Competition between General Practitioners and Specialists in the Primary Health Care Market ¤

Carine Brasseur
IRES Université catholique de Louvain ¥

January 2000

Abstract

In this paper, we study the optimal payment system for the primary health care market when general practitioners are not only in competition between themselves but also with specialists. We de…ne the copayment to impose in order to ensure a good allocation of patients among the two types of physicians. Further, we set the physician reimbursement system that guarantees an appropriate referral of patients to specialists. We prove that the GP's remuneration system is more prospective the larger the competition with specialists. Next, we show that the assumption of risk-averse patients precludes the optimal payment system from being a …rst-best solution. To conclude, we contrast the results of the analysis with systems of gatekeepers where all patients are required to go to a general practitioner before having access to specialized medicine.

Keywords : health care - competition - gatekeeping - payment system
JEL Classi…cation : I10, D60

¤The author ows many thanks to Professor M. Marchand for his suggestions and constant support. I am also especially grateful to Professors Marie-Christine Closon, Bart Cockx, Ines Macho-Stadler and Erik Schokkaert for their helpful comments.

¥Aspirant F.N.R.S., Institut de Recherches Economiques et Sociales, Université catholique de Louvain, Place Montesquieu 3, B-1348 Louvain-La-Neuve, Belgium. E-mail : brasseur@ires.ucl.ac.be

1 Introduction

The objective of this paper is to de…ne the optimal payment system for the primary health care market in a framework where general practitioners (GP) are not only in competition between themselves but also with specialists. Such a framework can be observed in the absence of a system of 'gatekeepers' where patients are forced to go …rst to a general practitioner before having access to specialized medicine.

It is usually recognized that specialized medicine is more expensive than general medicine. Yet, very often, we may observe that some patients systematically choose to go to a specialist, independently of the degree of severity of their illness. In this respect, it is sometimes suggested that the assignment of a gatekeeper function to general practitioners would contribute to center specialized health care services on more acute health problems, and so reduce the cost of health care expenditures. An example of application can be found in UK with the 'GP fundholders' where GPs receive their own budget for purchasing some hospital and other care services. But arguments against systems of gatekeeping are still discussed nowadays. A …rst objection concerns the e¢cacity of these systems in reaching their objective of reducing health care expenditures. It is argued that in the case of acute illnesses, a system of gatekeeping would have as sole consequence to multiply the number of medical visits since every patient would have to go …rst to a general practitioner before receiving the appropriate services from a specialist. Second, the GP's ability to refer his patients to a specialist when required by their health status becomes crucial in such a system if it is to guarantee a certain quality level for the health care services provided.

In the literature, two recent papers deal with questions related to this problem. Jelovac (1998), who is interested in the role of the general practitioner as gatekeeper for patients, compares the incentives o¤ered through the GP reimbursement system when referral is either compulsory or facultative. She studies under which conditions it is optimal to give GPs incentives to refer their patients to a specialist in frameworks of perfect and imperfect information. This model relies on the implicit assumption that patients are charged a higher copayment when visiting a specialist on their own. But it does not try to de…ne the optimal copayment to select in view of the patient behaviour. The latter is taken as given in the model. On the contrary, the paper of Bouck-

aert (1998) develops a horizontal di¤erentiation model where patients have to choose to get a treatment from an expert (the specialist) or a non expert (the general practitioner). The author studies the resulting price equilibrium for the two categories of medical treatments. He shows how the equilibrium solution depends on the degree of di¤erentiation of the two categories of services. In a welfare analysis, he …nds that too many consumers directly visit the expert if the latter su¤ers from a cost-disadvantage. But the study abstracts from physicians' incentives to provide the appropriate quality of services.

Our analysis di¤ers from these two studies in that it tries to give appropriate incentives to both the general practitioner and the patient. More precisely, it looks at the copayment to impose on patients for inducing them to visit a general practitioner (specialist) when they think they su¤er from a common (special) illness. Similarly, it studies the reimbursement system that favours a good referral behaviour for general practitioners. But contrary to Jelovac (1998), the search for appropriate incentives for GPs is not justi…ed here on the grounds of any opportunist behaviour. General practitioners are supposed to provide the most appropriate treatment given the illness severity they diagnose. Instead of …ghting opportunist behaviours, the incentives given to GPs aim to safeguard a certain quality of general medicine. This separates our study from the literature on experts (besides Jelovac, 1998, see for instance Wolinsky, 1993) where experts take advantage of having a better information to behave in an opportunist way.

The paper proceeds as follows. Section 2 describes the features of the model and provides the regulator's optimum in the case of full information, i.e. in the case where everything on the market is under the regulator's control. Section 3 develops the second-best solution where the regulator is constrained by decisions taken by the patients and general practitioners. It characterizes the optimal payment system that results from this situation. After a comparison of the two solutions, section 4 concludes.

2 The model

A central characteristic of the model comes from the two sorts of competition which general practitioners have to face. First, it is assumed that at the beginning of the period, before any illness appears, patients have to register at one GP's office. This induces general practitioners to compete among themselves for attracting a certain number of registered patients. Second, once registered, patients are supposed to be free to visit their GP[1] or a specialist. General practitioners are aware of this possibility and so take specialists as competitors.

Let two types of illness be distinguished according to their degree of severity : high severity illnesses (occuring with a probability p_h) and low severity ones (probability p_l). General practitioners and specialists are supposed to differ in their ability to treat high severity illnesses. We assume that these can only be cured by specialists. Hence, when a general practitioner diagnoses a high severity illness, he is supposed to refer his patients to a specialist. Let denote by e the effort provided by the GP in order to make a good diagnosis. We call D(e) his probability of diagnosing the correct severity of illness. By assumption, we set $D'(e) > 0$. When patients suffer from a major illness and their GP diagnoses a minor one, patients are supposed to be referred to a specialist with a certain delay that implies a disutility for the patient under the form of a health loss μ.

Patients have identical preferences but differ in the value of μ. This element of differentiation can be given two different interpretations. First, the health loss μ could be viewed as $\mu = kL$, with L measuring the extent of the health loss and k the weight the individual attaches to that loss. While L could be considered identical for all patients, k could differ across patients depending on their social status or the size of their family, for instance. Hence, following this interpretation, μ should be viewed as a parameter of preferences for patients that is defined ex-ante. On the contrary, a second interpretation would suggest that the health loss μ is revealed ex-post to patients, after illness has occured. In this case, all patients should be viewed as identical, with a similar distribution function $F(\mu)$, while the value of the health loss μ should now be consid-

[1] Physicians of the same type are supposed identical. Hence, patients who decide to be treated by a GP rather than by a specialist have no interest in asking a treatment from another GP than the one with whom they are registered. They are therefore supposed to visit the GP with whom they are registered.

ered random. The interpretation chosen for our analysis is the …rst one, but the approach would have been similar had the second interpretation been preferred.

To formalize this, we develop as in Bouckaert (1998) a simple horizontal di¤erentiation model where patients are distributed on interval $[\underline{\mu}; \bar{\mu}]$. The distribution function $F(\mu)$ (with $0 \cdot F(\mu) \cdot 1$) determines the allocation of patients between the general practitioners and the specialists present on the market. If we denote by $\hat{\mu}$ the level of μ that characterizes the patient who is indi¤erent between visiting a specialist or his GP, $F(\hat{\mu})$ de…nes the proportion of patients asking for a treatment from their general practitioner rather than from a specialist.

Neither $\hat{\mu}$ nor e are observable or veri…able by the regulator. Hence, the objective of our study is to evaluate how the regulator can intervene in the market in order to induce an appropriate distribution of patients $F(\hat{\mu})$ and GP's diagnosis e¤ort e. For this purpose, we develop a three-stage model. In the …rst stage, the regulator de…nes the optimal GP's reimbursement system and patient's insurance plan[2] anticipating their impact on the decision process of general practitioners and patients. In the second stage, general practitioners decide on their diagnosis e¤ort, given the payment system in force on the market and taking into account the competition they play with the other GPs and specialists. In the third stage, patients decide on the type of physician from whom to get a treatment anticipating the GP's diagnosis e¤ort and knowing the payment system.

To develop this study, we proceed backward and start analyzing the patient's decision process. But before this, it is important to note the role that is assigned to specialists in the model. These physicians only act through the competition that is supposed between GPs and specialists. It implies that we do not try to regulate the specialist's behaviour, but rather take it as given. And this is motivated by our wish to focus on the patient's and GP's decision process that are both central in the discussions on the potential interest of a system of gatekeeping.

[2] For the rest of the paper, we use the term 'payment system' to refer to the combination of GP's reimbursement system and patient's insurance plan.

2.1 The patient's problem

As announced, patients are here supposed to make two decisions. First, they have to choose, at the beginning of the period, the GP with whom to register. Second, once they become ill [3], they have to decide on whether to be treated by their GP or a specialist. The …rst decision will be formalized when analysing the GP's problem. In this section, we focus on the second decision.

To make this second decision, patients compare the utility level they get when selecting one type of provider instead of the other. We assume that the patient's utility function is additively separable in two terms : health status and income available for the consumption of other goods than health care. With respect to the health status, we suppose that once they ask for treatment, patients recover totally unless their general practitioner does not refer them immediately to a specialist although they su¤er from a high severity illness. We measure the patient's health status in monetary terms and normalize to zero the health status observed after full recovery. Hence, the only term related to health status that enters the patient's utility function is the health loss μ that occurs with a probability $p_h(1 - D(e))$ for patients that select a GP's treatment.

With regard to available income, we assume that a patient who chooses to be treated by a specialist rather than by his general practitioner is charged a copayment $\pm$. The copayment for a GP's treatment is supposed to be zero, whether the patient is referred or not to a specialist. In this way, the term $\pm$ can be interpreted as a penalty that is imposed on each patient that prefers to be treated by a specialist. Also central to the model is the assumption that all patients are risk averse. This implies that they are all willing to subscribe to an insurance. Suppose a public insurance system. We denote by $¼$ the social contribution imposed to all patients.

Keeping this in mind, we can now express the patients' utility function as following. Let denote by U_G and U_S the utility level patients may expect when being treated by their GP or a specialist, respectively. They can be expressed as

[3] For the sake of simplicity, we assume that all patients become ill during the period. The only uncertainty concerns the degree of severity of the illness. But this assumption does not alter our results.

$$
\begin{aligned}
U_G &= u(w - \pi) - p_h(1 - D(e))\,\mu \qquad (1)\\
U_S &= u(w - \pi - \pm)
\end{aligned}
$$

where w indicates the patient's gross income and function u(:) satisfies the property of $u'(:) > 0$ and $u''(:) < 0$ to reflect risk aversion. Note that the expression of U_G supposes that patients are able to observe the GP's probability of diagnosis error $(1 - D(e))$. This ability could be explained by some phenomenon of reputation. Thanks to reputation, patients are able to evaluate the effort level of each general practitioner. We assume here that this reputation phenomenon is sufficiently strong for general practitioners to be committed to supply the effort level that is expected given their reputation.

This allows us to define the patient who is indifferent between visiting his GP or a specialist through the threshold value $\hat{\mu}$ that equals U_G and U_S and depends upon π, $\pm$ and e :

$$
\hat{\mu} = \frac{u(w - \pi) - u(w - \pi - \pm)}{p_h(1 - D(e))} \qquad (2)
$$

Clearly, $\hat{\mu}$ gives the value of μ beyond which all patients prefer to go to a specialist rather than to their GP. From expression (2), it is straightforward to verify that $\hat{\mu} = \hat{\mu}(\pi; \pm; e)$ with

$$
\begin{aligned}
\frac{\partial \hat{\mu}}{\partial \pi} &= \frac{u'_S - u'_G}{p_h(1 - D(e))} > 0\\
\frac{\partial \hat{\mu}}{\partial \pm} &= \frac{u'_S}{p_h(1 - D(e))} > 0 \qquad (3)\\
\frac{\partial \hat{\mu}}{\partial e} &= \frac{\hat{\mu} D'(e)}{1 - D(e)} > 0
\end{aligned}
$$

where subscripts G and S refer to patients selecting a GP or a specialist, respectively, and the probability p_h is exogenous.

By definition, risk aversion implies $u'_S > u'_G$ if $\pm > 0$. Hence, expressions (3) indicate that the threshold value $\hat{\mu}$ increases with the insurance premium π as with the patient's copayment $\pm$ and the GP's effort e. The first positive relation results so directly from the assumption of risk aversion. An increase in the insurance premium π reduces the utility of available income u(:). Because of risk aversion, this utility reduction is higher, in marginal terms, for a patient who is treated by a specialist rather than by a general practitioner. As a consequence, more

patients are induced to prefer their GP to a specialist. This corresponds to a higher threshold value $\hat{\mu}$. The interpretation of the second positive result is straightforward. The higher the …nancial penalty imposed on patients selecting a treatment from a specialist, the higher the number of individuals choosing to be treated by their general practitioner rather than by a specialist, and so the larger the value of $\hat{\mu}$. As to the last result, this can be given the following intuitive interpretation. The higher the GP's e¤ort, the better the quality of his diagnosis and, so, the lower the risk of a health loss for a patient selecting a GP's treatment. Again, this induces additional patients to choose to be treated by their GP rather than by a specialist. And this is illustrated through a higher threshold value $\hat{\mu}$. Note again that this relation supposes that patients are able to observe the GP's probability of diagnosis error. As underlined above, this may be interpreted in view of some reputation.

2.2 The physician's problem

The general practitioner's problem consists in choosing the diagnosis effort e that maximizes his utility function. In the previous section, we have seen how the GP's e¤ort level in‡uences the patient's decision on selecting a treatment from his GP or a specialist. We now suppose that the GP decides on the value of e while anticipating this in‡uence. In other words, the GP makes his decision taking account of the competition he plays with specialists.

But general practitioners are also supposed to be in competition between themselves for the registration of patients. Assume that all GPs are identical. In a context of perfect information, this implies that patients are equally distributed among general practitioners. Let N be the number of patients registered at each GP's o¢ce. This number is considered as …xed by the general practitioner. The next section explicitly shows how N is determined by competition.

Subject to these two sources of competition, the general practitioner selects the e¤ort level that maximizes his utility function. The GP's utility function is supposed to be made of two components : the utility from his revenue and the disutility associated with his e¤ort. The size of both terms depends on the number of services provided. We normalize to one the number of health care services provided for each patient treated by the general practitioner. Given that competition with specialists implies that only a proportion $F(\hat{\mu})$ of the N registered pa-

tients is actually treated by their GP, the total number of treatments supplied by a general practitioner sums up to $NF(\hat{\mu})$. In return for his services, the general practitioner receives a mixed reimbursement. We assume here that his revenue combines a per unit payment β for each health care service provided with a salary γ that remunerates him for his participation on the market. Hence, the GP's utility function can be de…ned as

$$V = \gamma + NF(\hat{\mu})\beta - v(NF(\hat{\mu})e) \tag{4}$$

where the disutility function is increasing and convex ($v'(\cdot) > 0; v''(\cdot) > 0$). The general practitioner's problem reduces so to

$$\max_e V = \gamma + NF(\hat{\mu}(\alpha; \theta; e))\beta - v(NF(\hat{\mu}(\alpha; \theta; e))e) \tag{5}$$

with N and $\hat{\mu}(\alpha; \theta; e)$ resulting from the competition with the other GPs and specialists, respectively.

Note that this problem illustrates a characteristic we underlined in the introduction. No opportunist behaviour is considered here for the general practitioner with respect to the treatment provided. His only decision consists in choosing a diagnosis e¤ort. The treatment o¤ered by the GP follows then directly from his diagnosis. When the general practitioner diagnoses a high severity illness, he is supposed to refer his patient to a specialist whereas he provides him with a given treatment when he suspects a low severity illness. By determining so the degree of quality of general medicine, the GP's diagnosis e¤ort in‡uences the allocation of patients among general practitioners and specialists. Expression (5) explicitly shows that in selecting his e¤ort level, the GP trades o¤ the increased disutility and the larger workload, and so revenue, that both result from a higher diagnosis e¤ort. Note that no altruistic concern a¤ects the GP's decision.

>From the GP's optimisation problem, we obtain the following …rst-order condition

$$V_e = -NF(\hat{\mu})v'(\cdot) + f(\hat{\mu})\frac{\partial \hat{\mu}}{\partial e}N(\beta - ev'(\cdot)) = 0 \tag{6}$$

and if we di¤erentiate V_e, we may express the GP's e¤ort as

$$e = e(\beta; \alpha; \theta; N) \tag{7}$$

Note that the salary γ does not in‡uence the GP's choice of e¤ort. The exact role of this …nancial instrument will become clearer in the next section.

Under the assumption that the second-order condition of the GP's optimisation problem is satis…ed, it is easy to check that

$$\begin{aligned} \frac{@e}{@®} &> 0 \\ \frac{@e}{@N} &< 0 \end{aligned} \tag{8}$$

This means that the GP's remuneration ® gives him …nancial incentives to increase his diagnosis e¤ort. On the contrary, the e¤ect of competition between general practitioners acts in the opposite direction. This last result looks intuitive. The higher the number of patients registered at the GP's o¢ce, the lower the incentives for the general practitioner to produce a high diagnosis e¤ort in order to treat a large proportion of registered patients. The sign of the partial derivative of the GP's e¤ort e with respect to the insurance premium ¼ and the copayment ± is less obvious since it depends on the form of the distribution function $F(\mu)$. The only general conclusion that can be drawn from these partial derivatives is that

$$\frac{@e}{@±} = \frac{@\hat{\mu}=@±}{@\hat{\mu}=@¼}\frac{@e}{@¼} \tag{9}$$

Given that from (3), $\frac{@\hat{\mu}=@±}{@\hat{\mu}=@¼}$ is positive, it implies that the partial derivatives of e with respect to ± and ¼ have the same sign. Expression (9) will be used later on in the paper. Note that intuitively, this expression can be explained by relation (6) that shows how the insurance premium ¼ and copayment ± a¤ect the GP's diagnosis e¤ort e through their impact on the patient's threshold value of μ.

In section 2.1., we have seen that the patient's choice of $\hat{\mu}$ is contingent upon the insurance premium ¼ and the copayment ± in force on the market as well as on the GP's diagnosis e¤ort e. We have just seen how that this e¤ort depends itself on the payment system $(¾; ®; ¼; ±)$ and on the number of registered patients N that results from the competition played between general practitioners. In the next section, we analyze how the regulator decides on the payment system while anticipating its in‡uence on the patient and GP decision process.

2.3 The regulator's problem

As suggested in the introduction, the concern of the regulator is here twofold. He wishes …rst to control the number of patients having a direct access to specialized medicine. Second, he wants to guarantee a good referral behaviour of general practitioners through their diagnosis e¤ort. In formal terms, the regulator's objective is to in‡uence the patients' threshold value $\hat{\mu}$ and the GP's diagnosis e¤ort level e in order to maximize the patients' utility function. In this section, we will look at the problem in the …rst-best situation where the regulator can directly decide on the threshold value $\hat{\mu}$ and on the GP's e¤ort level e. The development of the second-best solution, where the regulator can only indirectly act on the patient's and GP's decision process through the choice of a particular payment system, will then follow in the next section. But before developing these solutions, we start to present the two constraints that the regulator has to face in his decision process.

2.3.1 The physician participation constraint

In solving his problem, the regulator must make sure, …rst, that his decision induces GPs to participate. In other words, the GP's utility level that results from the regulator's decision may not be lower than what general practitioners would get if choosing another occupation. Denoting by $\bar{V}$ this participation level, and recalling the GP's utility function given in (4), the GP participation constraint can be expressed as

$$V \acute{} \frac{3}{4} + NF(\hat{\mu})® ¡ v(NF(\hat{\mu})e) \, ¸ \, \bar{V} \qquad (10)$$

This constraint can be interpreted in view of the competition that is assumed to prevail between general practitioners. Suppose a …xed population of patients of size P on the primary health care market. This population has to be divided among the GPs such that for each general practitioner, the N registered patients are su¢cient for their participation constraint to be satis…ed. This condition determines the number M of general practitioners operating on the primary health care market. To see this, note that the GP's utility function V increases with N as, from (6), $® ¡ ev'(:) > 0$. This utility increase induces more physicians to be willing to participate on the market and so M increases also. Equilibrium on the primary health care market is reached when $N = P=M$. Hence, both N and M are determined by the physician

participation constraint (10). It can be shown that this constraint is satis…ed with equality.

As to the specialists, given that we do not try to regulate them, these are supposed to be willing to participate whatever the number of patients that select their practice.

2.3.2 The budget constraint

The assumption we make with regard to the regulator's budget constraint is rather usual. It consists in de…ning a fair insurance premium ¼ that fully compensates for the total cost of general and specialized medicine that is supported by the regulator.

To express this constraint formally, we de…ne …rst the cost of medical treatments. Note that the medical treatments that are here considered can go beyond the services o¤ered within the physician o¢ce. Nursing, technical sta¤, equipments, for instance, are all elements that can be included in the treatment cost. We denote by c_l and c_h the unit cost of a treatment ordered by general practitioners and specialists, respectively. Di¤erences in cost between the two categories of health care services could be imputed to di¤erences in the technology used, for example. We suppose here that GPs (specialists) supply health care services that are based on a low (high) technology. This re‡ects the idea that contrary to specialists, general practitioners are supposed to be able to cure only minor illnesses. To justify the interest of a regulation on the allocation of patients, we assume a cost-disadvantage of specialized medicine with $c_h > c_l$.

We know that the medical treatments provided by a general practitioner can include referrals to a specialist. Hence, the total cost of a GP's medical treatment depends on the severity of the patient's illness and on the GP's diagnosis. Recalling that p_h (p_l) indicates the probability of su¤ering from a major (minor) illness and that $D(e)$ measures the probability for the general practitioner to diagnose correctly the illness severity, the total cost of a medical treatment provided by a GP can be expressed as

$$\begin{aligned} C_G(e) &= p_h[D(e)c_h + (1 - D(e))(c_h + c_l)] \\ &+ p_l[D(e)c_l + (1 - D(e))c_h] \end{aligned} \tag{11}$$

Clearly, when a major illness is diagnosed, the patient is treated by a specialist, whatever the true intensity of his illness. Hence, the cost of

the treatment o¤ered to the patient amounts to c_h. When the general practitioner correctly diagnoses a minor illness, the cost reduces to c_l. But if the patient su¤ers from a major illness and this is diagnosed as minor, the patient has to face two treatments. After having provided his services, the GP is supposed to recognise his error and refer …nally his patients to a specialist. Hence, in this case, the total cost of treatment amounts to $c_l + c_h$. For patients selecting directly a specalist, the total technical cost of a treatment simply sums up to

$$C_S = c_h \tag{12}$$

The costs $C_G(e)$ and C_S do not yet totally de…ne the cost supported by the regulator. Besides these technical costs, the regulator still has to support the payment provided to the physicians in return for their services. For the general practitioner, we have seen that the reimbursement amounts to $¾ + NF(\hat{\mu})®$. It is then straightforward to deduce the reimbursement due to each individual patient. Concerning the specialist, since we are not interested in regulating him, we suppose his payment to be included in the unit amount c_h. Further, the regulator can deduct from all these costs the patient's …nancial participation. We know that a patient who selects a specialist is imposed a copayment ±. In all other cases, the copayment is equal to zero.

To sum up, the budget constraint implies that the insurance premium paid by every patient has to satisfy

$$¼ = F(\hat{\mu})[C_G(e) + ® + \frac{¾}{F(\hat{\mu})N}] + (1 ¡ F(\hat{\mu}))[C_S ¡ ±] \tag{13}$$

with $C_G(e)$ and C_S de…ned in (11) and (12), respectively. Having de…ned the two constraints, we can now express the regulator's optimisation problem in the …rst-best situation.

2.3.3 The …rst-best solution

In the …rst best, the regulator chooses the payment system $(¾; ¼)$, the allocation of patients $(N; \hat{\mu})$ and the GP's diagnosis e¤ort e that, under constraints (10) and (13), maximize the patients' utility functions. Note that in the …rst best, everything is under the regulator's control. Therefore, the …nancial instruments ± and ® that aim at giving appropriate incentives to patients and GPs are not introduced here.

Recalling the utility functions of patients selecting a GP or a specialist that were de…ned in (1), it is now straightforward to express the regulator's problem as

$$\begin{aligned}\max_{\gamma,\alpha,\mu^{\ddagger},N,e} \Psi = {} & u(w-\alpha) - p_h(1-D(e))\int_{\underline{\mu}}^{\mu^{\ddagger}} \mu dF(\mu) \\ & + \lambda[\alpha - F(\mu^{\ddagger})(C_G(e) + \gamma/(F(\mu^{\ddagger})N)) - (1-F(\mu^{\ddagger}))C_S] \\ & + \theta[\gamma - v(NF(\mu^{\ddagger})e) - \bar{V}]\end{aligned} \tag{14}$$

where λ and θ are the Lagrange multiplier associated with the budget and physician participation constraint, respectively, and the threshold value $\hat{\mu}$ is here denoted by $\mu^{\ddagger}$ to stress that it is chosen by the regulator rather than by patients. The …rst-order conditions of this optimisation problem are given by

$$\begin{aligned}
-\gamma: \quad & -\lambda/N + \theta = 0 \\
-\alpha: \quad & -u' + \lambda = 0 \\
-\mu^{\ddagger}: \quad & f(\mu^{\ddagger})[-p_h(1-D(e))\mu^{\ddagger} - u'(C_G(e) - C_S + ev'(\cdot))] = 0 \\
-N: \quad & \frac{u'}{N}[\frac{\gamma}{N} - F(\mu^{\ddagger})ev'(\cdot)] = 0 \\
-e: \quad & p_h D_e' \int_{\underline{\mu}}^{\mu^{\ddagger}} \mu dF(\mu) - u'F(\mu^{\ddagger})[\frac{dC_G}{de} + v'(\cdot)] = 0 \\
-\lambda: \quad & \alpha - F(\mu^{\ddagger})[C_G(e) + \gamma/(F(\mu^{\ddagger})N)] - (1-F(\mu^{\ddagger}))C_S = 0 \\
-\theta: \quad & \gamma - v(NF(\mu^{\ddagger})e) - \bar{V} = 0
\end{aligned}$$

These expressions are rather standard. For each parameter, the optimum is such that the marginal bene…t that would result from a change in the value of the parameter is equivalent to the marginal cost that this change would also imply. So, for instance, looking back at the …rst-order condition related to N, the …rst term at the right-hand side can be interpreted as the gain that results from an increase in N. This gain consists in a decrease in the per-registered patient cost of a medical treatment provided by a GP. Similarly, the second term at the right-hand side measures the loss, in terms of an increased disutility for the general practitioner, that results from an increase in N. Clearly, the …rst-order condition on the optimal number of registered patients trades o¤ the gain $\lambda\gamma/N^2$ against the loss $\theta F(\mu^{\ddagger})ev'(\cdot)$. The …rst-order condition related to the physician's diagnosis e¤ort e can be given a similar interpretation. The gain induced by an increased e¤ort e is here expressed through the

reduced health loss ($p_h D'_e \int_{\underline{\mu}}^{\mu^*} \mu dF(\mu)$) while the cost now combines an increased social cost ($\lambda F(\mu^*)\frac{dC_G}{de}$) with a larger disutility for the GP ($\circ N F(\mu^*)v'(:)$). Also particularly interesting is the …rst-order condition related to μ^*. This allows to state the next proposition.

Proposition 1 In the …rst best, the optimal allocation of patients among the general practitioners and the specialists present on the primary health care market is characterized by

$$\mu^* = \frac{u'}{p_h(1 - D(e))}[C_S - C_G(e) - \frac{\gamma}{F(\mu^*)N}]$$

Proof : The proof follows directly from the …rst-order condition related to μ^* where, from $-\alpha = 0$, λ is substituted by u' and where $ev'(:) = \frac{\gamma}{F(\mu^*)N}$ from $-N = -\gamma = 0$. ■

This proposition can be intuitively interpreted. The threshold value μ^*, and so the optimal proportion of patients selecting a GP, increases with the patient's marginal utility $u'(:)$ and with the cost-disadvantage of specialized medicine ($C_S - C_G(e) - \frac{\gamma}{F(\mu^*)N}$)[4]. The …rst result is linked to risk aversion whereas the latter justi…es the interest for a regulation on the allocation of patients when specialized services are found more expensive than general medicine. But both relations are tempered by the probability $p_h(1 - D(e))$ for a health loss to occur when the patients treated by their GP are not immediately referred to a specialist although their illness requires it. Given that the GP's diagnosis e¤ort e determines his referral behaviour, and so the quality of general medicine, this result con…rms the expectation that a regulation favouring …rst visits to GPs can be suggested only if we have some guarantee on the quality of general medicine.

Proposition 1 together with the other …rst-order conditions related to the …rst-best solution are taken as the benchmark to which the second-best solution can be compared. This solution is developed in the next section.

[4] Recall that the specialist's remuneration is supposed to be included in the treatment cost C_S.

3 The optimal payment system

In a second-best situation, the regulator cannot observe and verify the patient's health loss μ and the GP's diagnosis e¤ort e. The equilibrium values of $\hat{\mu}$ and e that are observed in the primary health care market result from decisions taken by the patient and the general practitioner. But from (3) and (7), we know that these decisions depend on the payment system in force on the market. In our model, the payment system combines the mixed GP's reimbursement system de…ned through $(\gamma; \theta)$ with the patient's insurance premium π and copayment $\pm$. Hence, we investigate now how the regulator de…nes both systems while anticipating their in‡uence on the patient's and GP's decision process.

The regulator's objective function and the constraints are equivalent to what was stated in the …rst best. Only the instruments di¤er. It is then straightforward to de…ne the second-best problem of the regulator as

$$\begin{aligned} \max_{\gamma;\theta;\pi;\pm;N} \; & - = F(\hat{\mu})u(w - \pi) - p_h(1 - D(e)) \int_{\underline{\mu}}^{\hat{\mu}} \mu dF(\mu) \\ & + (1 - F(\hat{\mu}))u(w - \pi - \pm) \\ & + \lambda[\pi - F(\hat{\mu})(C_G(e) + \theta + \gamma/(F(\hat{\mu})N)) - (1 - F(\hat{\mu}))(C_S - \pm)] \\ & + \rho[\gamma + NF(\hat{\mu})\theta - v(NF(\hat{\mu})e) - \bar{V}] \end{aligned} \tag{15}$$

$$\begin{aligned} \text{with } \hat{\mu} &= \hat{\mu}(\pi; \pm; e) \text{ and} \\ e &= e(\theta; \pi; \pm; N) \end{aligned}$$

where the two last expressions result from the patient's and GP's problems, respectively. Note that as in the …rst best, the regulator decides here on the number of registered patients at each GP's o¢ce. From the GP's participation constraint developed in section 2.3, we know the the value of N determines the number of physicians M operating in the primary health care market. Assuming that the regulator can choose the value of N amounts to supposing that he can decide on the number of general practitioners in practice.

The objective function - is the same as in the …rst best. We denote by W the regulator's objective function - that includes the relations de…ning $\hat{\mu}$ and e. Taking into account this notation, we can derive from (15) the following …rst-order conditions for the optimum :

$$W_{¼} \;\acute{}\; -_{¼} = ¡ ¸=N + ° = 0$$

$$W_{®} \;\acute{}\; F(\hat{¹})N -_{¼} + (-_{e} + -_{\hat{¹}}\frac{@\hat{¹}}{@e})\frac{@e}{@®} = 0$$

$$W_{¼} \;\acute{}\; -_{¼} + (-_{e} + -_{\hat{¹}}\frac{@\hat{¹}}{@e})\frac{@e}{@¼} + -_{\hat{¹}}\frac{@\hat{¹}}{@¼} = 0$$

$$W_{±} \;\acute{}\; -_{±} + (-_{e} + -_{\hat{¹}}\frac{@\hat{¹}}{@e})\frac{@e}{@±} + -_{\hat{¹}}\frac{@\hat{¹}}{@±} = 0$$

$$W_{N} \;\acute{}\; -_{N} + (-_{e} + -_{\hat{¹}}\frac{@\hat{¹}}{@e})\frac{@e}{@N} = 0$$

$$W_{¸} \;\acute{}\; -_{¸} = ¼ ¡ F(\hat{¹})(C_{G}(e) + ® + ¾=(F(\hat{¹})N) ¡ (1 ¡ F(\hat{¹}))(C_{S} ¡ ±) = 0$$

$$W_{°} \;\acute{}\; -_{°} = ¾ + NF(\hat{¹})® ¡ v(NF(\hat{¹})e) ¡ \bar{V} = 0$$

where $-_{¾}$ and $-_{e}$, evaluated here at the second-best equilibrium, are expressions similar to these appearing in the …rst-best problem and where

$$\begin{aligned} -_{N} &= ¸¾=N^{2} + °F(\hat{¹})(® ¡ ev^{0}) \\ -_{\hat{¹}} &= ¡¸f(\hat{¹})[C_{G}(e) + ® ¡ (C_{S} ¡ ±)] + °Nf(\hat{¹})(® ¡ ev^{0}) \\ -_{¼} &= ¡[F(\hat{¹})u_{G}^{0} + (1 ¡ F(\hat{¹}))u_{S}^{0}] + ¸ \\ -_{±} &= (1 ¡ F(\hat{¹}))(¸ ¡ u_{S}^{0}) \end{aligned}$$

>From the …rst-order condition (FOC) related to ¾, it is straightforward to verify that the …rst term on the right-hand side of $W_{®} = 0$ is equal to zero. Recalling that $\frac{@e}{@®} > 0$, it is then easy to check that the FOC for the optimal value of ® can be met if and only if

$$-_{e} + -_{\hat{¹}}\frac{@\hat{¹}}{@e} = 0 \tag{16}$$

>From the other …rst-order conditions, this implies in turn that

$$-_{N} = 0 \tag{17}$$

$$-_{¼} + -_{\hat{¹}}\frac{@\hat{¹}}{@¼} = 0 \tag{18}$$

$$-_{±} + -_{\hat{¹}}\frac{@\hat{¹}}{@±} = 0 \tag{19}$$

The next propositions all follow from these expressions. The …rst two propositions determine the optimal payment system at the second best while Proposition 4 characterizes the resulting GP's diagnosis e¤ort and patients' allocation.

Proposition 2 In the second best, where patients select one type of physician and where GPs decide on their diagnosis e¤ort, the optimal GP's remuneration system is characterized by :

$$\begin{aligned} \text{®} &= \frac{1}{\text{¸}\, f(\hat{\mu})}\left(\frac{@\hat{\mu}}{@e}\right)^{-1}\left[p_h D_e^{0} \int_{\underline{\mu}}^{\hat{\mu}} \mu dF(\mu) - \text{¸}\, F(\hat{\mu})\frac{dC_G}{de}\right] \\ &+ \left[(C_S - \pm) - C_G(e)\right] \\ \text{¾} &= -N F(\hat{\mu})(\text{®} - ev^{0}) \end{aligned}$$

Proof : It can be shown that the …rst part of Proposition 2 follows from (16) while the optimal level of salary ¾ is directly determined by $-_N = 0$. ■

Note furthermore that from the physician's problem (6),

$$\text{®} - ev^{0} = \frac{F(\hat{\mu})}{f(\hat{\mu})}\left(\frac{@\hat{\mu}}{@e}\right)^{-1} v^{0}$$

Hence, substituting this relation into the de…nition of ¾ given in Proposition 2, we …nd that at the second-best optimum, the physician's salary satis…es

$$\text{¾} = -N F(\hat{\mu})\frac{F(\hat{\mu})}{f(\hat{\mu})}\left(\frac{@\hat{\mu}}{@e}\right)^{-1} v^{0}$$

This last expression and Proposition 2 prove the role of competition on the de…nition of the optimal GP's remuneration system. Recall that this role can be measured through the derivative $\frac{@\hat{\mu}}{@e}$. Clearly, the larger the competition, the higher this derivative and therefore the smaller the fee-for-service ® and the larger the salary ¾. This suggests that the optimal GP's remuneration system is more prospective the larger the competition with specialists. Hence, this result would con…rm the idea that competition may reduce the need for regulation, with this one taking here the form of …nancial incentives.

At the second-best solution, the optimal prospective payment ¾ takes a negative value. As to the optimal fee-for-service reimbursement ®, Proposition 2 shows that it is function of the following factors. If it decreases with the e¤ects of competition (see the above argument), the optimal fee-for-service ® increases with $p_h D_e^{0} \int_{\underline{\mu}}^{\hat{\mu}} \mu dF(\mu) - \text{¸}\, F(\hat{\mu})\frac{dC_G}{de}$ and with the cost-di¤erence observed between specialized and general medicine. The …rst positive factor indicates the marginal bene…t for the

patient of an increase in the physician's e¤ort, taking into account the e¤ects it implies in terms of the cost of general medicine and neglecting its consequences with regard to the physician's utility level. Clearly, when these marginal bene…ts increase, the regulator is more interested in inducing the physician to produce a higher diagnosis e¤ort. Similarly, when the cost-disadvantage of specialized medicine becomes relatively higher, the regulator wants to increase the physician's e¤ort since, as a result of competition, this contributes to increase the proportion of patients selecting a general practitioner rather than a specialist. And from the physician's problem, the regulator knows that such a higher e¤ort can be induced through the adoption of a higher fee-for-service since $\frac{@e}{@®} > 0$.

Having characterized the optimal GP's reimbursement system, we can now look at the patient's copayment that satis…es the second-best optimum.

Proposition 3 In the second best, the optimal patient's copayment ± is de…ned by

$$\pm = [C_S - C_G(e) - ® - \frac{¾}{F(\hat{¹})N}] - \frac{F(\hat{¹})(1 - F(\hat{¹}))}{f(\hat{¹})} \frac{(u_S^0 - u_G^0)}{u_S^0 u_G^0} p_h(1 - D(e))$$

Proof : This de…nition results directly from the …rst-order conditions related to ± and ¼. To see this, start from the relation

$$W_¼ \frac{@\hat{¹}=@\pm}{@\hat{¹}=@¼} - W_\pm = 0$$

>From (9), it can be shown that a development of this relation yields

$$, = \frac{u_G^0 u_S^0}{(1 - F(\hat{¹}))u_G^0 + F(\hat{¹})u_S^0} \quad (20)$$

It remains then to substitute (20) into (18) to demonstrate Proposition 3. ■

Following Proposition 3, three elements a¤ect the value of the optimal patient's copayment ± : the relative cost-disadvantage of specialized medicine (…rst term at the right-hand side of Proposition 3), the GP's

diagnosis e¤ort e and the inequalities in marginal utility ($u_S^0 - u_G^0$) that are observed between the patients who choose to be treated by a general practitioner or a specialist. The regulator, who maximizes the patients' utility, does not wish to impose too large inequalities between the patients who visit a general practitioner or a specialist. This induces him to adopt a smaller value $\hat{\mu}$. On the contrary, the value of the patient's copayment ± increases with the …rst two factors. Clearly, the relatively more expensive the specialized medicine and the better the GP's diagnosis e¤ort (and so the quality of general medicine), the more the regulator is interested in inducing patients to visit …rst a GP rather than a specialist. And from (3), the regulator knows that an increase in the copayment ± provides such incentives.

Finally, note that from (3), Proposition 3 allows us to state that, as the optimal patient's copayment ±, the optimal number of patients that select a treatment from general practitioners increases with the cost-disadvantage of specialized medicine and the GP's diagnosis e¤ort e while decreasing with the inequalities in marginal utility ($u_S^0 - u_G^0$). Comparing this with Proposition 1, we may observe that the …rst two factors determine the optimal allocation of patients both in the …rst and second best. What mainly di¤erentiates the two solutions is the additional in‡uence of inequalities in marginal utilities that is reported in the second best. The next proposition compares further the …rst- and second-best solutions.

Proposition 4 The second-best optimum is characterized by a too small number of patients selecting a GP's treatment and by a too high GP's diagnosis e¤ort.

Proof : Proposition 4 follows from expression (16). To see this, recall the value of ¸ given in (20). It is then straightforward to verify that

$$-_{¼} < 0 \tag{21}$$

Further, it can be shown that risk aversion implies ¸ < u_S^0. Consequently,

$$-_{±} < 0 \tag{22}$$

>From (18) and (19), expressions (21) and (22) imply that

$$-_{\hat{\mu}} > 0 \tag{23}$$

and from (16),

$$-_e < 0 \qquad (24)$$

Expressions (23) and (24) demonstrate Proposition 4. ■

Proposition 4 proves that the regulator's choice of the optimal payment system at the second best does not allow to attain the …rst-best solution in terms of the patients' allocation and the GP's diagnosis e¤ort. Expressions (23) and (24) show that the size of the insurance premium ¼ and copayment ± is larger than what would be observed if their impact on μ was not considered. In fact, expressions (18) and (19) show that these higher values compensate for a smaller threshold value $\hat{\mu}$ and so for a too high number of patients choosing to be treated by a specialist. It is to counteract this inappropriate allocation of patients that higher copayment and insurance premium are adopted by the regulator. >From (3), it is expected that these higher levels will induce more patients to prefer to be treated by their GP rather than by a specialist. The larger diagnosis e¤ort e supplied by the general practitioner can be given a similar interpretation. To compensate for a too small $\hat{\mu}$, the regulator adopts a reimbursement system for the GP that gives him incentives to supply a higher diagnosis e¤ort. By increasing the quality of the GP's referral behaviour, it is expected that a larger number of patients will choose general medicine rather than specialized services. Hence, to sum up, if starting from the second best, we were able to directly control the patient allocation $\hat{\mu}$ and the GP e¤ort e, it would be optimal to increase $\hat{\mu}$ and, simultaneously, reduce the GP's diagnosis e¤ort e.

4 Conclusion

In this paper, we have studied the optimal payment system to impose on the primary health care market when competition prevails between general practitioners and specialists. To formalize such a framework, we have de…ned a three-stage model. In section 2.1, we have developed a horizontal di¤erentiation model to investigate the patient's selection of one category of physicians. In section 2.2, we have evaluated the diagnosis e¤ort chosen by the general practitioner when the latter takes into account the competition played with the other GPs and specialists. Finally, in section 3, we have examined how the regulator de…nes the patient copayment and the physician reimbursement system that guarantee

a good allocation of patients among the di¤erent types of physicians as well as an appropriate referral of patients to specialists.

The analysis proves that at the second-best solution, the GP's remuneration system is more prospective, the larger the competition with specialists. Next, it shows that the optimal value of fee-for-service increases with the marginal bene…ts patients get from a larger GP's diagnosis e¤ort and with the cost-disadvantage of specialized medicine. Since the latter calls for more patients visiting a general practitioner rather than a specialist, and given that, from competition, this can result from a higher GP's diagnosis e¤ort, both factors make the regulator more interested in inducing general practitioners to produce a higher diagnosis e¤ort. And from the physician's problem, he knows that a higher fee-for-service reimbursement o¤ers such an inducement.

Concerning the second-best optimal value of GP's diagnosis e¤ort and patients' allocation, we …nd that the inequalities in marginal utility that are observed between the patients who visit a general practitioner and those who select a specialist preclude the optimal payment system to achieve the …rst-best solution. But our study also con…rms the result of Bouckaert (1998) and Jelovac (1998) that it is not always optimal for the regulator to give …nancial incentives to restrict the use of specialized medicine. We …nd, indeed, that the optimal patient's copayment, and so the optimal patients' allocation among the two categories of physicians, does not only depend on these inequalities in marginal utility but also on the cost di¤erence reported between the two types of health care services. Next, we prove that the quality of the GP's referral behaviour is also crucial for evaluating the interest of controlling the access to specialized medicine.

In terms of policy implications, these results invite us to reconsider systems of gatekeepers where all patients are required to go to a general practitioner before having access to specialized medicine. They show that a less rigid system may be preferred in certain circumstances. Important factors are, notably, the cost di¤erence between the two types of health care services and the GP's referral behaviour. Our analysis con…rms that an alternative policy to compulsory gatekeeping could consist in adopting …nancial incentives. These might contribute to improve the allocation of patients on the primary health care market. Moreover, our analysis allows to assess the incentive power that can be assigned to the GP's reimbursement system and the competition in force on the market.

In order to re…ne these …rst results, it might be interesting to bring

some extensions to our analysis. First, we could introduce elements of imperfect information. If asymmetric information was supposed between patients and general practitioners, the impact of competition on the market equilibrium might di¤er. This might now be a¤ected by reputation concerns and by the patient's search for adequate treatments. Examples of theoretical studies devoted to this subject are notably the papers of Rochaix (1989) and Wolinsky (1993). Second, the GP's problem might be extended to a second decision variable. The general practitioner might not only have to decide on his diagnosis e¤ort but also on the treatment supplied. Such an extension could provide a potential source of opportunist behaviour for the general practitioner.

References

- Bouckaert, J. (1998), "Price competition between an expert and a non-expert", CEPR Discussion Paper Series, No 1905, Center for Economic Research, Tilburg University.
- Jelovac, I. (1998), "GP's payment contracts and their referral policy", in Incentive contracting and organisation in health services, International Doctorate in Economic Analysis, chapter 3, Departement d'Economia i d'Historia Economica, Universitat Autonoma de Barcelona, Spain.
- Rochaix, L. (1989), "Information asymmetry and search in the market for physicians' services", Journal of Health Economics, 8, 53-84.
- Wolinsky, A. (1993), "Competition in market for informed experts' services", Rand Journal of Economics, 24, 380-390.