GPs' payment contracts and their referral practice

Begoña Garcia Mariñoso, Izabela Jelovac

GPs' Payment Contracts and their Referral Practice[*]

Begoña Garcia Mariñoso[†] and Izabela Jelovac[††]

December 2002

Abstract

This paper compares the role of general practitioners in determining access to specialists in two types of health care systems: gate-keeping systems, where a GP referral is compulsory to visit a specialist, and non gate-keeping systems, where this referral is optional. We model the dependence between the GP's diagnosis effort and her referral behaviour, and identify the optimal contracts that induce the best behaviour from a public insurer's point of view, where there is asymmetry of information between the insurer and the GP regarding diagnosis effort and referral decisions. We show that gate keeping is superior wherever GP's incentives matter.

Keywords: health economics, referral, contracts and moral hazard.

JEL: D82, I18 and L51

---

[*] We would like to thank the following for their helpful comments: Matilde Machado, Inés Macho-Stadler, Maurice Marchand, Xavier Martínez-Giralt, Alistair Munro, Pau Olivella, David Pérez-Castrillo, Pedro Pita Barros, Diego Rodriguez-Palenzuela, Nancy Devlin, seminar participants at University College Dublin, University of East Anglia, Katholieke Universiteit Leuven, University of Santiago de Compostela, WZB at Berlin, University of Maastricht, University of Kent at Canterbury, participants at Jornadas de Economia Industrial 1999 and EARIE 1999, and two anonymous referees. Financial support from the DGES grant PB97-0181 from the Spanish Ministry of Education is gratefully acknowledged. The usual disclaimers apply.

[†] Department of Economics. City University London. Northampton Square. London EC1V 0HB. E-mail: b.garcia-marinoso@city.ac.uk

[††]Faculty of Health Sciences (HOPE/BEOZ). University of Maastricht. The Netherlands. E-mail: I.Jelovac@beoz.unimaas.nl

1. Introduction

This paper provides a theoretical framework to analyse the role of a general practitioner (GP) in determining access to specialists in contrasting health care systems. We compare systems where patients can only visit specialists after a general practitioner referral (gate-keeping systems) with systems where patients have free access to specialists (non gate-keeping systems). For example, the British National Health Service and the Spanish Servicio Nacional de Salud are gate-keeping systems, whereas the German and Belgium health care systems are not. Countries also differ in the methods used to remunerate GPs.[1] Given these differences, a theoretical framework enabling clear comparisons of the different health systems should constitute a valuable tool for health policy makers.

Within each of these institutional regimes we consider the provision of incentives to attain a cost-containment/quality-achievement objective, which is widely recognised as one of the main objectives for any health system. As GPs' referrals are an important determinant of health care costs we focus our attention on the role of the GP remuneration in providing incentives for the GPs to act as a filter for specialists.[2] Indeed, there is empirical evidence that financial incentives affect the referrals of GPs; e.g. see Hellinger (1996) on the referrals of physicians working in managed care plans in the US, and Croxson et al. (2001) on the referrals of GPs in the British fund-holding scheme.

---

[1] In Belgium and Germany GPs are paid according to a fee for service scheme. In Spain, they are paid according to a capitation contract or they receive a fixed salary. In the UK the former "GP fundholders" received a budget to pay for their patients' drugs and hospital services and could reinvest the "unspent" share of the budget in their own practice. Since 1997 until 2002, budgets were held by Primary Care Groups who commissioned services from hospitals on behalf of larger groups of the population. As a result, incentives might have been somewhat diluted.

[2] Competition for patients and co-insurance might also discipline GPs. We present a worst-case scenario, as we assume that there is no competition for patients (closed lists or high switching costs) and that patients are fully insured for medical expenses.

It is the special feature of the health services as expert services that allows a GP to adjust her referral strategy to the incentives she faces. In particular, it is hard to verify whether a referral decision by a GP is adequate or not. Therefore, when designing the payment scheme for a GP, it is important to consider its influence on the GP's referral practice. However, some perverse effects arise if the GP's contract only aims to induce the desired referral practice, since referring patients is not the only activity of a GP. In particular, prior to any referral or any other treatment decision, the GP should diagnose the patient. Moral hazard arises because the GP's diagnosis effort is hardly observable by outside parties while it is costly to the GP. The GP's effort in diagnosing also depends on her payment contract. We formalise these ideas in a model that recognises the dual role of the GP in performing a diagnosis and making referral decisions.

Since the pioneering contributions of Blomqvist (1991) and Ellis and McGuire (1990; 1993), the literature on the optimal design of payment systems for health care providers in the presence of agency problems has thrived. The contributions to this literature can be classified according to the form of agency problems considered. In his survey on physician agency, McGuire (2000) distinguishes between moral hazard with hidden information and moral hazard with hidden action. The former refers to the health care provider's private information concerning the outcome of a diagnosis, the benefits of a given treatment, or the convenience of referring the patient to a specialist. The latter refers to actions undertaken by a health care provider such as quality setting, cost-reducing effort, or actions necessary to identify the patient's condition.

Ma (1994), Chalkley and Malcomson (1998a) and Chalkley and Malcomson (1998b) consider a situation of double moral hazard with hidden actions. In their

contributions the cost-reduction effort and the quality-enhancement effort of a health care provider are non-contractible. In Ma (1994) and in Chalkley and Malcomson (1998a) patients' demand reflects the quality of care, while in Chalkley and Malcomson (1998b) it does not.[3] Ma (1994) derives some conditions under which supply-side cost sharing is optimal. Chalkley and Malcomson (1998a) show that if it is inefficient to treat all patients who require it, a pure prospective contract induces too low quality or excessive treatment. Finally, Chalkley and Malcomson (1998b) show that no contract can induce a self-interested health care provider to both contain costs and improve quality when patient's demand does not reflect quality.

The main difference between this paper and those referred to in the proceeding paragraph is that here the identification of the optimal payment contract is used to compare gate-keeping and non gate-keeping systems. Yet, our paper is similar to those in that it also finds an optimal contract in a double moral hazard situation. The difference is that in our case we deal with moral hazard with hidden action (the GP's diagnosis effort) and moral hazard with hidden information (the outcome of the diagnosis). These informational asymmetries are relevant for primary care, while quality-enhancement and cost-reducing efforts are more relevant for hospital provision. In this sense, the closest paper is Jelovac (2001), which considers a situation of moral hazard with hidden action (diagnosis effort of a physician) together with moral hazard with hidden information (diagnosis outcome). The optimal physician's payment includes a cost-sharing component. It yields the right diagnosis and recommendation incentives as the physician wishes to minimise the likelihood of a patient's second visit and its corresponding costs. This contract does not discipline a GP if there is provider specialisation (primary and specialist care) as the GP can refer

[3] The main difference between Ma (1994) and Chalkley and Malcomson (1998a) is that the former assumes that it is efficient to treat all patients who want treatment, while the latter does not.

the patient to the specialist without incurring any costs. This is the reason why the optimal GP's payment scheme in the presence of specialist care requires a more sophisticated combination of cost-sharing components: cost sharing of the GP's treatments and a bonus for not referring.[4]

In contrast to other models (e.g. Dionne and Contandropoulos, 1985), we do not recognise how a physician's professional ethics mitigate the inappropriate use of her discretion. We provide a benchmark for the worst-case scenario in which the GP does not take into consideration the patient's welfare when she makes decisions. To our knowledge, the extent to which ethical concerns influence physicians' behaviour has not been measured in empirical work. Still, it is worth acknowledging the empirical analysis of the effects of social norms (Encinosa et al, 1997) and non-pecuniary job characteristics (Scott, 2001) on physicians' behaviour.

We also neglect the role of litigation for clinical negligence as a means of disciplining the provider. The limited evidence suggests that patients and surrounding health professionals very rarely report malpractice.[5] Moreover, the fact that financial responsibility due to malpractice is normally shifted from individual providers to larger collectives means that the power of such incentives is diluted.

This paper is organised as follows. The model is presented in Section 2. Section 3 and Section 4 present the GP's problem and the insurer's problem, respectively. Section 5 compares a gate-keeping system with a non gate-keeping

[4] Hereafter, we define "specialist care" as services offered only by a specialist, as distinct from care offered by a GP.

[5] Reporting on previous studies, Towse and Danzon (1999) mention that in the UK only 7% of medical negligence result in a claim, and only 2% lead to payment (in the US the figures are 10% and 5%). In 1998, the costs of claims represented only 0.25% of the annual expenditure of the NHS (see Fenn et al., 2000).

system. The conclusions of the paper are included in Section 6 and proofs are provided in the Appendix.

2. The Model

We model the relationship between a public insurer, a GP, a specialist and a patient as a game with the following timing.[6] In stage 1, the public insurer sets the GP's payment contract. In stage 2, the GP either accepts or rejects the contract (in the latter case, the game ends). In stage 3, "Nature" determines the type of illness of the patient who then accordingly visits the relevant provider.[7] If the patient visits the specialist, the game ends. Otherwise, in stage 4, the GP exerts some level of effort in diagnosis. In stage 5, Nature draws the GP's signal about the patient's illness type. In stage 6, the GP decides either to provide primary care or to refer the patient to the specialist. If the patient returns to health, the game ends. Otherwise, the patient is referred to the specialist in stage 7.

Let us now detail the utility functions, decision variables and parameters that are relevant in the model, following the sequence of the game. Given the uncertainty that is inherent to the diagnosis outcome, the utility function of both the insurer and the GP must be described in expected terms. In turn, given the influence of the GP's effort on the diagnosis outcome, the expected utility of both the insurer and the GP depend on the level of effort exerted by the GP in stage 4.[8]

---

[6] For ease of exposition we refer to the GP as "she" and to the patient and insurer as "he". This convention is generally used in Principal-Agent models, where "she" relates to the Agent and "he" to the Principal.

[7] We do not address the issue of how to discipline patients who might strategically choose to visit a specialist or a GP. Garcia Mariñoso (1998) describes how the insurer can regulate access to expensive care by manipulating the patients' insurance contract and altering the types of patients accessing each provider.

[8] Section 3 analyses the GP's behaviour and provides a more detailed explanation of how a GP's effort affects the expected utilities.

In stage 1, the public insurer designs the GP's payment contract so as to maximise some measure of the expected patient's health, net of the expected insurer's costs. We make the assumption that it always pays for the public insurer to cure the patient.[9] In consequence, the health measure only reflects the insurer's valuation of some expected health loss, denoted by l, suffered by the patient when the first treatment he receives does not cure his illness. We interpret this loss as a waiting cost. It includes any disturbances caused by one unnecessary treatment and/or any health losses related to delays in receiving the adequate treatment. Therefore, the public insurer's objective is to minimise the sum of the patient's expected health loss and his own expected costs.

The insurer's costs first consist of the GP's payment. The payment contract, denoted by (D, B, S), is based on observable elements and consists of three non-negative components. First, the GP receives a payment, D, for performing a diagnosis. Second, she receives a bonus, B, when she does not refer the patient to the specialist. Third, the GP receives another bonus, S, when she does not provide primary care to the patient.[10] Furthermore, the public insurer fully reimburses the specialist for its costs h.[11] This latter costs includes the so-called hotel costs of a hospital and the costs of the diagnostic tests, which are generally run by a specialist for whatever type of illness.[12] The costs of treatment by a specialist are generally higher than the costs of

[9] By cure we mean that the condition of the patient improves to the point state-of-the-art medicine allows. This might not necessarily mean that the illness is effectively cured.

[10] Note that D can be interpreted as a capitation payment and S as the savings the GP makes on any cost-sharing scheme.

[11] Since we do not consider any agency problem on the specialist's side, there is no need for a more sophisticated payment scheme.

[12] The idea here is that the fixed cost component of hospital costs is large enough so that per patient costs do not depend on the type of illness.

treatment by a general practitioner, which we normalise to zero. To summarise stage 1, the insurer's problem is the following:

$$\underset{\{D,B,S\}}{Min} \quad \{EL(e) + EC(e,D,B,S)\},$$

where the functions *EL*( ), *EC*( ), and the variable *e* stand, respectively, for the patient's expected health loss, the insurer's expected costs, and the effort of the GP in stage 4.

In stage 2, the GP decides either to accept or to reject the contract, comparing her expected utility from accepting the contract with the utility of receiving some reservation wage, denoted *w*.[13] The GP's expected utility depends on her payment contract (*D, B, S*) and on the diagnosis effort she exerts. The dependence on effort is twofold: first, the GP bears some disutility of effort, denoted by $v(e) = (e - 1/2)^2/2$; second, the GP's effort influences her expected payment, denoted by *EI*. The GP is also assumed to be risk-neutral with respect to money. Her expected utility can thus be written as:

$$EU = EI(e,D,B,S) - (e - 1/2)^2/2 .$$

We say that the GP has negotiation power when she can require the contract to give her at least the same utility as from *w*, not only in expected terms but also in any contingency.

In stage 3, "Nature" determines the type of illness of the patient. The patient suffers from either of two types of condition *m*: common ($m = \underline{m}$), and special ($m = \overline{m}$). While all types of providers can cure the common condition, the special illnesses can only be cured with probability one by the specialist. In order to keep the model

[13] This reservation wage can be viewed as the average wage of a GP before a hypothetical reform takes place, the reform consisting in offering to the GP the possibility to choose between the former reservation wage, *w*, and the alternative contract (*D, B, S*).

realistic, we assume that the patient suffers from a common illness with probability $p = \Pr(\underline{m}) \geq 2/3$ and that this is common knowledge.[14]

The patient now visits either the GP or the specialist. In a gate-keeping system, the patient has no choice and has to visit the GP. If, instead, the system is not gate keeping, the patient can choose to visit either the GP or the specialist. To formalise this idea, we denote by α the probability that the patient visits the GP given that his illness is of a common type, $\underline{m}$, and therefore does not require a specialist's treatment: $\alpha = \Pr(GP|\underline{m})$.[15] Similarly, β denotes the probability that the patient visits the GP given that his illness is of a special type, $\overline{m}$, and therefore requires a specialist's treatment: $\beta = \Pr(GP|\overline{m})$. These probabilities are common knowledge and depend on the institutional restriction (whether the system is gate-keeping or not). In a gate-keeping system, the patient must visit the GP, and the probability that he visits the GP is one, no matter what his illness is. Therefore, $\alpha = \beta = 1$. In a non gate-keeping system, there is no such obligation for the patient and the probabilities α and β can take any value between zero and one. The only restriction we impose on these probabilities for the case of a non gate-keeping system is that $\beta \leq \alpha \leq 1$.[16] This implies that when his illness is of a common type, the patient is more likely to visit a GP (rather than a specialist) than a patient with a special illness.[17] Given the

---

[14] Horn, Sharkley and Gassaway (1996) report from the Managed Care Outcomes Project that, among 12,997 patients with at least one of the five most common illnesses, about 77% suffered from a condition classified as either light, normal, or needing some diagnostic test, and about 22% from a condition classified as either heavy and needing study, or catastrophic and requiring hospitalisation.

[15] The term "*GP*" in such an expression stands for the patient's choice to visit the GP at this stage.

[16] Note that this assumption can only go to the detriment of a gate-keeping system. However, our result is that, despite this assumption, the gate-keeping system is better whenever incentives matter.

[17] In order to focus on the GP's gatekeeper role and incentives, we do not explicitly model how the patient's beliefs and incentives affect the decision to visit either type of provider. However, we can easily imagine that the symptoms attached to a special illness are more likely to induce the patient to visit a specialist than symptoms attached to a minor illness.

probabilities of either type of illnesses (p and $1 - p$) and the conditional probabilities α and β, the probability that the patient visits the GP is: $\lambda = \Pr(GP) = \alpha p + \beta(1 - p)$.[18] Finally, we denote by q the probability that a patient has a common type illness given that he has first visited the GP rather than the specialist: $q = \Pr(\underline{m}|GP) = \alpha p/\lambda$.[19] In a gate-keeping system, the restriction $\alpha = \beta = 1$ implies that $\lambda = 1$ and $q = p$. In a non gate-keeping system, the weaker restriction $\beta \leq \alpha \leq 1$ implies that $\lambda \leq 1$ and $p \leq q$.

If the patient's choice is to visit the specialist the game ends because we assume that the specialist is able to treat any kind of illness and that the patient is then cured with probability one.

If the patient's choice is to visit the GP then, in stage 4, the GP exerts some level of effort, e, while performing the diagnosis. This effort yields a signal about the illness type to the GP: $m_d \in \{\underline{m}_d, \overline{m}_d\}$, which coincides with the true illness type m with some increasing probability on effort. For simplicity, we assume that the probability that the GP makes the right diagnosis is: $\Pr(\underline{m}_d|\underline{m} \cap GP) = \Pr(\overline{m}_d|\overline{m} \cap GP) = e$. We restrict the choice of effort to $e \in [1/2, 1]$. The reason for this is that the worst the GP could do is to toss a coin and decide on the patient's condition. Any diagnostic effort exerted would improve on this. Accordingly, the GP bears the disutility of effort, $v(e)$.

Moreover, the GP also knows by experience what the likelihood, q, is that the patient she is seeing has a minor condition (or the share of $\underline{m}$ types in her practice). The GP combines this piece of information with the diagnosis outcome to decide in

---

[18] If we consider a group of patients instead of one patient, the parameter λ can be interpreted as the (observable) proportion of patients visiting the GP.

[19] The parameter q can also be interpreted as the proportion of patients suffering from a common illness among the patients visiting the GP.

stage 6 whether to treat the patient or alternatively, to refer him to a specialist. If the patient recovers his health, that is, if he receives adequate treatment, the game ends. Otherwise, the patient suffers the health loss valued l by the insurer, and is referred to the specialist in stage 7.[20]

We solve the game by backward induction. For notational clarity, we will use the following notation:

$$X = qB - (2q\text{-}1)S, \qquad z = (1-q)l + qh, \qquad \text{and} \qquad m(y) = Min\{y;\ 1/2\}.$$

3. The GP's Problem

3.1. Ex Post Referral Decision

The GP faces a sample of patients composed of q patients of type $\underline{m}$ and $1 - q$ patients of type $\overline{m}$. Once diagnosis has been exerted the GP combines this sample information with her diagnosis outcome to update her belief on the true state of nature m. If the GP uses a Bayesian updating rule, she correctly diagnoses a common illness with probability:

$$\Pr(\underline{m}|\underline{m}_d \cap GP) = \frac{q.e}{q.e + (1-q).(1-e)} = 1 - \Pr(\overline{m}|\underline{m}_d \cap GP),$$

and she correctly diagnoses a special illness with probability:

$$\Pr(\overline{m}|\overline{m}_d \cap GP) = \frac{(1-q).e}{q.(1-e) + (1-q).e} = 1 - \Pr(\underline{m}|\overline{m}_d \cap GP),$$

(see Appendix 1).

We now detail the different decisions the GP can adopt. Referring the patient to the specialist results in a payment for the GP of $D + S$, while keeping the patient

[20] That is, if the patient actually suffers from an illness of the special type and has been treated by the GP.

for treatment pays the GP the capitation (or payment per visit) component D and, with some probability, the bonus B. The GP earns the bonus B only if the patient is not ultimately referred to the specialist, that is, if the true illness condition of the patient is common. Therefore, the probability of earning B is $\Pr(\underline{m}|\underline{m}_d \cap GP)$ if the GP's signal is $\underline{m}_d$, and $\Pr(\underline{m}|\overline{m}_d \cap GP)$ if the GP's signal is $\overline{m}_d$.

Comparing this payment with the payment achieved when the patient is immediately referred, we obtain that:

if $m_d = \underline{m}_d$, then the GP prefers to refer whenever $S \geq \dfrac{q.e}{q.e + (1-q).(1-e)} B$;

if $m_d = \overline{m}_d$, then the GP prefers to refer whenever $S \geq \dfrac{q(1-e)}{q.(1-e) + (1-q).e} B$.

Note first that, quite reasonably, the GP is more prone to refer patients that she diagnoses with a special illness.

Moreover, if q increases (maybe as a result of the patient making better decisions on whom to visit), the GP would be more reassured of a common illness diagnosis, ($\Pr(\underline{m}|\underline{m}_d \cap GP)$ would increase), and this would result in a reduction of her referral rate on common diagnosis outcomes. Given that in the non gate-keeping system $q \geq p$, there are less referrals of patients diagnosed with a common condition than in the gate-keeping system.[21]

Using the fact that $1/2 \leq e \leq 1$, we conclude with the following Lemma.

Lemma 1. *Ex Post Referral Decision (summary).*

** If $S \geq B$, the GP refers the patient, whatever her signal and level of effort are.*

[21] The same effect holds for a patient diagnosed with a special condition, as an increase in q makes the GP less reassured of her diagnosis outcome.

** If $B \geq S \geq qB$ and $e \leq (1-q)S / X$, the GP refers the patient whatever her signal is, but if $e \geq (1-q)S / X$, she refers the patient only when her signal is $\overline{m}_d$.*

** If $qB \geq S$ and $e \leq q(B-S) / X$, the GP does not refer the patient whatever her signal is, but if $e \geq q(B-S) / X$, she refers the patient only when her signal is $\overline{m}_d$.*

For all cases, if the effort exerted by the GP is below a certain threshold, the GP's decision to treat or refer is independent of the diagnosis outcome. In such cases, the bonus B is earned by the GP only with probability q since this is the probability that the patient has a common condition. Alternatively, S is earned whenever she refers the patient immediately (regardless of the patient's true condition). In consequence, qB and S drive the GP behaviour in different directions: if $S/B > q$, referring the patient becomes more profitable than treating him. Indeed, S/B determines the preference between keeping or referring patients and results in the three regimes, which are made manifest in Lemma 1.

This behaviour implies that before effort is exerted, the GP anticipates that ex post she might follow two strategies with respect to the information acquired during diagnosis. These strategies are: (1) to condition the referral decision to the information gained during diagnosis: a "most adequate" referral strategy, and (2) not to condition the referral decision on the information gained during diagnosis: a blind strategy. In this latter case, the GP can either systematically provide primary care whatever the diagnosis outcome indicates, a "keep all" referral strategy, or systematically refer the patient to the specialist whatever the diagnosis indicates, a "send all" referral strategy.[22] The referral strategy determines the effort the GP chooses, and also the patient's expected health loss, *EL*, the public insurer's expected

[22] Subscripts *a*, *k* and *s* refer to each of these strategies.

costs, *EC*, and the GP's expected utility, *EU* (see Appendix 2 for the derivation of *EL, EC* and *EU* under each of the above mentioned strategies).

3.2. Diagnosis

We begin this Section by mentioning a special feature of the blind referral strategies. If the GP adopts either of these, since the referral decision is independent of the information acquired during diagnosis, the GP's expected payment does not depend on her diagnosis effort. In the "send all" case, the GP's expected payment is $\lambda.(D + S)$, and in the "keep all" case it is $\lambda.(D + qB)$. Therefore, the GP always prefers to exert the lowest effort, be it contractible or not. This is why we describe these strategies as "blind strategies".

Lemma 2. *GP's effort under a blind strategy.*

Under blind referral strategies, the optimal diagnosis effort is minimal: $e_s = e_k = 1/2$.

On the other hand, if $B \geq S$ and $e \geq (X + |S - qB|) / 2X$, the GP adopts a most adequate referral strategy (i.e. referring only if her signal is $\overline{m}_d$, see Lemma 1). In that case, she exerts a level of effort e_a, which is the solution to $Max_{\{e\}} \{EU_a\}$ *s.t.* $e \geq (X + |S - qB|) / 2X$. This is:

$$e_a = \begin{cases} (X + |S - qB|) / 2X & if\ 2X^2 \leq |S - qB| \\ m(X) + 1/2 & if\ 2X^2 \geq |S - qB|. \end{cases}$$

The GP's expected utility evaluated at this level of effort is:

$$EU_a = \begin{cases} \frac{\lambda}{2}\left\{2D + S + qB + |S - qB| - \left((S - qB)/2X\right)^2\right\} & if\ 2X^2 \leq |S - qB| \\ \lambda\left\{D + qS + X\left(m(X) + 1/2\right) - \left(m(X)\right)^2 / 2\right\} & if\ 2X^2 \geq |S - qB|. \end{cases}$$

In order to know under which condition the GP exerts an effort e_a, e_k, or e_s, we compare her expected utility under each referral strategy evaluated at the corresponding optimal effort. We conclude with the following Lemma.

Lemma 3. *GP's ex ante referral practice and effort decision (summary).*

(i) *If* $m(X).(1 - m(X)) \geq 2q(B - S)$*, the GP refers the patient, exerting the lowest level of effort and her expected utility is* $EU_s = \lambda (D + S)$.

(ii) *If* $m(X).(1 - m(X)) \leq 2\ Min\{\ q(B - S);\ (1 - q)S\ \}$ *and* $B \geq S$*, the GP refers the patient only when her diagnosis outcome is* $\overline{m}_d$*, exerting a level of effort* $e_a = X + 1/2$*, and her expected utility is* $EU_a = \lambda\left\{D + qS + X\left(m(X) + 1/2\right) - \left(m(X)\right)^2 / 2\right\}$.

(iii) *If* $m(X).(1 - m(X)) \geq 2(1 - q)S$*, the GP does not refer the patient, exerting the lowest level of effort and her expected utility is* $EU_k = \lambda(D + qB)$.

When the GP adopts the most adequate referral strategy, her optimal effort is increasing in B and decreasing in S (see Lemma 3 (ii)). The rationale behind this dependence is that the likelihood of the right diagnosis is increasing in effort. Since B is a gain for the GP whenever she correctly diagnoses a common illness, as effort increases it is earned more frequently from illnesses of the common type, occurring with a probability q. The bonus S is earned whenever the GP diagnoses a special illness be it rightly or wrongly. As effort increases, S is earned more frequently from illnesses of the special type but less frequently from illnesses of the common type. While S acts as an incentive for effort for the special type illnesses (occurring with a probability $1 - q$), it acts as a disincentive for effort for the common type illnesses (occurring with a probability q). Since $q \geq p \geq 2/3$, the second effect dominates and the overall effect of S on effort is negative. Moreover, these effects are more prominent the larger is q. Therefore, B is a better incentive for effort and S is a better

disincentive for effort in the non gate-keeping case where $q \geq p$, than in the gate-keeping case where $q = p$.

On the other side, the incentive power of B is limited. If B is too high relative to S, then not referring the patient becomes very profitable to the GP, to the extent that the GP might systematically treat the patient by herself, whatever the diagnosis outcome is. This is reflected in the case (iii) of Lemma 3, with a "keep all" strategy and the effort being at its lowest level. Moreover, in the non gate-keeping case (where $q \geq p$), the GP's temptation to switch to a "keep all" strategy increases since there is a larger likelihood of earning B while exerting a low effort. Therefore, in this case, a lower B and/or a higher S would be required to avoid this systematic behaviour, although it would result in a lower effort.

Case (i) in Lemma 3 refers to the situation in which S is so high with respect to B that the GP systematically refers the patient, whatever the diagnosis is. In consequence, effort would be at its lowest level.

4. The Insurer's Problem

4.1. Contract Design

We first characterise the optimal GP's contract in a situation in which there is no asymmetric information between the public insurer and the GP, and the GP has no negotiation power. This first best contract minimises the insurer's expected costs and valuation of the patient's health loss. The only constraint it must satisfy is the participation constraint, ensuring that the GP accepts the contract in stage 2: $EU \geq w$. Since no other constraints matter, the participation constraint binds at the optimum since both the insurer's expected costs and the GP's expected utility are increasing in the expected GP's payment.

The following Lemma provides the optimal contract under symmetric information. In it, we do not a priori limit our attention to the most adequate referral strategy. However, it is easy to show that with blind strategies, the insurer prefers the effort to be minimal. Since the referral decision is independent of the GP's diagnosis, the expected payment and the patient's expected loss are independent of the GP's effort. In this case, inducing effort only increases the GP's remuneration and the insurer's costs at no gain for either the patient or the insurer.

Lemma 4. *First Best contract design.*

** If it is optimal to refer the patient whatever the GP's signal is, then any contract* (D, B, S) *such that* $D + S = w/\lambda$ *is an optimal contract, with* $e = 1/2$. *The insurer's value is* $(EC + EL)_s = h + w$.

** If it is optimal to refer the patient only when the GP's signal is* $\overline{m}_d$, *then any contract* (D, B, S) *such that* $D + qS + m(X).e - v(e) = w/\lambda$ *is an optimal contract, with* $e = 1/2 + m(z)$. *The insurer's value is* $(EC + EL)_a = h + w + \lambda\, m(z).(1 - m(z))/2 - \lambda\, qh$.

** If it is optimal not to refer the patient whatever the GP's signal is, then any contract* (D, B, S) *such that* $D + qB = w/\lambda$ *is an optimal contract, with* $e = 1/2$. *The insurer's value is* $(EC + EL)_k = h + w + \lambda\,\{(1 - q)l - qh\}$.

This first best solution is not only a full information benchmark, it also shows which allocation can be attained with asymmetric information on effort and diagnosis outcome if the GP has no negotiation power, that is, whenever the GP cannot impose any wealth constraint. Then, the problem is one of moral hazard with a risk neutral agent (the GP) and no wealth (or limited liability) constraint. Generally, in this type of situation, there exist contracts that attain the first best allocation at no further cost for

the Principal (with no informational rent for the agent). The reason is that with a risk neutral agent there is no problem of risk provision, and any kind of incentive can be provided at no premium. In our case, the insurer can induce the same effort as in the first best case, resulting in the same expected insurer's costs, patient's health loss and GP's utility.[23] In particular, for the contract to be first best, we need a very low payment in the worst contingency: $\lambda D < w$.

However, if the GP has negotiation power, the first best effort can not be attained without granting an informational rent to the GP since the contract must give her at least the same utility as from w in any contingency, and in particular in the worst one: $\lambda D \geq w$. As it is clear from Lemma 3, the capitation payment D has no incentive power while it is a cost for the insurer. Therefore, in a second best contract (when the GP has negotiation power), D will always be set at its lowest level, as stated in the next Lemma.

Lemma 5. *Second Best contract design (Capitation Payment).*

The Second Best capitation payment is such that the wealth constraint associated to the worst contingency binds: $D = w/\lambda$.

Considering the blind treatment strategies, we know from Lemma 1 that the GP's referral decisions are always independent of the diagnosis outcome, and hence her effort is the lowest one no matter what the bonuses in the contract are (Lemma 2). Therefore, if the public insurer wants the GP to adopt one blind strategy, it is better for him to economise on the GP's bonuses, provided that the GP has an incentive to

[23] An example of such a contract is the following. Suppose that the insurer wants the GP to refer the patient only when her signal is $\overline{m}_d$. Take the contract (D, B, S) with $D = (w/\lambda) - (3 - q)/8(1 - q)$, $B = 1/4q(1 - q)$ and $S = 1/4(1 - q)$. This contract satisfies the incentive compatibility constraints in Lemma 3 (ii). From Lemma 3 (ii), we find the optimal effort for the GP: $e_a = 1$, and the resulting GP's expected utility: $EU = w$.

adopt any of these blind strategies (see Lemma 3). The following Lemma gives the optimal bonuses for any of the blind strategies.

Lemma 6. *Second Best contract design* (*Blind strategies*).

If the insurer wants the GP either to refer the patient whatever her signal is, or not to refer the patient whatever her signal is, the optimal contract is characterised by $B = S = 0$. *The GP exerts the lowest level of effort:* $e = 1/2$, *and the insurer's value is defined as in Lemma 4.*

Considering the most adequate strategy, the bonuses now matter both for determining the level of effort and for providing the GP with incentives to adopt this strategy, following the incentive compatibility constraints in Lemma 3 (ii). The following Proposition provides us with the optimal incentive provision through bonuses in that case.

Proposition 1. *Second Best contract design* (*Incentive provision*).

If the insurer wants the GP to refer the patient only when her signal is $\overline{m}_d$, *the optimal contract is characterised by:*

* $B = S = 0$ *if* $z \leq (4 - 3q)/4(1 - q)$. *The GP exerts the lowest level of effort:* $e = 1/2$, *and the insurer's value is* $(EC + EL)_a = h + w + \lambda\{(1 - q)l - qh\}/2$.

* $B = (4 - 3q)/8q(1 - q)$ *and* $S = 1/8(1 - q)$ *if* $z \geq (4 - 3q)/4(1 - q)$. *The GP exerts the highest level of effort:* $e = 1$, *and the insurer's value is* $(EC + EL)_a = h + w + \lambda (4 - 3q)/8(1 - q) - \lambda qh$.

Proof. See Appendix 3.

Notice that the optimal contract is characterised by positive bonuses only when the specialist's costs as well as the patient's expected loss (both summarised in

z) are high enough. This happens because providing the GP with incentives through bonuses is costly for the insurer. Indeed, with these positive bonuses and because of the wealth constraint, the GP's expected utility is strictly higher than her reservation utility: $EU_a = w + \lambda(3 - 2q)/8(1 - q) > w$, meaning that the GP earns an informational rent. This is costly for the insurer and it is worthwhile only when both the specialist's costs and the patient expected loss are high enough.

In the case where it is worth giving incentives to the GP, the optimal incentive provision requires a positive bonus S, even though this bonus acts as a disincentive for effort. Still, any lower bonus S would make the GP go for a treat-all, blind strategy. Therefore, the incentive constraint in Lemma 3 (iii) that matters (and binds, see Appendix 3) is the one ensuring that the GP has no incentive to adopt the "treat all", blind strategy.

Moreover, we know that the GP's temptation to switch to the "keep all" strategy is higher when the referral is optional than when it is compulsory (gate-keeping system). This explains why S is increasing in q: the non gate-keeping system requires a higher bonus S to avoid this systematic behaviour. The bonus B is increasing in q as well: the non gate-keeping system also requires a higher bonus B to compensate for the disincentive effect of the higher S on effort. Ultimately, this results in the GP's contractual costs per patient being higher in the non gate-keeping system than in the gate-keeping system.

4.2. Contract and Referral Practice Choice

The former subsection provides, for both the first best and the second best situation, three optimal contracts, depending on whether the insurer prefers one or another referral strategy. To complete the analysis, the following Lemma presents the conditions under which the insurer prefers each of the referral strategies (and hence

offers the corresponding optimal contract and induces the corresponding level of effort).

Lemma 7. *Insurer's Choice.*

The insurer prefers the GP to adopt

* *the send all strategy, when* $m(z)(1 - m(z)) \geq 2\ qh$, *in the first best, and when* $qh \leq Min\{\ (1 - q)l; (4 - 3q)/8(1 - q)\ \}$, *in the second best;*

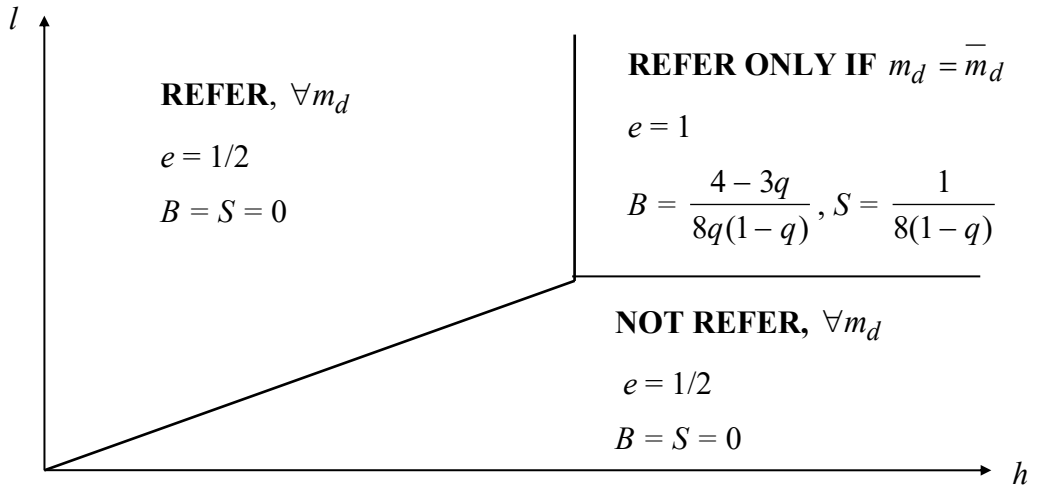
* *the most adequate strategy, when* $m(z)(1 - m(z)) \leq 2\ Min\{\ qh; (1 - q)l\ \}$, *in the first best, and when* $(4 - 3q)/8(1 - q) \leq Min\{\ (1 - q)l; qh\ \}$, *in the second best;*

* *the keep all strategy, when* $m(z)(1- m(z)) \geq 2\ (1 - q)l$, *in the first best, and when* $(1 - q)l \leq Min\{\ qh; (4 - 3q)/8(1 - q)\ \}$, *in the second best.*

Proof. Straightforward from Lemmas 4 and 6 and from Proposition 1.

The conditions presented in the former Lemma are intuitive (see Figure 1 for a graphical representation of these conditions in the second best case). To summarise, they state that it is worthwhile giving incentives to the GP to adopt the most adequate treatment strategy, when both the specialist's costs and the patient's health loss are high. If instead the specialist's costs are low, then the insurer is better off with a blind "send all" strategy since it economises on the GP's expected costs (the GP does not need to be compensated for any effort and she does not earn any informational rent). It also economises on the patient's expected loss, since the patient is always cured with probability one when he receives specialist care, while the specialist's costs do not heavily increase the insurer's costs. On the contrary, when the patient's health loss is low, then the insurer is better off with a blind treat all strategy. Again it economises on the GP's payment. It also limits the specialist's costs to the very necessary cases,

while the patient does not suffer much from delaying a specialist's treatment if necessary.



(*h* stands for the costs of specialist care. *l* denotes the insurer's valuation of the patient's health loss if he receives one unnecessary treatment. m_d is the signal received by the GP after diagnosis. This signal can be common type ($m_d = \underline{m}_d$) or special type ($m_d = \overline{m}_d$). In each area we indicate the optimal referral practice, the contract components *B* and *S*, and the effort exerted by the GP, *e*.)

Figure 1

5. Gate-keeping system or non gate-keeping system?

In this Section we focus on the comparison between the gate-keeping systems and the non gate-keeping ones, in the cases where the GP's incentives matter, that is, when the insurer wants the GP to adopt the most adequate strategy. Figure 2 provides a more complete comparison in the second best case, based on the results in Lemmas 6 and 7 and Proposition 1.

The following two elements are crucial for this comparison. First, the likelihood that the GP's patient needs primary care only is higher in the non gate-keeping system ($q \geq p$), than in the gate-keeping one ($q = p$). Second, the GP sees the

patient with a higher likelihood in the gate-keeping case ($\lambda = 1$) than in the non gate-keeping case ($\lambda \leq 1$).

In all cases, the insurer only cares about the patient's expected health loss and the expected service provision costs consisting in the specialist's costs and the GP's payment. The latter can be decomposed in reservation wage (w), effort compensation ($\lambda.v(e_a)$), and informational rent ($IR_a = EU_a - w$).[24]

In a first best situation, the GP earns no informational rent as her participation constraint always binds. Therefore, the comparison between having gate keeping or not only relies upon the expected specialist's costs and the GP's effort compensation when $z \geq 1/2$.[25] The expected specialist's costs are higher if the system is not gate-keeping as the patient might visit the specialist without needing to. The GP's effort compensation is higher in the gate-keeping system because the GP diagnoses the patient with a larger likelihood. Therefore, if the specialist's costs are low enough ($h < \underline{h} = (1 - \lambda)/8(p - \lambda q)$) the insurer prefers the non gate-keeping system, as the unnecessary visits to the specialist are not too costly. For higher levels of specialist's costs, the insurer rather prefers the gate-keeping system. Having $m(z).(1 - m(z)) = 1/4 \leq 2\ Min\{\ qh, (1 - q)l\ \}$ for the most adequate strategy to be optimal in the first best, allows for the specialist's costs h to be above or below $\underline{h}$.

To compare the gate-keeping and non gate-keeping systems in the second best case, one must not only consider the latter differences but also any difference between the relevant informational rents. In the second best case, the informational rent is: $IR_a = \lambda\,(3 - 2q)/8(1 - q)$. Whether this informational rent is higher in one system or the

[24] Using Lemma 5, we can write: $EC_a = w + \lambda v(e_a) + IR_a + (1 - \lambda q e_a)h$.

[25] This parameter restriction simplifies the intuitive argument without altering the result in Proposition 2.

other is not determined. Two opposite effects are at work here. First, there is a level effect: the informational rent increases with the probability of the patient visiting the GP, λ. The reason is that, as this probability increases, the GP earns the bonuses more often. Second, there is a marginal effect. The informational rent per patient is higher in the non gate-keeping system because both bonuses (S and B) are higher. Overall, the GP's informational rent is an additional cost for the insurer, whatever referral system (gate-keeping or not) is used. From Lemma 7, we know that it is worth paying this cost only if the specialist's costs are high enough: $h \geq (4 - 3q)/8q(1 - q)$. For such high levels of specialist's costs, the insurer always prefers the gate-keeping system, as the unnecessary visits to the specialist are too costly. The following Proposition summarises this comparison between systems.

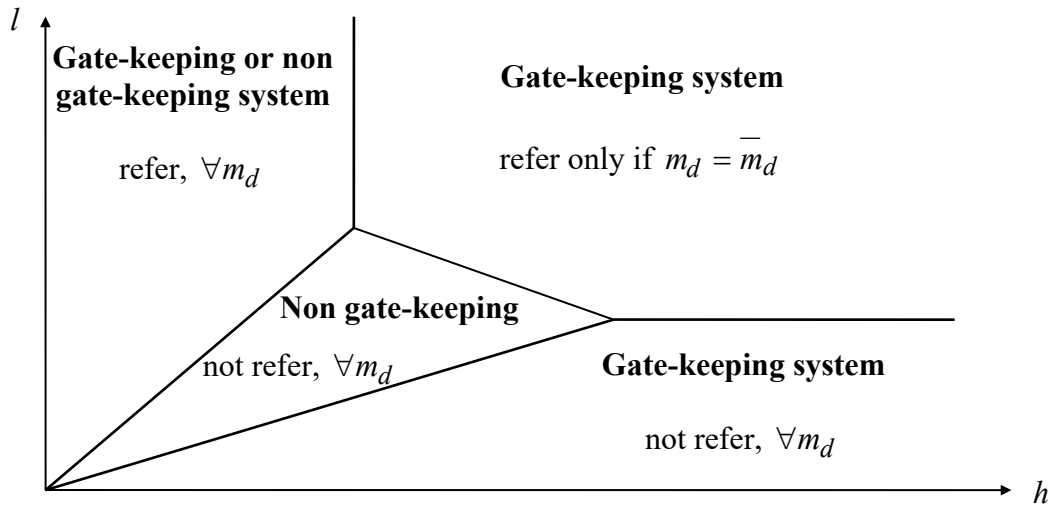
Proposition 2. *Gate-keeping system or non gate-keeping system?*

When the most adequate strategy is optimal:

in the first best situation, the insurer prefers the system to be gate-keeping or not, depending on the parameters' configuration;

in the second best situation, the insurer always prefers the gate-keeping system.

Proof. Straightforward from Proposition 1, and Lemmas 1, 4, 7 and 8.



(*h* stands for the costs of specialist care. *l* denotes the insurer's valuation of the patient's health loss when he receives one unnecessary treatment. m_d is the signal received by the GP after diagnosis. This signal can be common type ($m_d = \underline{m}_d$) or special type ($m_d = \overline{m}_d$). In each area we indicate the optimal referral system (gate-keeping or non gate-keeping) and strategy.

Figure 2

6. Conclusion

This paper compares the performance of health care systems in which access to a specialist requires the referral of a GP with the performance of systems where a GP referral is not compulsory. In the paper, performance is measured using an index of quality net of provision costs, where quality is related to the patients' waiting time.

We derive an optimal payment contract for the GP, which yields incentives for the GP to act as a filter for specialists. The contract combines two components: a bonus when the GP successfully treats a patient (and consequently, does not refer the patient elsewhere), and a cost-sharing component. While the bonus gives the GP incentives to exert effort whilst diagnosing, the cost-sharing parameter prevents her from exerting minimal effort and trying to treat all patients to benefit from the possibility of earning the bonus.

Taking into consideration the optimal payment contract, a gate-keeping referral system and a non gate-keeping referral system are compared. The main trade-

off underlying the comparison is the following: a system with gate keeping allows savings on the specialist's costs and results in a lower expected payment per patient for the GP. With a non gate-keeping system, expected waiting times for a patient are shorter, the payment for the GP per patient is larger, but fewer patients visit the GP.

The model shows that it only "pays" to provide incentives for the GP to exert maximal diagnosis effort and to refer the patient when both the specialist's costs and the waiting costs are large. In this case, the savings in costs out-weigh the informational rent. Moreover, in this case, the insurer prefers the gate-keeping system. On the one hand, it allows savings on the specialist's costs since the optimal contract gives incentives to the GP to avoid all the unnecessary visits to the specialist. On the other hand, although the expected waiting times for a patient are longer under the gate-keeping system, these times are kept at a low level because the GP's diagnosis is accurate and no patients wait for a specialist's visit without needing to.

In terms of policy recommendation, our result suggests that if GPs' payment contracts include proper incentives, then public insurers should go for a gate-keeping system. However, this recommendation should be taken cautiously since the optimal contracts derived in our work rely on assumptions. A delicate assumption concerns the utility function of the GP. Despite the empirical evidence that GPs' behaviour responds to financial incentives, it is likely that non-financial motives also enter the GPs' utility function. More empirical work estimating the determinants of GPs' behaviour is needed for this line of research to fearlessly recommend an incentive contract to be implemented. Nevertheless, this and other related papers offer arguments that can inform the debates on health policy.

Finally, two possible extensions of this piece of work are possible. The model proposed captures well the main features of a public health system in which there is

little or no competition between GPs. It would be interesting to investigate how the problem of access would be solved in a context of private medicine in which the objective function of insurers would be profit and not a measure of quality borne of costs. Equally, it would be good to know whether allowing for provider competition would alleviate the problem of GP incentive provision. However, in situations with good competition between insurers and high switching costs between GPs our results would continue to hold.

Appendix 1

In general:

$$\Pr(m \cap GP) = \Pr(GP).\Pr(m|GP) = \lambda \Pr(m|GP).$$

$$\Pr(m \cap m_d \cap GP) = \Pr(m \cap GP).\Pr(m_d|m \cap GP) = \lambda \Pr(m|GP).\Pr(m_d|m \cap GP).$$

$$\begin{aligned}\Pr(m_d \cap GP) &= \Pr(\underline{m} \cap m_d \cap GP) + \Pr(\overline{m} \cap m_d \cap GP) \\ &= \lambda \Pr(\underline{m}|GP).\Pr(m_d|\underline{m} \cap GP) + \lambda \Pr(\overline{m}|GP).\Pr(m_d|\overline{m} \cap GP).\end{aligned}$$

$$\Pr(m|m_d \cap GP) = \frac{\Pr(m \cap m_d \cap GP)}{\Pr(m_d \cap GP)} = \frac{\Pr(m|GP).\Pr(m_d|m \cap GP)}{q.\Pr(m_d|\underline{m} \cap GP) + (1-q).\Pr(m_d|\overline{m} \cap GP)}.$$

Therefore:

$$\Pr(\underline{m}|\underline{m}_d \cap GP) = \frac{q.e}{q.e + (1-q).(1-e)}; \qquad \Pr(\overline{m}|\underline{m}_d \cap GP) = \frac{(1-q).(1-e)}{q.e + (1-q).(1-e)};$$

$$\Pr(\underline{m}|\overline{m}_d \cap GP) = \frac{q.(1-e)}{q.(1-e) + (1-q).e}; \qquad \Pr(\overline{m}|\overline{m}_d \cap GP) = \frac{(1-q).e}{q.(1-e) + (1-q).e}.$$

Appendix 2.

$$EL_a = \Pr(\underline{m}_d \cap GP).\Pr(\overline{m}|\underline{m}_d \cap GP).l = \lambda(1-q).(1-e).l;$$

$$EC_a = \Pr(GP).D + \Pr(\underline{m}_d \cap GP).\Pr(\underline{m}|\underline{m}_d \cap GP).B + \Pr(\overline{m}_d \cap GP).S$$
$$+ \left\{1 - \Pr(GP) + \Pr(\overline{m}_d \cap GP) + \Pr(\underline{m}_d \cap GP).\Pr(\overline{m}|\underline{m}_d \cap GP)\right\}h$$
$$= h + \lambda\{D + qS + (X - qh)e\};$$

$$EU_a = \Pr(GP).(D - v(e)) + \Pr(\underline{m}_d \cap GP).\Pr(\underline{m}|\underline{m}_d \cap GP).B + \Pr(\overline{m}_d \cap GP).S$$
$$= \lambda\left\{D + qS + Xe - (e - 1/2)^2/2\right\};$$

$$EL_k = \left\{\Pr(\overline{m}_d \cap GP).\Pr(\overline{m}|\overline{m}_d \cap GP) + \Pr(\underline{m}_d \cap GP).\Pr(\overline{m}|\underline{m}_d \cap GP)\right\}.l$$
$$= \lambda(1 - q).l;$$

$$EC_k = \Pr(GP).D + \left\{\Pr(\overline{m}_d \cap GP)\Pr(\underline{m}|\overline{m}_d \cap GP) + \Pr(\underline{m}_d \cap GP).\Pr(\underline{m}|\underline{m}_d \cap GP)\right\}B$$
$$+ \left\{1 - \Pr(GP) + \Pr(\overline{m}_d \cap GP).\Pr(\overline{m}|\overline{m}_d \cap GP) + \Pr(\underline{m}_d \cap GP).\Pr(\overline{m}|\underline{m}_d \cap GP)\right\}h$$
$$= h + \lambda\{D + qB - qh\};$$

$$EU_k = \Pr(GP).(D - v(e))$$
$$+ \left\{\Pr(\overline{m}_d \cap GP)\Pr(\underline{m}|\overline{m}_d \cap GP) + \Pr(\underline{m}_d \cap GP).\Pr(\underline{m}|\underline{m}_d \cap GP)\right\}.B$$
$$= \lambda\left\{D + qB - (e - 1/2)^2/2\right\};$$

$$EL_s = 0;$$

$$EC_s = \Pr(GP).(D + S) + h = h + \lambda\{D + S\};$$

$$EU_s = \Pr(GP).(D + S - v(e)) = \lambda\left\{D + S - (e - 1/2)^2/2\right\}.$$

Appendix 3. Proof of Proposition 1.

When the contract design is submitted to the GP's wealth constraint and we consider the GP referring the patient only if her signal is $\overline{m}_d$, then the optimal GP's payment contract is the solution to:

$$\underset{\{B,S\}}{Min}\left\{h+w+\lambda\left[qS+(1-q)l+(X-z)(m(X)+1/2)\right]\right\}$$

$$s.t.\quad\begin{cases}B\geq S\geq 0\\ m(X).(1-m(X))\leq 2.Min\left\{q(B-S);(1-q)S\right\}\\ X=qB-(2q-1)S.\end{cases}$$

Consider first the case where $X \geq 1/2$. Then, $m(X) = 1/2$, and the insurer's program can be rewritten as:

$$\underset{\{X,S\}}{Min}\left\{h+w+\lambda\left[qS+(1-q)l+X-z\right]\right\}\quad s.t.\quad\begin{cases}X\geq 0\\ X-1/8\geq(1-q)S\geq 1/8.\end{cases}$$

The objective function is increasing in both X and S. Therefore, the optimal X and S are the lowest one such that the constraints hold: $X = 1/2$ and $S = 1/8(1 - q)$. The insurer's objective function is then:

$$(EC+EL)_1=h+w+\lambda\left(\frac{4-3q}{8(1-q)}-qh\right).$$

Consider now the case where $X \leq 1/2$. Then, $m(X) = X$, and the insurer's program can be rewritten as:

$$\underset{\{X,S\}}{Min}\left\{h+w+\frac{\lambda}{2}\left[(1-q)l-qh\right]+\lambda\left[qS+X^2-zX+X/2)\right]\right\}$$

$$s.t.\quad\begin{cases}1/2\geq X\geq(1-q)S\geq 0\\ X(1-X)\leq 2(1-q)S\leq X(1+X).\end{cases}$$

The objective function is increasing in S. Therefore, the optimal S is the lowest one such that the constraints hold: either $S = 0$, or $S = X.(1 + X)/2(1 - q)$.

If $S = 0$, then $X = 0$, otherwise the constraints could not hold together. The insurer's objective function would then be: $(EC + EL)_0 = h + w + \lambda\ \{(1 - q)l - qh\}/2$.

If $S = X.(1 + X)/2(1 - q)$, then the constraint $X.(1 - X) \leq 2(1 - q)S$ is always satisfied since $X.(1 + X) \geq X.(1 - X), \ \forall X$; the other constraints reduce to $0 \leq X \leq 1/2$, and the program can be rewritten as:

$$\underset{\{X\}}{Min}\left\{h + w + \frac{\lambda}{2}\left[(1-q)l - qh\right] + \lambda\left[\frac{X}{2(1-q)} - \frac{3q-2}{2(1-q)}X^2 - zX\right]\right\}$$
$$s.t. \quad 0 \leq X \leq 1/2.$$

The insurer's objective function is then concave in X: $\frac{\partial^2 EC_m}{\partial X^2} = -\lambda\frac{3q-2}{1-q} < 0.$

Therefore, it reaches its minimum either at $X = 0$ or at $X = 1/2$. Both candidates have already been considered above.

Comparing $(EC + EL)_0$ with $(EC + EL)_1$, and solving for B and e as functions of X and S, we find the optimum described in Proposition 1.

QED

References

Blomqvist, A., 1991. The doctor as a double agent: Information asymmetry, health insurance and medical care. Journal of Health Economics 10 (4), 411–432.

Chalkley, M., Malcomson, J.M., 1998a. Contracting for health services with unmonitored quality. Economic Journal 108 (449), 1093–1110.

Chalkley, M., Malcomson, J.M., 1998b. Contracting for health services when patient demand does no reflect quality. Journal of Health Economics 17 (1), 1–19.

Coxson, B., Propper, C., Perkins, A., 2001. Do doctors respond to financial incentives? UK family doctors and the GP fundholder scheme. Journal of Public Economics 79, 375–398.

Dionne, G., Contandriopoulos, A., 1985. Doctors and their workshops: A review article. Journal of Health Economics 4, 21–33.

Ellis, R.P., McGuire, T.G., 1993. Supply-side and demand-side cost sharing in health care. Journal of Economic Perspectives 7 (4), 131–151.

Ellis, R.P., McGuire, T.G., 1990. Optimal payment systems for health services. Journal of Health Economics 9 (4), 375–396.

Encinosa, W.E., Gaynor, M., Rebitzer, J.B., 1997. The sociology of groups and the economics of incentives: Theory and evidence on compensation systems. Working Paper 5953, National Bureau of Economic Research.

Fenn, P., Diacon, S., Gray, A., Hodges, R., Rickman, N., 2000. The current cost of medical negligence in NHS hospitals: analysis of claims database. British Medical Journal 320 (7249), 1567–1571.

Garcia Mariñoso, B., 1999. Optimal access to hospitalised attention from primary health care. Discussion paper 9.907, the Economics Research Center, University of East Anglia.

Hellinger, F., 1996. The impact of financial incentives on physician behaviour in managed care plans: A review of the evidence. Medical Care Research Review 53, 294–314.

Horn S.D., Sharkey P.D., Gassaway J., 1996. Managed Care Outcomes Project: Study design, baseline patient characteristics, and outcome measures. American Journal of Managed Care 2, 237–247.

Jelovac, I., 2001. Physicians' payment contracts, treatment decisions and diagnosis accuracy. Health Economics 10(1), 9–25.

Ma, C.t.A., 1994. Health care payment systems: Cost and quality incentives. Journal of Economics Management and Strategy 3, 93–112.

McGuire, T.M., 2000. Physician agency, in: Culyer, A.J., Newhouse, J.P. (Eds), Handbook of Health Economics, Vol. 1. Elsevier Science, Amsterdam, pp. 461–536.

Scott, A., 2001. Eliciting GPs' preferences for pecuniary and non-pecuniary job characteristics. Journal of Health Economics 20, 329–347.

Towse, A., Danzon, P., 1999. Medical negligence and the NHS: An economic analysis. Health Economics 8, 93–101.