral, so the gatekeeping would be completely ineffective. Added to the present model, it would increase the gains from providing gatekeepers with incentive contracts. Evidence on the effectiveness of gatekeeping is discussed in Section 5 below.

insurance contract. Conditional on additional effort e not being incurred, social welfare is

$$\int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{\bar{c}} (b-c)\, G(S(b)\,;c,b)\, h(c,b)\, dc\, db - (1+\alpha)\, \underline{U}, \tag{5}$$

the factor $(1+\alpha)$ appearing because the gatekeeper must be compensated for disutility $\underline{U}$ from public funds with premium α. (Recall that the cost c is defined to include this premium.)

For a gatekeeper with an incentive contract, the analysis depends on whether Assumption 1 or Assumption 2 holds. The sections that follow deal with each case in turn.

3 Gatekeeper chooses effort after receiving cost signal

This section considers the behaviour of a gatekeeper with an incentive contract under Assumption 1, when the gatekeeper observes the signal s and benefit b before deciding whether to incur effort e. It first investigates when it is socially efficient to have the gatekeeper put in this effort.

3.1 Efficient effort

Under Assumption 1, the decision whether to incur e can be made with knowledge of (s,b). If e is not incurred, it is efficient to refer a patient with $s \in S(b)$ at expected cost $c(s,b)$. Social welfare for given (s,b) is then

$$\max\left[0, b - c(s,b)\right] - (1+\alpha)\, \underline{U}. \tag{6}$$

If e is incurred, actual cost c is identified with probability π and the patient referred if $c \leq b$, but with probability $1-\pi$ no additional information is acquired and the patient is referred if $c(s,b) \leq b$. So social welfare is

$$\pi \int_{\underline{c}}^{b} (b-c)\, f(c;s,b)\, dc + (1-\pi)\max\left[0, b-c(s,b)\right] - (1+\alpha)\left(\underline{U}+e\right). \tag{7}$$

The social welfare *gain* from incurring e for given (s,b) is the difference between (7) and (6):

$$W(s,b) = \pi\left[\int_{\underline{c}}^{b} (b-c)\, f(c;s,b)\, dc - \max\left[0, b-c(s,b)\right]\right] - (1+\alpha)\, e. \tag{8}$$

It is convenient to consider separately values of s for which $c(s,b) \leq b$ (that is, $s \in S(b)$) and those for which $c(s,b) \geq b$. For the former case, $W(s,b)$ can be written

$$\begin{aligned} W(s,b) &= \pi\left[\int_{\underline{c}}^{b} (b-c)\, f(c;s,b)\, dc - \int_{\underline{c}}^{\bar{c}} (b-c)\, f(c;s,b)\, dc\right] - (1+\alpha)\, e, \\ &= -\pi \int_{b}^{\bar{c}} (b-c)\, f(c;s,b)\, dc - (1+\alpha)\, e, \text{ for all } s \in S(b)\,,\, b \in \left[\underline{b}, \bar{b}\right]. \end{aligned} \tag{9}$$

Because the integral in this is restricted to $c \geq b$, the integral term is positive. It corresponds to the welfare gain from not referring a patient whose expected cost of treatment given the signal s is less than the benefit but whose actual cost is revealed by additional effort to be greater than the benefit. That gain must, of course, be multiplied by the probability π that the actual cost is identified and have subtracted from it the social cost $(1+\alpha)\,e$ of providing the extra effort. It is socially efficient to incur effort e if $W(s,b) \geq 0$, that is if

$$\frac{\pi}{1+\alpha}\int_b^{\bar{c}} (c-b)\, f(c;s,b)\, dc \geq e, \text{ for } s \in S(b)\,, b \in \left[\underline{b}, \bar{b}\right]. \tag{10}$$

For $c(s,b) \geq b$, $W(s,b)$ defined in (8) can be written

$$W(s,b) = \pi \int_{\underline{c}}^{b} (b-c)\, f(c;s,b)\, dc - (1+\alpha)\, e, \text{ for all } s \text{ such that } c(s,b) \geq b, b \in \left[\underline{b}, \bar{b}\right]. \tag{11}$$

Again, the integral term is positive. It corresponds to the welfare gain from referring a patient whose expected cost of treatment given the signal s is greater than the benefit but whose actual cost is revealed by extra effort to be less than the benefit. Again, that gain must be multiplied by the probability π that the actual cost is identified and have subtracted from it the social cost $(1+\alpha)\,e$ of providing the extra effort.

These two terms for welfare gain correspond to the two ways mentioned in the Introduction in which making gatekeepers more cost conscious by using incentive contracts can potentially be beneficial. To provide some insight about when effort e is worthwhile, consider how a change in s affects the social welfare gain.

Proposition 2 *Suppose Assumption 1 holds. Then, for all $b \in \left[\underline{b}, \bar{b}\right]$, a change in s from s' to s'' for which $F\left(c;s'',b\right)$ stochastically dominates $F\left(c;s',b\right)$ in the first-order sense: (1) increases the welfare gain from incurring effort e if $c\left(s',b\right), c\left(s'',b\right) < b$ and $F\left(c;s'',b\right) < F\left(c;s',b\right)$ for some $c > b$; and (2) decreases the welfare gain if $c\left(s',b\right), c\left(s'',b\right) > b$ and $F\left(c;s'',b\right) < F\left(c;s',b\right)$ for some $c < b$.*

Proof. From (9) and integration by parts

$$\begin{aligned} W(s,b) &= -\pi\left\{\left[(b-c)\,F(c;s,b)\right]_{c=b}^{c=\bar{c}} + \int_b^{\bar{c}} F(c;s,b)\, dc\right\} - (1+\alpha)\, e \\ &= -\pi\left\{b - \bar{c} + \int_b^{\bar{c}} F(c;s,b)\, dc\right\} - (1+\alpha)\, e, \qquad \forall s \in S(b)\,. \end{aligned}$$

From (11) and integration by parts

$$\begin{aligned} W(s,b) &= \pi\left\{\left[(b-c)\,F(c;s,b)\right]_{c=\underline{c}}^{c=b} + \int_{\underline{c}}^{b} F(c;s,b)\, dc\right\} - (1+\alpha)\, e \\ &= \pi\int_{\underline{c}}^{b} F(c;s,b)\, dc - (1+\alpha)\, e, \qquad \forall s \text{ such that } c(s,b) \geq b. \end{aligned}$$

From these it follows directly that the change in welfare gain resulting from a change in s from s' to s'', $s', s'' \in S$, is given by

$$W(s'', b) - W(s', b) = -\pi \int_{b}^{\bar{c}} \left[F(c; s'', b) - F(c; s', b)\right] dc,$$
$$\text{for } s', s'' \text{ such that } c(s', b), c(s'', b) \leq b, b \in \left[\underline{b}, \bar{b}\right]; \quad (12)$$

$$W(s'', b) - W(s', b) = \pi \int_{\underline{c}}^{b} \left[F(c; s'', b) - F(c; s', b)\right] dc,$$
$$\text{for } s', s'' \text{ such that } c(s', b), c(s'', b) \geq b, b \in \left[\underline{b}, \bar{b}\right]. \quad (13)$$

By definition (see, for example, Laffont (1989, p. 32)), $F(c; s'', b)$ stochastically dominates $F(c; s', b)$ in the first-order sense if

$$F(c; s'', b) \leq F(c; s', b), \text{ for all } c \in \left[\underline{c}, \bar{c}\right]. \quad (14)$$

It then follows directly from (12) that $W(s'', b) - W(s', b) > 0$ for $c(s', b), c(s'', b) \leq b$, and from (13) that $W(s'', b) - W(s', b) < 0$ for $c(s', b), c(s'', b) \geq b$. ■

The result is intuitive. First-order stochastic dominance implies a higher expected cost of treatment resulting from referral. It is more worthwhile incurring the higher effort if the expected cost is closer to the benefit because there is then a higher probability that the actual cost is the opposite side of b from the expected cost. That applies whether the expected cost is above or below the benefit. From a statistical point of view, such a conclusion is neither surprising nor especially novel. It does, however, have useful implications for deriving a contract to induce a gatekeeper to incur the additional effort e for patient types for which that effort is efficient.

3.2 Gatekeeper decisions

Proposition 3 *Suppose Assumption 1 holds. Then (1) it is not optimal for a gatekeeper with an incentive contract to incur effort e for s such that $c(s, b) > b$, for any $b \in \left[\underline{b}, \bar{b}\right]$, and (2) no patient type at the referral stage strictly prefers a gatekeeper with an incentive contract to one without.*

Proof. Suppose a gatekeeper with an incentive contract observes s for which $c(s, b) > b$. By Proposition 1, she makes the referral decision unconstrained by the possibility of a second opinion. By not referring the patient, she receives payment P_0. With $P(c) \leq P_0$ for all c, it is not possible to increase her reward by incurring effort e and doing so involves disutility. Result (1) follows. Given result (1), a patient will be referred by neither a gatekeeper with an incentive contract nor one without if the signal s is such that $c(s, b) > b$. But a patient will certainly be referred by a gatekeeper without an incentive contract if s is such that $c(s, b) \leq b$. Thus the patient is never more likely to be referred by a gatekeeper with an incentive contract than by one without and, since $\underline{b} > 0$, never strictly prefers the former, as claimed in result (2). ■

An implication of this proposition is that the second potential benefit of making gatekeepers more cost-conscious discussed in the Introduction is never realised when Assumption 1 applies. A gatekeeper with an incentive contract has no interest in finding that actual cost is less than benefit when expected cost is greater than benefit and referral is thus reasonably refused. Note that this conclusion is independent of the incentive scheme for gatekeepers provided $P(c) \leq P_0$ for all c. It is simply the result of gatekeepers being made more cost-conscious. Note also that Proposition 3 applies to the referral stage. A privately-insured patient responsible for paying his own premium might, at the stage of choosing an insurer, prefer one that uses a gatekeeper with an incentive contract if that results in a lower premium.

Proposition 3 has established that gatekeepers with incentive contracts have no incentive to find out more about the cost resulting from referral for a patient with expected cost greater than benefit. Now consider a patient for whom the expected cost is less than the benefit, one with $s \in S(b)$. If no additional information is acquired, the patient is referred. The gatekeeper's expected reward for such a patient is

$$\int_{\underline{c}}^{\overline{c}} P(c)\, f(c; s, b)\, dc, \text{ for } s \in S(b)\,, b \in \left[\underline{b}, \bar{b}\right]. \tag{15}$$

If the gatekeeper incurs effort e, she discovers the actual cost c with probability π. If $c \leq b$, she then refers the patient and receives payment $P(c)$. If $c > b$, she does not refer the patient and receives payment P_0. With probability $1 - \pi$, she discovers only s, not c, from the additional effort e and refers the patient with the expected reward in (15). In both cases, she incurs disutility of effort e, so her expected utility is

$$\pi \left[\int_{\underline{c}}^{b} P(c)\, f(c; s, b)\, dc + \int_{b}^{\overline{c}} P_0 f(c; s, b)\, dc \right] + (1 - \pi) \int_{\underline{c}}^{\overline{c}} P(c)\, f(c; s, b)\, dc - e,$$
$$\text{for } s \in S(b)\,, b \in \left[\underline{b}, \bar{b}\right]. \tag{16}$$

It is incentive compatible to exert effort e given $s \in S(b)$ if (16) is no less than (15), that is, if

$$\pi \int_{b}^{\overline{c}} [P_0 - P(c)]\, f(c; s, b)\, dc \geq e, \text{ for } s \in S(b)\,, b \in \left[\underline{b}, \bar{b}\right]. \tag{17}$$

The intuition behind this condition is that the incentive to incur effort results from the increase in revenue $P_0 - P(c)$ from not referring patients (those with actual cost c greater than benefit b) who would have been referred had that effort not been incurred.

Proposition 4 *Suppose Assumption 1 holds and a gatekeeper has an incentive contract that induces effort e for patient type $s \in S(b)$ for $b \in \left[\underline{b}, \bar{b}\right]$. Then that patient type strictly prefers a gatekeeper without an incentive contract.*

Proof. The incentive compatibility condition (17) can be satisfied for $s \in S(b)$ only if $f(c; s, b) > 0$ for some $c > b$ such that $P(c) < P_0$. Then for patient type s there is strictly

positive probability that a gatekeeper with an incentive contract will find $c > b$ and, by Proposition 1, not refer the patient because $P(c) < P_0$. But, also by Proposition 1, a patient with $s \in S(b)$ is always referred by a gatekeeper without an incentive contract. ■

If the payment to a gatekeeper could be conditioned on the benefit b, it would be straightforward to induce the gatekeeper to incur effort for all signals $s \in S(b)$ for which it is efficient to do so. To see this, compare the incentive compatibility condition (17) with the condition (10) that specifies the $s \in S(b)$ for which it is efficient to incur effort e. It is apparent that the two conditions become identical if $P_0 - P(c)$ equals $(c - b)/(1 + \alpha)$ for all c. A first step to seeing what can be achieved with b unverifiable is the following result.

Proposition 5 *Suppose Assumption 1 holds, $P(c)$ is differentiable with $P'(c) \leq 0$ for all $c \in [\underline{c}, \bar{c}]$, and, for given $b \in [\underline{b}, \bar{b}]$, the gatekeeper incurs effort e for $s = s' \in S(b)$. Then, for $s'' \in S(b)$ for which $F(c; s'', b)$ stochastically dominates $F(c; s', b)$ in the first-order sense for given b, the gatekeeper also incurs effort e.*

Proof. The change in the left hand side of (17) as s changes from s' to s'' is

$$\begin{aligned}
&\pi \int_b^{\bar{c}} [P_0 - P(c)]\left[f(c; s'', b) - f(c; s', b)\right] dc \\
&\quad = \pi \left\{ \left[[P_0 - P(c)]\left[F(c; s'', b) - F(c; s', b)\right] \right]_{c=b}^{c=\bar{c}} \right. \\
&\quad \left. + \int_b^{\bar{c}} P'(c)\left[F(c; s'', b) - F(c; s', b)\right] dc \right\} \\
&\quad = \pi \left\{ - [P_0 - P(b)]\left[F(b; s'', b) - F(b; s', b)\right] \right. \\
&\quad \left. + \int_b^{\bar{c}} P'(c)\left[F(c; s'', b) - F(c; s', b)\right] dc \right\}, \qquad (18)
\end{aligned}$$

where the first inequality follows from integration by parts and the second from the fact that $F(\bar{c}; s, b) = 1$ for all (s, b). But, for $F(c; s'', b)$ that stochastically dominates $F(c; s', b)$ in the first-order sense, $\left[F(c; s'', b) - F(c; s', b)\right] \leq 0$, see (14). Thus, given $P_0 \geq P(c)$ for all c and $P'(c) \leq 0$, the left-hand side of (17) is no smaller for s'' than for s'. ■

This result is essentially the standard one that an agent prefers a wealth distribution that is stochastically dominant; with $P_0 - P(c)$ non-decreasing in c, the left-hand side of (17) does not decrease with a shift to a distribution of c that is stochastically dominant. It has a useful implication. Suppose every increase in s corresponds to a shift to a stochastically dominant distribution. Then, if for given b the gatekeeper incurs effort e for s', she also incurs effort e for all higher $s \in S(b)$. Of course, an increase in s corresponds to an increase in expected cost, so increasing s sufficiently will result in $c(s, b) > b$ and we know from Proposition 3 that a gatekeeper with an incentive contract can never be induced to incur effort e for s such that

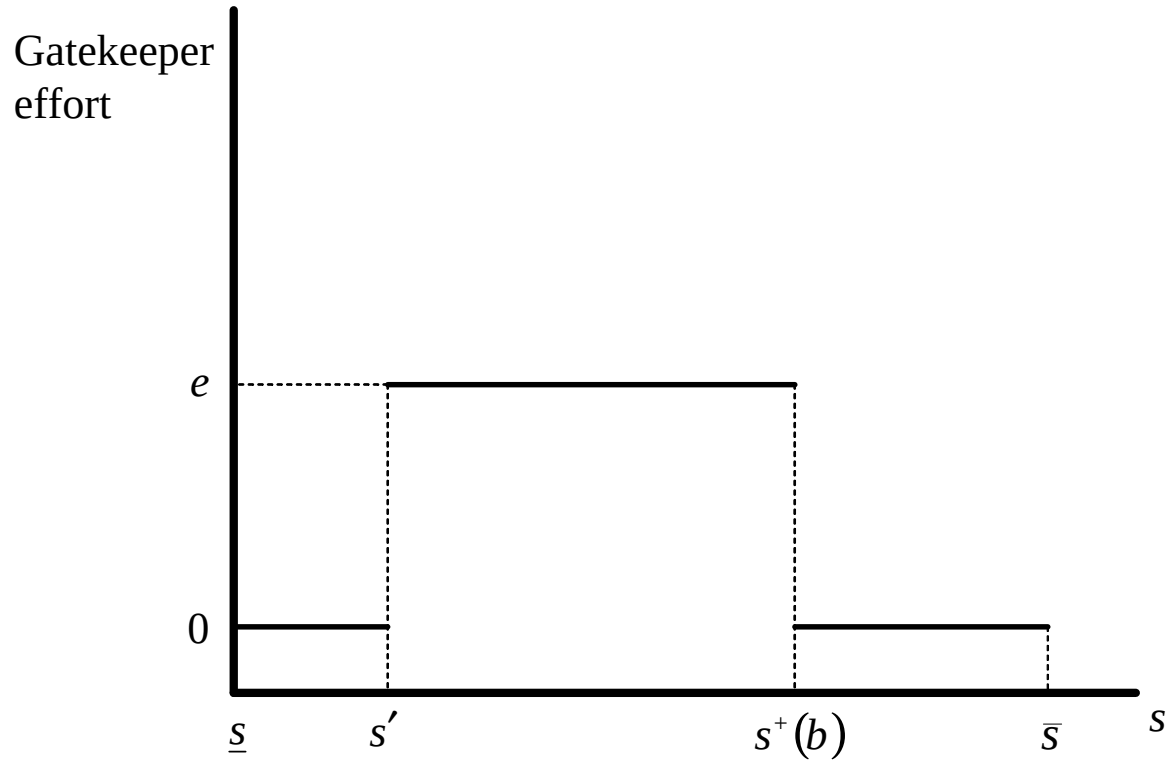

Figure 1: Gatekeeper effort with stochastic dominance

$c(s, b) > b$. But for all s up to $s^+(b)$ defined by $c\left(s^+(b), b\right) = b$, the gatekeeper will continue to incur that effort. This property is illustrated in Figure 1.

That property simplifies the problem of inducing the gatekeeper to exert effort for all $s \in S(b)$ for which it is socially efficient. For increasing s corresponding to a stochastically dominant distribution, define $s^-(b)$ by

$$s^-(b) = \min_{s \in S(b)} s \text{ such that } \frac{\pi}{1+\alpha} \int_b^{\bar{c}} (c - b)\, f(c; s, b)\, dc \geq e, \text{ for all } b \in \left[\underline{b}, \bar{b}\right]. \qquad (19)$$

This is the lowest s for which it is socially efficient to incur effort e, see (10). (The inequality in (19) may be strict for $s^-(b) = \underline{s}$.) Proposition 2 established that, with an increase in s corresponding to stochastic dominance, it is socially efficient to incur e for all $s \in \left[s^-(b), s^+(b)\right]$. But we also know from Proposition 5 that, if the gatekeeper is induced to incur effort for $s = s^-(b)$, then she will incur effort for all $s \in \left[s^-(b), s^+(b)\right]$. Thus, incentive compatibility (17) becomes consistent with social efficiency for all $s \in S(b)$ if the payment scheme ensures

$$\pi \int_b^{\bar{c}} \left[P_0 - P(c)\right] f\left(c; s^-(b), b\right) dc = e, \text{ for all } b \in \left[\underline{b}, \bar{b}\right]. \qquad (20)$$

(In this, "=" can be replaced by "$\geq$" for b such that $s^-(b) = \underline{s}$.) Then the incentive compatibility condition (17) for each (s, b) can be replaced by the single condition (20) for each b and the following result holds.

Proposition 6 *Suppose Assumption 1 holds and, for all $s', s'' \in S(b)$, $s'' > s'$ implies $F\left(c; s'', b\right)$ stochastically dominates $F\left(c; s', b\right)$ in the first-order sense for all $b \in \left[\underline{b}, \bar{b}\right]$. Then with a payment function $P(c) \leq P_0$ with $P'(c) \leq 0$ for all $c \in \left[\underline{c}, \bar{c}\right]$ that satisfies*

$$\pi \int_{\bar{b}}^{\bar{c}} \left[P_0 - P(x)\right] f\left(x; s^-\left(\bar{b}\right), \bar{b}\right) dx \;=\; e, \tag{21}$$

$$P_0 - P(c) = \int_c^{\bar{c}} \left[P_0 - P(x)\right] \frac{\frac{d}{dc} f\left(x; s^-(c), c\right)}{f\left(c; s^-(c), c\right)} dx, \qquad \text{for all } c \in \left[\underline{b}, \bar{b}\right], \tag{22}$$

it is incentive compatible for the gatekeeper to incur effort e for each type $s \in S(b)$ for all $b \in \left[\underline{b}, \bar{b}\right]$ if and only if that effort is socially efficient. For given $P_0 - P(c)$ for $c \in \left[\bar{b}, \bar{c}\right]$ that satisfies (21), there exists a unique function $P_0 - P(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$ that satisfies (22).

Proof. With $P(c) \leq P_0$ and $P'(c) \leq 0$ for all c, Proposition 5 implies that incentive compatibility is consistent with social efficiency if (20) holds. If $P(c)$ satisfies (21), it is immediate that it satisfies (20) for $b = \bar{b}$. It will then also satisfy (20) for all $b < \bar{b}$ if the derivative of the left-hand side of (20) with respect to b is zero for $b \leq \bar{b}$, that is, if

$$-\pi \left[P_0 - P(b)\right] f\left(b; s^-(b), b\right) + \pi \int_b^{\bar{c}} \left[P_0 - P(c)\right] \frac{d}{db} f\left(c; s^-(b), b\right) dc = 0, \qquad \text{for all } b \in \left[\underline{b}, \bar{b}\right], \tag{23}$$

which, with x substituted for c and c for b, can be re-written as (22).

Note that (22) can be written as

$$P_0 - P(c) = \int_c^{\bar{b}} \left[P_0 - P(x)\right] \frac{\frac{d}{dc} f\left(x; s^-(c), c\right)}{f\left(c; s^-(c), c\right)} dx + \int_{\bar{b}}^{\bar{c}} \left[P_0 - P(x)\right] \frac{\frac{d}{dc} f\left(x; s^-(c), c\right)}{f\left(c; s^-(c), c\right)} dx, \qquad \text{for all } c \in \left[\underline{b}, \bar{b}\right].$$

For given $P_0 - P(c)$ for $c \in \left[\bar{b}, \bar{c}\right]$ chosen to satisfy (21), the second integral on the right-hand side of this can be written as a known function $\zeta(c)$ of c independent of the choice of $P_0 - P(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$. Thus finding a function $P_0 - P(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$ that satisfies (22) is equivalent to finding an unknown function $\psi(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$ that satisfies

$$\psi(c) = \int_c^{\bar{b}} \psi(x)\, K(x, c)\, dx + \zeta(c) \tag{24}$$

for given functions $K(x, c)$ and $\zeta(c)$. But (24) is a Volterra integral equation that is known to have a unique solution for $\psi(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$, see Kolmogorov and Fomin (1975, p. 75).[6] ■

If the conditions of Proposition 6 are satisfied for some $P_0 - P(c)$, then P_0 can always be chosen so that the gatekeeper receives no rent from the contract because adding the same

[6] I am indebted to Philippe Choné for this observation.

constant to P_0 and $P(c)$ for all c leaves all the conditions in that proposition still satisfied but changes the payoff of the gatekeeper which can, therefore, be reduced to the reservation level. Then, even if $\alpha > 0$, the outcome for $s \in S(b)$ is efficient. (It is not necessarily efficient for all s because the gatekeeper does not exert effort e for any $s \notin S(b)$ even if that would be socially efficient.) Moreover, for given $P_0 - P(c)$ for $c \in \left[\bar{b}, \bar{c}\right]$ that satisfies (21), one can compute numerically the unique solution for $P_0 - P(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$ that satisfies (22) by the method of successive approximations applied to the Volterra integral equation (24), see Kolmogorov and Fomin (1975, pp. 75-76). However, without further restrictions on $f(c; s, b)$, one cannot guarantee that this solution will satisfy $P(c) \leq P_0$ and $P'(c) \leq 0$ for all c, in which case Proposition 5 does not apply. Of course, for different functions $P_0 - P(c)$ for $c \in \left[\bar{b}, \bar{c}\right]$ that satisfy (21), the unique solutions for $P_0 - P(c)$ for $c \in \left[\underline{b}, \bar{b}\right]$ that satisfy (22) will, in general, be different and some of these may satisfy $P(c) \leq P_0$ and $P'(c) \leq 0$ for all c even if others do not. Also, the requirement that $P'(c) \leq 0$ is a sufficient, not a necessary condition. All that is actually required is that $P'(c)$ is sufficiently small that the expression in (18) evaluated in the proof of Proposition 5 is non-negative. In some special cases, existence of a contract that induces efficient effort for all $s \in S(b)$ and $b \in \left[\underline{b}, \bar{b}\right]$ is easily demonstrated, as the following examples show.

Example 1 : *$s^-(b) = \underline{s}$ for all b. In this case, incentive compatibility is satisfied as long as the left-hand side of (20) is at least as great as the right-hand side. That condition can be satisfied with $P(c) = \bar{P}$ for $P_0 - \bar{P}$ sufficiently large, which certainly satisfies $P'(c) \leq 0$ for all c.*

Example 2 : *$f(c; s, b) = \phi(c - s)$ for all (c, s, b). In this case, knowledge of b provides no additional information about cost given s and s is a signal that, like the mean of a normal distribution, simply shifts the distribution. It is shown in an appendix that solutions to (19) then take the form $s^-(b) = b - k$ with the same constant k for all $b \in \left[\underline{b}, \bar{b}\right]$. It is also shown there that, for $\Phi(.)$ the distribution function associated with the density function $\phi(.)$, (20) is satisfied by $P(c) = \bar{P}$ for all c when $\bar{P}$ satisfies*

$$P_0 - \bar{P} = \frac{e}{\pi} / \left[1 - \Phi(k)\right] > 0, \tag{25}$$

which certainly implies $P(c) \leq P_0$ and $P'(c) \leq 0$ for all c.

A payment rule used in practice is the one specified in (3) making the gatekeeper residual claimant that can be implemented by having the gatekeeper herself meet any costs resulting from referral. With that rule, the incentive compatibility condition (17) can be written

$$\frac{\pi}{1 + \alpha} \int_b^{\bar{c}} c f(c; s, b)\, dc \geq e, \text{ for } s \in S(b)\,. \tag{26}$$

There are two things to note about this condition. The first is that it is independent of P_0, the payment the gatekeeper receives for seeing the patient. The second is given in the following proposition.

Proposition 7 *Suppose Assumption 1 holds and the payment rule to gatekeepers is* $P(c) = P_0 - c/(1+\alpha)$. *Then a gatekeeper incurs effort e for no fewer signals* $s \in S(b)$ *than is efficient.*

Proof. The result follows directly from comparison of (26) with (10). The left-hand side of the former is bigger than that of the latter by

$$\frac{\pi}{1+\alpha}\int_b^{\bar{c}} bf(c;s,b)\,dc \geq 0, \text{ for } s \in S(b)\,. \tag{27}$$

Thus, for any s for which (10) is satisfied, (26) is satisfied. ■

This result is an example of a case in which making an agent residual claimant for cost savings does not result in efficient choice of effort. The intuition behind it is that, in deciding whether to incur effort, the gatekeeper takes into account the reduction in cost from patients not treated as a result of that effort, but not the benefit lost. Thus the incentives to incur the additional effort are "too strong" from a social point of view. Note that the result applies whether or not increases in s correspond to stochastically dominant distributions. The gatekeeper's incentives would, of course, be aligned with social goals if the gatekeeper could be made residual claimant for benefit losses as well as cost savings. To do that would require deducting the left-hand side of (27) from the gatekeeper's cost savings for each (s, b). But that is not feasible with s and b not verifiable.

There are three main conclusions from this discussion for the use of incentive contracts for gatekeepers. First, for patients whose expected cost of treatment resulting from referral is less than the benefit (those with $s \in S(b)$), it is in some cases possible to devise a payment arrangement that induces a gatekeeper to incur effort to find out more about the cost when, and only when, efficient. Second, if instead of an optimal payment arrangement, payment takes the form of having the gatekeeper pay for the cost of any specialist treatment resulting from referral, the gatekeeper will incur that effort for more cost signals $s \in S(b)$ than is efficient because, in deciding for which patients to incur the effort, she takes no account of the benefit of treatment. Third, a gatekeeper with an incentive contract never incurs that effort for patients whose expected cost of treatment resulting from referral is greater than the benefit (those with $s \notin S(b)$), whereas it is in general efficient for it to be incurred for some of those patients. Thus the actual welfare gain from using an incentive contract is less than the potential social gain from incurring the effort to evaluate the cost of treatment resulting from referral because the incentives of gatekeepers are not perfectly aligned with social welfare.

4 Gatekeeper chooses effort before receiving cost signal

This section turns to the case of Assumption 2 in which the gatekeeper must decide whether to incur effort e without first observing the cost signal s and benefit b. In this case, if the additional effort e is incurred, it must be incurred for all (s, b).

Consider first when it is socially efficient for the additional effort e to be incurred. The benefit b is revealed in any case. With probability π, the additional effort reveals the actual cost c that would result from referral and the patient is referred if $c \leq b$ (whatever the signal s). With probability $1-\pi$, it reveals s but not c and the patient is referred if $c(s,b) \leq b$ or, equivalently, $s \in S(b)$. Social welfare with effort e is then

$$\pi \int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{b} (b-c)\, h(c,b)\, dc\, db$$
$$+ (1-\pi) \int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{\bar{c}} (b-c)\, G(S(b)\,;c,b)\, h(c,b)\, dc\, db - (1+\alpha)\left(\underline{U}+e\right). \qquad (28)$$

The social welfare *gain* from incurring effort e, denoted W (without arguments since it is not conditional on (s,b)), is the difference between the expression in (28) and that in (5). That is,

$$W = \pi \int_{\underline{b}}^{\bar{b}} \left[\int_{\underline{c}}^{b} (b-c)\, h(c,b)\, dc - \int_{\underline{c}}^{\bar{c}} (b-c)\, G(S(b)\,;c,b)\, h(c,b)\, dc \right] db - (1+\alpha)\, e. \qquad (29)$$

The first integral term is the welfare from referring a patient who is discovered to have actual cost less than benefit. The second integral term subtracts the welfare for a patient whose actual cost is discovered and who would have been referred anyway because $s \in S(b)$. For $c \leq b$, this welfare is included in the first integral. For $c > b$, the patient is not referred but would have been in the absence of effort e. Both these integral terms are multiplied by the probability π that actual cost is discovered. It is socially efficient to incur effort e if $W \geq 0$.

Now consider the incentive compatibility condition for effort e for a gatekeeper with an incentive contract. If the gatekeeper chooses not to exert that effort, she observes s and refers the patient if $s \in S(b)$ but not otherwise. Expected reward from not exerting effort e is thus

$$\int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{\bar{c}} \{P(c)\, G(S(b)\,;c,b) + P_0\,[1 - G(S(b)\,;c,b)]\}\, h(c,b)\, dc\, db$$
$$= P_0 - \int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{\bar{c}} [P_0 - P(c)]\, G(S(b)\,;c,b)\, h(c,b)\, dc\, db. \qquad (30)$$

If, on the other hand, she chooses to exert effort e, she discovers with probability π the actual cost c and refers the patient if $c \leq b$, whatever s would have been. With probability $1-\pi$, she does not discover the actual cost, only the signal s, refers the patient if $s \in S(b)$ and receives the expected reward in (30). For a patient not referred, she receives payment P_0. Expected

utility from an incentive contract that induces effort e is thus

$$P_0 - \int_{\underline{b}}^{\bar{b}} \left\{ \pi \int_{\underline{c}}^{b} [P_0 - P(c)]\, h(c,b)\, dc + (1-\pi) \int_{\underline{c}}^{\bar{c}} [P_0 - P(c)]\, G(S(b)\,;c,b)\, h(c,b)\, dc \right\} db - e. \tag{31}$$

The incentive compatibility condition for the gatekeeper to put in effort e is that the expression in (31) is greater than that in (30), or

$$\pi \int_{\underline{b}}^{\bar{b}} \left\{ \int_{\underline{c}}^{\bar{c}} [P_0 - P(c)]\, G(S(b)\,;c,b)\, h(c,b)\, dc - \int_{\underline{c}}^{b} [P_0 - P(c)]\, h(c,b)\, dc \right\} db - e \geq 0. \tag{32}$$

The individual rationality condition for a gatekeeper to accept the contract is given by the condition that the expected utility in (31) from exerting effort e is at least as great as $\underline{U}$. Specifically,

$$P_0 - \int_{\underline{b}}^{\bar{b}} \left\{ \pi \int_{\underline{c}}^{b} \int_{\underline{c}}^{\bar{c}} [P_0 - P(c)]\, h(c,b)\, dc + (1-\pi) \int_{\underline{c}}^{\bar{c}} \int_{\underline{c}}^{\bar{c}} [P_0 - P(c)]\, G(S(b)\,;c,b)\, h(c,b)\, dc \right\} db - e \geq \underline{U}. \tag{33}$$

As with Assumption 1, the incentive compatibility condition for incurring effort e differs from the condition for that effort to be efficient. The reason is again that, as is clear from (32), the gatekeeper takes no account of the benefit b, whereas the condition for efficiency that W defined in (29) is non-negative is affected by b. Despite that, effort can still be induced whenever it is socially efficient.

Proposition 8 *Suppose Assumption 2 holds and it is socially efficient that the gatekeeper incurs effort e. Then, the welfare gain from using an incentive contract for the gatekeeper is W defined in (29).*

Proof. Consider the contract with $P(c) = P_0$ for $c \in \left[\underline{c}, \bar{b}\right]$ and $P(c) = \bar{P}$, with $\bar{P} < P_0$, for $c \in (\bar{b}, \bar{c}]$. With $\bar{b} < \bar{c}$, the incentive compatibility condition (32) is then satisfied for $P_0 - \bar{P}$ sufficiently large. Changing P_0 for a given difference $P_0 - \bar{P}$ leaves that condition unaffected, so P_0 can be adjusted until the individual rationality condition (33) holds with equality. Then the gatekeeper exerts effort e but receives no rent. It follows that, even for $\alpha > 0$, the social welfare gain is W defined in (29). ■

By definition, W is the maximum potential social welfare gain under Assumption 2. Thus an immediate implication of Proposition 8 is that, under Assumption 2, a shift from a gatekeeper without an incentive contract to one with an incentive contract achieves the full social welfare gain that can be achieved from the effort level e. Then both of the potential benefits of making the gatekeeper more cost-conscious that were discussed in the Introduction are realised.

Moreover, this can be achieved by a simple contract with just two payment levels, the higher level for patients not referred or referred with cost below a specified threshold, the lower one for patients referred with cost above that threshold. This contrasts with the case of Assumption 1 for which an incentive contract can, at best, induce a gatekeeper to incur efficient effort only for $s \in S(b)$ for each b. The efficiency result under Assumption 2 should not, however, be overemphasised. It would no longer hold in the more realistic case of continuous effort and π an increasing function of effort. Then, because the gatekeeper ignores the benefit of treatment, the level of effort that maximises the gatekeeper's expected utility is different from that which maximises the social welfare gain.

The full welfare gain is not, however, necessarily achieved by the payment function $P(c) = P_0 - c/(1+\alpha)$ given in (3) by which the gatekeeper is residual claimant for cost savings. In that case, the incentive compatibility condition (32) takes the form

$$\frac{\pi}{1+\alpha}\int_{\underline{b}}^{\bar{b}}\left[\int_{\underline{c}}^{\bar{c}} cG(S(b)\,;c,b)\,h(c,b)\,dc - \int_{\underline{c}}^{b} ch(c,b)\,dc\right]db - e \geq 0. \tag{34}$$

Unlike with Assumption 1, it is not necessarily the case that the private gain to the gatekeeper from additional effort, given by the left-hand side of (34), is greater than the social gain, given by (29). The reason is that, although the gatekeeper does not take account of benefit lost by a patient not referred as the result of additional effort, she also does not take account of the off-setting benefit gained by a patient who would not have been referred without the additional effort. Either can dominate. To see this, recall that it is socially efficient to incur effort e if and only if W defined in (29) is non-negative or, equivalently since $\alpha \geq 0$, if and only if $W/(1+\alpha) \geq 0$. The difference between the left-hand side of (34) and $W/(1+\alpha)$ is

$$\frac{\pi}{1+\alpha}\int_{\underline{b}}^{\bar{b}}\left[\int_{\underline{c}}^{b} bh(c,b)\,dc - \int_{\underline{c}}^{\bar{c}} bG(S(b)\,;c,b)\,h(c,b)\,dc\right]db. \tag{35}$$

Suppose the benefit is the same for all types, so $\underline{b} = \bar{b}$. Then it is clear that the sign of this expression depends on the precise characteristics of the distributions $h(.)$ and $G(.)$.

Even though use of incentives can increase social welfare, patients will not necessarily choose to visit a gatekeeper with an incentive contract given the choice. With $\underline{b} > 0$, the patient wants to maximise the probability of being referred and will prefer a gatekeeper with an incentive contract only if that probability is higher. A gatekeeper without an incentive contract refers the patient if the expected cost of the resulting treatment given the cost signal is less than the benefit, that is, if $s \in S(b)$. For a patient who does not know his own type, the probability assessment of being referred by a gatekeeper without an incentive contract is thus

$$\int_{\underline{b}}^{\bar{b}}\int_{\underline{c}}^{\bar{c}} G(S(b)\,;c,b)\,h(c,b)\,dc\,db. \tag{36}$$

A gatekeeper with an incentive contract who incurs effort e refers the patient if she learns the

actual cost c and $c \leq b$, or if she learns only the signal s and $s \in S(b)$. For a patient who does not know his own type, the probability assessment of being referred by a gatekeeper with an incentive contract is thus

$$\pi \int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{b} h(c, b)\, dc\, db + (1 - \pi) \int_{\underline{b}}^{\bar{b}} \int_{\underline{c}}^{\bar{c}} G(S(b)\,;\, c, b)\, h(c, b)\, dc\, db. \tag{37}$$

The increase in the probability of being referred from shifting to a gatekeeper with an incentive contract is, therefore,

$$\pi \int_{\underline{b}}^{\bar{b}} \left[\int_{\underline{c}}^{b} h(c, b)\, dc - \int_{\underline{c}}^{\bar{c}} G(S(b)\,;\, c, b)\, h(c, b)\, dc \right] db. \tag{38}$$

This may be positive or negative. Whichever it is, there is no reason for its sign to be the same as that of the expression on the left-hand side of (29) that measures the social welfare gain.

This analysis, however, presumes that patients do not have information about the expected cost and benefit resulting from referral before they choose whether to visit a gatekeeper with an incentive contract. When they do, the following result applies.

Proposition 9 *Suppose Assumption 2 holds and all patients observe their own* (s, b) *before choosing whether to visit a gatekeeper with an incentive contract. Then a gatekeeper with an incentive contract does not incur the additional effort* e.

Proof. Suppose, contrary to the claim, that it is incentive compatible for a gatekeeper with an incentive contract to choose effort e. Patients who observe their own (s, b) before choosing which gatekeeper to visit choose a gatekeeper without an incentive contract if $c(s, b) \leq b$ and there exists some $c > b$ for which $P(c) < P_0$ and $f(c; s, b) > 0$ because that gatekeeper will certainly refer them, whereas there is strictly positive probability that a gatekeeper exerting effort e will not. Those with $c(s, b) > b$ choose a gatekeeper with an incentive contract because one without will certainly not refer them, whereas there is strictly positive probability that a gatekeeper incurring effort e will discover they have actual cost $c \leq b$ and so refer them. But then a patient choosing a gatekeeper with an incentive contract has $s \in S(b)$ only if $f(c; s, b) = 0$ for all $c > b$ for which $P(c) < P_0$. That corresponds to the first integral in the square brackets in the gatekeeper's incentive compatibility condition (32) being zero, so that condition cannot be satisfied. ■

The essential point here is that, if a gatekeeper with an incentive contract knows that all the patients who attend do so because they know they will not be referred by a gatekeeper without an incentive contract, then she has no incentive to put in effort to find out the cost resulting from referral. She knows that, without effort e, the evidence she will acquire will justify not referring the patient, so there can be no gain from the additional effort.

Assuming that patients know in advance exactly what a gatekeeper without an incentive contract will discover is clearly an extreme case. But it is certainly not unreasonable to suppose

that patients who can, as here, gain by finding out information about their type before choosing a gatekeeper will go to some trouble to do so. And if those who are likely to benefit from choosing a gatekeeper without an incentive contract actually do that, they reduce the incentive for gatekeepers with incentive contracts to incur effort to find out about the actual costs resulting from referral. The continued viability of incentive contracts thus depends on patient ignorance under circumstances in which patients have good reason not to remain ignorant.

5 Evidence and implications

A valuable source of empirical evidence on the effects of incentive contracts for gatekeepers on referrals comes from the experiment with GP fundholding in the publicly-funded British NHS. Non-fundholding GPs are paid a capitation fee for each patient on their list.[7] Patients they refer have their treatment costs met by a local health authority. These GPs are thus gatekeepers without an incentive contract. The model used here predicts that, in both scenarios, such gatekeepers refer a patient only if $b \geq c(s, b)$ whereas, without a gatekeeper, any patient with $b > 0$ would see a specialist. That is consistent with the widely-held view that the traditional GP arrangement reduces the usage of such services. See Gerdtham and Jönsson (2000, p. 46) for a summary of cross-country empirical results on the effect of gatekeepers.

Fundholding GPs received, in addition to the capitation fee, an allowance from which they were to meet the treatment costs of patients they referred for a range of (non-emergency) services up to a maximum for each patient in each year, with any excess being met by the health authority. For services in that range, they were paying gatekeepers with net reward given by the expression in (3) for c up to the specified maximum. The power of that reward system was, however, restricted because fundholders were not permitted to take any excess of their allowance over costs as additional personal income though, since they were allowed to use it for extra staff and for improvements to premises, they could personally gain indirectly from it. That reduced the effective reward from reducing referral costs and was thus like increasing the slope of the reward so that $P'(c) > -1/(1+\alpha)$.

Some researchers have questioned whether referral rates for fundholders were different from those for non-fundholders, see Coulter and Bradlow (1993). That is consistent with the scenario of Assumption 2 because the difference in the probability of referral given in (38) may be either positive or negative. It would be consistent with the scenario of Assumption 1 only if the signal s revealed an upper bound on the cost resulting from referral so that a gatekeeper with an incentive contract refers all patient who would be referred by one without. An example would be when the gatekeeper can easily establish the standard cost at the local hospital but may be able to achieve a lower cost by searching or bargaining. The formal model can be adapted to this example by defining $c(s, b)$ as the known cost in the absence of further investigation, so $f(c; s, b) = 0$ for $c > c(s, b)$. However, the weight of the empirical evidence surveyed by

[7] Gravelle (1999) studies some implications of capitation contracts but not for referrals or fundholding.

O'Donnell (2000) is that referral rates were lower for fundholders than for non-fundholders, as implied by the scenario of Assumption 1.

Patients could choose between a fundholding and a non-fundholding GP, though the former were not available in all geographical areas. In contrast to the results in Propositions 3 and 4 for the case of Assumption 1, many patients chose fundholder GPs, though it may have been that they simply continued to use the same GP when that GP became a fundholder. This would suggest that the scenario of Assumption 2 may have been more appropriate. There are, however, reasons to think that such a conclusion is premature because there were other aspects of the British system that made it attractive to be on the list of a fundholding GP but that were not inherent to gatekeeping with incentive contracts. First, the contracts used by health authorities to pay providers of specialist services for treating patients of non-fundholders did not, in many cases, involve any extra direct payment for treating an additional patient. In contrast, fundholders typically paid directly for each patient treated. Thus, not surprisingly, providers gave priority to the patients of fundholders who therefore had shorter waiting times for many treatments, see Propper, Croxson and Shearer (2002). Formally, that corresponds to the benefit b being higher for the patients of fundholding GPs. But this was the result of inappropriate contractual arrangements between health authorities and providers of specialist services for patients of non-fundholders, not something inherent to fundholding, so the comparison does not correspond to that made here. Second, fundholders had an incentive to treat patients at their own practice if that was cheaper than referring them to a specialist, whereas non-fundholders did not. Many fundholders did in fact provide more treatments themselves, which may have been more convenient for patients. That is a potential welfare gain not accounted for in the model but again not necessarily inherent to fundholding. When it comes to other aspects of services, a survey carried out by the Consumers Association (1995, p. 16), concluded: "Patients of fundholders are less satisfied with aspects of their GP service than patients of non-fundholders." That is consistent with the results for Assumption 1.

A number of practical considerations arise from the analysis. Proposition 6 specifies conditions for a payment rule that, under Assumption 1, ensure efficient effort for all patients for whom the initial cost signal indicates an expected cost less than the benefit. There are two complications with using such a payment rule in practice. First, the precise rule depends crucially on the distribution $f(c; s, b)$. This means that it will typically have to be specific to each medical condition. That could be an administrative nightmare. Second, an optimal rule depends on monitoring the actual cost of treatment which is, at least in some cases, unlikely to be costless as assumed in the model. The payment rule specified in (3) making the gatekeeper residual claimant (a paying gatekeeper) is both independent of the medical condition and avoids the need for monitoring cost because it can be implemented simply by having the gatekeeper meet any costs of treatment herself. As noted above, that rule corresponds to the reward to an HMO that both provides full insurance (without co-payments) and also acts as the gatekeeper because the HMO then receives the insurance premium (which corresponds to P_0) and must pay the costs of any treatment that is authorised. It also corresponds to the scheme used for

lower levels of expenditure on referrals by GP fundholders in the British NHS.

For that rule, Proposition 7 showed that, under Assumption 1, a gatekeeper incurs effort e for more signals $s \in S(b)$ than would be socially efficient. Moreover, when the effort is incurred, some patients are not referred because the cost is discovered to exceed the benefit. Thus, the gatekeeper refers fewer types s than is socially efficient. In this situation, restrictions such as those on how fundholders in the British NHS could spend cost savings may prove useful. Such restrictions reduce the value of cost savings below that of unrestricted cash, so they have the effect of multiplying the left-hand side of Eq. (17) by a factor less than one and thus make the incentives of the fundholder to incur effort closer to what is socially efficient.

This result also has implications for the structure of private health insurance. Suppose insurance is provided by an HMO that also acts as gatekeeper. If patients are required to make co-payments that result in the net reward to the HMO satisfying the conditions of Proposition 6, gatekeeping effort will be efficient for all signals s for which $b \geq c(s, b)$. But co-payments are a second-best solution that detract from efficient insurance. However, the same effect can be achieved without co-payments by separating the insurance function from the gatekeeping function, with the insurance premium paid to a separate insurance company that subcontracts the gatekeeping function to a third party using an optimal payment rule.

With Assumption 1, as Proposition 3 showed, a gatekeeper with an incentive contract incurs additional effort only for a patient who would be treated by a gatekeeper without an incentive contract, so a patient's probability of being treated is never higher from going to a gatekeeper with an incentive contract. Thus no patient prefers to use a gatekeeper with an incentive contract to one without. By Proposition 4, any for whom there are social gains to using a gatekeeper with an incentive contract actively prefer one without. So patients who do not know their own type always prefer a gatekeeper without an incentive contract. Giving them the choice between a gatekeeper with an incentive contract and one without will thus destroy the potential advantages of using incentive contracts. With the scenario of Assumption 2, that is no longer the case. As Proposition 9 showed, however, it remains the case that, if patients learn sufficient about their types for those likely to benefit from choosing a gatekeeper without an incentive contract to do so, the effectiveness of offering a choice between gatekeepers with different contracts breaks down. Thus, in neither case is it likely to be satisfactory to introduce gatekeepers with incentive contracts alongside those without and let patients choose between them, as in the fundholding arrangements in the British NHS. Gatekeeping with incentive contracts under these circumstances is an "all or nothing" system.

6 Concluding remarks

This paper has analysed some implications of gatekeeper arrangements for controlling access to specialist medical services. It has explored two scenarios. In both scenarios, a gatekeeper without an incentive contract receives a signal of the cost and benefit of referring a patient and uses this information to decide whether to do so. Also in both scenarios, it may be worthwhile

for a gatekeeper with an incentive contract to find out more about the cost before deciding whether to refer the patient. In the first scenario, the gatekeeper decides on the level of investigation after observing the initial signal of cost and benefit. In the second, that decision must be made before receiving any signal of cost or benefit.

Using incentive contracts for gatekeepers may or may not be socially worthwhile. Whether it is depends on the distribution of treatment costs and on the disutility associated with finding out further information about those costs. But two general messages come across clearly. First, in the first scenario, the arrangement by which the gatekeeper receives a fixed fee and must pay for the cost of treatment herself, as with HMOs that act as both full insurer and gatekeeper and with fundholding GPs in Britain, results in too strong incentives when expected cost based on the initial signal is less than the benefit. As a result, fewer patient types are referred than would be the case with efficient incentives. This contrasts with other agency problems in which making the agent residual claimant for costs ensures efficient decisions. This result has implications for the contractual arrangements for gatekeepers. It also has implications for the organisation of private health insurance. More efficient than having an HMO itself act as both gatekeeper and insurer is for insurance to be with a third party insurer who employs the HMO on a contractual basis that does not make the HMO residual claimant.

Second, it is not sensible to introduce incentive contracts for just some gatekeepers when patients have a choice between types of gatekeeper, as in the fundholding system in the British NHS. In the first scenario studied here, patients for whom there are potential social benefits from using a gatekeeper with an incentive contract always prefer one without. In the second, there are forces at work that should, with time, reduce the effectiveness of gatekeeper incentives. This implication of using incentive contracts may have been masked in the British NHS because there were other factors present not inherent to fundholding that made attending a fundholder attractive. But it is one that, in general, it would be unwise to ignore.

Appendix

Derivations for Example 2. For $f(c;s,b) = \phi(c-s)$ for all (c,s,b) and $s^-(b)$ strictly interior to $S(b)$ so that the weak inequality in (19) is always an equality, a solution $s^-(b)$ to (19) must satisfy

$$\int_b^{\bar{c}} (c-b)\,\phi\big(c-s^-(b)\big)\,dc = \frac{1+\alpha}{\pi}e, \text{ for all } b \in \big[\underline{b},\bar{b}\big]. \tag{A.1}$$

The left-hand side of this can be written

$$\int_b^{\bar{c}} \Big[\big(c-s^-(b)\big)\,\phi\big(c-s^-(b)\big) - \big(b-s^-(b)\big)\,\phi\big(c-s^-(b)\big)\Big]\,dc$$
$$= \Psi\big(b-s^-(b)\big) - \big(b-s^-(b)\big)\Big[1-\Phi\big(b-s^-(b)\big)\Big], \tag{A.2}$$

for some function $\Psi(.)$ and $\Phi(.)$ the distribution function associated with the density function $\phi(.)$. It is clear from (A.2) that solutions $s^-(b)$ to (A.1) for different values of b all take the form $b - s^-(b)$ equal to a constant. There may be multiple solutions of this form but, since $s^-(b)$ is defined as the lowest s that satisfies (A.1), the relevant solution for all b is always that for which $b - s^-(b)$ takes the largest value. Let k denote the corresponding value of $b - s^-(b)$ and note that $k < \bar{c}$ because otherwise the expression in (A.2) would be zero. Now suppose $P(c) = \bar{P}$, for all c. Then the left-hand side of (20), the condition that ensures incentive compatibility is consistent with social efficiency for all $s \in S(b)$, can be written

$$\begin{aligned}
\pi \int_b^{\bar{c}} \left(P_0 - \bar{P}\right) \phi\big(c - s^-(b)\big)\, dc &= \pi \left(P_0 - \bar{P}\right) \int_b^{\bar{c}} \phi\big(c - s^-(b)\big)\, dc \\
&= \pi \left(P_0 - \bar{P}\right) \left[1 - \Phi\big(b - s^-(b)\big)\right] \\
&= \pi \left(P_0 - \bar{P}\right) \left[1 - \Phi(k)\right].
\end{aligned}$$

Thus (20) is satisfied for all $b \in \left[\underline{b}, \bar{b}\right]$ by a value of $\bar{P}$ that satisfies (25).

References

Chalkley, M. and Malcomson, J. M. (2000), Government purchasing of health services, *in* A. J. Culyer and J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1A, Elsevier Science, Amsterdam, chapter 15, pp. 847–890.

Consumers Association (1995), 'Is your doctor a fundholder?', *Which?* pp. 16–19.

Coulter, A. and Bradlow, J. (1993), 'Effect of NHS reforms on general practitioners' referral patterns', *British Medical Journal* **306**, 433–437.

Cutler, D. M. and Zeckhauser, R. J. (2000), The anatomy of health insurance, *in* A. J. Culyer and J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1A, Elsevier Science, Amsterdam, chapter 11, pp. 564–643.

Ellis, R. P. and McGuire, T. G. (1986), 'Provider behavior under prospective reimbursement: Cost sharing and supply', *Journal of Health Economics* **5**, 129–151.

Ellis, R. P. and McGuire, T. G. (1990), 'Optimal payment systems for health services', *Journal of Health Economics* **9**(4), 375–396.

Gerdtham, U.-G. and Jönsson, B. (2000), International comparisons of health expenditure: Theory, data and econometric analysis, *in* A. J. Culyer and J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1A, Elsevier Science, Amsterdam, chapter 1, pp. 11–53.

Glennerster, H., Matsaganis, M., Owens, P. and Hancock, S. (1993), GP fundholding: Wild card or winning hand?, *in* R. Robinson and J. Le Grand, eds, 'Evaluating the NHS Reforms', King's Fund Institute, London, chapter 4, pp. 74–107.

Glied, S. (2000), Managed care, *in* A. J. Culyer and J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1A, Elsevier Science, Amsterdam, chapter 13, pp. 707–753.

Gravelle, H. (1999), 'Capitation contracts: Access and quality', *Journal of Health Economics* **18**(3), 315–340.

Kolmogorov, A. N. and Fomin, S. V. (1975), *Introductory Real Analysis*, Dover Publications, New York.

Laffont, J.-J. (1989), *The Economics of Uncertainty and Information*, MIT Press, Cambridge, MA.

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

Ma, C.-t. A. and Riordan, M. H. (2002), 'Health insurance, moral hazard, and managed care', *Journal of Economics and Management Strategy* **11**(1), 81–107.

Matsaganis, M. and Glennerster, H. (1994), 'The threat of 'cream skimming' in the post-reform NHS', *Journal of Health Economics* **13**(1), 31–60.

Newhouse, J. P. (1970), 'Toward a theory of nonprofit institutions: An economic model of a hospital', *American Economic Review* **60**(1), 64–74.

Newhouse, J. P. (1989), 'Do unprofitable patients face access problems?', *Health Care Financing Review* **11**(2), 33–42.

O'Donnell, C. A. (2000), 'Variation in GP referral rates: What can we learn from the literature?', *Family Practice* **17**(6), 462–471.

Propper, C., Croxson, B. and Shearer, A. (2002), 'Waiting times for hospital admissions: The impact of GP fundholding', *Journal of Health Economics* **21**, 227–252.

Scott, A. (2000), Economics of general practice, *in* A. J. Culyer and J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1B, Elsevier Science, Amsterdam, chapter 22, pp. 1175–1200.