The “Gatekeeping” Role of General Practitioners. Does Patients’ Information Matter?∗

Paula González†
CORE-Université catholique de Louvain

Abstract

We deal with a principal-agent model in which the health authority acts as a principal for both a patient and a General Practitioner (GP). In this framework, we study the role of GPs as filters for secondary care, emphasizing the implications that patients’ information may have for health authorities. We derive the GP’s payment contract that induces him to perform diagnosis and follow its recommendation, as well as the level of copayments that provide patients with incentives to select the appropriate medical provider. We show that when patients can freely choose their provider, the quality of their information has contradictory effects. The higher this quality is, the lower the expected losses the patient bears. A higher quality, however, worsens the GP’s agency problem, as GPs have more incentives to use patients’ information as a substitute for their own diagnosis. We also analyze the role of patients’ pressure for referral on the choice of the optimal system to access secondary care.

JEL classification: D82, H51, I18, L51.
Keywords: General Practice, Incentives, Patients’ beliefs, Patients’ pressure, Referrals.

∗I am indebted to Maurice Marchand and Nicolás Porteiro for useful discussions and valuable suggestions. I am also grateful to Louis Eeckhoudt, Izabela Jelovac and Motohiro Sato for their helpful comments. Financial support from Fundación Pedro Barrié de la Maza is gratefully acknowledged. The usual disclaimers apply.
†Address for correspondence: CORE-Université catholique de Louvain. 34, Voie du Roman Pays. B-1348 Louvain la Neuve (Belgium). Phone: 32-10/47.81.62. Fax: 32-10/47.43.01. E-mail: gonzalez@core.ucl.ac.be
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 may have for health authorities. Patients’ beliefs will be shown to determine not only patients’ behavior itself, but also GPs’ incentives to perform a diagnosis and make medical recommendations.

The gatekeeping role of GPs is in action when patients are charged with low or zero money prices when accessing the medical system, and primary care is the only point of contact for all kinds of non-urgent medical conditions. In other words, patients’ access to more specialized services is regulated by their GPs.

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 gatekeeping 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 generally leads to lower health care costs.1 With GP gatekeeping, specialists are said to be used more efficiently because patients who have problems more appropriately treated in primary care are screened out by GPs. This explains the interest in gatekeeping both in Eastern European countries, that 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)).

The gatekeeping role of the GPs also points out the relevance of the link between primary and specialist care in controlling costs and quality. Therefore, as Scott (2000) highlights, the regulation of the GPs and the incentive structures they face have significant implications for costs and health outcomes in health care systems.2 In this respect, although GPs make many different types of decisions that influence the care received by the patient, GP’s referral behavior is perhaps the most important decision in relation to overall costs in health care systems where GPs are gatekeepers.

Apart from GP’s incentives, there exists another key element when comparing the pros and cons of these two different types of health systems. This is the behavior of the patients and the beliefs they may have on the treatment they need. On the one hand, if patients have a

---

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, Coxon 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.
sufficiently accurate information on their problem, induce them to self-select themselves and visit either the GP or the specialist depending on their belief, may be more efficient than a compulsory visit to the GP. On the other hand, patients’ information may become patients’ pressure to obtain a referral to specialized treatment.

Since the pioneering work by Arrow (1963) there has been considerable attention in the literature devoted to whether physicians take advantage of the fact that patients are less well informed than they are (see McGuire (2000) for a recent survey on physician agency). However, there is also some 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, 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 main contribution of this paper is to stress the importance of the role of patients as one major determinant when choosing between gatekeeping or non-gatekeeping systems. We show that both the quality of the patients’ information as well as the patients’ pressure for referral, are the two key elements that drive the choice of the optimal system.

We derive the GP’s contract and the patient’s level of copayments that minimize social costs. The decision of the GP to perform a diagnosis, and the outcome of such diagnosis are hard to verify. At the same time, patients’ beliefs are private information. Hence, the level of copayments set to the patient together with the incentives included in the GP’s contract will be key factors in determining their behavior. We compute the GP’s contract that induces him to perform a diagnosis and follow its recommendation (either to treat or refer the patient), and the level of copayments that induce the patient to select the adequate medical provider.

Copayments are generally used to fight moral hazard and avoid patients’ overconsumption of medical services. In some countries, copayments have been introduced as financial incentives at the patients’ side to stimulate the gatekeeping role of the GPs.3 This is in line with the role copayments will play in our model. We introduce them to avoid a systematic utilization of specialized treatment by patients, and induce them to select the medical provider on the basis of their belief about the severity of their condition.

We analyze the impact of the patient’s information on social costs and, hence, on the choice of the optimal system to access secondary care. If we consider only patient’s disutility and specialized treatment costs, we find that the higher is the quality of the patient’s information the better a non-gatekeeping system will be. From the patient’s perspective, this is due not

---

3In Belgium, for instance, from 2002 onwards patients pay lower copayments if they register with one specific GP. Moreover, very recently, the copayments for those patients who go directly to the specialist, without having seen first a GP, have increased. See, Schokkaert and Van de Voorde (2003) for a detailed explanation of the Belgian reform.
only to a decrease in his expected health losses, but also to the fact that the equilibrium level of copayments the patient faces becomes negligible when the quality of his information is sufficiently high. Concerning secondary treatment costs, the more precise the patient’s information, the smaller the number of unnecessary visits to the specialist.

Focusing on primary care costs, on the contrary, we obtain that an increase in the accuracy of the patient’s beliefs generates higher costs for the health authority when the system is a non-gatekeeping one. The reason is that, since only those patients who think they are mild visit the primary provider, the patient’s belief is a source of pre-diagnosis information that reduces the incentives of the GP to incur costly diagnosis. In this sense, patient’s information acts as a substitute for GP’s diagnosis.

When choosing between the two kinds of health systems, however, one has to take the consequences of patients’ pressure to obtain a referral for specialized treatment also into account. In this respect, we show that if this pressure is sufficiently high, a gatekeeping system maybe unable to guarantee a successful process of diagnosis and treatment-referral to the patients, while this is not a problem under non-gatekeeping. Therefore, the optimal choice will depend on the relative importance of these effects. In particular, the lower is the patient’s pressure for referral, the more likely that a gatekeeping system dominates a non-gatekeeping one, and vice versa.

Although primary care is being recognized as the mainstay of many health care systems in developed countries, there has been little research by economists into general practice. The work by García-Mariñoso and Jelovac (2003) is the first that, focusing on GP’s incentives, provides an uniform theoretical framework in which the performance of gatekeeping systems is compared with the performance of those systems where GP’s referral is not compulsory. They find that, whenever GP’s incentives matter, a gatekeeping system is superior to a non-gatekeeping one. However, key issues of the modelization and, consequently, the results are different from ours. First, they do not explicitly address the issue of how to discipline patients who might strategically choose to visit a specialist or a GP. In their model it is implicitly assumed that patients are provided with incentives in order to avoid a systematic use of specialist treatment. In this paper, by explicitly modelling the patients’ behaviour, we are able to analyse the implications of both patients’ information and patients’ pressure to obtain a referral on the choice the optimal system. Secondly, in García-Mariñoso and Jelovac GP’s quality of the diagnosis is endogenously determined and is perfect under the optimal contract. This implies that the positive effect of non-gatekeeping systems, as they avoid unnecessary delays before accessing specialist care, is absent from their comparison (since GP’s diagnosis is totally accurate at equilibrium). In our model, on the contrary, even if the GP can strategically decide whether to perform or not the diagnosis, the accuracy of such diagnosis is not a choice variable to him. This allows us to show that when the precision of the GP’s diagnosis increases, gatekeeping is more likely to dominate
(what is in line with the results in García-Mariño and Jelovac). However, we find that even in the limit case with perfect diagnosis, non-gatekeeping may be superior if patients’ pressure for referral is sufficiently high.

To our best knowledge, only in García-Mariño (1999) a description of how the insurer can regulate access to specialized care by manipulating the patients’ insurance contract is provided. There are two main differences between her approach and ours. First, in our model the fraction of patients that visit directly the specialist is exogenously given (those who believe to suffer from a severe illness), while in García-Mariño the optimal screening of patients is endogenously determined. In exchange, 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. Secondly, the GP’s choice in her model is binary and extreme: either no diagnosis is made and the patient is directly referred to the hospital, or a perfect diagnosis is undertaken and the GP provides treatment only to low severe patients. As a consequence of this, the interaction between the patients’ information (together with the pressure for referral they may exert) and the GP’s opportunistic behaviour, which in our model is the main determinant when choosing the optimal system to access secondary care, is absent from her work.

The rest of the paper is organized as follows: In the following section, we present the model. Section 3 analyzes both the patient and the GP’s behavior. In Section 4 we derive the optimal patient’s copayment 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

The model in this paper is an adaptation of the set-up proposed by Jelovac (2001) and García-Mariño 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 attention, but we will consider him as a passive agent, as the analysis of his behavior is out of the scope of this article.

*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: $s$ and $\bar{s}$, which indicate whether the patient suffers from either a low or high severity of the illness. 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 how serious his illness is. His symptoms, however, provide him with a private signal or belief $(s^b)$ about the severity of his health problem. We assume that the
probability of his not mis-recognizing his symptoms to be $\beta \in (\frac{1}{2}, 1)$. Formally:

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

The patient, therefore, seeks health care from a medical provider. He will demand medical attention either from a GP or from a specialist. His decision on from whom demand treatment depends on the existing institutional framework. This means that, in a gatekeeping system, the patient has no choice and has to visit the GP. In a non-gatekeeping system, however, the patient can choose to visit either the GP or the specialist. As we will see later, the level of copayments set by the health authority, together with the patient’s belief about the severity of the illness, determine the patient’s choice of medical provider.

We consider the patient to be endowed with a utility function that depends on two components: his net monetary income ($Y$) and the health loss he suffers ($L$). We denote this utility function by $B(Y, L)$. It is quasi-linear in income but strictly decreasing and convex in health losses. Formally:

$$B(Y, L) = Y - K(L),$$

with $K(0) = 0$, $K'(\cdot) > 0$ and $K''(\cdot) > 0$. The convexity of $K(L)$ makes the patient risk averse with respect to health losses. We denote by $\bar{Y}$ the value of the patient’s initial income.

The utility of the patient is undermined by two main sources. First, from the health loss that the patient suffers when he receives primary care and a referral was necessary. These losses can be understood as the cost of waiting for specialized treatment. We consider that the patient knows the health loss he may suffer, which can take two values: $L \in \{l, \bar{l}\}$, with $0 < l < \bar{l}$. We define the probability that a patient, when bearing a health loss, incurs the cost $l$, as $r \in [0, 1]$.\(^4\)

Secondly, the patient also incurs a monetary cost. The health authority has to set certain copayments to induce the patient to enter the health care sector either on the primary level, or directly on the secondary one. These copayments, therefore, do not reflect the cost of the service and are only introduced to discipline patient’s behaviour.\(^5\) The set of copayments 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 directly specialized medical care.\(^6\)

As we have already claimed in the Introduction, we are interested in analyzing how the pressure of the patient to obtain a referral from the GP, affects the performance of the health

\(^4\)One could think that different health losses should be associated to different degrees of severity. In our model, however, low severe patients will never suffer from a health loss.

\(^5\)For this reason, copayments only make sense in non-gatekeeping systems, where patients can freely choose their medical provider.

\(^6\)As in our model the patient is endowed with a quasi-linear utility in money, copayments do not interfere with financial insurance issues. This problem is analysed in detail by García-Mariñoso (1999).
system. We model this pressure as the probability that the patient rejects GP’s treatment. In case of doing so, he will demand private specialized treatment at a cost $f$. This way, he avoids any potential health loss, although he bears the full cost of receiving specialist treatment. In particular, we assume that those patients with a high loss of waiting ($\bar{l}$), and who believe to be suffering from a high severity (i.e., $s^b = \pi$), will decide to reject GP’s treatment and pay for private services. Under this construction, the probability that the patient rejects a treatment prescription by the GP is:

$$\Pr(rej) = \begin{cases} r & \text{if } s^b = \pi \\ 0 & \text{if } s^b = \bar{s} \end{cases}.$$ 

Hence, implicitly, we are making the following assumption concerning the value of the private fee.

**Assumption 1** The costs of private specialized treatment $f$ is such that:

- It always pays for the patient entering the public health sector and not demand directly private treatment.
- If the patient is type $l$, he will always accept the GP’s recommendation.
- If the patient is type $\bar{l}$, he will accept a GP’s recommendation of treatment if and only if $s^b = \bar{s}$.

This assumption ensures that a problem of pressure for referral emerges. It requires that: On the one hand, the private fee is sufficiently high so that patients with a low waiting cost do never opt for the private alternative. On the other hand, the fee is not too high so that the private sector is an outside option that patients with a high cost of waiting may use. This will occur whenever the GP recommends treatment, but the patient believes he should be referred.

**The General Practitioner.**

We consider that the GP can cure a patient only if the severity of the condition is low. The specialist, on the contrary, can heal both levels of severity. If the patient visits the GP, he performs a diagnosis, which yields a signal ($s^d$) about the severity of the patient’s condition. We assume the probability of receiving a correct signal to be $\delta \in \left(\frac{1}{2}, 1\right]$. Formally:

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

We also consider that $\delta > \beta$, i.e., once the GP has performed the diagnosis, his level of knowledge about the true severity of the illness exceeds that of the patient.

In performing the diagnosis, the GP incurs a disutility, that we denote by $c_d$. As part of the diagnosis process, the physician also observes the patient’s belief about his true condition. Combining the diagnosis outcome with this piece of information, the GP decides
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.

Note that both the GP’s decision to perform diagnosis, and the outcome of such diagnosis are hard to verify. Then, the incentives included in the GP’s payment contract will crucially determine his behavior. We take the contract structure proposed by González (2003). 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 decides to treat the patient, he receives a payment \(T\). In this latter case, if the patient follows the GP’s recommendation and is cured, the GP receives a bonus \(B\). This bonus can be interpreted as a premium for cost-containment, since primary care treatment is cheaper than specialist one. Moreover, as will be shown later on, \(B\) will be an important instrument for offering incentives to the GP.

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 the specialist.

We denote by \(c_s\) the costs of the specialized services, which include not only the treatment costs but also the payments made to the specialist. We normalize, without loss of generality, the GP treatment costs to zero.

The health authority designs the GP’s contract and the patient’s level of copayments 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 payoffs to both the GP and the specialist) and the patient’s expected disutility (which include both his expected health losses and the level of copayments he pays).\(^7\)

From all the possible contracts the health authority may offer to both the GP and the patient, we concentrate on a particular one. On the one hand, the health authority designs the contract that induces the GP to perform the diagnosis and follow its recommendation, i.e., to treat the patient whenever the signal received from the diagnosis is \(s^d = s\) and refer him if \(s^d = \overline{s}\). On the other hand, the level of copayments will be designed in such a way that it ensures that a patient visits directly the specialist if \(s^b = \overline{s}\) and chooses to go to the GP when \(s^b = s\). The reader should note that, despite we consider this scenario as the most appealing one in terms of policy implications, it may not be, for certain parameter values, the most desirable one from a social point of view.

\(^7\)One should note that copayments are not introduced in the model as revenues for the health authority, and they only appear as costs for patients. All our results would remain valid if we consider copayments as income for the health authority, provided there exists a cost in raising public funds.
We denote by $C_{GP}$ the expected financial costs associated with primary care, $C_{Sp}$ those for specialized treatment and $C_{Pat}$ the patient’s expected disutility.\(^8\)

**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 copayments. 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 performs 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 or, in case he accepts, if the patient recovers his health, the game ends. Otherwise, the patient is referred to the specialist. Figure 1 summarizes the timing of the game for both the gatekeeping (gk) and the non-gatekeeping (Ngk) scenarios.

---

\(^8\)For convenience, we will work throughout the model with the patient’s expected costs, instead of with his expected utility. Hence, $C_{Pat}$ is simply the difference between the patient’s initial wealth $Y$ and his expected utility $E(B(Y,L))$. 

---

As usual, we solve the game by backward induction.

### 3 Agents’ Behavior

#### 3.1 Patient’s Behavior

In this subsection we analyze how the level of copayments determines the decision of the patient to either visit the GP, or directly request specialized medical treatment. This analysis only applies when considering systems in which patients are not obliged to compulsory visit the GP. As mentioned before, we concentrate on the restrictions on the copayments ensuring that a patient will decide to visit directly the specialist if $s^b = \bar{s}$ (i.e., when he believes to suffer from a high severity), and chooses to go to the GP when $s^b = \underline{s}$ (i.e., when his belief is that he has a
minor condition). Such level of copayments, moreover, ensures that the problem of pressure is completely absent in non-gatekeeping systems.

Throughout this subsection, it is implicitly assumed that the payments made to the GP are such that he decides to perform the diagnosis, and follow its recommendation. That is to say, the GP decides to treat the patient whenever the signal received from the diagnosis is $s^d = \bar{s}$ and refers him if $s^d = \bar{s}$.9

In our model, the patient can suffer either from a high severity or a low severity illness, 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 the severity of his illness are:10

$$\Pr(\bar{s} | s^b = \bar{s}) = \Pr(s | s^b = \bar{s}) = \beta.$$ 

Analogously, the patient’s probabilities to misrecognize the severity of his illness are:

$$\Pr(s | s^b = \bar{s}) = \Pr(\bar{s} | s^b = \bar{s}) = 1 - \beta.$$ 

The patient, whatever his beliefs on the severity or his health cost, has the choice between two alternatives: go directly to the specialist, or go first to the GP.

If the patient goes directly to the specialist, his disutility is given by the copayment he has to pay ($p_s$), but no health loss is borne.

If the patient goes first to the GP he always pays the copayment $p_g$ and, then, if he is eventually referred to the specialist, the copayment $p_{gs}$.

Moreover, he may also suffer from a health loss whenever he receives treatment from the GP that does not heal him. However, as mentioned in Section 2, type-$T$ patients (those with a high waiting cost) always reject GP’s treatment if they believe to be in a severe condition. In this case, therefore, they do not incur neither the copayment $p_{gs}$ nor the health loss $\bar{l}$, but they have to pay the private fee $f$.

Formally, when $s^b = \bar{s}$ the patient suffers from a cost $p_{gs}$ either if his belief is wrong and his condition is really severe (with a probability $\Pr(\bar{s} | s^b = \bar{s})$) or if his belief is right (with a probability $\Pr(s^d = \bar{s} | s^b = \bar{s})$) but the GP’s diagnosis is wrong (with a probability $\Pr(s^d = \bar{s} | \bar{s})$). Moreover, he suffers from the health loss (either $\bar{l}$ or $\bar{\bar{l}}$) if his condition is severe (with a probability $\Pr(\bar{s} | s^b = \bar{s})$) but the GP’s misrecognize his symptoms (with a probability $\Pr(s^d = \bar{s} | \bar{s})$).

Similarly, when $s^b = \bar{s}$, if the patient is type-$L$ (with probability $(1 - r)$) he suffers from a cost $p_{gs}$ either if his condition is really severe (with a probability $\Pr(\bar{s} | s^b = \bar{s})$) or if it is mild (with a probability $\Pr(s^d = \bar{s} | \bar{s})$) but the GP’s diagnosis is wrong (with a probability

---

9These values will be explicitly computed in Subsection 4.2.
10See Appendix A for a more detailed explanation.
\( \Pr(s^d = \bar{s} | g) \). Moreover, if his condition is really severe (with a probability \( \Pr(\bar{s} | s^b = \bar{s}) \)) but the GP’s misrecognize his symptoms (with a probability \( \Pr(s^d = \bar{s} | \bar{s}) \)) he will incur the health loss \( \ell \).

When \( s^b = \bar{s} \), if the patient is type-\( T \) (with probability \( r \)), he always rejects to be treated by the GP. Then, he will incur the cost of the private treatment \( f \), either if his condition is really severe (with a probability \( \Pr(\bar{s} | s^b = \bar{s}) \)) but the GP’s diagnosis is wrong (with a probability \( \Pr(s^d = \bar{s} | \bar{s}) \)) or if it is mild (with a probability \( \Pr(\bar{s} | s^b = \bar{s}) \)) and the GP’s diagnosis is right (with a probability \( \Pr(s^d = \bar{s} | \bar{s}) \)) Finally, the patient will pay \( p_{gs} \) if his condition is really severe (with a probability \( \Pr(\bar{s} | s^b = \bar{s}) \)) and the GP’s diagnosis is right (with a probability \( \Pr(s^d = \bar{s} | \bar{s}) \)) or if it is mild (with a probability \( \Pr(s^d = \bar{s} | \bar{s}) \)) but the GP’s diagnosis is wrong (with a probability \( \Pr(s^d = \bar{s} | \bar{s}) \)).

Hence, in comparing the patient’s expected utility when demanding first GP’s attention, or directly specialized care, we conclude the following:

If \( s^b = \bar{s} \) a type-\( L \) patient will choose to go directly to the specialist whenever:
\[
p_s \leq p_g + \beta (p_{gs} + (1 - \delta) K (\bar{l})) + (1 - \beta) (1 - \delta) p_{gs}.
\]

If \( s^b = \bar{s} \) a type-\( T \) patient will choose to go directly to the specialist whenever:
\[
p_s \leq p_g + \beta (\delta p_{gs} + (1 - \delta) f) + (1 - \beta) ((1 - \delta) p_{gs} + \delta f).
\]

If \( s^b = \bar{s} \) a type-\( L \) patient will choose to go first to the GP whenever:
\[
p_s \geq p_g + \beta (1 - \delta) p_{gs} + (1 - \beta) (p_{gs} + (1 - \delta) K (\bar{l})).
\]

If \( s^b = \bar{s} \) a type-\( T \) patient will choose to go first to the GP whenever:
\[
p_s \geq p_g + \beta (1 - \delta) p_{gs} + (1 - \beta) (p_{gs} + (1 - \delta) K (\bar{l})).
\]

From the conditions above, we see how for a type-\( L \) patient the cost of the private specialist plays a role, as the private sector is an outside option that he may actually use. For the type-\( L \) patient, this is not the case and, therefore, his decision is only determined by the copayments and his cost of delay.

It is worth noting that, even if we have two types of patients, we only have one set of copayments. Hence, we cannot construct a self-selection menu. This implies that, if we want the patients to select their provider according to their belief, we will have to ensure that the copayment levels provide appropriate incentives simultaneously to both types of patients. Taking this into account, we obtain the following lemma.

**Lemma 1** A patient, irrespectively of his type, will visit directly the specialist when \( s^b = \bar{s} \) and will go to the GP when \( s^b = \bar{s} \) if and only if:

- \( p_s - p_g \leq \beta (p_{gs} + (1 - \delta) K (\bar{l})) + (1 - \beta) (1 - \delta) p_{gs} \) and
- \( p_s - p_g \geq \beta (1 - \delta) p_{gs} + (1 - \beta) (p_{gs} + (1 - \delta) K (\bar{l})) \).
As this Lemma shows, each type of patient will be crucial to determine one of the restrictions on the optimal level of copayments. The patients with a low delay cost (type-$t$) are less interested in visiting directly the specialist. Hence, the condition ensuring that whenever \( s^b = s \) the patient visits the specialist is determined by these type-$t$ patients.\(^{11}\) On the contrary, type-$r$ patients are the ones who define the second condition, as they are more reluctant to visiting the GP.

It is worth noting that, the higher the quality of the patient’s information is (i.e., the higher \( \beta \) is) the milder both restrictions are. This is a natural result since, what the health authority is trying to induce through the copayments, is precisely that the patient uses his own information when selecting the medical provider. The more accurate this information is, therefore, the smaller the expected costs of his self-selection. The effect of an increase in the accuracy of the GP’s diagnosis on the conditions, however, is ambiguous. On the one hand, it makes more demanding the condition ensuring that whenever \( s^b = s \) the patient visits the specialist but, on the other hand, it relaxes the condition that makes the patient visits the GP when he believes to suffer from a low severity.

The conditions computed in Lemma 1 are the ones that the health authority will use to compute the optimal copayment levels, as we will see in Subsection 4.1.

### 3.2 General Practitioner’s Behavior

In this subsection we derive the conditions that the GP’s payment contract has to fulfill in order to ensure that the GP decides to perform the costly process of diagnosis, and afterwards he sticks to its recommendation.

Throughout this sub-section, it is implicitly assumed that the level of patient’s copayments is such that in a non-gatekeeping system a patient demands primary attention if \( s^b = s \) and specialized attention if \( s^b = \bar{s} \).\(^{12}\)

In our model, the GP faces a population of patients that can suffer from either a high severe or a low severe illness, with ex-ante the same probability. Once the patient visits the GP, he updates these probabilities. In order to do so, he uses two pieces of information: the patient’s beliefs and the signal received from the diagnosis.\(^{13}\) It is worth mentioning that we are ruling out the possibility that the physician wrongly observes the patient’s belief.

Once this information has been acquired, the probabilities of correctly diagnosing a low

---

\(^{11}\)It can be shown that this is always true under Assumption 1.

\(^{12}\)These values will be explicitly computed in Subsection 4.1.

\(^{13}\)We consider that the observation of the patient’s beliefs is part of the diagnosis process. Alternatively, we could have assumed that the GP acquires information about the patient’s belief, even if he does not perform the diagnosis. However, we consider this alternative less appealing as, constructing the model that way, a gatekeeping system would trivially be more costly than a non-gatekeeping one, as the latter would be simply a subset (when \( s^b = s \)) of the former.
severity of the illness are:\textsuperscript{14}
\[
\Pr \left( s^d = s \cap s^b = s \right) = \frac{\delta \beta}{\delta \beta + (1 - \delta)(1 - \beta)} = 1 - \Pr \left( \overline{s} | s^d = s \cap s^b = \overline{s} \right).
\]
\[
\Pr \left( s^d = s \cap s^b = \overline{s} \right) = \frac{\delta (1 - \beta)}{\delta (1 - \beta) + (1 - \delta) \beta} = 1 - \Pr \left( \overline{s} | s^d = s \cap s^b = \overline{s} \right).
\]

Analogously, the probabilities of wrongly diagnosing a low severity of the illness are:
\[
\Pr \left( s^d = \overline{s} \cap s^b = s \right) = \frac{(1 - \delta) \beta}{(1 - \delta) \beta + \delta (1 - \beta)} = 1 - \Pr \left( \overline{s} | s^d = \overline{s} \cap s^b = s \right).\]
\[
\Pr \left( s^d = \overline{s} \cap s^b = \overline{s} \right) = \frac{(1 - \delta) (1 - \beta)}{(1 - \delta) (1 - \beta) + \delta \beta} = 1 - \Pr \left( \overline{s} | s^d = \overline{s} \cap s^b = \overline{s} \right).
\]

Once the GP has diagnosed the true severity of the condition, he then decides on the best option for the patient. Regardless of the severity of the illness, however, the doctor always has the choice between two alternatives: treat the patient or refer him to the specialist.

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

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

1. If $s^d = \overline{s}$ and $s^b = \overline{s}$, the GP will treat the patient whenever $R - T \leq \frac{B(1 - \beta)}{\delta \beta + (1 - \delta)(1 - \beta)}$ and refer him otherwise.

2. If $s^d = \overline{s}$ and $s^b = s$, the GP will treat the patient whenever $R - T \leq \frac{B(1 - \beta)(1 - r)}{\delta(1 - \beta) + (1 - \delta) \beta}$ and refer him otherwise.

3. If $s^d = s$ and $s^b = \overline{s}$, the GP will refer the patient whenever $R - T \geq \frac{B(1 - \beta)}{(1 - \delta) \beta + \delta (1 - \beta)}$ and treat him otherwise.

4. If $s^d = s$ and $s^b = s$, the GP will refer the patient whenever $R - T \geq \frac{B(1 - \beta)(1 - r)}{(1 - \delta)(1 - \beta) + \delta \beta}$ and treat him otherwise.

We see how the conditions above are given by the relationship between the incentive payment $B$ and the difference between the two “safe” payments: $R$ for a direct referral to the specialist and $T$ for recommending treatment. In addition to this, it is worth mentioning that the conditions

\textsuperscript{14}See Appendix A for a more detailed explanation.
that the GP’s payment scheme have to fulfill in order to effectively induce him to follow the recommendation of the diagnosis are different for the two institutional frameworks. In a non-gatekeeping system, since only patients who believe to suffer from a mild severity visit the GP, the only relevant restrictions are (1) and (3). In a gatekeeping system, however, the four conditions have to be fulfilled. This leads to the following lemma.

Lemma 2 The GP always follows the recommendation given by the diagnosis if and only if:

- In a non-gatekeeping system:
  \[
  R - T \geq \frac{B (1 - \delta) \beta}{(1 - \delta) \beta + \delta (1 - \beta)} \quad \text{and} \quad (IC_{FD1}^{Ngk})
  \]
  \[
  R - T \leq \frac{B \delta \beta}{\delta \beta + (1 - \delta) (1 - \beta)}. \quad (IC_{FD2}^{Ngk})
  \]

- In a gatekeeping system:
  \[
  R - T \geq \frac{B (1 - \delta) \beta}{(1 - \delta) \beta + \delta (1 - \beta)} \quad \text{and} \quad (IC_{FD1}^{gk})
  \]
  \[
  R - T \leq \frac{B \delta (1 - \beta) (1 - r)}{\delta (1 - \beta) + (1 - \delta) \beta}. \quad (IC_{FD2}^{gk})
  \]

When there is a gatekeeping system, the GP faces all kind of patients, those who believe to be suffering from a low severity and those who believe to have a low severity illness. In order to ensure that the GP will always follow his diagnosis, we have to induce him to do so, even in those cases in which this is contrary to the patient’s beliefs. In one of these cases, when \( s^d = s \) and \( s^h = \pi \), following the diagnosis and, hence, recommending treatment, maybe rejected by the patient. As a result, the higher the pressure of the patient for a referral to the specialist (measured by the probability \( r \) that the patient has a high cost of waiting) the more difficult to induce the GP to stick to his diagnosis recommendation.

When there is non-gatekeeping, the GP always receives patients who think they suffer from a low severity. This implies that the risk of recommending a treatment that is rejected by the patient does not exist. The lack of this “pressure”, relaxes the restrictions.

It is relevant to analyze how the conditions are affected by the “accuracy” of the information the agents have. First, it can be shown that for both gatekeeping and non-gatekeeping systems, the higher the precision of the GP’s diagnosis (i.e., the higher is \( \delta \)) the milder the restrictions are. This effect has an intuitive interpretation as it implies that it will be easier to induce the GP to follow the diagnosis as it becomes more accurate.

Second, the effect of the quality of the patient’s information (\( \beta \)) is not so clear. On the one hand, it makes easier to induce the physician to provide treatment when \( s^d = s \) but, on the other hand, it makes the condition to recommend referral after observing \( s^d = \pi \) more demanding. As
we will see later, the interaction between these two effects will be crucial when comparing the costs borne by the health authority in gatekeeping systems versus non-gatekeeping ones.

At this point, we are able to make a partial comparison and see whether it is easier to induce the GP to follow the diagnosis recommendation under gatekeeping or under non-gatekeeping.

**Remark 1** *Even in the absence of pressure (i.e., \( r = 0 \)), the payments \( (R - T \) and \( B \) needed to induce the physician to follow the recommendation given by the diagnosis, are strictly higher in a gatekeeping system than in a non-gatekeeping one.***

We already know that patient’s pressure for referral is a problem arising only in gatekeeping systems. This remark shows that, even if we disregard this issue, it is still more difficult to induce the physician to follow the recommendation given by the diagnosis under gatekeeping than under non-gatekeeping. The reason is that, under gatekeeping, the physician is more likely to be confronted with a situation in which his diagnosis outcome contradicts the patient’s belief. Hence, as the patient’s belief is correlated with the true severity of the illness, the physician’s recommendation is more likely to be incorrect when it is not in line with the patient’s prior. This makes more expensive that the health authority induces the GP to always follow his diagnosis recommendation.

In our model, the GP does receive 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 perform the diagnosis or not, and what to do in case he does not perform it (either systematically treat or refer the patient). When the GP decides to perform the diagnosis, it could be the case that, afterwards, he decided not to follow the recommendation of such a diagnosis. When the conditions written in Lemma 2 hold, however, we can ensure that the GP will stick to the diagnosis recommendation.

The derivation of the GP’s expected utility when the GP performs the diagnosis and follows its recommendation \( (U) \), for both gatekeeping and non-gatekeeping systems, is detailed in Appendix B. The simplified structure of the GP’s expected utilities is given by:

- In a non-gatekeeping system:
  \[
  U^\text{Ngk} = T + (R - T) \left[ \delta + (1 - 2\delta) \beta \right] + B\delta\beta - c_d.
  \]

- In a gatekeeping system:
  \[
  U^\text{gk} = \frac{1}{2} \left[ R + T + \delta B \left( 1 - (1 - \beta) r \right) \right] - c_d.
  \]

Once the GP’s expected utility has been computed, we can obtain the restrictions that determine when he decides to perform the diagnosis. 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 \((T + \frac{1}{2}B (1 - (1 - \beta) r)\) if there is gatekeeping, or \(T + B\beta\) if the system is a non-gatekeeping one). These restrictions (together with the ones computed in Lemma 2) are the ones that the health authority will include later on in his optimization program as incentive constraints.

The following lemma summarizes the GP’s decision of performing the diagnosis.

**Lemma 3** The GP decides to perform the diagnosis if and only if:

- **In a non-gatekeeping system:**
  
  \[
  R - T \geq \frac{B (1 - \delta) \beta + c_d}{(1 - \delta) \beta + \delta (1 - \beta)} \quad \text{and} \quad (IC_{PD_1}^{N})
  \]
  
  \[
  R - T \leq \frac{B\delta\beta - c_d}{\delta\beta + (1 - \delta) (1 - \beta)}. \quad (IC_{PD_2}^{N})
  \]

- **In a gatekeeping system:**
  
  \[
  R - T \geq 2c_d + (1 - \delta) B (1 - (1 - \beta) r) \quad \text{and} \quad (IC_{PD_1}^{g})
  \]
  
  \[
  R - T \leq \delta B (1 - (1 - \beta) r) - 2c_d. \quad (IC_{PD_2}^{g})
  \]

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 \((\delta)\) makes the conditions less demanding. Also, analogously as in Lemma 2, the effect of an increase in the quality of the patient’s information is ambiguous. The impact of the patient’s information on the GP’s incentives to perform the diagnosis, however, is driven by a completely different effect for a gatekeeping and a non-gatekeeping system.

In a non-gatekeeping system, since only a patient with \(s^b = s\) visits the GP, the doctor acquires information simply by receiving the patient. The higher is the quality of the patient’s information, the more likely it is that his true severity is low. In this setting, therefore, the patient’s belief is a source of pre-diagnosis information that affects the incentives of the GP to incur in costly diagnosis.

In a gatekeeping system, since the patient visits the GP irrespectively of his beliefs, the visit does not convey any information by itself and there is no such acquisition of information. In exchange, the accuracy of the information of the patient plays an important role as it indirectly decreases the expected risk that the patient rejects a treatment that would successfully heal him.

Combining Lemmas 2 and 3 we find:

**Lemma 4** If the GP decides to perform the diagnosis:
In a non-gatekeeping system, he will always follow its recommendation.

In a gatekeeping system, he will always follow its recommendation if and only if $IC_{FD}^{gk}$ are fulfilled.

In a non-gatekeeping system we can ensure that, for every value of $c_d$, the conditions that have to be fulfilled so that the GP is willing to follow the diagnosis recommendation $IC_{FD}^{Ngk}$ are always milder than the ones that induces him to perform the diagnosis $IC_{PD}^{Ngk}$. This means that, once the GP has decided to perform the diagnosis, he will always follow its recommendation, so we can disregard the associated incentive constraints. In a gatekeeping system, on the contrary, we cannot ensure that for every value of $c_d$, $IC_{FD}^{gk}$ constraints are always implied by $IC_{PD}^{gk}$. Therefore, once the outcome of the diagnosis, which signals the severity of the patient, is received by the GP, he may decide not to make use of it. This is given by the interaction of the GP’s acquisition of the patient’s information, together with the risk of recommending a treatment that may be rejected. Neither of these effects appear in a non-gatekeeping system since, in it, all the patients that visit the GP have $s^b = s$. As a result, no information about the patient’s signal is acquired through the diagnosis and there is no risk of treatment rejection at all.

Figures 2 and 3 represent the set of “incentive compatible” payment schemes, both under non-gatekeeping and under gatekeeping. Figure 2 shows how with non-gatekeeping the only relevant conditions are the ones needed to induce the GP to actually perform a diagnosis. Figure 3 represents the situation with gatekeeping for a (relatively) high value of $r$. In this case, inducing the GP to follow the diagnosis recommendation becomes a relevant problem for the health authority.

Figure 2: GP’s “incentive compatible” payments in a non-gatekeeping system.
Figure 3: GP’s “incentive compatible” payments in a gatekeeping system if $r$ is sufficiently high.

The following proposition states how the pressure of the patient to obtain a referral can certainly be an unsolvable issue.

**Proposition 1** Designing a contract that induces the GP to treat the patient when his signal is $s^d = \underline{s}$ and to refer him if $s^d = \overline{s}$:

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

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

**Proof.** See Appendix C.

As mentioned before, in a non-gatekeeping system, once the GP has decided to perform the diagnosis, he will always follow its recommendation. Therefore, in order to induce the GP to treat the patient when the signal he receives is $s^d = \underline{s}$ and to refer him if $s^d = \overline{s}$, it is always enough to design a payment contract that induces him to perform diagnosis. The restrictions always determine a non-empty feasible set of contracts and, therefore, an incentive contract can always be successfully designed.

In a gatekeeping system, however, this is not the case, and the result is determined by the value of the “patient’s pressure”. If the patient’s pressure is sufficiently high, it is impossible for the health authority to find values of $R, T$ and $B$ that simultaneously fulfill the constraints guaranteeing that, whenever the outcome of the diagnosis is $s^d = \underline{s}$, the GP always treats the patient, while when $s^d = \overline{s}$ he always refers him to the specialist. The reason for it is the following: Given the high risk of rejection, the minimum value of the bonus $B$ that the health authority has to set to induce the GP to treat a patient whenever $s^d = \underline{s}$ is so high that the GP
will also be willing to treat a patient when \( s^d = \pi \). Conversely, if the difference \( R - T \) set by the health authority is high enough to induce the GP to refer a patient if \( s^d = \pi \), this will generate that 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 be unable to guarantee a successful process of diagnosis and treatment-referral to the patients, while this is not a problem in non-gatekeeping systems.

It is interesting to see how the threshold \( \tilde{\tau} \) depends on the quality of the information of the agents. It can be checked that \( \tilde{\tau} \) is increasing in \( \delta \) and decreasing in \( \beta \). This implies that, on the one hand, the higher the accuracy of the physician’s diagnosis, the more likely a gatekeeping system is sustainable. On the other hand, however, the accuracy of the patient’s belief has a completely opposite effect. As the belief becomes more precise, the maximum threshold of pressure compatible with a gatekeeping system decreases. As the patient knows more, the physician will be less willing to effectively recommend treatment when this is contrary to the patient’s belief. In spite of this, since \( \beta < 1 \) and \( \delta > \beta \), we can always ensure that \( \tilde{\tau} > 0 \). This means that even if the patient’s information is extremely accurate, there always exist levels of pressure for which a gatekeeping system guarantees a successful process of diagnosis and treatment/referral choice.

4 The Health Authority’s Problem

In this section we concentrate on the problem the health authority faces. The problem is two-fold. First, if the system is a non-gatekeeping one, the health authority has to design the set of copayments that induce the patient to visit directly a specialist if and only if he believes he suffers from a high severity. Secondly, the health authority has to design the optimal contract that provides the GP with incentives to perform (and follow) a process of diagnosis.

The health authority aims at minimizing the 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 disutility (\( C_{Pat} \)). \( C_{GP} \), \( C_{Sp} \) and \( C_{Pat} \) are derived in Appendix B. The simplified expressions are as follows:\(^{15}\)

- In a gatekeeping system:

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

\[
C^{gk}_{Sp} = \frac{c}{2} \left( 2 - \delta - r (1 - \delta) \beta \right).
\]

\[
C^{gk}_{Pat} = \frac{1}{2} \left[ (1 - \delta) \left( (1 - \beta) r K (\bar{l}) + (1 - r) K (\bar{l}) \right) + r f ((1 - \beta) \delta + (1 - \delta) \beta) \right].
\]

\(^{15}\)The reader should note that, as long as the GP performs and follows the diagnosis, \( C_{Sp} \) and \( C_{Pat} \) are independent from the GP’s contract. Analogously, \( C_{GP} \) is not altered by the copayment levels, provided they induce the patient to select the medical provider according to his belief about the severity.
- In a non-gatekeeping system:
  \[ C^N_{GP} = T + (R - T) \left[ \delta + (1 - 2\delta) \beta \right] + B\delta\beta. \]
  \[ C^N_{Sp} = \frac{c}{2} (2 - \delta)\beta. \]
  \[ C^N_{Pat} = \frac{1}{2} \left( p_s + p_g + p_{gs} (\beta (1 - \delta) + 1 - \beta) + \tilde{l} (1 - \beta) (1 - \delta) \right), \]
  with \( \tilde{l} = rK (\tilde{l}) + (1 - r) K (\tilde{l}) \).

In the following sub-sections we will analyze independently the two tasks of the health authority.

### 4.1 The Optimal Patient’s Copayment Levels

As mentioned before, copayments are introduced in the model to discipline the patient’s behavior, and prevent him from systematically visiting the specialist to avoid the risk of unnecessary second visits. As the patient’s beliefs are private information, these copayments are the only tool the health authority has to alter the patient’s choice.

As copayments are not introduced in the model with fund-raising purposes, the copayment levels set by the health authority will be the ones that minimize the patient’s expected disutility \( (C_{Pat}) \). The health authority has to take into account the constraints computed in Lemma 1, which ensure that the patient will visit directly the specialist when \( s_b = s_a \) and will go to the GP when \( s_b = s_g \), as well as the fact that the copayments have to be non-negative.

The health authority’s optimization program is as follows:

\[
\begin{align*}
\min_{p_g, p_{gs}, p_s} C_{Pat} &= \frac{1}{2} \left( p_s + p_g + p_{gs} (\beta (1 - \delta) + 1 - \beta) + \tilde{l} (1 - \beta) (1 - \delta) \right) \\
\text{s.t} & \quad \begin{cases}
    p_s - p_g \leq \beta (p_{gs} + (1 - \delta) K (\tilde{l})) + (1 - \beta) (1 - \delta) p_{gs} \\
    p_s - p_g \geq \beta (1 - \delta) p_{gs} + (1 - \beta) (p_{gs} + (1 - \delta) K (\tilde{l})) \\
    p_g \geq 0, p_{gs} \geq 0, p_s \geq 0,
\end{cases} \\
\text{with } \tilde{l} &= rK (\tilde{l}) + (1 - r) K (\tilde{l}).
\end{align*}
\]

The following proposition characterizes the optimal level of copayments.

**Proposition 2** If the health authority wants the patient to visit directly the specialist when his belief is \( s_b = \pi \) and to go first to the GP if \( s_b = s_g \), the optimal level of copayments \( (p_g, p_{gs}, p_s) \) is as follows:

- If \( \frac{K(\tilde{l})}{K(l)} \leq \frac{\beta}{1 - \beta} \), then \( p_g^* = 0, p_{gs}^* = 0 \) and \( p_s^* = (1 - \beta) (1 - \delta) K (\tilde{l}) \).
- If \( \frac{K(\tilde{l})}{K(l)} \geq \frac{\beta}{1 - \beta} \), then \( p_g^* = 0, p_{gs}^* = p_{gs} (\beta, \delta, \tilde{l}, \tilde{l}) > 0 \) and \( p_s^* = p_s (\beta, \delta, \tilde{l}, \tilde{l}) > 0 \).

**Proof.** See Appendix C. \( \square \)

The first result that we extract from this proposition is that the optimal level of copayments differ depending on the degree of heterogeneity of the two types of individuals.
If individuals’ health losses are not too different, then setting only \( p_s > 0 \) is enough to induce both types of patients to follow their belief and only use specialized care if they believe they suffer from a high severity of the illness.

If individuals suffer from high enough differences in their costs of waiting, setting a positive copayment only to those individuals who go directly to the specialist is not enough to induce both types of patients to follow their belief. In this case, the value of \( p_s \) that guarantees that a patient with a high cost visits the GP when he believes he suffers from a low severity of the illness, is so high that makes a patient with a low cost never visit directly the specialist. The health authority, therefore, has to increase the cost of visiting the GP by increasing the level of copayment the patient has to pay when visiting the specialist after being referred by the GP.\(^{16}\)

Qualitatively, our results can be summarized by the following relation: \( p_s^* \geq p_{gs}^* \geq p_g^* = 0. \)\(^{17}\) This copayment structure is in line with the very recent Belgian reform, aimed at enhancing the gatekeeping role of GPs. In the Belgian system, however, copayments also have a dissuasive purpose and, therefore, there is a positive level of copayment for visiting the GP. As in this model it is not considered the possibility of dealing with healthy individuals who make unnecessary visits to the system, copayments are not introduced to limit excess of demand at the medical system. Then, the optimal level of copayment for visiting the GP is shown to be null.

Once the copayments needed to discipline the patient in a non-gatekeeping system have been computed, we can analyze, both in gatekeeping and non-gatekeeping systems, the effect of an increase in the accuracy of the patient’s information over the patient’s disutility.

**Corollary 1** *In both gatekeeping and non-gatekeeping systems, the patient’s expected disutility is decreasing in the accuracy of his belief.*

This corollary shows how, irrespectively of the health care system, the patient always benefits from a higher accuracy in the belief about his severity. In a non-gatekeeping system, the reason is two-fold: First, the health losses he bears are lower, as his self-selection of medical provider is less likely to be incorrect. Secondly, the monetary expenses he faces also diminish, as the copayments needed to induce him an appropriate selection of medical provider are decreasing in the accuracy of his belief. In a gatekeeping system, despite the effect is the same, the reason is different. A type-I patient does not benefit from an increase in \( \beta \), as his health losses are completely determined by the accuracy of the GP’s diagnosis. A type-I patient, on the contrary, does benefit from a more accurate belief, as this reduces the likelihood of incurring unnecessary expenses in a private specialist.

\(^{16}\)It can be shown that these equilibrium copayment levels are such that there exist values of the private fee \( f \) for which Assumption 1 is fulfilled.

\(^{17}\)The condition \( p_s \geq p_{gs} \) is fulfilled, except when the difference in the health costs of the patients is too extreme. Formally \( p_s \geq p_{gs} \) if and only if \( \frac{\kappa(\ell)}{\kappa(l)} \leq \frac{\beta^2}{(1-\beta)} \).
Finally, we can compare a gatekeeping system with a non-gatekeeping one focusing only on the patient’s side of the problem. It leads to the following proposition:

**Proposition 3** *Focusing only on the patient’s expected disutility, there exists a threshold $\beta^* < 1$, such that:*

- If $\beta \leq \beta^*$, the health authority prefers a gatekeeping system to a non-gatekeeping one.
- If $\beta > \beta^*$, the health authority prefers a non-gatekeeping system to a gatekeeping one.

This proposition shows how, 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 a gatekeeping system, 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.

### 4.2 The Optimal GP’s Payment Contract

The diagnosis of the severity of a patient’s ailment and the choice of the best alternative that he should be given can only be done by a qualified physician. This implies that it may not be possible for the health authority to control GP’s decisions in such activities. In this subsection, hence, we compute the GP’s optimal contract under asymmetric information, i.e., when neither his diagnosis nor the treatment strategy he employs are observable.

The payments the health authority offers to the GP will be the ones that minimize the health authority’s expected primary care costs ($C_{GP}$). 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 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 existent 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 in the health authority’s optimization program the GP’s incentive compatibility constraints ($IC$). These are the restrictions that induce the GP to perform the diagnosis and follow its recommendation (defined in Lemmas 2 and 3).
The health authority’s optimization program is as follows:

\[
\begin{align*}
\min_{D,T,R} & \quad C_{GP} \\
\text{s.t} & \quad \begin{cases} U \geq 0 & PC \\ T \geq c_d & LLC_1 \\ R \geq c_d & LLC_2 \\ B \geq 0 & LLC_3 \\ IC & \end{cases}
\end{align*}
\]

(2)

With \( C_{GP} \in \left\{ C_{GP}^{Ngk}, C_{GP}^{gk} \right\} \) and \( IC \in \left\{ \left( IC_{PD1}^{Ngk}, IC_{PD2}^{Ngk} \right), \left( IC_{PD1}^{gk}, IC_{PD2}^{gk}, IC_{FD1}^{gk}, IC_{FD2}^{gk} \right) \right\} \), depending on whether we are dealing with a non-gatekeeping system or with a gatekeeping one.

As mentioned before, in a non-gatekeeping system the constraints that induce the GP to perform diagnosis are always tougher than the ones that induce him to follow its recommendation. This means that, in this framework, the only relevant restriction for the health authority is \( IC_{PD}^{Ngk} \). On the contrary, in a gatekeeping system, both \( IC_{PD}^{gk} \) and \( IC_{FD}^{gk} \) are necessary to induce the GP to perform the diagnosis and follow its recommendation.

By Proposition 1 we know that, in a gatekeeping system, designing a contract that induces the GP to perform diagnosis and to treat a patient when his signal is \( s^d = s \) and to refer him when \( s^d = \overline{s} \), is only possible provided the probability that the patient rejects to be treated by the GP is not too high. Hence, hereinafter, we restrict our analysis to values of \( r \) such that \( r \leq \tilde{r} \).

Let us define \( \tilde{r} \equiv \frac{\delta - \beta}{(1 - \beta)(2\delta - 1)} \in (0, \tilde{r}) \). This threshold will determine two regions in which the impact of the patient’s pressure for referral affects differently the costs borne by the health authority.

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

**Proposition 4** If the health authority wants the GP to treat the patient when his signal is \( s^d = s \) and to refer him if \( s^d = \overline{s} \), the optimal contract \((R, T, B)\) is as follows:

- In a non-gatekeeping system:
  \[
  \begin{align*}
  R_{Ngk} &= \frac{(1 + (2\delta - 1)(1 - \beta))c_d}{(2\delta - 1)(1 - \beta)} \\
  T_{Ngk} &= c_d \\
  B_{Ngk} &= \frac{c_d}{(2\delta - 1)(1 - \beta) \tilde{r}}.
  \end{align*}
  \]

The health authority’s expected primary care costs are:

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

\[ R^{gk} = \frac{cd(3(2\delta - 1) + 4(1 - \delta)\Gamma(\delta, \beta, r))}{2\delta - 1} \]

\[ T^{gk} = cd \]

\[ B^{gk} = \frac{4cd\Gamma(\delta, \beta, r)}{(2\delta - 1)(1 - (1 - \beta)r)} \]

The health authority’s expected primary care costs are:

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

with \( \Gamma(\delta, \beta, r) = \begin{cases} 1 & \text{if } r \leq \tilde{r}, \\ \frac{(2\delta - 1)(1 - (1 - \beta)r)(\delta(1 - \beta) + (1 - \delta)\beta)}{2(1 - (1 - \beta)r)(\delta - (1 - \beta)(\delta(1 - \beta) + (1 - \delta)\beta) - \delta \beta)} & \text{otherwise.} \end{cases} \)

**Proof.** See Appendix C. ■

The first result extracted from this proposition is that, in both scenarios, the payment made to the GP for recommending treatment only covers the cost of performing the diagnosis and, therefore, does not play any role in providing incentives to the GP. Again in both situations, we can observe that \( T + B > R \). As \( R \) is a riskless payment (the GP receives it whenever a patient is directly referred), whereas \( T + B \) is gained if the GP treats but also the true condition of the patient was really mild, it is necessary that \( T + B > R \) in order the GP to be willing to treat a patient when \( s = s \).

Let us now move to independently study the two scenarios. In a non-gatekeeping system both \( R \) and \( B \) are decreasing in \( \delta \). The reason is clear: the higher the accuracy of the GP’s signal is, the cheaper to induce him to perform the diagnosis will be. Health authority’s costs from primary care are also decreasing in \( \delta \): the increase in the costs that the health authority incurs, since \( B \) is paid more often, is compensated enough by the fact that both \( R \) and \( B \) are lower at equilibrium.

Concerning the quality of the patient’s information, we see how in non-gatekeeping systems both \( R \) and \( B \) are increasing in \( \beta \). The higher the precision of the patient’s information is, the more likely that a patient who visits the GP suffers from a low severity of the illness. This has a perverse effect over the GP’s incentives. The information of the patients trades-off GP’s incentives to diagnose. In this respect, the higher the accuracy of the patient’s belief the more willing the GP will be to rely only on this information and skip his own diagnosis. The health authority’s costs derived from providing incentives in primary care, then, are unambiguously increasing in \( \beta \) in non-gatekeeping systems.

In a sense, what happens is that, for high values of \( \beta \), the HA is trying to induce the GP to perform an almost wasteful activity. As the accuracy of the patient’s belief increases, the extra information acquired through the diagnosis becomes smaller. The problem for the HA is that
the cost of ensuring that the GP performs a diagnosis increases as this task becomes more and more redundant.

When we analyze the gatekeeping scenario, we see how the degree of the patient’s pressure determines two regions with important differences. First, there exists a threshold \( r > 0 \) such that, for values of pressure below it, the patient’s interest in obtaining a referral, has no effects on the health authority’s costs. In this region, the marginal increase in the bonus for cost-containment due to the pressure, is compensated by the fact that \( B \) is paid less often at equilibrium. However, for values beyond \( r \), the incentive problem caused by the pressure is so severe that appears reflected in primary care costs. These costs are higher the larger the value of \( r \).

We can also assess the effects of the accuracy of the agents’ information on the optimal contract. On the one hand, an increase in the quality of the GP’s diagnosis has, analogously to the non-gatekeeping scenario, a positive impact on the costs, as it directly implies more information on the system and does not worsen the incentive problem. On the other hand, a higher accuracy of the patient’s belief has contradictory effects. For low levels of pressure (below \( r \)), it has no effect on costs. Otherwise, it implies higher government’s expenditures, as it fosters the patient’s pressure for referral. The more accurate the patient’s information is, the more difficult that the GP follows the diagnosis, when the recommendation of such diagnosis is contrary to the patient’s will.

By combining the discussion above with the one provided for the non-gatekeeping situation it follows that:

**Corollary 2** Expected costs from primary care are always (weakly) increasing in the accuracy of the patient’s belief.

Corollary 2 highlights the negative impact of the quality of the patient’s information on the GP’s incentive problem the health authority faces. Despite the effect is qualitatively analogous for both non-gatekeeping and gatekeeping scenarios, the reason is of a different nature. In a non-gatekeeping system the patient’s information generates a problem of “diagnosis substitution”. GPs have more incentives to use the patient’s belief as a substitute of their own diagnosis process and, hence, inducing the GP to perform the diagnosis becomes very expensive. In a gatekeeping system, on the contrary, the patient’s informations does not affect the GP’s incentives to perform diagnosis. In exchange, however, it worsens the problem of patient’s pressure for referral, as it makes more difficult for the health authority to avoid an excessive number of referrals by the GP.

After all the partial effects have been identified, we can now compare a gatekeeping system with and non-gatekeeping one, focusing only on the GP’s side of the problem. It leads to the following proposition.

25
Proposition 5  Focusing only on primary care expected costs, there exists a threshold $r^* > \tilde{r}$ such that:

- If $r \leq r^*$, the health authority prefers a gatekeeping system to a non-gatekeeping one.
- If $r > r^*$, the health authority prefers a non-gatekeeping system to a gatekeeping one.

With $r^* = \frac{\delta(1-\beta)-(\delta(1-\beta)+(1-\delta)\beta)^2}{(1-\beta)[\delta-(\delta(1-\beta)+(1-\delta)\beta)^2]}$.

Proof.  See Appendix C. 

Proposition 5 shows that only in those cases in which the patient’s pressure for referral is sufficiently high, a non-gatekeeping system generates lower primary health care costs for the health authority. If this pressure does not exceed the given threshold, however, the negative effect of the patient’s information on the GP’s incentives to skip the diagnosis, makes a gatekeeping system to be less costly.

It is interesting to study how the threshold $r^*$ depends both on the accuracy of the GP’s signal ($\delta$) and on the quality of the patient’s information ($\beta$).

Corollary 3  $r^*$ is always increasing in $\delta$.

This means that the higher the precision of the GP’s diagnosis, the more likely to be in the region where gatekeeping dominates. However, it should be noted that, even in the limit case with $\delta \to 1$, there exists a positive range of values for $r$ in which non-gatekeeping yields lower primary care costs.

The effect of $\beta$ on the threshold $r^*$ is not as clear.

Corollary 4  There exists a threshold $\delta < 1$ such that:

- If $\delta < \delta$, then $\frac{\partial r^*}{\partial \beta} < 0$.
- If $\delta > \delta$, then $\frac{\partial r^*}{\partial \beta} > 0$.

Thus, Corollary 4 makes clear the ambiguous effect that the quality of the patient’s belief has on the choice of the optimal system. If the precision of the GP’s diagnosis is low, an increase in the quality of the patient’s information decreases the value of $r^*$ and therefore, focusing only on primary health care costs, the more likely to be in the region where non-gatekeeping dominates. However, as the value of $\delta$ increases, an increase in $\beta$ makes more likely to be in the region in which gatekeeping is preferred by the health authority.

In the following section we provide a discussion on the global problem the health authority faces when designing the optimal system to access health care.
5 On the Choice of the Optimal System

This section provides some insights on the choice of the optimal health care system by the health authority. In order to perform the comparisons, we integrate the analyses of the two preceding sections, and we also take the expected costs of specialized treatment into account. This way, both financial costs (GP’s and specialist’s costs) and patient’s disutility are simultaneously considered.

As it has become clear throughout the paper, the quality of the patient’s beliefs as well as their pressure for referral are the two key elements that drive the health authority’s choice between gatekeeping and non-gatekeeping systems.

We start by considering the two extreme situations concerning the accuracy of the patient’s information. If $\beta \rightarrow 1$, i.e., if the patient’s information is almost completely accurate, then, as we have already seen in Corollary 1, patient’s expected disutility converges to zero in a non-gatekeeping system. As patients make no mistakes when selecting their medical provider, at equilibrium they bear null expected health losses. Moreover, the monetary costs the patient faces, which are given by the copayments set by the health authority to induce him to select the adequate provider, become negligible. On top of this, there is not an over-utilization of specialized services, as no low severe patients mis-interpret their symptoms. Therefore, the smaller specialized treatment costs that we observe under gatekeeping ($C_{Sp}^{N_{gk}} > C_{Sp}^{gk}$ if $\beta \rightarrow 1$), are only due to the existence of patients who leave the public system at the primary level (as $r > 0$). These larger specialized costs under non-gatekeeping are compensated enough by the higher expected disutility that patients bear, under gatekeeping, due to the cost of demanding private health care.

The discussion above provides strong arguments in favor of non-gatekeeping systems, as they allow health authorities to use and benefit from patients’ information.

A high accuracy of the patient’s information, however, has perverse effects on GP’s behavior. In particular, we have shown in Proposition 3 that, in a non-gatekeeping system, the agency costs borne by the health authority are increasing in $\beta$. Moreover, when $\beta \rightarrow 1$ inducing the GP to perform and follow the diagnosis becomes prohibitively expensive. This is because the higher is the quality of the patient’s information, the more likely that the patient does not make mistakes when he visits the GP. In this sense, the patient’s belief is a source of pre-diagnosis information that reduces the incentives of the GP to incur a costly diagnosis. Patient’s information, therefore, acts as a substitute for GP’s diagnosis. In a gatekeeping system, on the contrary, the accuracy of the patient’s information does not have this perverse substitution effect. It only affects costs, indirectly, through the patient’s pressure. In this sense, combining Proposition 4 and Corollary 2 it follows that when $\beta \rightarrow 1$, a gatekeeping system dominates from the GP’s incentives point of view. However, we should also recall that, as the quality of the patient’s belief increases, the
set of values for the pressure that make impossible to sustain diagnosis and treatment/referral in a gatekeeping system also increases.\footnote{In spite of this, we have already shown that for every value of $\beta < 1$, it holds that $\bar{r} > 0$. This means that there always exist levels of pressure compatible with a gatekeeping system.}

The last paragraphs seem to generate a puzzle. First, non-gatekeeping systems are very interesting for the health authority when focusing on the patient’s side of the problem. But, when GP’s incentives are taken into account, this system may be unsustainable. The message that seems to emerge from this result, then, is the following: \textit{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 the health authority wants the GP to perform costly diagnosis and follow its recommendation. If we considered a system in which patients have free choice of their medical provider but the GP always treats without diagnosing, such a system would dominate a gatekeeping system. It would allow the health authority to profit completely from the patient’s information, eliminating at the same time the GP’s incentive problem. Skipping the GP’s diagnosis process becomes optimal because it does not improve much over the patient’s belief, as this information is sufficiently accurate by itself.

Let us now move to the other extreme situation with $\beta \to \frac{1}{2}$. In this case, both patient’s expected health losses and copayments are higher in a non-gatekeeping system and, then, from the patient’s point of view a gatekeeping system dominates. From the GP’s incentives point of view, however, a gatekeeping system also dominates, provided the patient’s pressure is not too high ($r \leq r^*$). On top of this, a gatekeeping system saves with respect to a non-gatekeeping one in terms of unnecessary visits to the specialist ($C_{gk}^{Sp} < C_{Ngk}^{Sp}$ if $\beta \to \frac{1}{2}$). Therefore, only in those cases with a very severe problem of patient’s pressure for referral, systems where patients freely choose their medical provider would dominate.

For intermediate parameter values, the optimal choice will depend on the relative strength of two opposite effects. On the one hand, a non-gatekeeping system, even if it allows to successfully use patient’s information, it will generate a substitution of GP’s diagnosis by patient’s information. On the other hand, a gatekeeping system suffers the problem of patient’s pressure for referral, that may even make impossible a successful process of diagnosis and treatment/referral choice. The optimal system would depend on the quantitative relevance of the problem each system faces.

We summarize the discussion above in the following corollary. We restrict our attention to the case in which both systems are feasible alternatives for the health authority (i.e. when $r \leq \bar{r}$). For values of pressure above this threshold, non-gatekeeping is the only alternative for the health authority.

\textbf{Corollary 5} If the health authority wants the GP to perform costly diagnosis and follow its...
recommendation, and the patient to adequately select his medical provider, then:

- When $\beta \to \frac{1}{2}$ a gatekeeping system generically dominates.

- For intermediate values of $\beta$ there exist a threshold in the level of patient’s pressure such that, for values below it a gatekeeping system dominates whereas, for values above it, the optimal system is a non-gatekeeping one.

- When $\beta \to 1$ a gatekeeping system dominates. This scenario, however, is always dominated by a non-gatekeeping system in which the health authority forces the GP to always treat the patient (without performing a diagnosis).

Finally, it would be also interesting to study how the choice of the optimal system depends on the GP’s diagnosis accuracy ($\delta$). In general, what one would expect is that the higher the precision of the GP’s diagnosis, the more efficient is a system with compulsory visits to the GP, 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 by the following effect: the more accurate the GP’s diagnosis is, the cheaper for the health authority to induce him to perform the diagnosis and follow its recommendation. This is because, at equilibrium, the informational rent paid to the GP is decreasing in the precision of his diagnosis.

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. The main contribution of the paper is to stress the importance of the role of patients as one major determinant when choosing between gatekeeping and non-gatekeeping systems.

For these two alternative systems of accessing medical care, we have derived the GP’s payment contract that induces him to perform a diagnosis and follow its recommendation. Moreover, when patients can freely choose their medical provider, we have computed the level of copayments that provide them with incentives to select appropriately their provider.

In this setting, we have shown that both the quality of the patients’ information, as well as the patients’ pressure for referral, are the key elements that drive the choice of the optimal system. On the one hand, patients’ beliefs determine not only patients’ behavior itself, but also GPs’ incentives to perform a diagnosis and make medical recommendations. On the other hand, patients’ pressure for referral has been shown to be a relevant problem only in gatekeeping
systems, where it may be impossible to guarantee a successful process of diagnosis and treatment-referral to patients, if pressure is sufficiently high.

We have found that a higher quality of the patients’ information has contradictory effects on health authority’s expected costs. Whereas a higher quality reduces patient’s expected losses in both gatekeeping and non-gatekeeping systems, it increases the expected costs from providing primary care. When the accuracy of the patient’s information is sufficiently high, patients’ expected losses are always lower if patients can freely choose their medical provider. However, a gatekeeping system has been shown to lead to lower primary health costs (provided the patient’s pressure for referral is not too high).

Our model, hence, generates an, a priori, surprising prediction: a more accurate patients’ information may be negative from the health authority’s point of view. Contrary to what one could think, this implies that if patients have a sufficiently accurate information on their problem, induce them to self-select themselves may not necessarily be more efficient than a compulsory visit to the GP.

In terms of policy recommendations, our analysis suggests that when choosing between the two kinds of health systems, the consequences of patients’ pressure to obtain a referral and the quality of patients’ information have to be taken into account (as well as the interaction among them). The optimal choice will depend on the relative importance of these effects. In particular, for low levels of patient’s pressure, a gatekeeping system is more likely to dominate, and vice versa. Concerning the quality of the patients’ information, however, there is not a clear prediction. Still, in general, one could expect that if we want to provide the GP with incentives to diagnose, a non-gatekeeping system will be optimal only if there is a sufficiently high pressure for referral, and the quality of the patient’s information is not extreme (neither too bad nor to good).

Finally, we would like to highlight that, despite 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 as a contribution to this scarce literature, as well as to the ongoing debate over the pros and cons from enhancing the gatekeeping role of General Practitioners. Certainly, more research, both theoretical and empirical is needed to assess the relevance of the relationship between patients’ information, pressure for referral and GP’s incentives, that we have spotted in this work.
References


Appendixes:

Appendix A. GP’s and Patient’s updated probabilities.
Let us consider three events $s$, $s^d$ and $s^b$, such that $s$, $s^d$, $s^b \in \{\pi, \xi\}$.
Both $s^d$ and $s^b$ are correlated with $s$. However, we consider $s^d$ and $s^b$ to be independent events.

In general, $\forall i, j \in \{\pi, \xi\}$ 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:

$$\Pr\left(\overline{s} | s^b = \pi\right) = \Pr\left(s | s^b = \pi\right) = \beta \quad \text{and} \quad \Pr\left(\overline{s} | s^b = \xi\right) = 1 - \beta.$$

It is also true that $\forall i, j \in \{\pi, \xi\}$:

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

Moreover,

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

Therefore:

$$\Pr\left(s | s^d = \pi \cap s^b = \pi\right) = \frac{\delta \beta}{\delta \beta + (1 - \delta) (1 - \beta)} = 1 - \Pr\left(s | s^d = \pi \cap s^b = \pi\right).$$

$$\Pr\left(s | s^d = \pi \cap s^b = \xi\right) = \frac{\delta (1 - \beta)}{\delta (1 - \beta) + (1 - \delta) \beta} = 1 - \Pr\left(s \cap \pi \cap s^b = \xi\right).$$

$$\Pr\left(s | s^d = \pi \cap s^b = \xi\right) = \frac{(1 - \delta) \beta}{(1 - \delta) \beta + \delta (1 - \beta)} = 1 - \Pr\left(s | s^d = \pi \cap s^b = \xi\right).$$

$$\Pr\left(s | s^d = \xi \cap s^b = \pi\right) = \frac{(1 - \delta) (1 - \beta)}{(1 - \delta) (1 - \beta) + \delta \beta} = 1 - \Pr\left(s | s^d = \xi \cap s^b = \pi\right).$$

Appendix B. GP’s expected utility, health authority’s expected financial costs and patient’s expected disutility.

Under Gatekeeping:

GP’s expected utility:

$$U^{gk} = \Pr\left(s^d = \pi \cap s^b = \pi\right) \left(T + B\right) + \Pr\left(s^d = \pi \cap s^b = \xi\right) \left(T + (1 - \Pr\left(\text{rej}\right) B)\right) + \left(\Pr\left(s^d = \xi \cap s^b = \xi\right) + \Pr\left(s^d = \xi \cap s^b = \pi\right) R\right) \left[(\Pr\left(s^d = \pi \cap s^b = \pi\right) + \Pr\left(s^d = \pi \cap s^b = \xi\right) + \Pr\left(s^d = \pi \cap s^b = \xi\right) + \Pr\left(s^d = \pi \cap s^b = \xi\right) + \Pr\left(s^d = \xi \cap s^b = \pi\right) + \Pr\left(s^d = \xi \cap s^b = \xi\right) + \Pr\left(s^d = \xi \cap s^b = \xi\right)] - c_d = \frac{1}{2} [R + T + \delta B (1 - (1 - \beta) r)] - c_d.$$
Health authority’s expected primary care costs:

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

Health authority’s expected specialized care costs:

\[ C_{Sp}^{\text{gh}} = \left[ \Pr (\overline{s}) (1 - r \Pr (s^d = \overline{s} \cap s^b = \overline{s} \mid \overline{s})) + \Pr (s) \Pr (s^d = \overline{s} \mid \overline{s}) \right] c_s = \frac{c_s}{2} (2 - \delta - r (1 - \delta) \beta). \]

Finally, patient’s expected disutility:

\[ C_{Pat}^{\text{gh}} = \Pr (s) \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) r f + \Pr (\overline{s}) \left[ \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) (r f + (1 - r) K (\overline{l})) + \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) \frac{1}{3} \right] = \frac{1}{2} \left[ (1 - \delta) ((1 - \beta) r K (\overline{l}) + (1 - r) K (\overline{l})) + r f ((1 - \beta) \delta + (1 - \delta) \beta) \right], \]

with \( \overline{l} = r K (\overline{l}) + (1 - r) K (\overline{l}). \)

Under Non-Gatekeeping:

GP’s expected utility:

\[ U^{\text{Ngk}} = \Pr (s^b = \overline{s}) \left[ \Pr (s^d = \overline{s} \mid \overline{s}) (T + B) + \Pr (s^d = \overline{s} \mid \overline{s}) R \right] + \Pr (\overline{s}) \left[ \Pr (s^d = \overline{s} \mid \overline{s}) R + \Pr (s^d = \overline{s} \mid \overline{s}) T \right] - c_d = T + (R - T) [\delta + (1 - 2\delta) \beta] + B \delta \beta - c_d. \]

Health authority’s expected primary care costs:

\[ C_{GP}^{\text{Ngk}} = U^{\text{Ngk}} + c_d = T + (R - T) [\delta + (1 - 2\delta) \beta] + B \delta \beta. \]

Health authority’s expected specialized care costs:

\[ C_{Sp}^{\text{Ngk}} = \left[ \Pr (\overline{s}) + \Pr (s) \left( \Pr (s^d = \overline{s} \cap s^b = \overline{s} \mid \overline{s}) + \Pr (s^b = \overline{s} \mid \overline{s}) \right) \right] c_s = \frac{c_s}{2} (2 - \delta \beta). \]

Finally, patient’s expected disutility:

\[ C_{Pat}^{\text{Ngk}} = \Pr (s) \left[ \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) p_g + \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) (p_g + p_{gs}) + \Pr (s^b = \overline{s} \mid \overline{s}) p_s \right] + \Pr (\overline{s}) \left[ \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) (p_g + p_{gs}) + \Pr (s^b = \overline{s} \cap s^d = \overline{s} \mid \overline{s}) (p_g + p_{gs}) + \Pr (s^b = \overline{s} \mid \overline{s}) p_s \right] = \frac{1}{2} \left[ p_g + p_g + p_{gs} (\beta (1 - \delta) + 1 - \beta) + \tilde{l} (1 - \beta) (1 - \delta) \right], \]

with \( \tilde{l} = r K (\overline{l}) + (1 - r) K (\overline{l}). \)

Appendix C.

Proof of Proposition 1

In a non-gatekeeping system it is easy to check that, for any value of \( \beta \) and \( \delta \), there exist values of \( R, T \) and \( B \), such that \( IC_{PD}^{\text{Ngk}} \) and \( IC_{PD}^{\text{Ngk}} \) are simultaneously fulfilled.
In a gatekeeping system, 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 $r \leq \bar{r}$, with $\bar{r} = 1 - \frac{(1 - \delta)\beta}{\delta (1 - \beta)}$. It can be shown then that, for any $r \leq \bar{r}$, there exist values of $R, T$ and $B$, such that $IC_{PD}^{gk}$ and $IC_{FD}^{gk}$ are simultaneously fulfilled.

This completes the proof.

**Proof of Proposition 2**

The optimal level of copayments is the solution to the program given by (1). 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. We find two solutions depending on the value of the parameters:

- If $\frac{K(l)}{K(l)} \leq \frac{\beta}{1 - \beta}$, then $p_g^* = 0, p_{gs}^* = 0$ and $p_s^* = (1 - \beta) (1 - \delta) K(l)$.
- If $\frac{K(l)}{K(l)} \geq \frac{\beta}{1 - \beta}$, then $p_g^* = 0, p_{gs}^* = \frac{1 - \delta}{\delta(2\beta - 1)} ((1 - \beta) K(l) - \beta K(l)) > 0$ and $p_s^* = p_{gs}^* (1 - \delta\beta) + (1 - \beta) (1 - \delta) K(l) > 0$.

This completes the proof.

**Proof of Proposition 3**

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

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

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

- If $\beta \leq \beta^*$ then $C_{Pat}^{Ngk} \geq C_{Pat}^{gk}$.
- If $\beta > \beta^*$ then $C_{Pat}^{Ngk} < C_{Pat}^{gk}$.

This completes the proof.

**Proof of Proposition 4**

We compute the optimal payment contract separately for a non-gatekeeping system and for a gatekeeping one.

a) In a non-gatekeeping system:
The program the HA faces is as follows:

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

\[
s.t \quad \begin{cases}
U \geq 0 & \text{PC} \\
T \geq c_d & \text{LLC}_1 \\
R \geq c_d & \text{LLC}_2 \\
B \geq 0 & \text{LLC}_3 \\
R - T \geq \frac{B(1-\delta)\beta + c_d}{(1-\delta)\beta + (1-\beta)} & \text{IC}_{PD1}^{Ngk} \\
R - T \leq \frac{B(1-\delta)\beta + c_d}{(1-\delta)\beta + (1-\beta)} & \text{IC}_{PD2}^{Ngk} \\
\end{cases}
\]

First of all, it is straightforward to see that LLC_1, LLC_2 and LLC_3 imply the PC. Therefore, the HA 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_{PD1}^{Ngk} and IC_{PD2}^{Ngk} would be mutually incompatible. Moreover, LLC_2 and IC_{PD1}^{Ngk} binding cannot be a solution as IC_{PD2}^{Ngk} would not be fulfilled. A similar reasoning rules out LLC_2 and IC_{PD2}^{Ngk} binding as a potential solution.

The optimal solution of the problem, hence, has to be such that IC_{PD1}^{Ngk} and IC_{PD2}^{Ngk} are binding. From here we obtain that:

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

The health authority’s expected primary care costs are:

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

b) In a gatekeeping system:

The problem the HA faces is as follows:

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

\[
s.t \quad \begin{cases}
U \geq 0 & \text{PC} \\
T \geq c_d & \text{LLC}_1 \\
R \geq c_d & \text{LLC}_2 \\
B \geq 0 & \text{LLC}_3 \\
R - T \geq 2c_d + (1 - \delta) B (1 - (1 - \beta) r) & \text{IC}_{PD1}^{gk} \\
R - T \leq \delta B (1 - (1 - \beta) r) - 2c_d & \text{IC}_{PD2}^{gk} \\
R - T \geq \frac{B(1-\delta)\beta}{(1-\delta)\beta + (1-\beta)} & \text{IC}_{FD1}^{gk} \\
R - T \leq \frac{B(1-\delta)\beta}{(1-\delta)\beta + (1-\beta)} & \text{IC}_{FD2}^{gk} \\
\end{cases}
\]
First of all, it is straightforward to see that LLC₁, LLC₂ and LLC₃ imply PC. Moreover, by a similar reasoning as in the non-gatekeeping case, \( T^g_k = c_d \) at the optimum.

Let us define \( \tilde{r} \equiv \frac{\delta - \beta}{(1-\beta)(\delta - \beta + 2\delta \beta)} \in (0, \bar{r}) \). We solve the program by distinguishing two cases:

- If \( r \leq \tilde{r} \), then \( IC^g_{P_{D_1}} \) and \( IC^g_{P_{D_2}} \) imply both \( IC^g_{F_{D_1}} \) and \( IC^g_{F_{D_2}} \).

A completely analogous reasoning to the one followed for a non-gatekeeping system shows that the solution of the problem is such that both \( IC^g_{P_{D_1}} \) and \( IC^g_{P_{D_2}} \) are binding. The optimal values, hence, are given by:

\[
R^{g_k} = \frac{(2\delta + 1)c_d}{2\delta - 1}, \quad B^{g_k} = \frac{4c_d}{(2\delta - 1)(1 - (1 - \beta)r)}.
\]

- If \( r \in (\tilde{r}, \bar{r}] \), then it is not true that \( IC^g_{F_{D_1}} \) and \( IC^g_{F_{D_2}} \) are implied by \( IC^g_{P_{D_1}} \) and \( IC^g_{P_{D_2}} \).

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:

- \( IC^g_{P_{D_1}} \) and \( IC^g_{P_{D_2}} \) binding violates \( IC^g_{F_{D_2}} \).
- \( IC^g_{F_{D_1}} \) and \( IC^g_{F_{D_2}} \) binding violates both \( IC^g_{P_{D_1}} \) and \( IC^g_{P_{D_2}} \).
- \( IC^g_{P_{D_1}} \) and \( IC^g_{P_{D_2}} \) binding violates \( IC^g_{F_{D_1}} \).
- \( IC^g_{F_{D_1}} \) and \( IC^g_{F_{D_2}} \) binding violates \( IC^g_{P_{D_2}} \).
- Finally, \( IC^g_{P_{D_1}} \) and \( IC^g_{F_{D_1}} \) binding, as well as \( IC^g_{P_{D_1}} \) and \( IC^g_{F_{D_2}} \) 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 \( IC^g_{P_{D_1}} \) and \( IC^g_{F_{D_2}} \) binding are smaller and, hence, this constitutes the optimal contract. Some algebraic manipulations yield:

\[
R^{g_k} = 3c_d + \frac{4c_d(1 - \delta)}{2\delta - 1} \left[ \frac{(2\delta - 1)(1 - (1 - \beta)r)(\delta - (1 - \delta)(1 - (1 - \beta)r) - \delta \beta)}{2(1 - (1 - \beta)r)(\delta - (1 - \delta)(\delta - (1 - \beta)r) - \delta \beta)} \right]
\]

\[
B^{g_k} = \frac{4c_d}{(2\delta - 1)(1 - (1 - \beta)r)} \left[ \frac{(2\delta - 1)(1 - (1 - \beta)r)(\delta - (1 - \delta)(1 - (1 - \beta)r) - \delta \beta)}{2(1 - (1 - \beta)r)(\delta - (1 - \delta)(\delta - (1 - \beta)r) - \delta \beta)} \right].
\]

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

\[
R^{g_k} = \frac{c_d(3(2\delta - 1) + 4(1 - \delta)\Gamma(\delta, \beta, r))}{2\delta - 1}, \quad T^{g_k} = \frac{c_d\Gamma(\delta, \beta, r)}{(2\delta - 1)(1 - (1 - \beta)r)}, \quad B^{g_k} = \frac{4c_d\Gamma(\delta, \beta, r)}{(2\delta - 1)(1 - (1 - \beta)r)}.
\]

with \( \Gamma(\delta, \beta, r) = \begin{cases} 1 & \text{if } r \leq \tilde{r}, \\ \frac{(2\delta - 1)(1 - (1 - \beta)r)(\delta - (1 - \delta)(\delta - (1 - \beta)r) - \delta \beta)}{2(1 - (1 - \beta)r)(\delta - (1 - \delta)(\delta - (1 - \beta)r) - \delta \beta)} & \text{otherwise}. \end{cases} \)
The health authority’s expected primary care costs are:

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

This completes the proof.

**Proof of Proposition 5**

By comparing \( C_{GP}^{Ngk} \) and \( C_{GP}^{gk} \), as defined in Proposition 4, we find that:

- If \( r \leq \tilde{r} \), then it is straightforward that \( C_{GP}^{gk} < C_{GP}^{Ngk} \).
- If \( r > \tilde{r} \), then \( C_{GP}^{gk} > C_{GP}^{Ngk} \iff 2 [\Gamma (\delta, \beta, r) - 1] > \frac{\beta}{1-\beta} - 1. \)

The inequality above holds if and only if:

\[ r > \frac{\delta (1 - \beta) - (\delta (1 - \beta) + (1 - \delta) \beta)^2}{(1 - \beta) \left[ \delta - (\delta (1 - \beta) + (1 - \delta) \beta)^2 \right]} \equiv r^*. \]

This completes the proof.