Working papers series

WP ECON 06.09

The Gatekeeping Role of General Practitioners. Does Patients' Information Matter?

Paula González (U. Pablo de Olavide)

JEL Classification numbers: D82, H51, I18, L51.
Keywords: General Practice, Moral hazard, Incentives, Patients' beliefs, Patients' pressure, Referrals.

Department of Economics

The Gatekeeping Role of General Practitioners. Does Patients' Information Matter?*

Paula González†
Universidad Pablo de Olavide (Sevilla)

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 investigate the relative merits of gatekeeping and non-gatekeeping systems and to analyze the role of the quality of patient information and referral pressure in determining which model dominates. We find that, whenever GPs incentives matter, non-gatekeeping is better only if there is a sufficiently high pressure for referral. At the same time, for a non-gatekeeping system to dominate, the quality of the patient information should not be extreme: neither too bad (patient's self-referral would be very inefficient) nor too good (the GP's agency problem would be very costly).

JEL classification: D82, H51, I18, L51.

Keywords: General Practice, Moral hazard, Incentives, Patients' beliefs, Patients' pressure, Referrals.

*This work was done 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 93 39. E-mail: pgonrod@upo.es

1 Introduction

This work is a contribution to the current debate over the benefits and drawbacks from enhancing the gatekeeping role of General Practitioners (henceforth, GPs). The paper aims at studying the role of GPs as filters for secondary care, emphasizing the implications that patients' information and referral pressure have for health authorities.

Currently, two main types of health care systems can be observed in most European countries. In some of them, like Italy, the Netherlands, Norway, Spain, or the United Kingdom, a gatekeping system exists and, hence, GPs control access to other levels of health care. There are other countries, like Belgium, Finland, France or Germany, where the gatekeeping role of the GP is very limited, as patients have a free choice of GPs and specialists.

Despite this heterogeneity, it has often been argued that a system in which GPs act as gatekeepers to specialist care leads to lower health care costs.[1] This explains the interest in gatekeeping both in Eastern European countries, where they are reforming their health care systems (Hebing (1997)), and in the United States, where gatekeeping has been a central strategy in the cost-containment initiatives of managed care organizations (Wolf and Gorman (1996)).

It has been widely recognized in the literature that the regulation of the GPs and the incentive structures they face have significant implications for costs in health care systems (see, for instance, Scott (2000)). In particular, the incentives provided to the GPs determine their two main decisions in relation to overall health costs: diagnosis and referral.[2]

The work by García-Mariñoso and Jelovac (2003) is the first that provides a uniform theoretical framework in which the identification of the optimal payment system is used to compare gatekeeping and non-gatekeeping systems. They find that the optimal GP's payment scheme requires a combination of cost-sharing components: cost sharing of the GP's treatment and a bonus for not referring. This contract yields the right diagnosis and recommendation incentives to the GP. They show that, when GPs' incentives matter, gatekeeping is superior to non-gatekeeping.

However, there exist other elements apart from GPs' incentives that should be considered when comparing the pros and cons of these two different systems to access secondary care. First, the issue of how to discipline patients who might strategically choose to visit a specialist or a GP should be addressed. Secondly, the quality of the patients' information must also play a role, since if patients have a sufficiently accurate information on their problem, allowing them to freely choose their medical provider, may be more efficient than a compulsory visit to the GP. Thirdly, patients' information may turn into patients' pressure to obtain a referral to specialist treatment.

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

[2] There is empirical evidence that GPs' 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.

In this respect, there is evidence that patients do state their preferences and expectations to GPs about whether they want to be referred or prescribed medication (Armstrong et al. (1991)), and also that in many occasions, this may alter GPs' decisions. For instance, 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.

The goal of this paper is to investigate the relative merits of gatekeeping and non-gatekeeping in a model that embeds all these elements. Following the approach that GPs' contracts should include appropriate diagnosis and referral incentives, we analyze the role of patient information and referral pressure in determining which system dominates.

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: First, moral hazard (as neither GP's diagnosis decision nor diagnosis outcome are verifiable). Second, the selection of medical provider among types of patients is ex-ante unknown to the principal (as patients' belief about their severity is private information).

Co-payments are generally used to avoid patients' overconsumption of medical services. In some countries, co-payments have been introduced as financial incentives at the patients' side to stimulate the gatekeeping role of GPs.[3] This is in line with the role co-payments play in our model. We introduce them to solve the selection problem and avoid a systematic utilization of specialist treatment by patients in non-gatekeeping.

We find that if we want to provide the GP with diagnosis and recommendation incentives, non-gatekeeping is optimal only if there is a sufficiently high pressure for referral. The reason is that when this pressure is high enough, a gatekeeping system may be unable to provide the GP with right incentives, while this is not a problem under non-gatekeeping.

In addition to this, for a non-gatekeeping to be the best system, the quality of the patient's information should not be extreme. If patients' signal is highly uninformative, patients' self-referral is very inefficient. Patients' expected health losses are very high, and so are specialist costs, due to the high proportion of unnecessary visits to the specialist. When patients' information is extremely accurate, non-gatekeeping is convenient both from the perspective of patients and specialists costs. However, this advantage is outweighed by the fact that in non-gatekeeping the accuracy of the patient's beliefs fosters primary care costs. Since only those patients who think they are mild cases visit the primary provider, the patient's belief is a source of pre-diagnosis information that reduces the incentives of the GP to make a costly diagnosis.

Although primary care is being recognized as the mainstay of many health care systems in

[3] In Belgium, for instance, from 2002 onwards patients pay lower co-payments if they register with one specific GP. Moreover, very recently, the co-payments for those patients who go to the specialist directly, without having seen first a GP, have increased. See, Schokkaert and Van de Voorde (2005) for a detailed explanation of the Belgian reform.

developed countries, there has been little theoretical research by economists into general practice.

Malcomson (2004) discusses which contractual agreements are most effective at inducing gatekeepers to exert effort in diagnosis. He shows that implementing incentive contracts is not worthwhile when patients are allowed to choose between a gatekeeper with an incentive contract and one without.

García-Mariñoso and Jelovac (2003) advocate in favour of gatekeeping systems, whenever GP's incentives matter. In the present paper we show that, when the role of patient information and referral pressure is also considered, this general prevalence of gatekeeping is no longer true.

Brekke et al. (2005) contributes to the discussion on gatekeeping by analizing the competition effects amongst secondary care providers that arise when GPs are equipped with a gatekeeping role. Also, from this perspective, the implementation of a gatekeeping system need not be always socially desirable.

Finally, to the best of our knowledge, only García-Mariñoso (1999) provides a description of how the insurer can regulate access to specialist care by manipulating the patients' insurance contract. There are two main differences between her approach and ours. First, she does not take the quality of the information of the patients into account, as the patient's signal is always perfectly correlated with the true probability of facing a given severity. In exchange, in our paper, the fraction of patients that visit the specialist directly is exogenously given (those who believe to be high-severity), while in García-Mariñoso the optimal screening of patients is endogenously determined.

The rest of the paper is organized as follows: In the following section, we present the model. Section 3 analyzes both the patient's and the GP's behavior. In Section 4 we derive the optimal patient's co-payment levels and the optimal GP's payment contract. Section 5 compares the two institutional frameworks. Finally, in Section 6, we present our conclusions.

2 The Model

Our model is inspired by Jelovac (2001) and 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 specialist medical attention, but we will consider him as a passive agent, as the analysis of his behavior is out of the scope of this article.[4]

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 he is ill but does not know just

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

how serious his illness is. His symptoms, however, provide him with a private signal or belief about the severity of his 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. He 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 his available income are the same in all contingencies (minor or major illness). Therefore, in this model, maximizing the patient's expected utility is equivalent to minimizing the value of his expected costs. These costs come from two main sources. First, from the health loss (l) that the patient suffers when he receives primary care and a 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 specialist medical care directly.[5]

As we have already claimed in the Introduction, there exists empirical evidence suggesting that the pressure of the patient to obtain a referral may alter GP's behavior. We model this pressure as the probability that the patient rejects GP's treatment. In case of doing so, he will demand private specialist treatment at a cost f.[6] This way, he avoids any potential health loss, although he bears the full cost of receiving specialist 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's treatment and pay for private services even if, as we will see later, it is worthwhile for them to follow GPs' recommendation.

The General Practitioner.

We consider that the GP is able to cure a patient only if the severity of the condition is low, while the specialist can heal both levels of severity. Whenever the GP receives a patient, he is

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

[6] The reader should note that in this work we are ruling out the existence of a set of potential patients who decide to directly access specialist private treatment.

required to make a diagnosis, which yields a signal about the severity of the patient's condition ($s^d \in \{\underline{s}^d, \bar{s}^d\}$). We assume the probability of receiving a correct signal to be $\delta \in (\frac{1}{2}, 1)$. 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 also 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. For simplicity, we focus on the case in which both s^d and s^b are correlated with s, but patient's and GP's errors are conditionally independent. In making a diagnosis, the GP incurs a disutility, that we denote by c_d.

As part of the diagnosis, the physician also observes the patient's belief about his true condition.[7] This implies that, although patient's and GP's errors are conditionally independent, GPs' posterior beliefs are positively correlated with patient information. Combining the diagnosis with this piece of information, the GP should decide on treating the patient or referring him 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 him directly.

As both the GP's decision to diagnose a patient and the diagnosis are hard to verify, the incentives included in the GP's payment contract will crucially determine his behavior. We take a contract structure similar to the one proposed by González (2004). GP's contract, hence, consists of three non-negative components (R, T, B). R is the amount of money that the GP receives when the patient is referred directly to the specialist. If, instead, the GP proposes treatment to the patient, he receives a payment T. In this latter case, if the patient accepts GP's treatment, the GP receives a bonus B provided the patient is eventually not referred to specialist treatment. This payment structure contains: (i) a capitation component or payment per visit ($\min\{T, R\}$), (ii) a cost-sharing of the GP's treatment (whenever $R > T$ in equilibrium) and (iii) a bonus for not referring (B) that can be interpreted as a premium for cost-containment.[8]

The Health Authority.

The third agent involved in the model is the health authority. The health authority pays the

[7] 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 system would trivially be more costly than non-gatekeeping, as the latter would be simply a subset (when $\underline{s}^b$) of the former.

[8] 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. In the UK the former GPs fundholders were 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. In Italy GPs' 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.

costs of the treatment provided to the patient, and also the payments made to both the GP and 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 the latter to zero.

The health authority designs the GP's contract and the patient's level of co-payments so as to minimize expected social costs. Such costs are the sum of the financial costs both from primary and specialist health care (i.e. expected treatment costs and payoffs to both the GP and the specialist) and the patient's expected costs (which includes both his expected health losses and his monetary expenses).[9]

Our aim is to study whether it is socially useful to use patients' information as a mechanism for provider selection. Hence, the level of co-payments will be designed to ensure that patients use their own 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 recommendation incentives, we focus 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 him if $\bar{s}^d$.

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

Timing.

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

As usual, we solve the game by backward induction.

[9] One should note that: (i) the health authority internalizes the cost of the private treatment, through patients' 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 Agents' 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 in which patients are not obliged to compulsorily visit the GP. Second, we set out to derive the conditions that the GP's payment contract has to fulfill in order to ensure that the GP 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's 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 his own symptoms and is aware of his personal circumstances, he is able to update these probabilities. Then, the probabilities that the patient recognizes/misrecognizes the severity of his illness are:[10]

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

In non-gatekeeping the patient has the choice between two alternatives: go to the specialist directly, or go first to the GP. If the patient goes to the specialist directly, his cost is given by the co-payment he has to pay p_s, but no health loss is borne. If the patient goes first to the GP he always pays p_g and, then, if he is eventually referred to the specialist p_{gs}. Moreover, he may also suffer from a health loss whenever he receives treatment from the GP that does not heal him. Those patients who are obstinate always reject GP's 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 patients' costs in any circumstance can be easily computed.[11] Take first belief $\underline{s}^b$. First, if the patient is low-severity ($\Pr\left(\underline{s}|\underline{s}^b\right)$) he 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 ($\Pr\left(\bar{s}|\underline{s}^b\right)$) he 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 ($\Pr\left(\underline{s}^d|\bar{s}\right)$).

Take now belief $\bar{s}^b$. First, if the patient suffers from a major illness ($\Pr\left(\bar{s}|\bar{s}^b\right)$) he 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 ($\Pr(\underline{s}^d|\bar{s})$),

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

[11] 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's diagnosis is right ($\Pr\left(\underline{s}^d|\underline{s}\right)$), the patient incurs $p_g + f$ if he 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 the patient's expected costs when demanding first GP's attention, or direct specialist care, we conclude the following:

(i) If $\underline{s}^b$, a patient chooses to go 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 chooses to go to 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 chooses to go to 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.[12]

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 quality 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, for the patient to use his own information when selecting the medical provider. The more accurate this information is, therefore, the smaller the expected costs of his self-referral.

3.2 General Practitioner's 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. It is worth mentioning that, in order to avoid dealing with information revelation issues, we rule out the possibility that the physician wrongly observes the patient's belief.

[12] In order to ensure that non-obstinate patients always find it optimal to accept GP's treatment, we are implicitely considering that the private alternative is sufficiently costly. In particular f must exceed the patient's expected costs associated with accepting GPs' treatment in the most demanding contingency, i.e. when the patient receives signal $\bar{s}^b$ 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:[13]

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

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 condition, he then decides on the best option for the patient. The doctor always has two alternatives: treat the patient or refer him to the specialist.[14]

If the GP refers the patient to the specialist he always receives R. If the GP recommends treatment he gains T and, with a certain probability, B. If $\underline{s}^d$ and $\underline{s}^b$, the GP receives B if the patient's condition is really mild ($\Pr\left(\underline{s}|\underline{s}^d \cap \underline{s}^b\right)$). Likewise, if $\bar{s}^d$ and $\underline{s}^b$, the GP receives B with a probability $\Pr\left(\underline{s}|\bar{s}^d \cap \underline{s}^b\right)$. When $\underline{s}^d$ and $\bar{s}^b$, the GP receives B if the patient has a minor illness ($\Pr\left(\underline{s}|\underline{s}^d \cap \bar{s}^b\right)$) and he does not reject the treatment (with a probability $(1-r)$). Finally, when $\bar{s}^d$ and $\bar{s}^b$, the GP receives B with a probability of $\Pr\left(\underline{s}|\underline{s}^d \cap \bar{s}^b\right)(1-r)$.

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 $R - T \leq \frac{B\delta\beta}{\delta\beta+(1-\delta)(1-\beta)}$ and refers him otherwise.

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

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

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

From these conditions we see, first, that the difference between the two "safe" payments: R for a direct referral to the specialist and T for recommending treatment, has to be strictly positive. This means that the contract should include some cost sharing of the GP's treatment. Moreover, the premium B plays an important role in avoiding a systematic referral of patients.

It is also worth mentioning that the conditions that the GP's 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 believe to be low-severity visit the GP,

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

[14] 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.

the only relevant restrictions are (i) and (iii). In gatekeeping, however, the four conditions have to be fulfilled. This leads to the following lemma.

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

- *In non-gatekeeping:*

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

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

- *In gatekeeping:*

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

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

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. In non-gatekeeping, the GP always receives patients who think they are low-severity. This implies that there is no pressure for referral, what makes the restrictions less demanding.

It can be shown that for both 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 it becomes more accurate.

In our model, the GP receives neither his signal of the patient's severity 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. When the conditions written in Lemma 2 hold, however, we can ensure that the GP will stick to the 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:

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

- In gatekeeping:

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

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 (R) or from systematically treating him: $\left(T + \frac{1}{2}B\left(1 - (1-\beta)\, r\right)\right)$ in gatekeeping, or $(T + B\beta)$ in non-gatekeeping.

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:*

$$R - T \geq \frac{B\left(1-\delta\right)\beta + c_d}{\left(1-\delta\right)\beta + \delta\left(1-\beta\right)} \ \ and \qquad (IC^{Ngk}_{Pd_1})$$

$$R - T \leq \frac{B\delta\beta - c_d}{\delta\beta + \left(1-\delta\right)\left(1-\beta\right)}. \qquad (IC^{Ngk}_{Pd_2})$$

- *In gatekeeping:*

$$R - T \geq 2c_d + \left(1-\delta\right)B\left(1-(1-\beta)\, r\right) \ \ and \qquad (IC^{gk}_{Pd_1})$$

$$R - T \leq \delta B\left(1-(1-\beta)\, r\right) - 2c_d. \qquad (IC^{gk}_{Pd_2})$$

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 always follows the diagnosis.*
- *In gatekeeping he always follows the diagnosis if and only if IC^{gk}_{Fd} are fulfilled.*

In non-gatekeeping we can ensure that, for every value of c_d, the conditions that have to be fulfilled for the GP to follow the diagnosis IC^{Ngk}_{Fd} are always milder than the ones that induce him to make a diagnosis IC^{Ngk}_{Pd}. 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 c_d IC^{gk}_{Fd} constraints are always implied by IC^{gk}_{Pd}. 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. 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* (R, T, B) *that induces the GP to treat the patient when* $\underline{s}^d$ *and to refer him 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 shows how 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 result is determined by the "patients' pressure". If the referral pressure is sufficiently high, it is impossible for the health authority to find values of R, T and B that simultaneously fulfill all the constraints. The reason for it is the following: Given the high risk of treatment rejection, the minimum value of the bonus B that induces the GP to treat a patient whenever $\underline{s}^d$, is so high that the GP will also be willing to treat a patient when $\bar{s}^d$. Conversely, if the cost-sharing $R-T$ is high enough to induce the GP to refer a patient if $\bar{s}^d$, the GP ends up referring patients for which the diagnosis recommended a treatment.

Proposition 1 has shown an important implication of the presence of patient's pressure for referral. 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 the patient's belief becomes more precise, the maximum threshold of pressure compatible with gatekeeping decreases. As the patient gets to know more, the physician will be less willing to effectively recommend treatment when this is contrary to the patient's will.[15]

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 the patient's expected costs (C_{Pat}). C_{GP}, C_{Sp} and C_{Pat} are derived formally

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

in Appendix C. The simplified expressions are as follows:

- In gatekeeping:

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

- In non-gatekeeping:

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

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 induce 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.[16]

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 the co-payments have to be non-negative.

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 $p_s^* = \left(1-\beta\right)\left(1-\delta\right) l$.*

[16] 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's contract; (ii) C_{GP} is not altered by the co-payment levels, provided they induce the patient to select the medical provider according to his belief about the severity; (iii) C_{Sp} depends only on the institutional framework.

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 reform, aimed at enhancing the gatekeeping role of GPs. In the Belgian system, however, 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 GPs' referral. This is not necessary in our model as we do not deal with healthy individuals who make unnecessary visits to the system.

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.[17]

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

[17] This problem is analysed in detail by García-Mariñoso (1999).

The health authority's optimization program is as follows:

$$\begin{array}{ll} \min\limits_{T,R,B} & C_{GP} \\ s.t & \begin{cases} U \geq 0 & PC \\ T \geq c_d & LLC_1 \\ R \geq c_d & LLC_2 \\ B \geq 0 & LLC_3 \\ IC & \end{cases} \end{array} \tag{5}$$

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 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 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 him if $\bar{s}^d$ the optimal contract (R, T, B) is as follows:*

- *In non-gatekeeping:*

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

  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:*

$$\begin{aligned} R^{gk} &= \frac{c_d\left(3(2\delta-1) + 4(1-\delta)\Gamma(\delta,\beta,r)\right)}{2\delta-1} \\ T^{gk} &= c_d \\ B^{gk} &= \frac{4c_d\Gamma(\delta,\beta,r)}{(2\delta-1)(1-(1-\beta)r)}. \end{aligned}$$

  The health authority's expected primary care costs are:

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

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

Proof. See Appendix D. ■

The structure of the GP's optimal payment contract shares some features with the one in García-Mariñoso and Jelovac (2003). 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 c_d. Second, to avoid systematic treatment of the patient, the contract includes some cost sharing of the GP's treatment ($R - T > 0$). Finally, this premium has to be smaller than the bonus for not referral ($R - T < B$) to compensate for the disincentive effect that a positive $R - T$ has on the decision of making diagnosis.

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.

The referral pressure (only present in gatekeeping) raises the health authority's expected primary costs, but only beyond a certain threshold ($\widetilde{r} > 0$). For values of pressure below $\widetilde{r}$ the marginal increase in the bonus B to avoid an excessive number of referrals due to pressure, is compensated by the fact that such a bonus is paid less often at equilibrium. For values beyond $\widetilde{r}$, the increase in the bonus is so high that it always appears reflected in primary care costs, what makes these costs 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. In non-gatekeeping patient information generates a problem of "diagnosis substitution". The GP acquires information simply by receiving the patient and, then, he has more incentives to use the patient's belief as a substitute of his own diagnosis. Hence, inducing the GP to make a diagnosis becomes more 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 problem the health authority faces when choosing the best system to access health care. As it has become clear throughout the paper, the quality of the patients' beliefs, as well as their referral pressure, are the two key elements that will drive the health authority's 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: patients' costs, primary care costs and specialist 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 patients' expected costs than non-gatekeeping.*
- *If $\beta > \beta^*$, non-gatekeeping generates lower patients' expected costs than gatekeeping.*

Proof. See Appendix D. ■

When we concentrate on the patient's 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 information.

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 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 is problematic in non-gatekeeping systems. This is reinforced by the following corollary.

Corollary 1 *r^* is increasing in β.*

Patients' belief generates a problem of informational substitution that is present only in non-gatekeeping. As the accuracy of this belief 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 specialist costs in both systems.*
- *The higher is r, the lower the expected specialist costs in gatekeeping relative to non-gatekeeping.*

Proof. See Appendix D. ■

Even if, in general, gatekeeping allows savings in specialist costs, the higher the quality of patient information the more similar 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 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 the overall comparison of gatekeeping and non-gatekeeping when both financial costs (GP's and specialist's costs) and patient's costs are simultaneously considered.

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

When patients' signal is extremely uninformative $\left(\beta \to \frac{1}{2}\right)$ both the patient's expected health losses and the 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 specialist treatment is sufficiently more costly than primary care. Therefore, in general, systems where patients freely choose their medical provider would not dominate.

When patients' 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 patients misinterpret their symptoms. A high accuracy of the patient's information, however, has perverse effects on GP's 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.[18]

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 information. On the other hand, gatekeeping suffers from the problem of patient's pressure for referral, 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 his medical provider, then:*

- *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's 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 be also 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.

[18] We should recall that, as quality of patient 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.

5.3 Discussion

It is important to pause to discuss how gatekeeping and non-gatekeeping compare. The general lesson that we can draw from our analysis is that if we want to provide the GP with incentives to diagnose a patient (and follow the diagnosis), non-gatekeeping is optimal if there is a sufficiently high pressure for referral. At the same time, for a non-gatekeeping to be the best system, the quality of the patient's information should not be extreme: neither too bad (patient's self-selection would be very inefficient) nor to good (the GP's incentives to skip the diagnosis would be very strong).

The fact that when patients' information is very precise gatekeeping dominates creates an apparent puzzle: *When the information of the patients is very accurate, it is not worthwhile using it.* This paradoxical recommendation, however, is true because we have restricted our analysis to those situations in which GPs' incentives matter. In this case it is not possible to design a contract that combines this information with the posterior physicians' diagnosis. In this sense, in non-gatekeeping, patients' information becomes a substitute, rather than a complement, for GPs' diagnosis.

The last paragraph has an interesting implication. When patients' information is very precise, and the health authority wants to profit from it, it is not worthwhile giving incentives to the GP. Patients' information should be used instead of, and not in addition to, GPs' diagnosis. In fact, it can be shown that a system in which patients have free choice of their medical provider but the GP systematically treats without diagnosing, would dominate any other alternative. It would allow the health authority to benefit completely from the patient's information, eliminating at the same time the GP's incentive problem.

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. In such a model, we have analyzed the role of GPs as filters for secondary care. We have followed the conventional wisdom in the literature that GPs' contracts should include appropriate diagnosis and referral incentives. In this setting, we have shown that patient information and referral pressure alter the choice of the system to access specialist care. These two elements drive the results through their direct impact on patients' expected costs, and also, indirectly, altering GPs' behavior and, hence, expected primary and secondary costs.

In terms of policy recommendations, our analysis suggests that whenever GPs incentives matter, the choice of the system to access secondary care depends on the relative importance of two features: referral pressure in gatekeeping and GPs' diagnosis substitution in non-gatekeeping. In general, non-gatekeeping is optimal only if there is a sufficiently high pressure for referral,

and if the quality of the patient's information is not extreme (neither too bad nor too good). If patients' signal is highly uninformative, patients' self-referral is very inefficient. When patients' information is extremely accurate any benefit from non-gatekeeping is outweighed by the increase that the GPs' diagnosis substitution generates in primary costs.

One insight that emerges from this analysis is that when health authorities want to effectively use patients' own information, it may not be worthwhile giving incentives to the GP. This opens a potentially fruitful path of research: the study of the substitutability/complementarity relationship between patients and GPs' information. Such analyses would allow us to go further in the discussion about the shape of optimal contractual agreements.

One potential criticism to our work is the fact that the quality of the patients' information might be hardly observable by health authorities. Still, with chronic or inherited conditions, repeated illness episodes, or illnesses with symptoms which are easy to recognize, one should expect a higher accuracy of the patients' information than that of other types of pathologies.

Finally, we would like to highlight that, although primary care is recognized as the basis of health care systems in many developed countries, there has been little research by economists into general practice. We believe this work is a contribution to this scarce literature, as well as to the ongoing debate over the pros and cons of enhancing the gatekeeping role of General Practitioners. Certainly, further research, both theoretical and empirical, is needed to assess the relevance of the relationship between patients' information, pressure for referral and GP's 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] Brekke, K.R., Nuscheler, R. and Straume, O.R. (2005). "Gatekeeping in Health Care". CESIFO Working Paper No. 1552.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] González, P. (2004) "Should Physicians' Dual Practice Be Limited? An Incentive Approach". *Health Economics* 13 (6): 505-524.

[9] 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.

[10] 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.

[11] Jelovac, I. (2001) "Physicians' Payment Contracts, Treatment Decisions and Diagnosis Accuracy". *Health Economics* 10, 9-25.

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

[13] 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.

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

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

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

[17] 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.
$s^b \in \{\underline{s}, \bar{s}\}$	Patients' belief.
$\beta \in \left(\frac{1}{2}, 1\right)$	Accuracy of patients' 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.
(R, T, B)	Physician's payments: R if the patient is directly referred to the specialist. T if the GP recommends treatment. B bonus if the patient is cured in primary care.
$c_s > 0$	Cost of the specialist services.
C_{Pat}	Patients' expected costs.
C_{GP}	Expected costs associated with primary care.
C_{Sp}	Expected costs associated with specialist treatment.

Appendix B. GP's and Patient's 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's and GP's 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 $\forall i,j \in \{\bar{s},\underline{s}\}$:

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

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's expected costs.

In gatekeeping:

GP's expected utility:

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

Health authority's expected primary care costs:

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

Health authority's expected specialist care costs:

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

Finally, patient's expected costs:

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

In non-gatekeeping:

GP's expected utility:

$$\begin{array}{l}U^{Ngk}=\Pr\left(\underline{s}|\underline{s}^b\right)\left[\Pr\left(\underline{s}^d|\underline{s}\right)\left(T+B\right)+\Pr\left(\bar{s}^d|\underline{s}\right)R\right]+\\ \Pr\left(\bar{s}|\underline{s}^b\right)\left[\Pr\left(\bar{s}^d|\bar{s}\right)R+\Pr\left(\underline{s}^d|\bar{s}\right)T\right]-c_d=\\ T+\left(R-T\right)\left[\delta+\left(1-2\delta\right)\beta\right]+B\delta\beta-c_d.\end{array}$$

Health authority's expected primary care costs:

$$C_{GP}^{Ngk}=\Pr\left(\underline{s}^b\right)\left(U^{Ngk}+c_d\right)=\frac{1}{2}\left[T+\left(R-T\right)\left[\delta+\left(1-2\delta\right)\beta\right]+B\delta\beta\right].$$

Health authority's expected specialist care costs:

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

Finally, patient's expected costs:

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

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:

$$\frac{B(1-\delta)\beta}{(1-\delta)\beta + \delta(1-\beta)} \leq \frac{B\delta(1-\beta)(1-r)}{\delta(1-\beta) + (1-\delta)\beta}$$

This holds if and only if $r \leq \overline{r}$, with $\overline{r} = 1 - \frac{(1-\delta)\beta}{\delta(1-\beta)}$. It can be shown then that, for any $r \leq \overline{r}$, there exist values of R, T and B, 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, it is well-known that 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 $p_s - p_g \geq \beta(1-\delta)p_{gs} + (1-\beta)(p_{gs} + (1-\delta)l)$ must be binding at the optimum. The solution, therefore, is given by $p_g^* = p_{gs}^* = 0$ and $p_s^* = (1-\beta)(1-\delta)l$. 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:

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

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

First of all, it is straightforward to see that LLC_1, LLC_2 and LLC_3 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 T. In addition to this, from the IC_{Pd}^{Ngk} we see that necessarily $R > T$ and that the minimum value of R compatible with the restriction is increasing in T. As a result $T^{Ngk} = c_d$.

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:

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

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:

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

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

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

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 $IC_{Pd_1}^{gk}$ and $IC_{Pd_2}^{gk}$ imply both $IC_{Fd_1}^{gk}$ and $IC_{Fd_2}^{gk}$.

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

$$R^{gk} = \frac{(2\delta + 1)\, c_d}{2\delta - 1},\ B^{gk} = \frac{4c_d}{(2\delta - 1)\left(1 - (1 - \beta)\, r\right)}.$$

- 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 $B \geq 0$ nor $R \geq c_d$ 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:

$$\begin{aligned} R^{gk} &= 3c_d + \frac{4c_d(1-\delta)}{2\delta-1}\left[\frac{(2\delta-1)(1-(1-\beta)r)(\delta(1-\beta)+(1-\delta)\beta)}{2[(1-(1-\beta)r)(\delta-(1-\delta)(\delta(1-\beta)+(1-\delta)\beta))-\delta\beta]}\right] \\ B^{gk} &= \frac{4c_d}{(2\delta-1)(1-(1-\beta)r)}\left[\frac{(2\delta-1)(1-(1-\beta)r)(\delta(1-\beta)+(1-\delta)\beta)}{2[(1-(1-\beta)r)(\delta-(1-\delta)(\delta(1-\beta)+(1-\delta)\beta))-\delta\beta]}\right]. \end{aligned}$$

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

$$\begin{aligned} R^{gk} &= \frac{c_d(3(2\delta-1)+4(1-\delta)\Gamma(\delta,\beta,r))}{2\delta-1} \\ T^{gk} &= c_d \\ B^{gk} &= \frac{4c_d\Gamma(\delta,\beta,r)}{(2\delta-1)(1-(1-\beta)r)}, \end{aligned}$$

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

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

This completes the proof.

Proof of Proposition 5

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

If $\beta < \frac{1+4\delta}{2+4\delta}$ then $C_{GP}^{gk} > C_{GP}^{Ngk}$, 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.

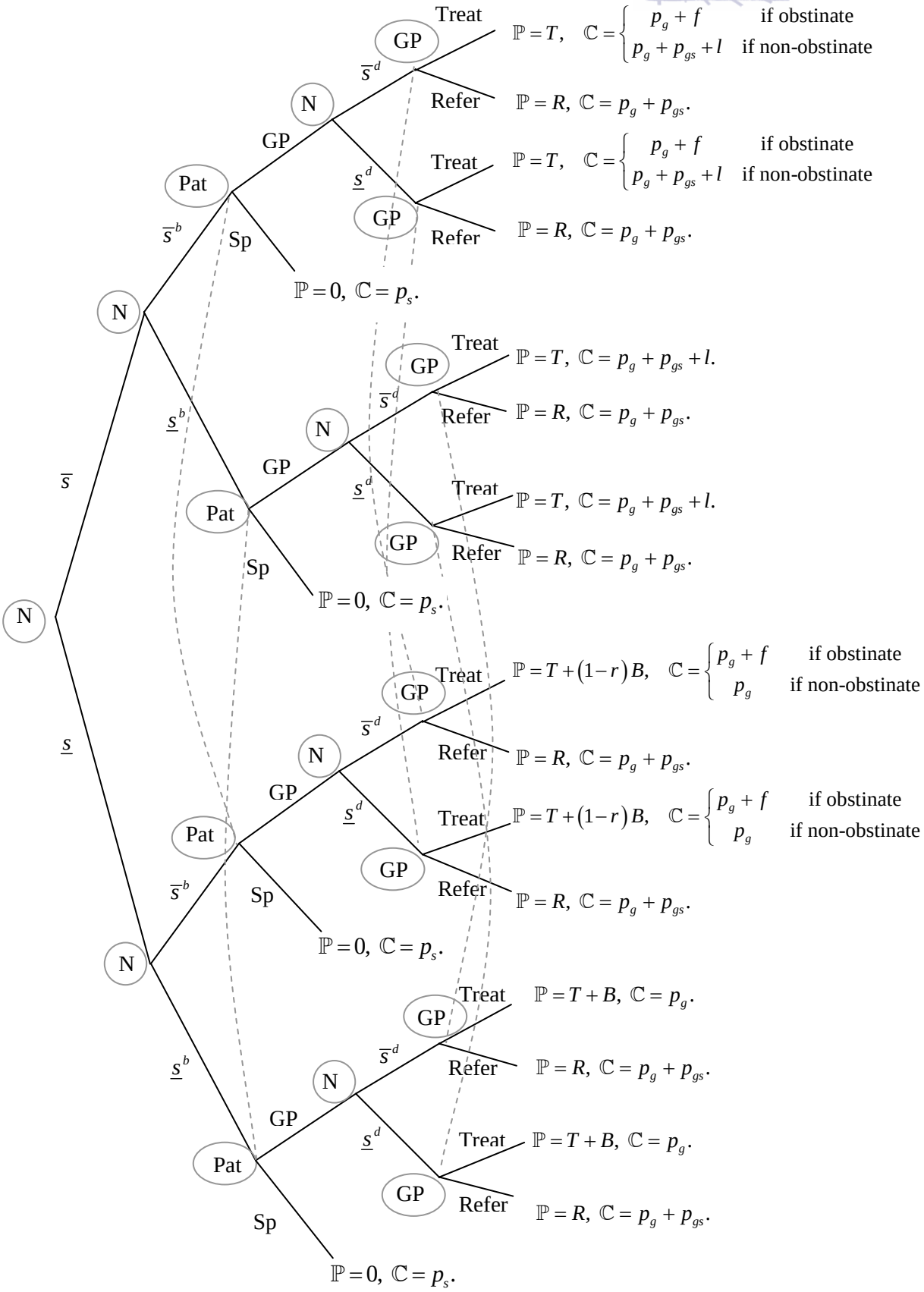

Figure 1: GP's and patient's decision tree: GP's payoffs ($\mathbb{P}$) and patient's costs ($\mathbb{C}$).