fedea

Fundación de
Estudios de
Economía Aplicada

Gatekeeping versus Direct-Access when Patient Information Matters

by

Paula González*

DOCUMENTO DE TRABAJO 2008-05

Serie Economía de la Salud y Hábitos de Vida
CÁTEDRA Fedea – la Caixa

January 2008

* Universidad Pablo de Olavide (Sevilla) and FEDEA

ISSN:1696-750X

Jorge Juan, 46
28001 Madrid -España
Tel.: +34 914 359 020
Fax: +34 915 779 575
infpub@fedea.es

N.I.F. G-78044393

Gatekeeping versus Direct-Access when Patient Information Matters*

Paula González†
Universidad Pablo de Olavide (Sevilla) and FEDEA

Abstract

We develop a principal-agent model in which the health authority acts as a principal for both a patient and a general practitioner (GP). The goal of the paper is to weigh the merits of gatekeeping versus non-gatekeeping approaches to health care when patient self-health information and patient pressure on GPs to provide referrals for specialized care are considered. We find that, when GPs incentives matter, a non-gatekeeping system is preferable only when (i) patient pressure to refer is sufficiently high and (ii) the quality of the patient's self-health information is neither highly inaccurate (in which case the patient's self-referral will very inefficient) nor highly accurate (in which case the GP's agency problem will be very costly).

JEL classification: D82, H51, I18, L51.

Keywords*:* General Practice, Moral hazard, Incentives, Patient self-health information, Patient pressure, Referrals.

*This work was initiated while I was visiting CORE-UCL (Belgium) whose hospitality is gratefully acknowledged. I am indebted to Maurice Marchand and Nicolás Porteiro for useful discussions and valuable suggestions. I also would like to thank Louis Eeckhoudt, Karen Eggleston, Begoña García-Mariñoso, Javier Hualde, Izabela Jelovac, Robert Nuscheler, Motohiro Sato as well as seminar participants at XXVIII Simposio de Análisis Económico (Sevilla), 5_{th} European Health Economics Workshop (York), Sixth Biennial Conference on the Industrial Organization of Health Care (Hyannis, Massachusetts), Universidad de Navarra, Universidad Carlos III de Madrid and University of Copenhagen for their helpful comments. Financial support from Fundación Pedro Barrié de la Maza is gratefully acknowledged. The usual disclaimers apply.

†Dpto. Economía, Métodos Cuantitativos e Historia Económica. Universidad Pablo de Olavide. Carretera de Utrera, km 1. 41013 Sevilla (Spain). Phone: +34 954 34 8380. Fax: +34 954 34 9339. E-mail: pgonzalez@upo.es

1 Introduction

This paper weighs the relative benefits of gatekeeping versus non-gatekeeping healthcare systems, within the context of the current debate over the merits of giving general practitioners (henceforth, GPs) control over patient access to specialized care. In particular, it studies how patient self-health information and patient pressure on GPs to provide them with referrals affect the efficacy of each system, and thus contribute to our assessment of its desirability.

At present, two main types of health care systems can be observed in most European countries. In countries like Italy, the Netherlands, Norway, Spain, and the United Kingdom a gatekeeping system prevails and, hence, GPs control patient access to specialized care. There are other countries, like Belgium, Finland, France, Germany and Sweden, where the gatekeeping role of the GP is very limited, as patients are free to choose between GPs and specialists.

Researchers have debated the pros and cons of these two systems for the past several years. On the one hand, it has often been argued that appointing GPs as gatekeepers to specialist care can help reduce overall health care costs.[1] For many countries in Eastern Europe, where health care is now under reform (Hebing, 1997), this perceived financial advantage seems to be driving policymakers' current interest in designing systems based on a gatekeeping strategy. Cost-containment concerns also seem to have encouraged some direct-access countries to switch over to gatekeeping-like systems in recent years.[2]

Nevertheless, there seems to be little evidence that gatekeeping actually serves to lower health care expenditures (see, e.g., Barros, 1998 and Kroneman et al., 2006). In fact, in countries such as the United States, where gatekeeping has been a central strategy in the cost-containment initiatives of managed care organizations (Wolf and Gorman, 1996), some HMOs are now relaxing the restrictions on access to specialized care (Ferris et al., 2001).

Closely related to the debate over the relative merits of these organizational models is that concerning the role of GP incentives. It has been widely recognized in the literature that the regulation of the GPs and the payment structures they face have significant implications for costs in health care systems (see, for instance, Scott, 2000). Because such incentives tend to significantly influence patient diagnosis and referral –the two decision-making processes of

[1] See, for instance, Delnoij et al. (2000), Franks et al. (1992), Martin et al. (1989) and Starfield (1994, 1996).

[2] For instance, Belgium, France and Germany recently showed initiatives to introduce a gatekeeping system. In Belgium, to stimulate the gatekeeping role of the GP patients pay lower co-payments if they register with one specific GP and accept that their individual 'global medical file' is kept by that GP. This measure was first introduced for older patients and extended to the whole population in 2002 (Schokkaert and Van deVoorde, 2005). Similarly, since July 2005 all those benefiting in France from health insurance coverage must choose their main physician ('médecin traitant'). As a result, it will cost them more to consult a specialist directly, without being referred by their 'médecin traitant' (Sandier et al., 2004). In Germany since 2004 sickness funds are obliged to offer the option to enrol in a "family physician care model" with financial incentives for patients who comply with the gatekeeping rules (Busse and Riesberg, 2004).

greatest potential impact as far as overall costs are concerned—they must be taken into account whenever gatekeeping and non-gatekeeping systems are compared.[3]

To the best of our knowledge, García-Mariñoso and Jelovac (2003) are the only to provide a uniform theoretical framework for identifying the optimal GP payment system, which they then use to compare the desirability of gatekeeping versus non-gatekeeping systems. They find that the optimal GP payment scheme would require a combination of cost-sharing components: a cost sharing of the GP treatment and a bonus for non-referral. This contract yields the right diagnosis and referral incentives to the GP. They conclude that, when GP incentives matter, gatekeeping must be viewed as superior to non-gatekeeping.

Despite the novelty of their analysis and its obvious relevance to this discussion, García-Mariñoso and Jelovac focus only on GP incentives when defining the optimal system. In particular, they disregard the role played by patient self-awareness and the effect of patient pressure on GPs for referrals in gatekeeping versus non-gatekeeping systems.

This paper aims to fill this gap by proposing a model that considers not only GP incentives but also patient behaviour as determinant of each system's relative efficacy and desirability. To do so, we consider three aspects of the patient's role. First, we acknowledge that the patient's self-health information will certainly influence our view of which system is preferable, since it may be more efficient to allow a patient with an accurate understanding of her condition to freely choose her own medical provider than it would be to require her to obtain a referral.[4] Second, we take into account that when patients can freely choose their medical provider, the design of a co-payment system to discipline patients who might strategically choose to visit a specialist or a GP should be addressed, since the extra cost to patients of co-payments may reduce the attractiveness of a direct-access system. Third, we recognize that patient pressure on GPs to refer them for specialized treatment may ultimately undermine the efficiency of a gatekeeping system. In this regard, it has been suggested that GPs are highly responsive to patient pressure as far as their referral behaviour is concerned (see Virji and Britten, 1991; Armstrong et al., 1991 and O'Donnell, 2000, among others).[5] There is also some evidence that patient expectations may play a role in many referral decisions and that GPs who feel pressured by their patients or who sense that the latter may expect a referral may, in fact, be more likely to provide one (Webb and Lloyd, 1994).

[3] There is empirical evidence that GP behavior is influenced by economic motives. See, for instance, Croxson et al. (2001) on the referrals of GPs in the UK, and Iversen and Lurås (2000) on the volume of services provided by GPs in Norway.

[4] Another issue that we do not address in this paper but that can also be a relevant criterium to assess the performance of a health care system is patient satisfaction. In this respect, a European study on patient satisfaction with GP services in 18 European countries finds that countries with gatekeeping systems show less patient satisfaction compared to direct-access countries (Kroneman et al. 2006).

[5] Fleming (1992) in a European study of referrals, reported that pressure from patients about whether they should be referred "influenced" between 30 percent and 60 percent of referrals.

In accordance with the approach that GP contracts should include appropriate diagnosis and referral incentives, we analyze how patient information and patient pressure on GPs to refer them to secondary care may affect our choice of one system over the other. The model used here is a principal-agent model, in which the health authority acts as a principal for both a patient and a GP. Two types of informational asymmetries arise in this context. First, the health authority faces a problem of moral hazard in its relationship with the GP, since neither the diagnosis nor its outcome are verifiable. Second, the patient has an informational advantage when selecting a physician, as her belief about her severity is private information.

We find that if the regulator wants to give GP incentives to making correct diagnoses and referrals, non-gatekeeping is optimal only when patient pressure on GPs is sufficiently high. This is because gatekeeping systems may not be able to provide adequate GP incentives in such high-pressure contexts. In addition, we find that non-gatekeeping systems are optimal only when the patient's information regarding her own condition is neither too bad nor too good. If a patient's signal is very uninformative, her self-referral will be very inefficient. Patient expected health losses will be very high, and so will be specialist costs, due to the high proportion of unnecessary visits to the specialist. When patient information is extremely accurate, however, non-gatekeeping is convenient for the patient and also saves some of the costs of specialized care. Nevertheless, this advantage is outweighed by the fact that in a non-gatekeeping system the accuracy of the patient's self-diagnosis fosters primary care costs. Stated more clearly, since only patients who think that their health condition is mild will visit the primary provider, this self-diagnosis is a source of information that will dissuade the GPs from making a costly diagnosis.

Although primary care is recognized as the mainstay of many healthcare systems in developed countries, theoretical work by economists into general practice is still scarce. The study that comes closest to our own –the aforementioned article by García-Mariñoso and Jelovac (2003)—diverges from ours in its conclusion that gatekeeping systems are always better than non-gatekeeping ones when GP incentives matter. By contrast, we argue that the optimality of gatekeeping systems should be qualified when the role of patient information and patient pressure on GPs to refer are also taken into account.

Brekke et al. (2007) contributes to this discussion by analyzing the competition effects amongst secondary care providers that arise when GPs are assigned a gatekeeping role. From this perspective, he also sustains that gatekeeping systems may not always be socially desirable.

Finally, the behaviour of GPs has been explored from a number of other angles. Thus, Malcomson (2004) discusses which contractual agreements are most effective to induce GPs to exert effort in diagnosis, and finds that it is counterproductive to offer GPs incentive-based contracts in contexts where patients are allowed to choose between a gatekeeper with an incentive contract and one without. Karlsson (2007), on the other hand, analyzes the desirability of

competition among GPs as an instrument for assuring the quality of primary care services. He finds that capitation systems have trouble providing GPs with appropriate incentives, since the search activity of patients offsets the direct effects of a change in the capitation rate.

The rest of the paper is organized as follows. In the following section, we present our model. In Section 3, we analyze both patient and GP behaviour. In Section 4, we derive the optimal patient co-payment levels and the optimal contractual payment scheme for GPs. In section 5 we compare the two institutional frameworks discussed above. Finally, in Section 6 we conclude.

2 The Model

Our basic set-up is in line with García-Mariñoso and Jelovac (2003). There are three agents in our economy: a patient, a GP and the regulator or health authority. In fact, there is implicitly a fourth agent: a provider of specialized medical care, but we will consider him as a passive agent, as the analysis of his behavior is out of the scope of this article.

We will now go through the three players of the model and outline their objectives, their choice variables and their knowledge.[6] Finally we will set out the timing of the game.

2.1 The patient

The patient suffers from a certain illness. The severity of the illness is measured by a random variable s. We assume that s can only take two values: $\underline{s}$ and $\bar{s}$, which indicate whether the patient is either low or high-severity. For the sake of simplicity, we assume that both types of illnesses are equally likely. The patient is perfectly aware that she is ill but does not know just how serious her illness is. Her symptoms, however, provide her with a private signal or belief about the severity of her health problem ($s^b \in \left\{\underline{s}^b, \bar{s}^b\right\}$). We assume that the probability that a patient receives a correct signal is $\beta \in \left(\frac{1}{2}, 1\right)$. Formally:

$$\Pr\left(\bar{s}^b|\bar{s}\right) = \Pr\left(\underline{s}^b|\underline{s}\right) = \beta \text{ and } \Pr\left(\bar{s}^b|\underline{s}\right) = \Pr\left(\underline{s}^b|\bar{s}\right) = 1-\beta.$$

The patient, therefore, seeks health care from a medical provider. She will demand medical attention either from a GP or from a specialist. This decision depends on the existing institutional framework. In gatekeeping the patient has no choice and has to visit the GP. In non-gatekeeping, however, the patient can choose to visit either the GP or the specialist.

We consider the patient to be endowed with a utility function that is separable in health and income. The patient's health status and her available income are the same in all contingencies (minor or major illness). Therefore, in this model, maximizing patient expected utility is equivalent to minimizing the value of her expected costs. These costs come from two main sources. First, from the health loss (l) that the patient suffers when she receives primary care and a

[6] Appendix A provides a summary of all the relevant variables of the model.

referral is necessary. These losses can be understood as the cost of waiting for specialist treatment. Secondly, the patient may also incur a monetary cost. In non-gatekeeping, where patients can freely choose their medical provider, the health authority has to set certain co-payments to induce the patient to enter the health care sector either on the primary level, or directly on the secondary one. The set of co-payments is denoted by $(p_g, p_{gs}, p_s) \in \mathbb{R}^3_+$, where p_g measures the monetary cost of visiting the GP, p_{gs} represents the cost of visiting the specialist with a GP referral and, finally, p_s is paid in case the patient decides to access specialized medical care directly.[7]

As claimed in the Introduction, there exists empirical evidence suggesting that patient pressure on GPs for referrals may alter GP behavior. We model this pressure as the probability that the patient rejects GP treatment. In case of doing so, she will demand private specialist treatment at a cost f.[8] This way, she avoids any potential health loss, although she bears the full cost of receiving specialized treatment. In particular, we assume that there is a fraction r of patients that are *obstinate*, in the sense that the GP can not convince them that they have a minor illness when they believe they have a major one. These patients always decide to reject GP treatment and pay for specialized services even if, as we will see later, it would be worthwhile for them to follow the GP's recommendation.[9]

2.2 The General Practitioner

We consider that the GP is able to cure a patient only if the severity of the health condition is low, while the specialist can heal both levels of severity. The GP, upon receiving a patient, is required to make a diagnosis in order to assess whether the patient requires specialized treatment or not. The diagnosis yields a signal about the severity of the patient's condition ($s^d \in \left\{\underline{s}^d, \bar{s}^d\right\}$) that is correct with probability $\delta \in \left(\frac{1}{2}, 1\right)$. Formally:

$$\Pr\left(\bar{s}^d|\bar{s}\right) = \Pr\left(\underline{s}^d|\underline{s}\right) = \delta \text{ and } \Pr\left(\bar{s}^d|\underline{s}\right) = \Pr\left(\underline{s}^d|\bar{s}\right) = 1 - \delta.$$

We consider that $\delta > \beta$, i.e. once the GP has made a diagnosis, his level of knowledge about the true severity of the illness exceeds that of the patient.[10] In making a diagnosis, the GP incurs

[7] These co-payments are only introduced to discipline patient behavior and, hence, they do not reflect the real cost of the service.

[8] Observe that we are ruling out the existence of a set of potential patients who decide to directly access specialist private treatment.

[9] We have chosen this modelization with obstinate patients purely for the purposes of expositional clarity. The same qualitative results hold in a more complex model where all patients are fully rational but there is heterogeneity in their cost of waiting, in such a way that those with a high waiting costs find it optimal to reject the GP's treatment recommendation. The details of this alternative are available upon request and for the convenience of the referees are attached in an Appendix not intended for publication.

[10] For simplicity, we focus on the case in which both s^d and s^b are correlated with s, but patient and GP errors are conditionally independent.

a disutility, that we denote by c_d.

As part of the diagnosis, the physician also observes the patient's belief regarding her own health condition.[11] This implies that, although patient and GP errors are conditionally independent, the GP's posterior beliefs are positively correlated with the information of the patient. Combining the diagnosis with this piece of information, the GP should decide on treating the patient or referring her to the specialist. If the GP prescribes a treatment that cures the patient, the game ends. Otherwise, the patient is referred to the specialist, bearing a health loss in those cases where the GP has not referred her directly.

As both the GP's decision to diagnose a patient and the diagnosis are hard to verify, the incentives included in the payment contract will crucially determine GP behavior.[12] Our payment contract is based on observable variables and consists of four non-negative components (D, B, T, R). First, the GP receives a fixed payment D. Secondly, he receives a budget B to purchase a range of specialist services and to cover his prescribing. Third, he incurs a cost T when providing treatment to the patient. Finally, the GP pays R when the patient is referred to the specialist. This payment structure contains: (i) a capitation component or payment per visit (D), (ii) the savings the GP makes on any cost-sharing scheme, either $(B - T)$ if he does not refer the patient to specialized care or $(B - R)$ if he does not prescribe treatment.[13] We restrict our analysis to situations where the budget allocated to the GP is sufficiently large so as to cover any potential service the patient may require. Formally, this amounts to considering that $(B - T - R) \geq 0$.[14]

2.3 The Health Authority

The third agent involved in the model is the health authority. The health authority pays the costs of the treatment provided to the patient, and also the payments made to both the GP and

[11] Alternatively, we could have assumed that the GP acquires information about the patient's belief, even if he does not make a diagnosis. However, we consider this alternative less appealing as, constructing the model that way, gatekeeping would trivially be more costly than non-gatekeeping, as the latter would be simply a subset (when $\underline{s}^b$) of the former.

[12] The remuneration methods for GPs differ across countries and experimentation with their contractual arrangements abounds. In general, the reforms depart from strict capitation or fee-for-service payments and introduce additional components aimed at containing costs and reducing referrals to hospital.

[13] This contract is in the same spirit of the fundholding scheme for GPs reintroduced in the UK in April 2005. In the UK, GPs fundholders are allocated a budget to provide primary health care and purchase some of the specialist services for which they referred patients. The "unspent" share of their budget could be reinvested in their own practice (Dusheiko et al., 2007). Also, in some Italian ASLs ("Aziende Sanitarie Locali") GP contractual arrangements combine capitation with an additional payment that rewards GPs with a proportion of the savings generated from meeting expenditure targets, including the cost of pharmaceuticals, laboratory tests and therapeutic treatments prescribed by the GP (France et al., 2005).

[14] In the UK, budgets for GPs fundholders were intended to be sufficient to buy the bundle of services which the GP's patients had consumed before the GP became a fundholder (Dusheiko et al. 2007).

the specialist.

We denote by c_s the costs of the specialist services, which include not only the treatment costs but also the payments made to the specialist. As the costs of treatment by a specialist are generally higher than the costs of treatment by a GP we normalize, without loss of generality, the latter to zero.

The health authority designs the GP contract and the patient level of co-payments so as to minimize expected social costs. Such costs are the sum of the financial costs both from primary and specialized health care (i.e. expected treatment costs and payments to both the GP and the specialist) and the patient's expected costs (which includes both her expected health losses and her monetary expenses).[15]

Our aim is to study whether it is socially useful to use patient self-awareness as a mechanism for provider selection. Hence, the level of co-payments will be designed to ensure that patients use their self-health information and visit the specialist directly only if they believe a GP will not heal them. As we are interested in providing the GP with right diagnosis and referral incentives, we focus exclusively on contracts that induce the GP to diagnose and follow the diagnosis, i.e. to treat the patient whenever the signal received from the diagnosis is $\underline{s}^d$ and refer her if $\bar{s}^d$.

We denote by C_{GP} the expected financial costs associated with primary care, C_{Sp} those for specialized treatment and C_{Pat} the patient expected costs.

2.4 Timing

The timing of the game consists of the following stages. First, the health authority sets the GP payment contract, which the GP can either accept or reject (in which case the game ends), and also sets the patient level of co-payments. Secondly, the severity of the illness is realized, and the patient seeks health care from a medical provider. If she visits the specialist the game ends. If she visits the GP, then the doctor makes a diagnosis, which provides him with a signal about the patient's severity. In the third stage, after observing the signal, the GP decides whether to treat the patient himself or to refer her to the specialist. If he decides to refer the patient, the game ends. In case he decides to treat her, she may accept or reject this treatment. If she rejects it or, in case she accepts, if the patient recovers her health, the game ends. Otherwise, the patient is referred to the specialist.

As usual, we solve the game by backward induction.

[15] Observe that: (i) the health authority internalizes the cost of the private treatment, through patient expected costs, and (ii) although co-payments appear in the model only as costs for patients, all our qualitative results would remain valid if we also include co-payments as revenues for the health authority, provided there is a cost of raising public funds.

3 Agent Behavior

In this section we characterize the behavior of the patient and the GP in our model. First, we analyze how the level of co-payments determines the decision of the patient to either visit the GP, or directly request specialist medical treatment. This analysis only applies when considering systems where patients are not obliged to compulsorily visit the GP. Second, we set out to derive the conditions that the GP payment contract has to fulfill in order to ensure that he decides to costly diagnose the patient and, afterwards, follow the diagnosis. Figure 1 provides the extensive form representation of the game under analysis.

[Insert Figure 1]

3.1 Patient Behavior

In our model, the patient can be either high-severity or low-severity, with an ex-ante equal probability. However, once the patient observes her own symptoms and is aware of her personal circumstances, she is able to update these probabilities. Then, the probabilities that the patient recognizes/misrecognizes the severity of her illness are:[16]

$$\begin{aligned} &\Pr\left(\bar{s}|\bar{s}^b\right) = \Pr\left(\underline{s}|\underline{s}^b\right) = \beta \\ &\Pr\left(\underline{s}|\bar{s}^b\right) = \Pr\left(\bar{s}|\underline{s}^b\right) = 1-\beta. \end{aligned} \tag{1}$$

In non-gatekeeping the patient has the choice between two alternatives: visit the specialist directly, or go first to the GP. If the patient visits the specialist directly, her cost is given by the co-payment she has to pay p_s, but no health loss is borne. If the patient visits first the GP she always pays p_g and, then, if she is eventually referred to the specialist p_{gs}. Moreover, she may also suffer from a health loss whenever she receives treatment from the GP that does not heal her. Those patients who are obstinate always reject GP treatment if they believe to be in a severe condition. In this case, they do not incur either p_{gs} or the health loss l, but they have to pay the private fee f.

With the help of Figure 1 patient costs in any circumstance can be easily computed.[17] Consider first belief $\underline{s}^b$. First, if the patient is low-severity ($\Pr\left(\underline{s}|\underline{s}^b\right)$) she incurs p_g if the GP's diagnosis is right ($\Pr\left(\underline{s}^d|\underline{s}\right)$), and $p_g + p_{gs}$ if the GP's diagnosis is wrong ($\Pr(\bar{s}^d|\underline{s})$). Secondly, if the patient is high-severity ($\Pr\left(\bar{s}|\underline{s}^b\right)$) she incurs $p_g + p_{gs}$ if the GP's diagnosis is right ($\Pr(\bar{s}^d|\bar{s})$) and $p_g + l + p_{gs}$ if the GP's diagnosis is wrong ($\Pr\left(\underline{s}^d|\bar{s}\right)$).

Consider now belief $\bar{s}^b$. First, if the patient suffers from a major illness ($\Pr\left(\bar{s}|\bar{s}^b\right)$) she incurs $p_g + p_{gs}$ if the GP's diagnosis is correct ($\Pr\left(\bar{s}^d|\bar{s}\right)$). If the GP's diagnosis is wrong ($\Pr(\underline{s}^d|\bar{s})$),

[16] See Appendix B for a more detailed explanation.

[17] Throughout this sub-section it is considered that the GP behaves optimally, i.e. makes a diagnosis and follows it. The payments that ensure this behaviour are computed in Sub-section 4.2.

the cost is $p_g + p_{gs} + l$ for a non-obstinate patient, and $p_g + f$ for an obstinate one. Secondly, if the patient is low-severity ($\Pr\left(\underline{s}|\bar{s}^b\right)$) but the GP correctly diagnoses ($\Pr\left(\underline{s}^d|\underline{s}\right)$), the patient incurs $p_g + f$ if she is obstinate, and p_g otherwise. Finally, if the patient is low-severity and the GP's diagnosis is wrong ($\Pr\left(\underline{s}|\bar{s}^b\right)\Pr\left(\bar{s}^d|\underline{s}\right)$) the patient incurs $p_g + p_{gs}$.

Hence, in comparing patient expected costs when demanding first GP attention, or direct specialist care, we conclude the following:

(i) If $\underline{s}^b$, a patient goes first to the GP whenever:

$p_s \geq p_g + \beta\,(1-\delta)\,p_{gs} + (1-\beta)\,(p_{gs} + (1-\delta)\,l)$.

(ii) If $\bar{s}^b$, a *non-obstinate* patient visits the specialist directly whenever:

$p_s \leq p_g + \beta\,(p_{gs} + (1-\delta)\,l) + (1-\beta)\,(1-\delta)\,p_{gs}$.

(iii) If $\bar{s}^b$, an *obstinate* patient visits the specialist directly whenever:

$p_s \leq p_g + \beta\,(\delta p_{gs} + (1-\delta)\,f) + (1-\beta)\,((1-\delta)\,p_{gs} + \delta f)$.

Taking into account that the co-payment levels have to provide appropriate incentives to any patient, we obtain the following lemma.[18]

Lemma 1 *A patient visits the specialist directly when $\bar{s}^b$ and goes to the GP when $\underline{s}^b$ if and only if:*

- $p_s - p_g \leq \beta\,(p_{gs} + (1-\delta)\,l) + (1-\beta)\,(1-\delta)\,p_{gs}$ *and*
- $p_s - p_g \geq \beta\,(1-\delta)\,p_{gs} + (1-\beta)\,(p_{gs} + (1-\delta)\,l)$.

This lemma shows that the higher the accuracy of the patient information is the milder both restrictions are. This is a natural result since what the health authority is trying to induce through the co-payments is, precisely, that the patient use her self-health information when selecting the medical provider. The more accurate the understanding of her condition is, therefore, the smaller the expected costs of her self-referral.

3.2 General Practitioner Behavior

In our model, the GP faces a population of patients that can be either high or low-severity, with ex-ante the same probability. In order to update these probabilities, the GP uses two pieces of information: the patient's beliefs and the signal received from the diagnosis.

[18] In order to avoid that no patient accepts GP treatment, the private alternative must be sufficiently costly. In particular f must exceed the patient's expected costs associated with accepting GP treatment when the patient receives signal $\bar{s}^b$ (i.e., when the patient is more interested in being referred to a specialist). Formally, this requires that $f > \frac{\beta(1-\delta)}{\beta(1-\delta)+(1-\delta)\beta}\,(l + p_{gs})$.

Once this information has been acquired, the probabilities of correctly diagnosing a low-severity are:[19]

$$\begin{array}{l} \Pr\left(\underline{s}|\underline{s}^d \cap \underline{s}^b\right) = \frac{\delta\beta}{\delta\beta+(1-\delta)(1-\beta)} = 1 - \Pr\left(\bar{s}|\underline{s}^d \cap \underline{s}^b\right). \\ \Pr\left(\underline{s}|\underline{s}^d \cap \bar{s}^b\right) = \frac{\delta(1-\beta)}{\delta(1-\beta)+(1-\delta)\beta} = 1 - \Pr\left(\bar{s}|\underline{s}^d \cap \bar{s}^b\right). \end{array} \tag{2}$$

Analogously, the probabilities of wrongly diagnosing a low-severity are:

$$\begin{array}{l} \Pr\left(\underline{s}|\bar{s}^d \cap \underline{s}^b\right) = \frac{(1-\delta)\beta}{(1-\delta)\beta+\delta(1-\beta)} = 1 - \Pr\left(\bar{s}|\bar{s}^d \cap \underline{s}^b\right). \\ \Pr\left(\underline{s}|\bar{s}^d \cap \bar{s}^b\right) = \frac{(1-\delta)(1-\beta)}{(1-\delta)(1-\beta)+\delta\beta} = 1 - \Pr\left(\bar{s}|\bar{s}^d \cap \bar{s}^b\right). \end{array} \tag{3}$$

Once the GP has diagnosed the true severity of the health condition, he then decides on the best option for the patient. The doctor always has two alternatives: treat the patient or refer her to the specialist.[20]

If the GP refers the patient to the specialist he receives the capitation payment D and keeps the unspent budget $(B-R)$. If the GP recommends treatment he always incurs the cost-sharing T on the treatment and, with a certain probability (i.e., if the patient has a major illness and needs a referral) he also incurs R. If $\underline{s}^b$, the GP receives $D+(B-T)$ if patient condition is really mild (with a probability $\Pr\left(\underline{s}|\underline{s}^d \cap \underline{s}^b\right)$ if $s^d = \underline{s}^d$, and $\Pr\left(\underline{s}|\bar{s}^d \cap \underline{s}^b\right)$ otherwise). The GP receives $D+(B-T-R)$ if the patient has a major illness. When $\bar{s}^b$, there is a probability r that the patient rejects the GP's treatment, in which case the GP loses the budget associated to this patient and receives the capitation payment D. If the patient does not reject the treatment (with a probability $(1-r)$), the GP receives $D+(B-T)$ if the patient has a minor illness (with a probability $\Pr\left(\underline{s}|s^d \cap \bar{s}^b\right)$) and $D+(B-T-R)$ if she has a major illness.

In comparing the different payments that the GP receives from prescribing either treatment or referral, we can conclude that:

(i) If $\underline{s}^d$ and $\underline{s}^b$, the GP treats the patient whenever $T \leq \frac{\delta\beta}{\delta\beta+(1-\delta)(1-\beta)}R$ and refers her otherwise.

(ii) If $\underline{s}^d$ and $\bar{s}^b$, the GP treats the patient whenever $T \leq \frac{\delta(1-\beta)(1-r)}{\delta(1-\beta)+(1-\delta)\beta}R - r\left(B-T-R\right)$ and refers her otherwise.

(iii) If $\bar{s}^d$ and $\underline{s}^b$, the GP refers the patient whenever $T \geq \frac{(1-\delta)\beta}{(1-\delta)\beta+\delta(1-\beta)}R$ and treats her otherwise.

(iv) If $\bar{s}^d$ and $\bar{s}^b$, the GP refers the patient whenever $T \geq \frac{(1-\delta)(1-\beta)(1-r)}{(1-\delta)(1-\beta)+\delta\beta}R - r\left(B-T-R\right)$ and treats her otherwise.

From these conditions we see that both T and R have to be strictly positive, which means that the GP contract should include some cost sharing both on treatment and specialist services.

[19] See Apendix B for a more detailed explanation.

[20] Throughout this sub-section it is considered that patients behave optimally, i.e. in non-gatekeeping the patient demands primary attention only if $\underline{s}^b$. The co-payments that ensure this behaviour are computed in Sub-section 4.1.

Observe also that the conditions that the GP payment scheme has to fulfill in order to effectively induce him to follow the diagnosis are different for the two institutional frameworks. In non-gatekeeping, since only patients who think that their health condition is mild visit the GP, the only relevant restrictions are (i) and (iii). In gatekeeping the four conditions matter, but it can be easily checked that the most demanding ones that actually determine GP behaviour are (ii) and (iii) This leads to the following lemma.

Lemma 2 *The GP always follows the diagnosis if and only if:*

- *In non-gatekeeping:*

$$T \geq \frac{(1-\delta)\,\beta}{(1-\delta)\,\beta + \delta\,(1-\beta)} R \text{ and} \qquad (IC^{Ngk}_{Fd_1})$$

$$T \leq \frac{\delta\beta}{\delta\beta + (1-\delta)\,(1-\beta)} R. \qquad (IC^{Ngk}_{Fd_2})$$

- *In gatekeeping:*

$$T \geq \frac{(1-\delta)\,\beta}{(1-\delta)\,\beta + \delta\,(1-\beta)} R \text{ and} \qquad (IC^{gk}_{Fd_1})$$

$$T \leq \frac{\delta\,(1-\beta)\,(1-r)}{\delta\,(1-\beta) + (1-\delta)\,\beta} R - r\,(B - T - R)\,. \qquad (IC^{gk}_{Fd_2})$$

In gatekeeping the GP faces all kind of patients. In order to ensure that the GP always follows his diagnosis, we have to induce him to do so even in those cases in which this is contrary to the patient's beliefs. As a result, the higher the referral pressure (measured by r) the more difficult to induce the GP to stick to his diagnosis as he will be more tempted to over-refer patients. In non-gatekeeping, the GP always receives patients who think they are low-severity. This implies that there is no pressure for referral, which makes the restrictions less demanding.

It can be shown that both for gatekeeping and non-gatekeeping systems, the higher the precision of the GP's diagnosis the milder the restrictions are. This effect has an intuitive interpretation as it implies that it is easier to induce the GP to follow the diagnosis as this becomes more accurate.

In our model, the GP receives neither his signal nor the patient's one until Stage 3 of the game. Before this stage, therefore, the GP has to decide whether to make a diagnosis or not, and what to do in case he does not make it (either systematically treat or refer the patient). When the GP decides to diagnose the patient, it could be the case that, afterwards, he might decide not to follow the diagnosis. The conditions written in Lemma 2 ensure that the GP will stick to his diagnosis.

The derivation of the GP's expected utility when the GP diagnoses the patient and follows the diagnosis (U), for both gatekeeping and non-gatekeeping systems, is detailed in Appendix C. The simplified structure of the GP's expected utilities is given by:

- In non-gatekeeping:

$$U^{Ngk} = D + (B - T - R) + T\left[\delta\left(1-\beta\right) + \left(1-\delta\right)\beta\right] + R\delta\beta - c_d.$$

- In gatekeeping:

$$U^{gk} = D+$$
$$\frac{1}{2}\left[(B-R) + (B-T)\,\delta\left(\beta + (1-r)\left(1-\beta\right)\right) + (B-T-R)\left(1-\delta\right)\left(1-\beta+(1-r)\,\beta\right)\right] - c_d$$

Once the GP's expected utility has been computed, we can obtain the restrictions that determine when he decides to diagnose the patient. These restrictions come from ensuring that the above stated utility is higher than both the utility the GP would obtain from systematically referring the patient ($U_R^{Ngk} = U_R^{gk} = D + (B - R)$) or from systematically treating him:

- In non-gatekeeping:

$$U_T^{Ngk} = D + (B - T - R) + R\beta$$

- In gatekeeping:

$$U_T^{gk} = D + \frac{1}{2}\left[(B-T)\left(\beta + (1-r)\left(1-\beta\right)\right) + (B-T-R)\left(1-\beta+(1-r)\,\beta\right)\right].$$

The following lemma summarizes the GP's decision of making a diagnosis.

Lemma 3 *The GP decides to make a diagnosis if and only if:*

- *In non-gatekeeping:*

$$T \geq \frac{R\left(1-\delta\right)\beta + c_d}{\left(1-\delta\right)\beta + \delta\left(1-\beta\right)} \text{ and} \qquad (IC_{Pd_1}^{Ngk})$$
$$T \leq \frac{R\delta\beta - c_d}{\delta\beta + \left(1-\delta\right)\left(1-\beta\right)}. \qquad (IC_{Pd_2}^{Ngk})$$

- *In gatekeeping:*

$$T \geq 2c_d + \left(1-\delta\right)R\left(1-\left(1-\beta\right)r\right) - \left(B-T-R\right)\left(\left(1-\delta\right)\left(1-\beta\right)+\delta\beta\right)r \text{ and} \quad (IC_{Pd_1}^{gk})$$
$$T \leq \delta R\left(1-\left(1-\beta\right)r\right) - \left(B-T-R\right)\left(\left(1-\delta\right)\beta + \delta\left(1-\beta\right)\right)r - 2c_d. \qquad (IC_{Pd_2}^{gk})$$

The conditions to induce diagnosis are, as predictable, more demanding as the cost of the diagnosis increases. As it is the case in Lemma 2, an increase in the accuracy of the diagnosis makes the conditions less demanding.

Combining Lemmas 2 and 3 we find:

Lemma 4 *If the GP decides to diagnose the patient:*

- *In non-gatekeeping he will always follow the diagnosis.*
- *In gatekeeping he will follow the diagnosis if and only if* IC_{Fd}^{gk} *are fulfilled.*

In non-gatekeeping we can ensure that the conditions that have to be fulfilled for the GP to follow the diagnosis IC_{Fd}^{Ngk} are always milder than the ones that induce him to make a diagnosis IC_{Pd}^{Ngk}. This means that, once the GP has decided to diagnose the patient, he will always follow the diagnosis. In gatekeeping, on the contrary, we cannot ensure that for every value of the cost of diagnosing, IC_{Fd}^{gk} constraints are always implied by IC_{Pd}^{gk}. Therefore, once the GP has made a diagnosis, he may decide not to use it. This is due to the referral pressure of the patient on GPs. In non-gatekeeping, since all the patients that visit the GP have $\underline{s}^b$, there is no problem of pressure at all.

The following proposition states how the referral pressure can be an unsolvable problem.

Proposition 1 *Finding a contract* (D, B, T, R) *that induces the GP to treat the patient when* $\underline{s}^d$ *and to refer her if* $\bar{s}^d$*:*

- *In non-gatekeeping it is always possible.*
- *In gatekeeping it is possible provided* $r \leq \overline{r}$*.*

  With $\overline{r} = 1 - \frac{(1-\delta)\beta}{\delta(1-\beta)}$.

Proof. See Appendix D. ■

This proposition ensures that in non-gatekeeping it is always possible to design a payment contract that induces the GP to diagnose a patient and follow the diagnosis. In gatekeeping, however, this is not the case, and the existence of such a contract depends on the intensity of patient pressure. If the referral pressure is sufficiently high, it is impossible for the health authority to find values of (D, B, T, R) that simultaneously fulfill all the constraints. The reason for this is that patient pressure on GPs for referrals generates a problem of over-referral of patients with a mild diagnosis. The value of the cost-sharing on the specialist services that would induce the GP to treat a patient whenever $\underline{s}^d$ is so high that it would eventually give incentives to the GP to treat also patients with a severe diagnosis.

Proposition 1 has shown an important implication of the presence of patient pressure for referrals. A gatekeeping system maybe unsustainable, since it may not be able to provide the GP with proper incentives to diagnose the patient and follow the diagnosis, while this is not a problem in non-gatekeeping.

It is interesting to see how the threshold $\overline{r}$ depends on the quality of the information of the agents. It can be checked that $\overline{r}$ is increasing in δ and decreasing in β. This implies that, on the one hand, the higher the accuracy of the GP's diagnosis, the more likely a gatekeeping system

is sustainable. On the other hand, as patient self-health information becomes more accurate, the maximum threshold of pressure compatible with gatekeeping decreases. As the patient has a more accurate understanding of her condition, the physician will be less willing to effectively recommend treatment when this is contrary to the patient's will.[21]

4 The Health Authority's Problem

The health authority aims at minimizing total expected social costs, computed as the sum of the financial costs: both expected costs associated with primary and secondary care (C_{GP} and C_{Sp} respectively), and patient expected costs (C_{Pat}). C_{GP}, C_{Sp} and C_{Pat} are derived formally in Appendix C. The simplified expressions are as follows:

- In gatekeeping:

$$C_{GP}^{gk} = D + \tfrac{1}{2}\left[(B-R) + (B-T)\,\delta\,(\beta + (1-r)\,(1-\beta)) + (B-T-R)\,(1-\delta)\,(1-\beta+(1-r)\,\beta)\right].$$
$$C_{Sp}^{gk} = \tfrac{c_s}{2}\,(2-\delta-r\,(1-\delta)\,\beta).$$
$$C_{Pat}^{gk} = \tfrac{1}{2}\left[(1-\delta)\,((1-\beta)\,rl + (1-r)\,l) + rf\,((1-\beta)\,\delta + (1-\delta)\,\beta)\right].$$

- In non-gatekeeping:

$$C_{GP}^{Ngk} = \tfrac{1}{2}\left[D + (B-T-R) + T\left[\delta + (1-2\delta)\,\beta\right] + R\delta\beta\right].$$
$$C_{Sp}^{Ngk} = \tfrac{c_s}{2}\,(2-\delta\beta).$$
$$C_{Pat}^{Ngk} = \tfrac{1}{2}\,(p_s + p_g + p_{gs}\,(\beta\,(1-\delta) + 1 - \beta) + l\,(1-\beta)\,(1-\delta)).$$

The problem of the health authority can be analyzed in two steps. First, if the system is a non-gatekeeping one, the health authority has to design the set of co-payments that induces the patient to visit a specialist directly if and only if he believes he is high-severity. Secondly, the health authority has to design the contract that provides the GP with incentives to make (and follow) a diagnosis.[22]

4.1 The Optimal Co-payment Levels

The co-payment levels set by the health authority will be the ones that minimize C_{Pat}. The health authority has to take into account the constraints computed in Lemma 1, which ensure that the patient only visits the specialist directly when $\bar{s}^b$, as well as the fact that co-payments have to be non-negative.

[21] In spite of this, since $\beta < 1$ and $\delta > \beta$, we can always ensure that $\overline{r} > 0$. This means that even if patient information is extremely accurate, there always exist levels of pressure compatible with gatekeeping.

[22] The problem can be solved in two steps since: (i) As long as the GP diagnoses the patient and follows the diagnosis C_{Pat} is independent from the GP contract; (ii) C_{GP} is not altered by the co-payment levels, provided they induce the patient to select the medical provider according to her belief about the severity; (iii) C_{Sp} depends only on the institutional framework.

The health authority's optimization program is as follows:

$$\min_{p_g,p_{gs},p_s} C_{Pat} = \tfrac{1}{2}\left(p_s + p_g + p_{gs}\left(\beta\left(1-\delta\right)+1-\beta\right)+l\left(1-\beta\right)\left(1-\delta\right)\right)$$
$$s.t \quad \begin{cases} p_s - p_g \leq \beta\left(p_{gs}+\left(1-\delta\right)l\right)+\left(1-\beta\right)\left(1-\delta\right)p_{gs} \\ p_s - p_g \geq \beta\left(1-\delta\right)p_{gs}+\left(1-\beta\right)\left(p_{gs}+\left(1-\delta\right)l\right) \\ p_g \geq 0,\ p_{gs} \geq 0,\ p_s \geq 0, \end{cases} \tag{4}$$

The following proposition characterizes the optimal level of co-payments.

Proposition 2 *If the health authority wants the patient to visit the specialist directly when* $\bar{s}^b$ *and to go first to the GP if* $\underline{s}^b$*, the optimal level of co-payments* $\left(p_g^*, p_{gs}^*, p_s^*\right)$ *is such that* $p_g^* = p_{gs}^* = 0$ *and* $p_s^* = \left(1-\beta\right)\left(1-\delta\right)l$*.*

Proof. See Appendix D. ■

This proposition shows that setting only $p_s > 0$ is enough to induce patients to follow their belief when choosing their medical provider. This co-payment structure is in line with the very recent Belgian and French reforms, aimed at enhancing the gatekeeping role of GPs. In both countries the co-payments are larger for those patients who go to the specialist directly, without a referral.[23]

It is worth noting that this simple structure for the optimal co-payments relies on the fact that in our model the patient is endowed with a linear utility in money, so income effects are absent and co-payments do not interfere with financial insurance issues.[24]

Finally, it is straightforward to see that the patient always benefits from a higher accuracy in the understanding of her condition. The reason is two-fold: First, the health losses she bears are lower, as her self-selection of medical provider is more likely to be correct. Secondly, the monetary expenses she faces also diminish, as the co-payments needed to induce her an appropriate selection of medical provider are decreasing in the accuracy of her belief.

4.2 The Optimal Payment Contract

The payments offered to the GP will be the ones that minimize C_{GP}. The health authority has to consider the fact that the GP's expected utility (U) cannot be lower than his reservation utility (normalized to zero) (PC), and also that his liability constraints have to be fulfilled (LLC). We do the analysis within this framework with limited liability constraints for the doctor, i.e. we impose that, under any circumstance, the doctor must receive a positive payment. Such a

[23] In these countries co-payments also have a dissuasive purpose and, therefore, there is a positive (though small) level of co-payment for visiting the GP or the specialist with a GP's referral. This is not necessary in our model as we do not deal with healthy individuals who make unnecessary visits to the system.

[24] García-Mariñoso (1999) provides a description of how the insurer can regulate access to specialist care by manipulating the patients' insurance contract in a model with income effects.

restriction reflects the existing limitations on the public liabilities that can be imposed on a doctor in the execution of his professional duties, which arise from the fact that the result of any medical treatment is, to a certain extent, unpredictable.

On top of this, we must include the GP's incentive compatibility constraints (IC) in the health authority's optimization program. These are the restrictions that induce the GP to diagnose the patient and follow the diagnosis (defined in Lemmas 2 and 3).

The health authority's optimization program is as follows:

$$\min_{D,B,T,R} \; C_{GP}$$
$$s.t \quad \begin{cases} U \geq 0 & PC \\ D \geq c_d & LLC_1 \\ T \geq 0 & LLC_2 \\ R \geq 0 & LLC_3 \\ B \geq T + R & LLC_4 \\ IC & \end{cases} \tag{5}$$

With $C_{GP} \in \left\{C_{GP}^{Ngk}, C_{GP}^{gk}\right\}$ and $IC \in \left\{\left(IC_{Pd_1}^{Ngk}, IC_{Pd_2}^{Ngk}\right), \left(IC_{Pd_1}^{gk}, IC_{Pd_2}^{gk}, IC_{Fd_1}^{gk}, IC_{Fd_2}^{gk}\right)\right\}$, depending on whether we are in non-gatekeeping or gatekeeping.

From Proposition 1 we know that, in gatekeeping, designing a contract that induces the GP to diagnose the patient and to follow the diagnosis, is only possible provided the referral pressure is not too high. Hence, hereinafter, we restrict our analysis to values of r such that $r \leq \overline{r}$.

Let us define $\widetilde{r} \equiv \frac{\delta-\beta}{(1-\beta)(\delta-\beta+2\delta\beta)} \in (0, \bar{r})$. This threshold will determine two regions in which the impact of the referral pressure on GPs affects the costs borne by the health authority differently.

The following proposition characterizes the GP's optimal payment contract.

Proposition 3 *If the health authority wants the GP to treat the patient when $\underline{s}^d$ and to refer her if $\bar{s}^d$ the optimal contract (D, B, T, R) is as follows:*

- *In non-gatekeeping:*

$$D^{Ngk} = c_d$$
$$B^{Ngk} = T^{Ngk} + R^{Ngk}$$
$$T^{Ngk} = \frac{c_d}{(2\delta - 1)(1 - \beta)}$$
$$R^{Ngk} = \frac{c_d}{(2\delta - 1)(1 - \beta)\beta}.$$

The health authority's expected primary care costs are:

$$C_{GP}^{Ngk} = \frac{c_d}{2(2\delta - 1)}\left[4\delta + \left(\frac{\beta}{1-\beta} - 1\right)\right].$$

- *In gatekeeping:*

$$\begin{aligned} D^{gk} &= c_d \\ B^{gk} &= T^{gk} + R^{gk} \\ T^{gk} &= \frac{2c_d\left((2\delta-1)+2\left(1-\delta\right)\Gamma\left(\delta,\beta,r\right)\right)}{2\delta-1} \\ R^{gk} &= \frac{4c_d\Gamma\left(\delta,\beta,r\right)}{\left(2\delta-1\right)\left(1-\left(1-\beta\right)r\right)}. \end{aligned}$$

The health authority's expected primary care costs are:

$$C_{GP}^{gk} = \frac{c_d}{2\delta-1}\left[4\delta+2\left(\Gamma\left(\delta,\beta,r\right)-1\right)\right].$$

With $\Gamma\left(\delta,\beta,r\right) = \begin{cases} 1 \text{ if } r \leq \widetilde{r}. \\ \frac{(2\delta-1)(1-(1-\beta)r)(\delta(1-\beta)+(1-\delta)\beta)}{2[(1-(1-\beta)r)(\delta-(1-\delta)(\delta(1-\beta)+(1-\delta)\beta))-\delta\beta]} > 1 \text{ otherwise.} \end{cases}$

Proof. See Appendix D. ■

The structure of the GP's optimal payment contract is similar in gatekeeping and non-gatekeeping. First, in the worst possible contingency, the GP receives a capitation payment or a payment per visit that covers the cost of making a diagnosis ($D^{Ngk} = D^{gk} = c_d$). Second, to avoid systematic treatment of the patient, the contract includes some cost-sharing of the GP's treatment (both T^{Ngk} and T^{gk} are strictly positive). Third, to avoid systematic referral the contract also imposes some cost-sharing of the specialist costs (both R^{Ngk} and R^{gk} are strictly positive). Finally, the budget allocated to the physician (B) should be just enough to cover the expenses in which the GP may incur.[25]

We focus now on analyzing how the health authority's expected primary costs are affected by our relevant variables (δ, r and β). As expected, both in gatekeeping and non-gatekeeping these costs are decreasing in the accuracy of the GP's diagnosis. As far as patient pressure on GPs for referrals is concerned**,** we find that such pressure raises the payments the health authority has to allocate to the physician in order to avoid an excessive number of referrals in a gatekeeping system. To be more precise, the savings the GP makes if he does not refer the patient to specialized care ($B^{gk} - T^{gk}$), are increasing in r. Nevertheless, we find that, for values of pressure below $\widetilde{r}$, the referral pressure does not affect the health authority's expected primary costs. The reason is that the increase in the payments to the GP is compensated by the fact that such payments are paid less often at equilibrium. For values beyond $\widetilde{r}$, the increase in the payments needed to provide the GP with proper incentives is so high that it always affects primary care costs. As a result, these costs will be higher the larger the value of r.

Finally, the quality of the patient's information unambiguously raises expected costs from primary care. The reason, however, is of a different nature in non-gatekeeping and gatekeeping.

[25] The qualitative insights of the optimal contract are similar to the ones in García-Mariñoso and Jelovac (2003).

In non-gatekeeping patient information generates a problem of "diagnosis substitution". The fact that only those patients who think that their health condition is mild visit the GP fosters GP's incentives to use the patient's self-diagnosis as a substitute of his own diagnosis and systematically treat the patient. Hence, inducing the GP to make a diagnosis becomes extremely expensive. In gatekeeping, patient information increases health costs through the referral pressure. For levels of pressure above $\widetilde{r}$, the more accurate the patient's information is, the more difficult that the GP decides to follow the diagnosis. This makes it more costly for the health authority to avoid an excessive number of referrals.

5 On the Choice of the Optimal System

In this section we provide a discussion on the global picture the health authority faces when choosing the best organizational system to access health care. As it has become clear throughout the paper, the role played by patient self-awareness, as well as their pressure on GPs for referrals, are the two key elements that will drive the health authority choice between gatekeeping and non-gatekeeping systems.

5.1 Partial Comparisons

As a first step we confront gatekeeping and non-gatekeeping focusing separately on each of the components of the health authority's expected costs: patient costs, primary care costs and specialized costs.

First, focusing only on the patient's side of the problem, we find:

Proposition 4 *There exists a threshold $\beta^* < 1$, such that:*

- *If $\beta \leq \beta^*$, gatekeeping generates lower patient expected costs than non-gatekeeping.*
- *If $\beta > \beta^*$, non-gatekeeping generates lower patient expected costs than gatekeeping.*

Proof. See Appendix D. ■

When we concentrate only on patient expected losses, non-gatekeeping may be the optimal system to access medical care. The reason is clear as when patients can freely choose their medical provider, the health authority relies on their information. As the quality of the patient's belief increases, the self-selection becomes perfect and the costs associated with this system converge to zero. In gatekeeping, on the contrary, as we force patients to disregard their own belief and always access primary care we do not profit completely from their more accurate understanding.

Considering only the GP's side of the problem, we get:

Proposition 5 *Focusing only on primary care expected costs:*

- *If* $\beta < \frac{1+4\delta}{2+4\delta}$ *non-gatekeeping is preferred to gatekeeping.*
- *If* $\beta \geq \frac{1+4\delta}{2+4\delta}$ *there exists a threshold* $r^* \in [\tilde{r}, \overline{r}]$ *such that:*
  - *If* $r \leq r^*$ *gatekeeping is preferred to non-gatekeeping.*
  - *If* $r > r^*$ *non-gatekeeping is preferred to gatekeeping.*

Proof. See Appendix D. ■

Proposition 5 illustrates the trade-off the health authority faces in terms of primary care costs. It can be checked that, whenever a patient visits the GP, primary care costs are always smaller in gatekeeping systems, provided the problem of pressure is not severe. However, these costs are incurred less often in non-gatekeeping, as not all the patients visit the GP. For this reason, we find that there exist values of β for which non-gatekeeping is less costly, even in the absence of pressure.

As patient self-health information becomes more accurate, the GP's incentives to skip the diagnosis increase, which raises the costs of non-gatekeeping. When β is sufficiently high, then, gatekeeping is superior, unless there is a sufficiently important problem of pressure for referral, i.e. if $r > r^*$.

The above discussion provides some hints suggesting that a more accurate patient belief can be problematic in non-gatekeeping systems. This is reinforced by the following corollary.

Corollary 1 r^* *is increasing in* β.

Patient self-health information generates a problem of informational substitution that is present only in non-gatekeeping. As the accuracy of her information increases it is more likely to be in the region where gatekeeping is superior.

Finally, in terms of expected secondary costs it is direct to see that:

Proposition 6 *Expected secondary costs are never lower in non-gatekeeping. Moreover:*

- *The higher is* β*, the closer the expected specialized costs in both systems.*
- *The higher is* r*, the lower the expected specialized costs in gatekeeping relative to non-gatekeeping.*

Proof. See Appendix D. ■

Even if, in general, gatekeeping allows savings in specialist costs, the more accurate the patient's understanding of her condition, the closer the costs in the two systems. The improvement in the accuracy of self-referrals reduces the over-utilization of specialist treatment in non-gatekeeping. On the contrary, a higher referral pressure on GPs implies more patients leaving the public sector, what reduces the expected specialist costs in gatekeeping.

5.2 Global Comparison

This sub-section combines the previous results and provides an overall comparison of gatekeeping and non-gatekeeping when both financial costs (GP and specialist costs) and patient costs are simultaneously considered.

We start by considering the two extreme situations concerning the accuracy of the patient's information.

When the patient's signal is extremely uninformative $\left(\beta \to \frac{1}{2}\right)$ both patient expected health losses and co-payments are higher in non-gatekeeping and, hence, from the patient's point of view gatekeeping dominates. From the GP's incentives point of view, however, non-gatekeeping generates lower costs ($C_{Gp}^{Ngk} < C_{Gp}^{gk}$ if $\beta \to \frac{1}{2}$), but only because the GP is visited less often. Nevertheless, gatekeeping saves with respect to non-gatekeeping in terms of unnecessary visits to the specialist ($C_{Sp}^{gk} < C_{Sp}^{Ngk}$ if $\beta \to \frac{1}{2}$). Moreover, this difference in specialist costs will outweigh any saving that non-gatekeeping may yield in terms of primary care costs, provided specialized treatment is sufficiently more costly than primary care. Therefore, in general, systems where patients freely choose their medical provider would not dominate.

When patient information is extremely accurate ($\beta \to 1$) there are strong arguments in favor of non-gatekeeping. First, since patients make no mistakes when selecting their medical provider, at equilibrium they bear no health losses and the co-payments they pay become negligible. Secondly, there is not an over-utilization of specialist services, as no low severe patient misinterpret her symptoms. A high accuracy of the patient's information, however, has perverse effects on GP behavior and these are more severe in non-gatekeeping. In particular, when $\beta \to 1$ inducing the GP to diagnose the patient and follow the diagnosis becomes prohibitively expensive when patients can freely choose their provider. This makes that, overall, gatekeeping dominates.[26]

For intermediate parameter values, the optimal choice will depend on the relative strength of two opposite effects. On the one hand, non-gatekeeping, even if it allows to successfully use patient information, will generate a substitution of GP's diagnosis by patient self-health information. On the other hand, gatekeeping suffers from the problem of patient pressure on GPs for referrals, that may even make a successful process of diagnosis and treatment/referral choice impossible.

We summarize the discussion above in the following corollary.

Corollary 2 *If the health authority wants the GP to diagnose the patient and follow the diagnosis, and the patient to adequately select her medical provider, then:*

[26] We should recall that, as quality of the patient's belief increases, the set of values for the pressure that make it impossible to sustain diagnosis and treatment/referral in gatekeeping also increases. However, we have already shown that for every value of $\beta < 1$, it holds that $\bar{r} > 0$, i.e. there always exist levels of pressure compatible with a gatekeeping system.

- *If $r > \overline{r}$, gatekeeping is unsustainable.* ***Non-gatekeeping*** *dominates.*
- *If $r \leq \overline{r}$, then:*
  - *When $\beta \to \frac{1}{2}$* ***gatekeeping*** *dominates.*
  - *For intermediate values of β there exists a threshold in the level of patient pressure such that, for values below it,* ***gatekeeping*** *dominates whereas, for values above it, the optimal system is a* ***non-gatekeeping*** *one.*
  - *When $\beta \to 1$* ***gatekeeping*** *dominates.*

Finally, it would also be interesting to study how the choice depends on the GP's diagnosis accuracy. In general, what one would expect is that the higher the precision of the GP's diagnosis, the more efficient a system with compulsory visits to the GP is, as GP's information is socially more valuable and allows to decrease the expected number of unnecessary visits to the specialist ($C_{Sp}^{gk} < C_{Sp}^{Ngk}$ if $\delta \to 1$). This argument is reinforced in our model as the more accurate the diagnosis is, the milder the agency problem the health authority faces when contracting with the GP.

5.3 Discussion

At this point, it seems appropriate to briefly review the case at hand. Our analysis suggests that, if the goal is to provide GPs with incentives to diagnose patients and refer them only when necessary, a non-gatekeeping system is optimal only when (i) patient pressure on GPs to refer them for specialized care is sufficiently high, and (ii) the quality of the patient's self-health information is neither very inaccurate (in which case the patient's self-referral will be very inefficient) nor very accurate (in which case the GP will be strongly tempted to skip the diagnosis and systematically treat).

This result points to an apparent contradiction: when patient self-health information is very accurate, it is not worth using. However, such a paradoxical recommendation can only be made because we have restricted our analysis to those situations for which GP incentives matter, which makes it impossible to design a contract that combines both the patients' prior knowledge and the physician's eventual diagnosis. As a result of this impossibility, in non-gatekeeping the patient's self-diagnosis substitutes that of the GP instead of complementing it.

An alternative reading of the above result is that it is impractical to provide GP incentives when patient knowledge is highly accurate, and when the health authority wishes to profit from it. In fact, the best approach in this case would be to use the patient's self-health information instead of (and not merely in addition to) the GP's diagnosis, for two reasons. First, when a patient knows a great deal about her own state of health, the added value of the physician's diagnosis is small because the probability that the patient will misinterpret her symptoms is

very low. Second, providing the GP with incentives to diagnose is very costly. In fact, it can be shown that, when compared against any other alternative, a system in which patients are free to choose their medical provider and only receive the GP's referral after a failed treatment will always dominate. It would allow the health authority to benefit completely from the patient's knowledge, while eliminating the financial burden of giving incentives to GPs.

6 Concluding Remarks

We have developed a principal-agent model in which the health authority acts as a principal for both a patient and a general practitioner, and have employed it to evaluate the appropriateness of using the latter as a gatekeeper to secondary care. We have followed the conventional wisdom in the literature that GP contracts should include appropriate diagnosis and referral incentives. In this setting, we have shown that patient self-health information and referral pressure on GPs alter our outcome as to which system offers the best strategy for accessing specialist care. Specifically, we have demonstrated that these two factors directly affect the expected health care costs of patients and indirectly affect the diagnosis and referral behaviour of the GP, and thus have an impact on the expected primary and secondary costs of care.

In terms of policy recommendations, our analysis suggests that whenever GP incentives matter, which system we choose depends on the relative importance of two factors: patient pressure on GPs to provide referrals for secondary care (in the gatekeeping system) and the use of patient self-health information as a substitute for the GP's diagnosis (in the non-gatekeeping one). In general, non-gatekeeping is optimal only if the pressure to refer is sufficiently high and if the patient's information is neither very accurate nor very inaccurate. Two factors play a role here. On the one hand, if the patient's signal is highly uninformative, her self-referral will be very inefficient. However, if her self-health information is extremely accurate, the benefits of using a non-gatekeeping system will be outweighed by the high expense of providing GPs with incentives to diagnose. In this case, a health authority that wishes to put patient self-health information to efficient use may not find it worthwhile to encourage the GP to screen out the less serious cases. On the contrary, the authority may find it more efficient to allow patients to freely choose their medical providers, calling on the GP's referral only if treatment fails.

This insight opens up a potentially fruitful path of research centred on the potential substitutability and/or complementarity of the health information provided by patients and GPs, respectively, which would further our knowledge about the shape of optimal contractual agreements for GPs.

One potential criticism to our work is that the quality of the patient's self-health information may be difficult for health authorities to discern. However, one would expect such information to be more accurate among patients whose condition is chronic or inherited, recurrent, or has easily

recognizable symptoms than among those with other types of pathologies. And it is precisely for these types of illnesses with clear symptoms where the comparison between gatekeeping and non-gatekeeping becomes relevant. For diseases with more diffuse symptoms, where patients are much more likely to choose the wrong specialist, gatekeeping would clearly offer the additional advantage of favouring a correct match between patients and specialists.

For this article, we chose to focus on a single-physician framework, avoiding contexts in which GPs compete over patients. For such a setting, an alternative to patient pressure would be the patient's option to obtain the desired referral from a competing GP. The above analysis ruled out this possibility in order to keep our model analytically tractable. Nevertheless, all of our qualitative insights regarding the effects of patient pressure on GPs for referrals would remain valid in a model characterized by GP competition provided: (i) the patient bears a cost of switching to another GP (such as a searching cost), and (ii) the GP's revenues diminish whenever he loses a patient.

Finally, we would like to highlight that, although primary care is recognized as the basis of health care systems in many developed countries, to date economists have done little theoretical research into general practice. We hope this study will open the door to further research into that area, and that it will also stimulate the ongoing debate over the pros and cons of encouraging general practitioners to take on a gatekeeping role. Both theoretically and empirically, further research is clearly needed before we can accurately assess the relevance of the relationship between patient self-health information, patient pressure on GPs to refer and GP incentives.

References

[1] Armstrong, D., Fry, J. and Armstrong, P. (1991) "Doctors' Perception of Pressure from Patients for Referral". *British Medical Journal* 302: 1186-1188.

[2] Barros, P.P., (1998) "The black box of health care expenditure growth determinants". *Health Economics* 7 (6), 533-544.

[3] Bradley, C.P. (1992) "Uncomfortable prescribing decisions: a critical study". *British Medical Journal* 304: 294-296.

[4] Brekke, K.R., Nuscheler, R. and Straume, O.R. (2007). "Gatekeeping in Health Care". *Journal of Health Economics*, 26 (1): 149-170.

[5] Busse, R. and Riesberg,A. (2004). Health Care Systems in Transition: Germany. Copenhagen, WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

[6] Croxson, B., Propper, C. and Perkins, A. (2001). "Do Doctors Respond to Financial Incentives? UK Family Doctors and the GP Fundholder Scheme". *Journal of Public Economics* 79: 375-398.

[7] Delnoij, D., Van Merode, G., Paulus, A. and Groenewegen, P. (2000) "Does General Practitioner Gatekeeping Curb Health Care Expenditure". *Journal of Health Services Research and Policy* 5(1): 22-26.

[8] Dusheiko M., Gravelle, H., Yu, N. and Campbell, S. (2007)."The impact of budgets for gatekeeping physicians on patient satisfaction: Evidence from fundholding". *Journal of Health Economics* 26:742-762.

[9] Ferris, T.G, Chang, Y., Blumenthal, D. and Pearson, S.D. (2001). "Leaving gatekeeping behind - effects of opening access to specialists for adults in a Health Maintenance Organization". *New England Journal of Medicine* 345: 1312-1317.

[10] Fleming, D. (1992) "The Interface between General Practice and Secondary Care in Europe and North America". In: Roland, M. and A. Coulter. Hospital Referrals. Oxford General Practice Services 22. Oxford University Press. Oxford.

[11] France, G., Taroni, F., and Donatini, A. (2005). The Italian health-care system. *Health Economics*, 14, 187-202.

[12] Franks, P., Clancy, C.M. and Nutting, P.A. (1992) "Gatekeeping Revisited-Protecting Patients from Overtreatment". *New England Journal of Medicine* 327:424-429.

[13] García-Mariñoso, B. and Jelovac, I. (2003) "GPs' Payment Contracts and their Referral Practice". *Journal of Health Economics,* 842: 1-19.

[14] García-Mariñoso, B. (1999) "Optimal Access to Hospitalized Attention from Primary Health Care". Discussion Paper 9.907, the Economics Research Center, University of East Anglia.

[15] Hebing, E.H. (1997) "Health Care Reforms in Central and Eastern Europe: Dutch Contributions". In: Schrijvers AJP editor. Health and Health Care in the Netherlands. Utrecht: De Tijdstrom.

[16] Iversen, T. and Lurås, H. (2000) "Economic Motives and Professional Norms: the Case of General Medical Practice". *Journal of Economic Behavior and Organization* 43: 447-470.

[17] Karlsson, M. (2007)."Quality Incentives for GPs in a Regulated Market". Journal of Health Economics 26(4): 699-720.

[18] Kroneman, M.W, Maarse, H. and van der Zee, J. (2006) "Direct Access in Primary Care and Patient Satisfaction: A European Study". *Health Policy* 76. 72-79.

[19] Malcomson, J.M. (2004) "Health Service Gatekeepers". *RAND Journal of Economics* 35-2: 401-421.

[20] Martin, D., Marinker, M. and Pereira Gray, D. (1989) "Effect of a Gatekeeper Plan on Health Services Use and Charges: A Randomized Trial". *American Journal of Public Health* 79: 1628-1632.

[21] O' Donnell, C.A. (2000) "Variation in GP referral rates: what can we learn form the literature" Family Practice 17(6): 462-471.

[22] Sandier S., Paris, V. and Polton, D. (2004) "Health care systems in transition: France". Copenhagen, WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

[23] Scott, A. (2000) "Economics of General Practice". In: Culyer, A.J. and Newhouse, J.P., eds., Handbook of Health Economics (Elsevier, Amsterdam). Chapter 22.

[24] Schokkaert, E. and Van de Voorde, C. (2005) "Health Care Reform in Belgium". *Health Economics* 14: S25-S39.

[25] Starfield, B. (1994) "Is Primary Health Care Essential?" *The Lancet* 344:1129-1133.

[26] Starfield, B. (1996) "Is Strong Primary Care Good for Health Outcomes?" In: Griffin J., editor. The Future of Primary Care. London: Office of Health Economics.

[27] Virji, A. and Britten, N. (1991) "A study of the relationship between patients' attitudes and doctors' prescribing". *Family Practice* 8: 314-319.

[28] Webb, S. and Lloyd, M. (1994) "Prescribing and Referral in General Practice: a study of patients expectations and doctors' actions". *British Journal of General Practice* 44: 165-169.

[29] Wolf, L.F. and Gorman, J.K. (1996) "New Directions and Developments in Managed Care Financing". *Health Care Financing Review* 17(3): 1-5.

Appendixes:

Appendix A. Summary of Notation.

$s \in \{\underline{s}, \bar{s}\}$	True severity of the illness.
$s^b \in \{\underline{s}, \bar{s}\}$	Patient belief.
$\beta \in \left(\frac{1}{2}, 1\right)$	Accuracy of patient belief.
$l > 0$	Health loss if patient receives primary care and a referral is necessary.
$f > 0$	Cost of the private treatment alternative.
$r \in (0, 1)$	Proportion of *obstinate* patients (rate of referral pressure).
(p_g, p_{gs}, p_s)	Set of co-payments: p_g when visiting the GP. p_{gs} when visiting the specialist with a GP's referral. p_s when visiting directly the specialist.
$s^d \in \{\underline{s}, \bar{s}\}$	GP's diagnosis outcome.
$\delta \in (\beta, 1)$	Accuracy of GP's diagnosis.
$c_d > 0$	Cost of GP's diagnosis.
(D, B, T, R)	GP's contract: D Capitation (or fee for service) payment. B Budget allocated to purchase services for the patient. T Cost borne when treating the patient. R Cost borne when referring the patient.
$c_s > 0$	Cost of the specialist services.
C_{Pat}	Patient expected costs.
C_{GP}	Expected costs associated with primary care.
C_{Sp}	Expected costs associated with specialist treatment.

Appendix B. GP and Patient updated probabilities.

Let us consider three random variables s, s^d and s^b, such that s, s^d, $s^b \in \{\bar{s}, \underline{s}\}$.

Both s^d and s^b are correlated with s. However, we consider that patient and GP errors are not correlated.

In general, $\forall i, j \in \{\bar{s}, \underline{s}\}$ it is true that:

$$\Pr\left(s = i | s^b = i\right) \frac{\Pr\left(s^b = i | s = i\right) \Pr\left(s = i\right)}{\Pr\left(s^b = i | s = i\right) \Pr\left(s = i\right) + \Pr\left(s^b = i | s = j\right) \Pr\left(s = j\right)}.$$

Then, (1) follows directly.

It is also true that $\forall i, j \in \{\bar{s}, \underline{s}\}$:

$$\Pr\left(s = i | s^d = i \cap s^b = j\right) = \frac{\Pr(s = i)\Pr\left(s^d = i \cap s^b = j | s = i\right)}{\Pr(s = i)\Pr\left(s^d = i \cap s^b = j | s = i\right) + \Pr(s = j)\Pr\left(s^d = i \cap s^b = j | s = j\right)}.$$

Moreover,

$$\Pr\left(s^d = i \cap s^b = j | s = i\right) = \Pr\left(s^d = i | s = i\right)\Pr\left(s^b = j | s = i\right).$$

From here it is straightforward to derive the expressions in (2) and (3).

Appendix C. GP's expected utility, health authority's expected financial costs and patient expected costs.

In gatekeeping:

GP's expected utility:

$$\begin{aligned} &U^{gk} = D + \Pr(\underline{s})\left[\Pr\left(\underline{s}^d \cap \underline{s}^b | \underline{s}\right)(B - T) + \Pr\left(\underline{s}^d \cap \bar{s}^b | \underline{s}\right)(1 - r)(B - T) + \right. \\ &\left.\left(\Pr\left(\bar{s}^d \cap \underline{s}^b | \underline{s}\right) + \Pr\left(\bar{s}^d \cap \bar{s}^b | \underline{s}\right)\right)(B - R)\right] + \Pr(\bar{s})\left[\left(\Pr\left(\bar{s}^d \cap \underline{s}^b | \bar{s}\right) + \Pr\left(\bar{s}^d \cap \bar{s}^b | \bar{s}\right)\right)(B - R)\right. \\ &\left. + \Pr\left(\underline{s}^d \cap \underline{s}^b | \bar{s}\right)(B - T - R) + \Pr\left(\underline{s}^d \cap \bar{s}^b | \bar{s}\right)(1 - r)(B - T - R)\right] - c_d = \\ &D + \tfrac{1}{2}\left[(B - R) + (B - T)\,\delta\left(\beta + (1 - r)(1 - \beta)\right) + (B - T - R)(1 - \delta)\left(1 - \beta + (1 - r)\beta\right)\right] - c_d \end{aligned}$$

Health authority's expected primary care costs:

$$\begin{aligned} &C_{GP}^{gk} = U^{gk} + c_d = \\ &D + \frac{1}{2}\left[(B - R) + (B - T)\,\delta\left(\beta + (1 - r)(1 - \beta)\right) + (B - T - R)(1 - \delta)\left(1 - \beta + (1 - r)\beta\right)\right]. \end{aligned}$$

Health authority's expected specialized care costs:

$$\begin{aligned} &C_{Sp}^{gk} = \left[\Pr(\bar{s})\left(1 - r\Pr\left(\underline{s}^d \cap \bar{s}^b | \bar{s}\right)\right) + \Pr(\underline{s})\Pr\left(\bar{s}^d | \underline{s}\right)\right] c_s = \\ &\tfrac{c_s}{2}\left(2 - \delta - r(1 - \delta)\beta\right). \end{aligned}$$

Finally, patient expected costs:

$$\begin{aligned} &C_{Pat}^{gk} = \Pr(\underline{s})\Pr\left(\bar{s}^b \cap \underline{s}^d | \underline{s}\right) rf + \\ &\Pr(\bar{s})\left[\Pr\left(\bar{s}^b \cap \underline{s}^d | \bar{s}\right)(rf + (1 - r)\,l) + \Pr\left(\underline{s}^b \cap \underline{s}^d | \bar{s}\right) l\right] = \\ &\tfrac{1}{2}\left[(1 - \delta)\left((1 - \beta)\,rl + (1 - r)\,l\right) + rf\left((1 - \beta)\,\delta + (1 - \delta)\,\beta\right)\right]. \end{aligned}$$

In non-gatekeeping:

GP's expected utility:

$$\begin{aligned} &U^{Ngk} = D + \Pr\left(\underline{s} | \underline{s}^b\right)\left[\Pr\left(\underline{s}^d | \underline{s}\right)(B - T) + \Pr\left(\bar{s}^d | \underline{s}\right)(B - R)\right] + \\ &\Pr\left(\bar{s} | \underline{s}^b\right)\left[\Pr\left(\bar{s}^d | \bar{s}\right)(B - R) + \Pr\left(\underline{s}^d | \bar{s}\right)(B - T - R)\right] - c_d = \\ &D + (B - T - R) + T\left[\delta(1 - \beta) + (1 - \delta)\beta\right] + R\delta\beta - c_d. \end{aligned}$$

Health authority's expected primary care costs:

$$C_{GP}^{Ngk} = \Pr\left(\underline{s}^b\right)\left(U^{Ngk} + c_d\right) = D + (B - T - R) + T\left[\delta\left(1-\beta\right) + \left(1-\delta\right)\beta\right] + R\delta\beta.$$

Health authority's expected specialized care costs:

$$C_{Sp}^{Ngk} = \left[\Pr\left(\bar{s}\right) + \Pr\left(\underline{s}\right)\left(\Pr\left(\bar{s}^d \cap \underline{s}^b | \underline{s}\right) + \Pr\left(\bar{s}^b | \underline{s}\right)\right)\right] c_s = \frac{c_s}{2}\left(2 - \delta\beta\right).$$

Finally, patient expected costs:

$$\begin{aligned} &C_{Pat}^{Ngk} = \Pr\left(\underline{s}\right)\left[\Pr\left(\underline{s}^b \cap \underline{s}^d | \underline{s}\right) p_g + \Pr\left(\underline{s}^b \cap \bar{s}^d | \underline{s}\right)\left(p_g + p_{gs}\right) + \Pr\left(\bar{s}^b | \underline{s}\right) p_s\right] + \\ &\Pr\left(\bar{s}\right)\left[\Pr\left(\underline{s}^b \cap \underline{s}^d | \bar{s}\right)\left(p_g + p_{gs} + l\right) + \Pr\left(\underline{s}^b \cap \bar{s}^d | \bar{s}\right)\left(p_g + p_{gs}\right) + \right. \\ &\left.\Pr\left(\bar{s}^b | \bar{s}\right) p_s\right] = \tfrac{1}{2}\left(p_s + p_g + p_{gs}\left(\beta\left(1-\delta\right) + 1 - \beta\right) + l\left(1-\beta\right)\left(1-\delta\right)\right). \end{aligned}$$

Appendix D.

Proof of Proposition 1

In non-gatekeeping it is easy to check that, for any value of β and δ, there exist values of R, T and B, such that $IC_{Pd}^{Ngk}$and IC_{Fd}^{Ngk} are fulfilled simultaneously.

In gatekeeping, however, $IC_{Fd_1}^{gk}$ and $IC_{Fd_2}^{gk}$ are mutually compatible if and only if:

$$\frac{\left(1-\delta\right)\beta}{\left(1-\delta\right)\beta + \delta\left(1-\beta\right)} R \leq \frac{\delta\left(1-\beta\right)\left(1-r\right)}{\delta\left(1-\beta\right) + \left(1-\delta\right)\beta} R - r\left(B - T - R\right)$$

Since $B \geq T + R$ the best case for this inequality to be fulfilled is when $B = T + R$. Taking this into account, it is direct to check that the inequality can be true if and only if $r \leq \overline{r}$, with $\bar{r} = 1 - \frac{(1-\delta)\beta}{\delta(1-\beta)}$. It can be shown, then, that for any $r \leq \overline{r}$, there exist values of B, T and R, such that $IC_{Pd}^{gk}$and IC_{Fd}^{gk} are fulfilled simultaneously.

This completes the proof.

Proof of Proposition 2

The optimal level of co-payments is the solution to the program given by (4). The problem is one of linear programming. Hence, the solution lies on a vertex of the restricted domain of the program. It can be shown that the restrictions $p_g \geq 0$, $p_{gs} \geq 0$ and $p_s - p_g \geq \beta\left(1-\delta\right) p_{gs} + \left(1-\beta\right)\left(p_{gs} + \left(1-\delta\right) l\right)$ must be binding at the optimum. The solution, therefore, is given by $p_g^* = p_{gs}^* = 0$ and $p_s^* = \left(1-\beta\right)\left(1-\delta\right) l$. This completes the proof.

Proof of Proposition 3

We compute the optimal payment contract for non-gatekeeping and gatekeeping separately. To determine the relevant incentive constraints we use Lemmas 2, 3 and 4.

a) In non-gatekeeping, the program the health authority faces is:

$$\min_{D,B,T,R} \frac{1}{2}\left[D + (B - T - R) + T\left[\delta + (1 - 2\delta)\beta\right] + R\delta\beta\right]$$

$$s.t \quad \begin{cases} U \geq 0 & PC \\ D \geq c_d & LLC_1 \\ T \geq 0 & LLC_2 \\ R \geq 0 & LLC_3 \\ B \geq T + R & LLC_4 \\ T \geq \frac{R(1-\delta)\beta + c_d}{(1-\delta)\beta + \delta(1-\beta)} & IC_{Pd_1}^{Ngk} \\ T \leq \frac{R\delta\beta - c_d}{\delta\beta + (1-\delta)(1-\beta)} & IC_{Pd_2}^{Ngk} \end{cases}$$

First of all, it is straightforward to see that LLC_1, LLC_2, LLC_3 and LLC_4 imply the PC. Therefore, the health authority chooses the cheapest contract compatible with the LLC and the IC_{PD}^{Ngk}. It can be checked that LLC_1 has to be binding at the optimum. The reasoning is the following: the health authority's costs are increasing in D and D does not appear in any of the incentive constraints. As a result $D^{Ngk} = c_d$. An analogous reasoning allows us to ensure that LLC_4 has to be also binding and that, hence, $B^{Ngk} = T^{Ngk} + R^{Ngk}$

It is easy to see that LLC_3 binding cannot be a solution as $IC_{Pd_1}^{Ngk}$ and $IC_{Pd_2}^{Ngk}$ would be mutually incompatible. Moreover, LLC_2 and $IC_{Pd_1}^{Ngk}$ binding cannot be a solution as $IC_{Pd_2}^{Ngk}$ would not be fulfilled. A similar argument rules out LLC_2 and $IC_{Pd_2}^{Ngk}$ binding as a potential solution.

The optimal solution of the problem, hence, has to be such that $IC_{Pd_1}^{Ngk}$ and $IC_{Pd_2}^{Ngk}$ are binding. From here we obtain that:

$$T^{Ngk} = \frac{c_d}{(2\delta - 1)(1 - \beta)} \text{ and } R^{Ngk} = \frac{c_d}{(2\delta - 1)(1 - \beta)\beta}.$$

The health authority's expected primary care costs are:

$$C_{GP}^{Ngk} = \frac{c_d}{2(2\delta - 1)}\left[4\delta + \left(\frac{\beta}{1 - \beta} - 1\right)\right].$$

b) In gatekeeping the problem the health authority faces is:

$$\min_{D,B,T,R} \; D+$$
$$\frac{1}{2}\left[(B-R)+(B-T)\,\delta\,(\beta+(1-r)\,(1-\beta))+(B-T-R)\,(1-\delta)\,(1-\beta+(1-r)\,\beta)\right]$$

$$s.t \quad \begin{cases} U \geq 0 & PC \\ D \geq c_d & LLC_1 \\ T \geq 0 & LLC_2 \\ R \geq 0 & LLC_3 \\ B \geq T+R & LLC_4 \\ T \geq 2c_d+(1-\delta)\,R\,(1-(1-\beta)\,r)-(B-T-R)\,(\delta\beta+(1-\delta)\,(1-\beta))\,r & IC^{gk}_{PD_1} \\ T \leq \delta R\,(1-(1-\beta)\,r)-2c_d-(B-T-R)\,((1-\delta)\,\beta+\delta\,(1-\beta))\,r & IC^{gk}_{PD_2} \\ T \geq \frac{R(1-\delta)\beta}{(1-\delta)\beta+\delta(1-\beta)} & IC^{gk}_{FD_1} \\ T \leq \frac{R\delta(1-\beta)(1-r)}{\delta(1-\beta)+(1-\delta)\beta}-r\,(B-T-R) & IC^{gk}_{FD_2} \end{cases}$$

First of all, it is straightforward to see that LLC_1, LLC_2, LLC_3 and LLC_4 imply PC. Moreover, by a similar reasoning as in the non-gatekeeping case, $D^{gk}=c_d$ at the optimum.

Secondly, a simple, but tedious process allows to check that LLC_4 has to be binding at the optimum. If $B>T+R$ this would increase the equilibrium values of T and R resulting, hence, in higher costs for the HA. Therefore, $B^{gk}=T^{gk}+R^{gk}$

In order to proceed, let us define $\widetilde{r} \equiv \frac{\delta-\beta}{(1-\beta)(\delta-\beta+2\delta\beta)} \in (0,\bar{r})$. We solve the program by distinguishing two cases:

- If $r \leq \widetilde{r}$, then $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Pd_2}$ imply both $IC^{gk}_{Fd_1}$ and $IC^{gk}_{Fd_2}$.

A completely analogous reasoning to the one followed for non-gatekeeping shows that the solution of the problem is such that both $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Pd_2}$ are binding. The optimal values, hence, are given by:

$$T^{gk}=\frac{2c_d}{2\delta-1},\; R^{gk}=\frac{4c_d}{(2\delta-1)\,(1-(1-\beta)\,r)}.$$

- If $r \in (\widetilde{r},\bar{r}]$, then it is not true that $IC^{gk}_{Fd_1}$ and $IC^{gk}_{Fd_2}$ are implied by $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Pd_2}$.

First of all, it is straightforward to see that neither $T \geq 0$ nor $R \geq 0$ can be binding at equilibrium. Therefore, the optimal contract has to be on one of the vertexes determined by the set of IC constraints.

By pairwise crossing all the IC constraints we find:

- $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Pd_2}$ binding violates $IC^{gk}_{Fd_2}$.
- $IC^{gk}_{Fd_1}$ and $IC^{gk}_{Fd_2}$ binding violates both $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Pd_2}$.
- $IC^{gk}_{Pd_2}$ and $IC^{gk}_{Fd_1}$ binding violates $IC^{gk}_{Pd_1}$.
- $IC^{gk}_{Pd_2}$ and $IC^{gk}_{Fd_2}$ binding violates $IC^{gk}_{Pd_1}$.
- Finally, $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Fd_1}$ binding, as well as $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Fd_2}$ binding, are shown to be vertexes of the domain and, hence, potential solutions of the program.

It can be checked that the equilibrium values of T and R obtained from $IC^{gk}_{Pd_1}$ and $IC^{gk}_{Fd_2}$ binding are smaller and, hence, this constitutes the optimal contract. Some algebraic manipulations yield:

$$\begin{aligned} T^{gk} &= 2c_d + \frac{4c_d(1-\delta)}{2\delta-1}\left[\frac{(2\delta-1)(1-(1-\beta)r)(\delta(1-\beta)+(1-\delta)\beta)}{2[(1-(1-\beta)r)(\delta-(1-\delta)(\delta(1-\beta)+(1-\delta)\beta))-\delta\beta]}\right] \\ R^{gk} &= \frac{4c_d}{(2\delta-1)(1-(1-\beta)r)}\left[\frac{(2\delta-1)(1-(1-\beta)r)(\delta(1-\beta)+(1-\delta)\beta)}{2[(1-(1-\beta)r)(\delta-(1-\delta)(\delta(1-\beta)+(1-\delta)\beta))-\delta\beta]}\right]. \end{aligned}$$

Summarizing the results obtained in the two regions, we can write the GP's optimal contract in a gatekeeping system as follows:

$$\begin{aligned} D^{gk} &= c_d \\ B^{gk} &= T^{gk} + R^{gk} \\ T^{gk} &= \frac{2c_d((2\delta-1)+2(1-\delta)\Gamma(\delta,\beta,r))}{2\delta-1} \\ R^{gk} &= \frac{4c_d\Gamma(\delta,\beta,r)}{(2\delta-1)(1-(1-\beta)r)}. \end{aligned}$$

with $\Gamma(\delta,\beta,r) = \begin{cases} 1 \text{ if } r \leq \widetilde{r}. \\ \frac{(2\delta-1)(1-(1-\beta)r)(\delta(1-\beta)+(1-\delta)\beta)}{2[(1-(1-\beta)r)(\delta-(1-\delta)(\delta(1-\beta)+(1-\delta)\beta))-\delta\beta]} > 1 \text{ otherwise.} \end{cases}$

The health authority's expected primary care costs are:

$$C^{gk}_{GP} = \frac{c_d}{2\delta-1}\left[4\delta + 2(\Gamma(\delta,\beta,r)-1)\right].$$

This completes the proof.

Proof of Proposition 4

First, evaluating the equilibrium levels of C^{Ngk}_{Pat} and C^{gk}_{Pat} for one extreme of the domain $\beta = \frac{1}{2}$, it can be checked that $C^{Ngk}_{Pat} > C^{gk}_{Pat}$.

Conversely, when $\beta \to 1$ it is easy to check that $C^{gk}_{Pat} > C^{Ngk}_{Pat}$.

Moreover $\frac{\partial C^{Ngk}_{Pat}}{\partial \beta} < 0$, and C^{gk}_{Pat} is also decreasing (and linear) in β. All the conditions above ensure us that there exists a unique threshold $\beta^* < 1$ such that:

$$\begin{aligned} \text{If } \beta &\leq \beta^* \text{ then } C^{Ngk}_{Pat} \geq C^{gk}_{Pat}. \\ \text{If } \beta &> \beta^* \text{ then } C^{Ngk}_{Pat} < C^{gk}_{Pat}. \end{aligned}$$

This completes the proof.

Proof of Proposition 5

By comparing C^{Ngk}_{GP} and C^{gk}_{GP}, as defined in Proposition 3, we find that:

If $\beta < \frac{1+4\delta}{2+4\delta}$ then $C^{gk}_{GP} > C^{Ngk}_{GP}$, for every value of r.

If $\beta \geq \frac{1+4\delta}{2+4\delta}$ then:

- If $r \leq \widetilde{r}$, then it is straightforward that $C_{GP}^{gk} < C_{GP}^{Ngk}$.

- If $r > \widetilde{r}$, then $C_{GP}^{gk} > C_{GP}^{Ngk} \Leftrightarrow \Gamma(\delta, \beta, r) > \frac{3-2\beta}{4(1-\beta)} - \delta$.

It can be checked that $\Gamma(\delta, \beta, r)$ is monotonically increasing in r. Therefore, if for $r = \bar{r}$, $\Gamma(\delta, \beta, \bar{r}) > \frac{3-2\beta}{4(1-\beta)} - \delta$, then there exists a threshold $r^* \in [\tilde{r}, \overline{r})$ such that:

If $r \leq r^*$ then $C_{GP}^{gk} < C_{GP}^{Ngk}$.

If $r > r^*$ $C_{GP}^{gk} > C_{GP}^{Ngk}$.

If for $r = \bar{r}$, $\Gamma(\delta, \beta, \bar{r}) \leq \frac{3-2\beta}{4(1-\beta)} - \delta$, then for every value of $r \in [0, \bar{r}]$ it holds that $C_{GP}^{gk} < C_{GP}^{Ngk}$.

Substituting the value of $\bar{r}$ and checking the inequality, it can be shown that there exists a value $\tilde{\beta} \in \left(\frac{1+4\delta}{2+4\delta}, 1\right)$ such that $\Gamma(\delta, \beta, \bar{r}) \leq \frac{3-2\beta}{4(1-\beta)} - \delta$ if and only if $\beta \geq \tilde{\beta}$.

Summarizing:

- If $\beta < \frac{1+4\delta}{2+4\delta}$ then $C_{GP}^{gk} > C_{GP}^{Ngk}$ for every value of $r \in [0, \bar{r}]$.
- If $\beta \geq \frac{1+4\delta}{2+4\delta}$ there exists a threshold $r^* \in [\tilde{r}, \overline{r}]$ such that:
  - If $r \leq r^*$ then $C_{GP}^{gk} < C_{GP}^{Ngk}$.
  - If $r > r^*$ then $C_{GP}^{gk} > C_{GP}^{Ngk}$.
    With $r^* \in [\tilde{r}, \overline{r})$ if $\beta \leq \tilde{\beta}$, and $r^* = \bar{r}$ if $\beta > \tilde{\beta}$.

This completes the proof.

Proof of Proposition 6

Comparing C_{Sp}^{Ngk} and C_{Sp}^{gk}, as computed in Appendix C, we get that:

$$C_{Sp}^{Ngk} - C_{Sp}^{gk} = \frac{c_s}{2}\left[\delta(1-\beta) + r(1-\delta)\beta\right] > 0.$$

Moreover, it is direct to check that $\frac{\partial\left[C_{Sp}^{Ngk} - C_{Sp}^{gk}\right]}{\partial\beta} < 0$, while $\frac{\partial\left[C_{Sp}^{Ngk} - C_{Sp}^{gk}\right]}{\partial r} > 0$. This completes the proof.

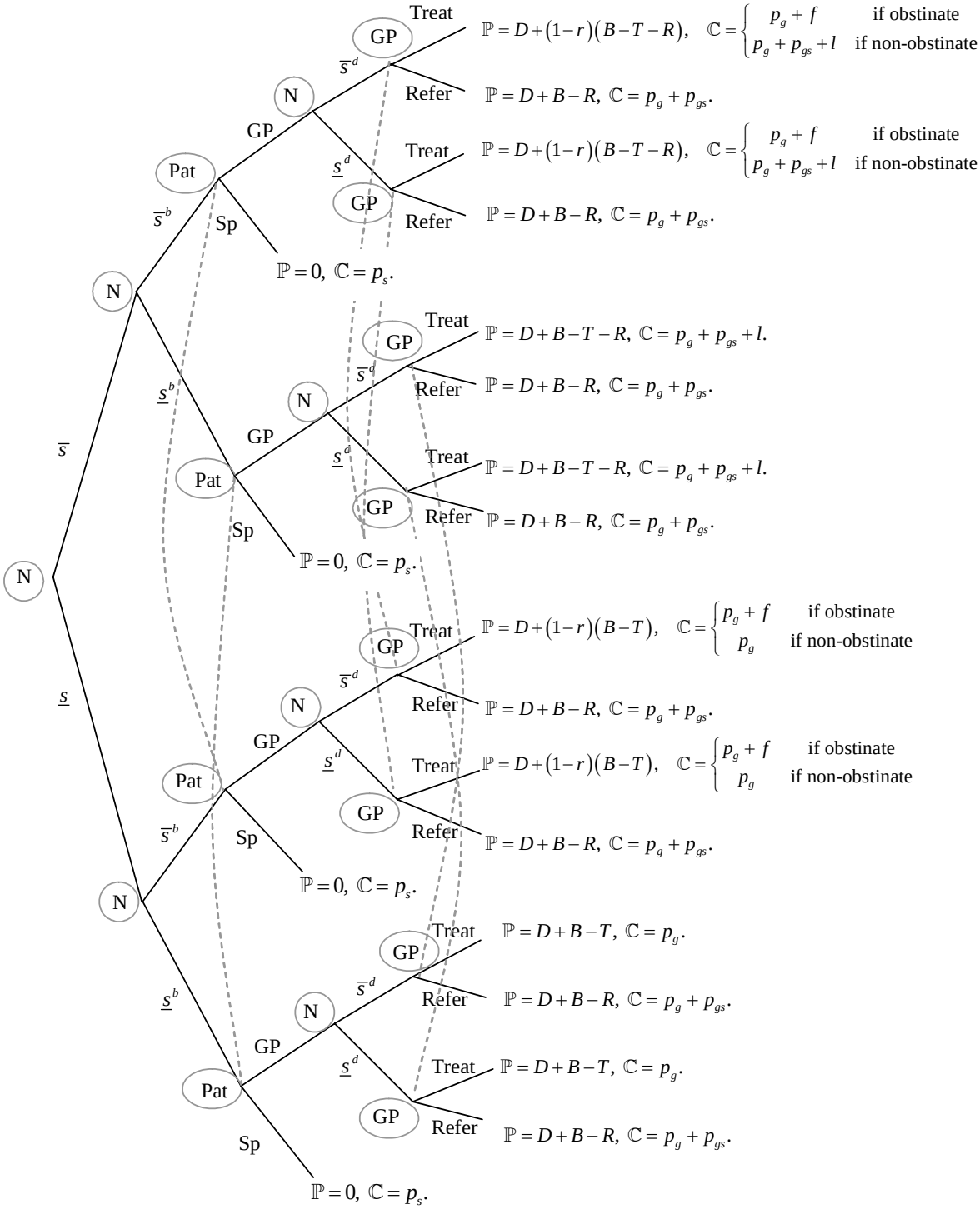

Figure 1: GP and patient decision tree: GP payoffs ($\mathbb{P}$) and patient costs ($\mathbb{C}$).