Treatment and referral decisions under different physician payment mechanisms

Marie Allard, Izabela Jelovac, Pierre-Thomas Léger

Treatment and referral decisions under different physician payment mechanisms.*

Marie Allard
HEC Montréal

Izabela Jelovac
CNRS-GATE Lyon St Etienne

Pierre Thomas Léger
HEC Montréal, CIRANO and CIRPÉE†

June 2011

Abstract

This paper analyzes and compares the incentive properties of some common payment mechanisms for GPs, namely fee for service (FFS), capitation and fundholding. It focuses on gatekeeping GPs and it specifically recognizes GPs heterogeneity in both ability and altruism. It also allows inappropriate care by GPs to lead to more serious illnesses. The results are as follows. Capitation is the payment mechanism that induces the most referrals to expensive specialty care. Fundholding may induce almost as much referrals as capitation when the expected costs of GPs care are high relative to those of specialty care. Although driven by financial incentives of different nature, the strategic behaviours associated with fundholding and FFS are very much alike. Finally, whether a regulator should use one or another payment mechanism for GPs will depend on (i) his priorities (either cost-containment or quality enhancement) which, in turn, depend on the expected cost difference between GPs care and specialty care, and (ii) the distribution of profiles (diagnostic ability and altruism levels) among GPs.

*Léger thanks HEC Montréal, FQRSC and SSHRC for funding. Jelovac thanks Banque Nationale Belge for funding. We thank two anonymous referees as well as seminar and conference participants at the University of Lausanne, GREQAM (Marseille), Maastricht University, Oslo University, GATE (Lyon), PSE (Paris), Université Catholique de Lille, University of Lund, CHESG (Ottawa), AHEC (Arizona State), CIRPÉE Conference in Health Economics at l'Université Laval (Québec), Health Economics Worskhop at the Norwegian School of Economics (Bergen), SSES Meetings (Lausanne), CEA (Vancouver) and ASHE (Duke) for helpful comments and suggestions. Previously titled: Physician Payment Mechanisms: Dynamics, Diagnostic Ability and Altruism.

†Corresponding author: 3000 Côte-Sainte-Catherine, Montréal, Quebec, Canada H3T 2A7. E-mail: ptl@hec.ca

1 Introduction

An important part of the literature on the industrial organization of health care has focused on examining patient and physician behavior in different institutional frameworks in order to ultimately suggest policies that will encourage the efficient provision of care. Institutional frameworks are most often characterized by their insurance system and their physician payment scheme. As patients, physicians and insurance providers are likely to have private information and competing objectives, deriving optimal insurance plans and physician payment schemes is complicated. Although much has been written on the topic, the previous literature has mostly analyzed patient and physician behavior in a static framework without gatekeeping.[1]

In practice, some of the most commonly used and studied payment mechanisms for physicians are capitation, fee for service (FFS) and fundholding. The FFS mechanism, similar to the hospitals' retrospective reimbursement contracts, reimburses health care providers for their full costs of providing health services. Although it has become less popular, FFS remains the predominant payment system in the United States, Canada and many European countries. The capitation scheme, similar to prospective reimbursement for hospitals, shifts cost responsibilities to physicians as it gives the physician or a group of physicians a periodical payment for each patient they enlist, and makes the physician responsible for all physician costs. Capitation is well known for being used in Health Maintenance Organizations in the US, but it is also used in other countries including Spain and the United Kingdom. With fundholding, the principle of capitation is extended to make physicians responsible for costs of specialized and hospital care as well. This

[1] In these models (Blomqvist (1991), Ma and McGuire (1997), Blomqvist and Léger (2005)), patients generally face a distribution of illness severity and, conditional on a particular ilness severity, decide on the type of provider and the quantity of care to consume. Once the treatment has been provided, the patient's ex post health is revealed, payments are made and the analysis ends. One exception to this is Allard, Léger, and Rochaix (2009) who examine patient and doctor interaction in a repeated-game framework.

payment system was introduced in the UK in 1991 and abolished in 1998.

It is usually admitted that FFS gives no incentive to limit medical costs as the only way that a physician can increase his income under such a system is to increase volume; it is, however, expected to preserve the quality of physicians' services. Conversely, capitation is suspected to induce a lower quality and/or quantity of care as physicians may wish to control costs in order to increase their income. Chalkley and Malcomson (1998) are among the authors concluding so when analyzing the case of self-interested hospitals. Fundholding is expected to contain costs even more than capitation as it is designed as a generalization of the capitation scheme.

This paper brings important nuances to these commonly shared beliefs on physicians' payment mechanisms. The distinctive features of our approach are the following. First, we focus on gatekeeping GPs rather than considering health care providers in general or hospitals specifically. This focus is motivated by the recent trend to force patients to seek a referral by a GP before accessing specialty care. This pivotal role of the GPs is expected to reduce overall specialty care costs. Second, we recognize that GPs are heterogeneous in terms of both ability and altruism. We believe that this is not specific to the medical profession. However, it should not be ignored when analyzing GPs' responses to the incentives associated with their payment mechanisms. Third, we introduce dynamics in the health production function by allowing inappropriate care by the GP to lead to more serious illnesses.

In our model, GPs decide whether to treat or to refer a patient according to the outcome of the diagnosis. Certain types of illnesses are diagnosed with potential error – where the diagnostic precision depends on the GP's diagnostic ability.[2] The most able GPs have the

[2]Both Garcia-Marinoso and Jelovac (2003) and Malcomson (2004) consider incentive contracts for gatekeeping GPs in a setting where the diagnostic precision is endogenous. Although we do not consider endogenous precision, we acknowledge that diagnostic precision may come in part from costly effort. In this sense, the results presented here can be considered complementary to those presented by the aforementioned

most accurate diagnosis so that they should ideally trust their judgment when deciding between a GPs' treatment and a referral to a specialist. The less able GPs should in turn adopt a systematic behavior, depending on the priority of the regulator. If the regulator's main concern is cost containment, then the less able GPs should systematically treat their patient before eventually referring them to the specialist, to avoid the financial costs of wasteful referrals. If the regulator's main concern is to guarantee patients' expected health, then the less able GPs should systematically refer patients to more expensive and able specialists, to avoid the health and utility costs of inappropriate treatment.

A regulator cannot realistically tell a GP how to behave because both the GP's type (i.e., his level of altruism and diagnostic ability) and the outcome of his diagnosis are the GP's private information. However, as usually recognized, improving efficiency is an essential part of any strategy for reconciling rising demands for health care with the need for public budget restraint. Therefore, providing the right incentives for key actors in the system is a pre-requisite. In that sense, a regulator may propose different financial incentives to induce GPs' optimal behaviors. We focus here on the incentives associated with some of the most common GPs' payment methods and present our main results on these incentives hereafter.

The implication of focusing on gatekeeping GPs is that the incentives associated with capitation give opposite results to those usually admitted. We show that a GP paid by capitation is better off systematically referring patients to more expensive specialized care to avoid paying the costs of directly treating the patients, no matter his level of ability and altruism. Overall, this referral behavior increases costs without limiting quality, even though the costs of treatment by GPs are kept at the lowest level. This result is in line with one already obtained by Garcia-Marinoso and Jelovac (2003) for a self-interested

authors.

representative GP. This result is also consistent with empirical evidence, which suggested that physicians paid on a FFS basis were less likely to refer patients than those paid by capitation (Grembowski, Cook, Patrick, and Roussel (1998)).

Furthermore, we show that fundholding surprisingly results in GPs' behaviors that can be similar to the ones induced by FFS. To understand this similarity, it is important to identify the link between GPs' response to incentives and their diagnostic ability and altruism profile. When paid on a FFS basis, the less altruistic yet able GPs have an incentive to systematically treat their patients on their own to maximize their revenue from fees without risking the patient's health in the process. Under fundholding, those GPs behave in the same way to economize on the expected specialist costs for which they are held responsible. Conversely, sufficiently altruistic and less able GPs have an incentive to systematically refer their patients to specialized care because they know that specialists are better at providing the good care that they wish for their patients, no matter whether they are paid on a FFS or fundholding basis. The remaining GPs, i.e. those with a high diagnostic ability and a moderate sense of altruism, have an incentive to base their decision (either treat or refer the patient) on the outcome of their rather accurate diagnosis.

Health economists are increasingly incorporating physicians' altruism into their models, as well as the positive influence of altruism on the endogenous quality of care (see Chalkley and Malcomson (1998), Jack (2005), Biglaizer and Ma (2007), Deflgaauw (2007), and Choné and Ma (2011)). We complement this approach by considering that the quality of health services has an exogenous component as well, i.e. GP's ability. Therefore, good quality care in our model can result either from a high level of altruism (systematic referrals to good quality specialists) or a high level of GP's ability (adequate decisions based on an accurate diagnosis). This distinction is worth noticing because of its important implications in terms of costs.

Whether a regulator should use one or another mechanism to pay GPs depends on his priorities (either cost containment or quality enhancement) as well as on the distribution of GPs profiles. The same conclusion appears in Chalkley and Malcomson (1998) despite the important differences that we have now identified between hospitals' and gatekeeping GPs' optimal responses to payment incentives. However, having no observation about the distribution of GPs profiles makes it difficult to provide a clear-cut policy recommendation. To tackle this problem, Jack (2005) proposes to design a menu of non-linear contracts relating payment to costs, to induce physicians to reveal their profile. However, in a companion paper (Allard, Jelovac, and Léger (2010)), we show that self selection of a payment mechanism is not always optimal for gatekeeping GPs who differ in both altruism and ability, again depending on GPs profiles and on the regulator's priorities. A thorough empirical analysis would be very useful to elicit the distribution of GPs profiles and ultimately provide reliable policy recommendations.

The remainder of the paper is organized as follows. In Section 2 we present the model. In Section 3 we derive the ideal (First-Best) GP's treatment and referral decisions as a function of his diagnostic ability. In Section 4 we analyze how different types of GPs (with respect to their altruism and diagnostic ability) are likely to respond in terms of treatment and referral decisions to three common forms of payment mechanisms: (i) FFS, (ii) capitation, and (iii) fundholding. In this section, we also compare the GP's behavior under each of these payment mechanisms to those derived in the First Best. In Section 5, we compare the behavior of each type of GP across the three types of payment mechanisms in order to identify if and when a particular payment mechanism yields the First-Best outcome and/or provides better incentives than any other. Conclusions are drawn in Section 6.

2 The Model

In this section we introduce a simple model where GPs act as gatekeepers to specialty care and must decide whether to treat their patient or refer them to specialty care. In our model, individuals may suffer from different illness severities. The treatment protocol, however, may not be obvious for all types of illnesses. That is, individuals with different types of illnesses requiring different types of treatments may exhibit the same set of symptoms (for example, skin spots which may or may not be cancerous). If a GP fails to refer a severely ill patient to the specialist (for example, fails to refer a patient with skin cancer to the oncologist), then the patient's health may deteriorate, potentially requiring more invasive and expensive care in the future. If the GP refers a relatively mild case to the specialist (for example, refers a patient with simple skin spots to the oncologist), unnecessary expenses will be incurred without improving the patient's health or utility. To allow for such a situation, we consider two types of illnesses which are diagnosed with potential error and where the physician's ability to differentiate between them depends on his diagnostic ability. The decision to refer patients to specialty care will obviously depend on the GP's diagnostic ability (i.e., the ability to differentiate between illnesses with similar symptoms), the GP's level of altruism, and the institutional framework.[3]

We next present the timing of a two-period model while specifying the patients' and physicians' objective functions.

2.1 The Timing

The First Period:

Stage 1 - Physician Payments and Patient Insurance:

[3] Other papers on GP contracting and gatekeeping have also considered the role of GPs in recommending on particular type of specialty care among many possibilities (Brekke, Nuscheler, and Straume (2007), Levaggi and Rochaix (2007)).

All physician payments and insurance parameters are contracted upon, where α denotes the actuarially fair insurance premium. In our model, we assume that patients are fully insured and always follow their physician's recommendations.[4,5]

Stage 2 - Nature Plays:

The patient suffers from an illness severity M which takes on two values with corresponding probabilities: a minor condition M^L with probability p or a severe condition M^H with probability $(1-p)$.

Stage 3 - The Diagnostic Signal:

The GP is assumed to receive a costless yet noisy *diagnostic signal* which is a function of his diagnostic ability $a \in \left[\frac{1}{2}, 1\right]$.[6] More specifically, we assume that the quality of the *diagnostic signal,* S^L or S^H, is increasing in the GP's diagnostic ability a, where:

$$\Pr(S^L|M^L) = a$$
$$\Pr(S^H|M^L) = 1-a$$
$$\Pr(S^H|M^H) = a$$
$$\Pr(S^L|M^H) = 1-a.$$

That is, given a diagnostic ability a and a true illness severity M^L, the GP receives the correct diagnostic signal S^L with probability a while receiving the incorrect diagnostic signal S^H with probability $(1-a)$. Similarly, given a diagnostic ability a and a true illness severity M^H, the GP receives the correct diagnostic signal S^H with probability a while receiving the incorrect diagnostic signal S^L with probability $(1-a)$. Notice that when

[4] Ex-post moral hazard is non-existent and co-payments are unnecessary.

[5] We make this simplifying assumption (that patients are passive) to abstract from demand-side incentives and highlight supply-side ones. See Gonzalez (2010) for an example where this assumption is relaxed and the role of patients' information about their own health is recognized in the analysis of GP decisions.

[6] For simplicity, we assume that diagnosis is effortless for GPs. It is however without loss of generality since the accuracy of the diagnosis is exogenous here.

$a = \frac{1}{2}$, the diagnostic signal contains no information, whereas when $a = 1$, the diagnostic signal is perfect (i.e., physicians work in a perfect information setting).

Further notice that by using Bayes' rule we can define the probability that the patient suffers from a particular illness severity given a specific diagnostic signal as a function of the physician's diagnostic ability:

$$\begin{aligned}
\Pr(M^L|S^L) &= \tfrac{pa}{pa+(1-p)(1-a)} \\
\Pr(M^H|S^L) &= \tfrac{(1-p)(1-a)}{pa+(1-p)(1-a)} \\
\Pr(M^H|S^H) &= \tfrac{(1-p)a}{p(1-a)+(1-p)a} \\
\Pr(M^L|S^H) &= \tfrac{p(1-a)}{p(1-a)+(1-p)a}.
\end{aligned}$$

Stage 4 - Treatment and Referral Decisions:

Following his diagnosis which may be more or less accurate the GP must decide on a course of action. We assume that each illness severity is associated with an appropriate treatment T, where the appropriate treatment for M^L (respectively, M^H) is T^L (respectively, T^H). We further assume that the GP can provide the treatment T^L whereas specialists can provide both treatments T^L and T^H.

Given the diagnostic signal (S^L or S^H), the GP must decide on whether to treat his patient with T^L or refer her to the specialist. If the patient is referred to specialty care, the specialist can either treat her with T^L or T^H. Finally, we assume that the specialist's diagnostic ability is perfect and that he always provides the appropriate treatment. That is, the specialist is assumed to behave non-strategically.

The Second Period:

We introduce dynamics into the health production function by allowing for inappropriate treatment by the GP to lead to more serious illnesses. More specifically, we assume that an individual suffering from illness severity M^H but treated inappropriately by her

GP with T^L sees her illness severity worsen in the following period (an illness severity which can only be treated by the specialist).

Formally, if the patient received the appropriate treatment in Period 1 (i.e., T^L for M^L or T^H for M^H), then the patient's health remains the same and no action is taken in the second period. If, however, the patient received an inappropriate treatment in Period 1 (i.e., T^L for M^H), then she enters the second period with a deteriorated health and is referred to the specialist for treatment. More specifically, the inappropriate treatment implies a health loss L for the patient and further treatment costs. L may include pain and suffering, and can also be thought of as the "expected" health loss under inappopriate care.

2.2 The Patient and Physicians' Preferences:

The patient's one period utility function is given by:

$$U = u(h - M + T) + I - \alpha,$$

where $u(.)$ denotes the utility the patient derives from health, h denotes the initial health status, I denotes the state-independent income, and α denotes the actuarially fair insurance premium.[7] We assume that $u'(.) > 0$ and $u''(.) < 0$.

The GP's one period utility function is given by:[8]

$$V_{GP}(T) = R_{GP}(T) - c_{GP}(T) + \beta u(h - M + T),$$

[7] We assume a separable utility function to make things more tractable. However, the qualitative results do not depend on this assumption.

[8] Because of the full-insurance assumption, the patient's income is invariant to the illness severity and treatment. Consequently, adding the patient's entire utility into the physician's utility function (instead of only the patient's utility from health) would not change the results.

where $R_{GP}(T)$ denotes the GP's payment for treating the patient with T and where $c_{GP}(T)$ denotes the GP's cost of providing treatment T. Furthermore, $\beta \in [0,1]$ denotes the weight the GP puts on the patient's utility from health (i.e., the altruism parameter).[9]

We assume that the GP can treat the patient with T^L at cost $c_{GP}(T^L)$ whereas specialists are assumed to provide T^L and T^H at costs $c_{SP}\left(T^L\right)$ and $c_{SP}\left(T^H\right)$, respectively.[10] We also assume that $c_{GP}(T^L) < c_{SP}(T^L)$ and that $c_{SP}(T^L) = c_{SP}(T^H)$. The first assumption is made to reflect the fact that specialists use more sophisticated and expensive technologies to diagnose and treat patients than their GP counterparts.[11] The second assumption is made to reflect the fact that for a variety of treatments (T^L and T^H), the differences in treatment costs for specialty care are likely to be small given the large up-front costs (such as diagnostic testing). We further assume that all agents face a common discount factor given by δ. Finally, we make the natural assumption that it is better to be appropriatly treated than to be inappropriatly treated and suffer from a health loss L. Collectively, these assumptions translate into conditions:

$$(1+\delta)u(h-M^L+T^L) \geq (1+\delta)u(h-M^H+T^H) > u(h-M^H+T^L)+\delta u(h-M^H-L+T^H).$$

3 First Best

In this section, we derive the GPs' treatment and referral decisions which maximize the patient's expected utility subject to the GPs' participation constraint and assuming that the GP's diagnostic ability a is known - what we call the First Best. Doing so is consis-

[9] Thus, a physician with $\beta = 1$ weighs equally his patient's utility from health and his own revenue. Although a physician could weigh his patient's health greater than his own income (i.e., have a $\beta > 1$) we do not consider such a case.

[10] We allow costs of treatment to include the physician's time and effort.

[11] This assumption is consistent with empirical evidence provided by the studies of Greenfield (1992) and Carey (1995).

tent with a model where patients can write physician-type-and-state-contingent contracts, essentially forcing the hand of their physicians based on their diagnostic ability and diagnosis. Although such a contract is infeasible, it can serve as a benchmark to which we can compare physician behaviour (i.e., treatment and referral decisions) under different payment mechanisms (i.e., feasible contracts). In order to just satisfy the GPs' participation constraint, GPs' payment will be set equal to the cost of providing the treatment (i.e., $R_{GP}(T^L) = c_{GP}(T^L)$).

Next we derive the First-Best treatment and referral strategies for a GP who observes a diagnostic signal S that can either be S^L (signaling a relatively mild illness) or S^H (signaling a relatively severe illness), respectively. These cases will be discussed in Lemmas 1 and 2, respectively.

Lemma 1: To maximize the patient's expected utility subject to the GP's participation constraints, a GP who observes a diagnostic signal S^L should treat the patient with T^L (should refer the patient to the specialist) if and only if his diagnostic ability:

$$a \geq (<) \frac{(1-p)B}{pA + (1-p)B},$$

where

$$A \equiv c_{SP}(T^L) - c_{GP}(T^L),$$

and

$$\begin{aligned} B \equiv\ & (1+\delta)u(h - M^H + T^H) - c_{SP}(T^H) \\ & - \left[u(h - M^H + T^L) + \delta u(h - M^H - L + T^H) - c_{GP}(T^L) - \delta c_{SP}(T^H)\right]. \end{aligned}$$

Proof: see Appendix 1.

Note that A is the cost difference between having the GP and the specialist treat with

T^L when the patient suffers from M^L (i.e., the relatively mild illness). B, on the other hand, is the difference in the patient's net expected utility between seeking care directly from the specialist and seeking care from a GP, when she suffers from M^H (i.e., the relatively severe illness), which is increasing in L. Thus pA can be thought of as the weighted marginal benefit of treatment by a GP when suffering from M^L (relative to treatment by a specialist) whereas $(1-p)B$ can be thought of as the weighted marginal benefit of seeking specialty care when suffering from M^H (relative to treatment by a GP) net of expected costs. We have that $A > 0$ because we reasonably assume that specialists costs are higher than GPs costs. Furthermore, we assume that $B > 0$ to account for the fact that an expensive and appropriate treatment is preferred to an inexpensive and inappropriate treatment.

From the above, we can show that as the probability p (i.e., the probability that the patient suffers from M^L) increases (decreases), the minimum diagnostic ability level for which GPs should treat with T^L rather than refer the patient to specialty care will decrease (increase). Furthermore, as the cost differential associated with the treatment with T^L by a GP and a specialist decreases (increases), the minimum diagnostic ability level for which GPs should treat the patient with T^L rather than refer to the specialist will increase (decrease).

Lemma 2: To maximize the patient's expected utility subject to the GP's participation constraints, a GP who observes a diagnostic signal S^H should refer the patient to the specialist (treat with T^L) if and only if his diagnostic ability:

$$a \geq (<)\frac{pA}{pA + (1-p)B}$$

Proof: see Appendix 2.

We can also show that as the probability $(1-p)$ (i.e., the probability that the patient suffers from M^H) increases (decreases), the minimum diagnostic ability level for which

GPs should refer the patient to specialty care rather than treat with T^L will decrease (increase). Furthermore, as the marginal benefit of seeking specialty care when suffering from M^H relative to treatment by the GP decreases (increases), the minimum diagnostic ability for which GPs should refer the patient to specialty care rather than treat with T^L will increase (decrease).

The decision rule:

Given the results of *Lemmas 1* and *2*, we can now present the GP's First-Best treatment and referral decisions as a function of the diagnostic signal and exogenously given values of A, B, p and $(1-p)$. The GP's First-Best treatment and referral decisions are, however, dependent on whether $pA > (1-p)B$ or $pA < (1-p)B$ which is also exogenously determined. Recall that A is the marginal benefit of treatment by a GP (relative to treatment by a specialist) when the patient suffers from a relatively mild illness (M^L), and that B is the marginal benefit of seeking specialty care (relative to care by a GP) when the patient suffers from a relatively severe illness (M^H). Consequently, if a patient who suffers from M^L is sent to specialty care rather than being treated by the GP, what we define as a Type I error, she suffers a loss of A. Furthermore, if a patient who suffers from M^H is not sent to specialty care, what we define as a Type II error, then she suffers a loss of B. Thus, if $pA > (<)\ (1-p)B$, it is worse in expected utility terms to make a Type I (Type II) error than a Type II (Type I) error. We henceforth refer to the cases where $pA > (1-p)B$ (a situation where concerns of wasteful referrals dominate concerns of under referrals) and $pA < (1-p)B$ (a situation where concerns of under referrals dominate concerns of wasteful referrals) as Scenarios I and II, respectively.

Scenario I (Scenario II) is likely to occur, when: (i) the costs associated with specialty care are relatively large (small), and/or (ii) the likelihood of a relatively severe illness (M^H) is low (high). Given that the results derived throughout depend on whether $pA > (1-p)B$

or $pA < (1-p)B$, we summarize the different scenarios in the following table (Table 1) for easy reference.

Table 1:

Scenario I $\mathbf{pA > (1-p)B}$	**Concerns of wasteful referrals dominate concerns of under referrals**	Better to make a Type II error than a Type I error
Scenario II $\mathbf{pA < (1-p)B}$	**Concerns of under referrals dominate concerns of wasteful referrals**	Better to make a Type I error than a Type II error

Proposition 1: Under Scenario I:

(i) a GP with a diagnostic ability $a \in \left[\frac{1}{2}, \frac{pA}{pA+(1-p)B}\right]$ *should always treat with* T^L *irrespective of his diagnostic signal (i.e., he should ignore his diagnostic signal);*

(ii) a GP with a diagnostic ability $a \in \left[\frac{pA}{pA+(1-p)B}, 1\right]$ *should always follow his diagnostic signal (i.e., should treat the patient with* T^L *when he receives a diagnostic signal* S^L *and refer the patient to specialty care when he receives a diagnostic signal* S^H*).*

Proof: See Appendix 3.

Proposition 2: Under Scenario II:

(i) a GP with a diagnostic ability $a \in \left[\frac{1}{2}, \frac{(1-p)B}{pA+(1-p)B}\right]$ *should always refer the patient to the specialist irrespective of his diagnostic signal (i.e., he should ignore his diagnostic signal);*

(ii) a GP with a diagnostic ability $a \in \left[\frac{(1-p)B}{pA+(1-p)B}, 1\right]$ *should always follow his diagnostic signal (i.e., should treat the patient with* T^L *when he receives a diagnostic signal* S^L *and refer the patient to specialty care when he receives a diagnostic signal* S^H*).*

Proof: See Appendix 4.

Hereafter, a^* denotes the first-best threshold values of the GP's ability under scenarios I and II, i.e., $a^* = \frac{Max\{pA,(1-p)B\}}{pA+(1-p)B}$.

From *Propositions 1* and *2*, we can see that low-ability GPs should adopt a systematic strategy of either always treating the patient with T^L or always referring the patient to the specialist, whereas high-ability GPs should follow their diagnostic signal. Which systematic strategy the low-ability GP should adopt will depend on whether it is worse in expected utility terms for a GP to refer a patient who suffers from M^L rather than treat her with T^L (a Type I error) or treat with T^L a patient who suffers from M^H rather than refer her to specialty care (a Type II error).[12]

Notice that, irrespective of the scenario, a GP with perfect diagnostic ability (i.e., $a = 1$) should always follow his diagnostic signal. Consequently, when the GP's diagnostic ability is perfect and he follows his diagnostic signal, the patient always receives the appropriate treatment.

4 Physician Payment Mechanisms

In the three subsections below, we derive the GP's treatment and referral behavior which maximizes his expected utility under three common payment mechanisms: (i) fee-for-service, (ii) capitation, and (iii) fundholding, assuming that the patient is fully insured. We then compare the results derived under these physician payment mechanisms with full insurance to those derived in the First Best.[13,14]

[12] If the costs of specialty care were not higher than the costs of treatment by GPs ($C_{GP}(T^L) = C_{SP}(T^L)$, implying $A = 0$ and $a^* = 1$), then Scenario II holds and all GPs should systematically refer their patient to a specialist. In this extreme case, there would be no rationale for GP gatekeeping.

[13] Although mixed-payment systems (i.e., payment mechanisms which are part capitation, part fee for service) are both common in theory and in practice, we do not consider them here because care is considered uni-dimensional in our model. Generally, mixed-payment systems are considered in a multi-tasking environment (where care depends on both an observable component and unobservable physician effort). See Allard, Léger, and Rochaix (2009) and Ma and McGuire (1997) for examples of multitasking in healthcare and the use of mixed-payment systems.

[14] One should recognize that differences that are uncovered reflect both: (i) the strategic behavior induced by the payment mechanism, and (ii) the fact that GPs' utility does not include the costs of treatment borne by the patient (i.e., we assume a paternalistic form of altruism where the GP only cares about his patient's health and not the externality that his behavior may have on the patient's insurance premium at

In the following subsections, we continue to assume that the GP acts as the patient's gatekeeper to specialty care. We further assume that the patient is passive and that all care decisions are taken by the patient's physicians.

4.1 Fee for service

In the traditional fee-for-service (FFS) system, physicians are paid a fixed rate for each service they provide. These treatment-dependent rates are typically above their marginal costs. As a result, we assume that the GP's payment under FFS is given by:[15]

$$R_{GP}(T^L) > c_{GP}(T^L).$$

We now solve for the GP's expected utility maximizing treatment and referral decisions as a function of his private signals, his diagnostic ability as well as his level of altruism.

The GP's expected utility maximizing decisions as a function of the diagnostic signal:

(i) S^L :

If the GP receives a diagnostic signal S^L, he knows that with probability $pa/(pa+(1-p)(1-a))$ the patient suffers from M^L (the diagnosis is correct) while with probability $(1-p)(1-a)/(pa+(1-p)(1-a))$ the patient suffers from M^H (the diagnosis is incorrect). Whether the GP will follow his signal S^L and treat the patient with T^L or ignore it and refer the patient to specialty care, will depend on the expected benefits of following each of these strategies - which in turn will depend on his diagnostic ability a, and his altruism parameter β.

If the GP follows his diagnostic signal (or, equivalently, treats with T^L), his expected

equilibrium).

[15] We assume that the GP only earns the fee when he provides a treatment T^L. We implicitely assume that the GP's diagnostic activity is effortless and that the GP is not compensated for it.

utility is:

$$\frac{pa}{pa+(1-p)(1-a)}\{R_{GP}(T^L)-c_{GP}(T^L)+\beta(1+\delta)u(h-M^L+T^L)\}+$$
$$\frac{(1-p)(1-a)}{pa+(1-p)(1-a)}\{R_{GP}(T^L)-c_{GP}(T^L)+\beta(u(h-M^H+T^L)+$$
$$\delta u(h-M^H-L+T^H))\}.$$

If instead the GP ignores his diagnostic signal and refers the patient to specialty care, his expected utility is:

$$\frac{pa}{pa+(1-p)(1-a)}\beta(1+\delta)u(h-M^L+T^L)+$$
$$\frac{(1-p)(1-a)}{pa+(1-p)(1-a)}\beta(1+\delta)u(h-M^H+T^H).$$

Thus, the GP will follow his signal and treat with T^L rather than refer the patient to specialty care if his diagnostic ability[16]

$$a>\frac{(1-p)\left\{\beta(n-m)-\left[R_{GP}(T^L)-c_{GP}(T^L)\right]\right\}}{(1-p)\left\{\beta(n-m)-\left[R_{GP}(T^L)-c_{GP}(T^L)\right]\right\}+p\{R_{GP}(T^L)-c_{GP}(T^L)\}}\equiv\widetilde{a}_2,$$

where

$$n\equiv(1+\delta)u(h-M^H+T^H),$$

and

$$m\equiv u(h-M^H+T^L)+\delta u(h-M^H-L+T^H).$$

[16]Note that $\widetilde{a}_2$ is increasing in β and concave where $\widetilde{a}_2<1$ for all values of β and $\widetilde{a}_2=\frac{1}{2}$ when $\beta=\frac{(R_{GP}(T^L)-c_{GP}(T^L))}{(1-p)(n-m)}\equiv\widetilde{\beta}$.

Notice that more altruistic GPs will require more precision in their diagnosis before following their diagnostic signal rather than referring their patients to specialty care. This is simply because referring their patients to specialty care is always weakly preferred by the patient while treating is always more lucrative for the GP. Thus, a more altruistic GP will be less willing to trade his own welfare for his patient's and thus requires a more precise diagnosis in order to follow it.

(ii) S^H :

It can also be shown that, if the GP receives a diagnostic signal S^H, the GP will follow his diagnostic signal and refer the patient to specialty care rather than treat with T^L if his diagnostic ability[17,18]

$$a > \frac{p\{R_{GP}(T^L) - c_{GP}(T^L)\}}{(1-p)\{\beta(n-m) - [R_{GP}(T^L) - c_{GP}(T^L)]\} + p\{R_{GP}(T^L) - c_{GP}(T^L)\}} \equiv \widetilde{a}_3.$$

Recall that under a FFS system, treating patients is more lucrative than referring them to specialty care. As a result, less altruistic GPs require more precision in their diagnosis to follow their diagnostic signal, that is, to refer their patients to specialty care when they receive a diagnostic signal S^H.

Using the above results and the properties of $\widetilde{a}_2$ and $\widetilde{a}_3$, we summarize the GP's expected utility maximizing strategies in Figure A.

[Insert Figure A Here]

Notice that when the GP receives a particular diagnostic signal S (either S^L or S^H), he will systematically treat the patient if $\widetilde{a}_2 < a < \widetilde{a}_3$, he will follow his diagnostic signal if $a > \max\{\widetilde{a}_2, \widetilde{a}_3\}$, and he will systematically refer the patient if $\widetilde{a}_3 < a < \widetilde{a}_2$. In general,

[17] Note that $\widetilde{a}_3$ is decreasing in β and convex where $\widetilde{a}_3 < 1$ iff $\beta > \frac{(R_{GP}(T^L) - c_{GP}(T^L))}{(n-m)}$, and $\widetilde{a}_3 = \frac{1}{2}$ when $\beta = \frac{(R_{GP}(T^L) - c_{GP}(T^L))}{(1-p)(n-m)} \equiv \widetilde{\beta}$.

[18] $R_{GP}(T^L) - c_{GP}(T^L) > 0$ assures that $\beta > 0$ when $\widetilde{a}_3 = 1$ is satisfied under both scenarios.

very selfish FFS GPs will wish to always treat their patients as this strategy is income maximizing. On the other hand, relatively altruistic FFS GPs who also have relatively low levels of diagnostic ability will systematically refer their patients to specialty care as they care enough for them not to risk a wrong diagnosis and consequently a potentially bad outcome.

Further notice that, under the FFS system, $R_{GP}(T^L) - c_{GP}(T^L)$ is an important financial incentive which also determines the GPs equilibrium strategies. As $R_{GP}(T^L) - c_{GP}(T^L)$ increases, income becomes more important to the GP relative to the patient's utility. As a result, GPs will be more inclined to treat their patients and less inclined to refer them to specialists. Graphically, this increase in $R_{GP}(T^L) - c_{GP}(T^L)$ shifts curves $\widetilde{a}_2$ and $\widetilde{a}_3$ to the right. Conversely, as $R_{GP}(T^L) - c_{GP}(T^L)$ decreases, GPs will be less inclined to treat their patients and more inclined to refer them to specialists. Graphically, this decrease in $R_{GP}(T^L) - c_{GP}(T^L)$ shifts curves $\widetilde{a}_2$ and $\widetilde{a}_3$ to the left.

Finally, notice that very selfish GPs will not be willing to follow their diagnostic signal even in the presence of a perfect diagnostic ability (i.e., $a = 1$). This is simply because systematically treating with T^L is income maximizing. On the other hand, very altruistic GPs will not all systematically refer their patients to specialty care as the ones with a relatively high level of diagnostic ability will follow their diagnostic signal. This is simply because in our model a GP with $\beta = 1$ weighs equally the utility from his income and his patient's utility from health.

Comparing the FFS equilibrium to the First Best:

Scenario I: $pA > (1-p)B$

In Figure B, we provide the FFS equilibrium strategies derived above and the First-Best treatment and referral decisions for the scenario where it is worse, in expected utility terms, to send a patient who suffers from a relatively mild illness (M^L) to the specialist

than it is to fail sending a patient who suffers from a relatively severe illness (M^H) to the specialist (or equivalently, it is worse in expected utility terms to make a Type I error than a Type II error). Notice that the FFS equilibrium strategies coincide with the First Best in two regions. In the first region, when the GP is relatively selfish and has a generally low diagnostic ability (i.e., $a < \min\{a^*, \widetilde{a}_3\}$), both the FFS and First-Best strategies are to systematically treat the patient with T^L. The FFS equilibrium coincides with the First Best in this region simply because: (i) very selfish GPs will wish to hoard all patients (i.e., always treat with T^L) as this maximizes their income, and (ii) GPs with relatively low diagnostic ability should always treat with T^L irrespective of their diagnostic signal. Furthermore, this region becomes bigger as the financial reward of treatment $R_{GP}(T^L) - c_{GP}(T^L)$ increases. In the second region, when the GP's altruism and diagnostic ability are both relatively high (or, equivalently, $a > \max\{a^*, \widetilde{a}_2, \widetilde{a}_3\}$), both the FFS and First-Best strategies are to follow the diagnostic signal. The FFS equilibrium coincides with the First Best in this region simply because: (i) relatively altruistic GPs with a relatively high diagnostic ability can trust their diagnostic signal, and (ii) GPs with relatively precise diagnostic ability should always follow their diagnostic signal. However, this region may become smaller as the financial reward of treatment $R_{GP}(T^L) - c_{GP}(T^L)$ increases.

It is also interesting to consider the regions where the FFS outcomes do not coincide with the First Best. In the area labelled by (-), the GP's First-Best strategy is to always follow his diagnostic signal. However, in this region the FFS GP will always treat with T^L as he is relatively selfish. Thus, such FFS GPs under-refer their patients to specialty care. Furthermore, this area becomes bigger as $R_{GP}(T^L) - c_{GP}(T^L)$ increases. In the area labelled by (+) and where $a < a^*$, the GP's First-Best strategy is to always treat the patient with T^L; however, the FFS GP will always follow his diagnostic signal. Thus, in this area, the GP will over-refer his patients to specialty care. In the area labelled by (+)

and where $a > a^*$, the GP's First-Best strategy is to follow his diagnostic signal whereas such a FFS GP will always refer his patients to specialty care. Thus, in this area, the GP will over-refer his patients to specialty care. Finally, in the area labelled by (++), the GP's First-Best strategy is to always treat with T^L (because his diagnostic ability is imprecise), while the FFS GP will always refer the patient to the specialist (because his diagnostic ability is imprecise and he is relatively altruistic). Thus, in this area, GPs will greatly over-refer their patients to the specialist. However, the areas labelled (+) and $a > a^*$ or (++) become smaller as $R_{GP}(T^L) - c_{GP}(T^L)$ increases or may even vanish if $R_{GP}(T^L) - c_{GP}(T^L)$ is large enough.

[Insert Figure B Here]

Scenario II: $pA < (1-p)B$

In Figure C, we provide a similar analysis as in Figure B but for the scenario where it is better, in expected utility terms, to send a patient who suffers from a relatively mild illness (M^L) to the specialist than to treat with T^L a patient who suffers from a relatively severe illness (M^H) (or equivalently, it is worse, in expected utility terms, to make a Type II error than a Type I error). Notice that the FFS and the First-Best outcomes coincide in two different regions. As in Scenario I, the FFS GP with a relatively high altruistic parameter and a relatively high diagnostic ability (i.e., $a > \max\{a^*, \widetilde{a}_2, \widetilde{a}_3\}$) will follow his diagnostic signal, and this strategy coincides with the First Best. However, again, this area may become smaller as $R_{GP}(T^L) - c_{GP}(T^L)$ increases. In the second region, where the GP is relatively altruistic and has a relatively low diagnostic ability (i.e., $a < \min\{a^*, \widetilde{a}_2\}$), the FFS GP will always refer the patient to specialty care. This coincides with the First Best as GPs who have relatively low diagnostic ability should systematically refer patients to specialty care (because it is worse from an expected utility standpoint to make a Type II error rather than a Type I error). However, again, this area becomes

smaller as $R_{GP}(T^L) - c_{GP}(T^L)$ increases or may even disappear if $R_{GP}(T^L) - c_{GP}(T^L)$ is large enough.

Under Scenario II, compairing the FFS equilibrium strategies to the First-Best treatment and referral decisions also provides the regions where the outcomes do not coincide. In the two areas labelled (-) the FFS GP under-refers the patient to specialty care, whereas he greatly under-refers in the area labelled (- -) and over-refers in the area labelled (+).

[Insert Figure C Here]

Recall that under Scenario I, it is worse to make a Type I error than it is to make a Type II error. As such, over referrals should especially be avoided where the over referrals are associated with relatively unable yet relatively altruistic GPs. Furthermore recall that under Scenario II, it is worse to make a Type II error than it is to make a Type I error. As such, insufficient referrals should especially be avoided where under referrals are associated with relatively selfish GPs.

4.2 Capitation

In this section, we turn our attention to deriving the GP's expected utility maximizing treatment and referral decisions assuming that he is paid by capitation. In such a system, the GP receives a fixed payment K_{GP} for each patient he treats without any marginal reimbursement. We continue to assume that the specialist behaves non-strategically.

We now solve for the GP's expected utility maximizing treatment and referral decisions as a function of his diagnostic signals, his diagnostic ability as well as his level of altruism.

The GP's expected utility maximizing decisions as a function of the diagnostic signal:

The GP will always refer the patient to specialty care irrespective of the diagnostic signal because $K_{GP} > K_{GP} - c_{GP}(T^L)$. Also, notice that the GP's strategy is invariant

to his patient's true illness, his diagnostic signal and his level of altruism. This is simply because (i) fully insured patients always weakly prefer to be sent to the specialist and (ii) sending the patient to the specialist is both income and utility (through the altruism effect) maximizing. Thus, even a selfish GP treats and refers in a manner which is in line with the patient's preferences. Furthermore, recall that a patient who suffers from M^L (M^H), will receive the appropriate treatment T^L (T^H) from the specialist.

Comparing the capitation equilibrium to the First Best:

We now compare the GP's expected utility maximizing treatment and referral decisions under capitation to those derived in the First Best. As before, we must distinguish between the two different scenarios.

Scenario I: $pA > (1-p)B$

Recall that under this scenario, GPs with a relatively low diagnostic ability (i.e., $a < a^*$) should in the First Best systematically treat their patients with T^L (because it is worse from an expected utility standpoint to make a Type I error rather than a Type II error), whereas GPs with a relatively high diagnostic ability (i.e., $a > a^*$) should follow their diagnostic signal. Under capitation, GPs will systematically refer their patients to the specialist. Thus, under this scenario, the GP's strategy never coincides with the First Best. More specifically, capitated GPs with high diagnostic ability ($a > a^*$) will over-refer their patients to specialty care, while capitated GPs with low diagnostic ability ($a < a^*$) will greatly over-refer their patients to specialty care.

Scenario II: $pA < (1-p)B$

Recall that under this scenario, a GP with a relatively low diagnostic ability (i.e., $a < a^*$) should in the First Best systematically refer their patients to specialty care (because it is worse from an expected utility standpoint to make a Type II error rather than a Type I error), whereas GPs with a relatively high diagnostic ability (i.e., $a > a^*$) should follow

their diagnostic signal. Under capitation, GPs will systematically refer their patients to the specialist. Thus, under this scenario, the strategy of GPs with a relatively low diagnostic ability ($a < a^*$) will coincide with the First Best, whereas GPs with a relatively high diagnostic ability ($a < a^*$) will over-refer their patients to specialty care.

In general, capitation payment systems provide too much incentive for GPs to refer patients to specialty care. Nonetheless, this tendency to over-refer is particularly important (and should be avoided) when it is worse to make a Type I error than a Type II error, i.e., when concerns of wasteful referrals dominate concerns of under referrals.

4.3 Fundholding

In this section we examine a more comprehensive form of the capitation payment system, sometimes known as a fundholding system, whereby the GP receives a fixed-payment K_{GP} for enlisting a patient into his practice but is then responsible for providing the patient with care "as needed" while also paying for all other care consumed by the patient, including specialty care, without any marginal reimbursement.[19]

We next solve for the GP's expected utility maximizing treatment and referral decisions as a function of his private signals, his diagnostic ability as well as his level of altruism.

The GP's expected utility maximizing decisions as a function of the diagnostic signal:

(i) S^L :

If the GP receives a diagnostic signal S^L, it can be shown that the GP will follow his diagnostic signal and treat with T^L rather than refer the patient to specialty care if his

[19] Fundholding refers to a payment scheme that existed in the United Kingdom in which practices received a budget for each patient they enlisted that covered non-emergency elective surgical procedures, all laboratory tests and out-patient visits, drugs prescribed by the practice, and staff costs (Matsagaris and Glennerster (1994)).

diagnostic ability

$$a > \frac{(1-p)B'}{pA + (1-p)B'} \equiv \widehat{a}_2.$$

where $B' = \beta(n-m) - \left\{c_{SP}(T^H) - \left[c_{GP}\left(T^L\right) + \delta c_{SP}(T^H)\right]\right\}$.[20]

(ii) S^H :

If the GP receives a diagnostic signal S^H, it can be shown that the GP will refer the patient to specialty care (i.e., follow his diagnostic signal in this case) rather than treat with T^L if his diagnostic ability [21,22]

$$a > \frac{pA}{pA + (1-p)B'} \equiv \widehat{a}_3.$$

Using the above results and the properties of $\widehat{a}_2$ and $\widehat{a}_3$, we summarize the GP's expected utility maximizing strategies in Figure D.

[Insert Figure D Here]

When the GP receives a diagnostic signal S, he will systematically treat the patient if $\widehat{a}_2 < a < \widehat{a}_3$, he will follow his diagnostic signal if $a > \max\{\widehat{a}_2, \widehat{a}_3\}$, and he will systematically refer the patient if $\widehat{a}_3 < a < \widehat{a}_2$. Similarities with the FFS system appear and are discussed below.

Notice that, under the fundholding system, the expected cost difference $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$ is an important financial incentive for the GP which also determines the GPs equilibrium strategies. It represents the cost difference between having the patient referred to the specialist and treated with T^H, and having the patient treated by

[20] Note that $\widehat{a}_2$ is increasing in β and concave where $\widehat{a}_2 < 1$ for all values of β and $\widehat{a}_2 = 1/2$ when $\beta = \frac{c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]}{(1-p)(n-m)} \equiv \widehat{\beta}$.

[21] Note that $\widehat{a}_3$ is decreasing in β and convex where $\widehat{a}_3 < 1$ iff $\beta > \frac{c_{SP}(T^H) - \left[c_{GP}(T^L) + \delta\rho c_{SP}(T^H)\right]}{n-m}$, and $\widehat{a}_3 = \frac{1}{2}$ when $\beta = \frac{c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]}{(1-p)(n-m)} \equiv \widehat{\beta}$.

[22] Note that as in the Capitation case, the GP's utility maximizing behavior is independent of the up-front payment.

the GP with T^L and then by the specialist with T^H. As $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$ increases, the cost of referring becomes relatively more important for the GP than the expected cost of treating. As a result, GPs will be more inclined to treat their patients (and thus pay the expected cost $\left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$) than to refer their patients (and thus pay $c_{SP}(T^H)$). Graphically, this increase in $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$ shifts both curves $\widehat{a}_2$ and $\widehat{a}_3$ to the right. This is similar to what we obtain in the FFS case, although the financial incentive is of a different nature. Conversely, as $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$ decreases, GPs will be less inclined to treat their patients and more inclined to refer them to specialists. Graphically, this decrease in $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$ shifts both curves $\widehat{a}_2$ and $\widehat{a}_3$ to the left. Again, this is similar to what we obtain in the FFS case. Finally, if $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$ becomes negative (for example, if c_{GP} becomes very high and/or the difference between c_{SP} and c_{GP} is non-significant), GPs have very little financial incentive to treat their patients. Consequently, the financial incentive added to the GP's altruism will lead to a case where the GPs equilibrium strategies are very similar to those of the capitation system. Notice, however, that no incentive could give rise to such a case under the FFS system.

Not surprizingly, if the FFS financial incentives ($R_{GP} - c_{GP}$) are similar to the fundholding financial incentives ($c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right]$), both payment systems lead to the same GPs equilibrium strategies. As a result, Figures A and D coincide most of the time. However, if the FFS financial incentives are much greater (smaller) than the fundholding financial incentives, then the curves $\widetilde{a}_2$ and $\widetilde{a}_3$ which summarize the GPs equilibrium strategies under the FFS system will be more to the right (left) than the curves $\widehat{a}_2$ and $\widehat{a}_3$ which summarize the GPs equilibrium strategies under the fundholding system.

Finally, recall that $\widehat{\beta}$ is the value of β such that $\widehat{a}_2 = \widehat{a}_3 = 1/2$, i.e., the value of β where the curves $\widehat{a}_2$ and $\widehat{a}_3$ intersect in Figure D. The fundholding equilibrium strategies

can also be characterized with respect to the value of $\widehat{\beta}$. It can easily be shown that $\widehat{\beta} \gtrless 0$ iff $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right] \gtrless 0$. Furthermore, if $\widehat{\beta}$ is positive, it can also be shown that $\widehat{\beta} \gtrless 1$ iff $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right] \gtrless (1-p)(n-m)$. As a result, we must consider the cases where $\widehat{a}_2$ or $\widehat{a}_3$ may not figure in the $(\beta, a) - space$ equal to $[0,1] \times [1/2, 1]$ (i.e., where no physician will always treat or where no physician will always refer).

Comparing the fundholding equilibrium to the First Best:

Scenario I: $pA > (1-p)B$

Under Scenario I, $pA > (1-p)B$, which is equivalent to the following condition: $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right] > (1-p)(n-m)$. That is, the cost concern associated with wasteful referrals dominates the patient's expected utility concern, or equivalently, it is worse to make a Type I error than a Type II error. The scenario I condition is more likely to be satisfied when the expected cost difference is high, i.e., when c_{SP} is relatively large and/or the probability of a relatively severe illness (M^H) is low. This condition holds when comparing the fundholding equilibrium strategies to the First Best in Figure E. Accordingly, under Scenario I, the only relevant equilibrium strategies are those where no GP will ever wish to always refer (i.e., $\hat{a}_2$ does not figure in the $(\beta, a) - space$ equal to $[0,1] \times [1/2, 1]$.

The fundholding equilibrium strategies coincide with the First Best in two regions. In the first region, when the GP has a generally low diagnostic ability (i.e., $a < a^*$), both the fundholding and First-Best strategies are to systematically treat the patient with T^L. In the second region, when the GP's altruism and diagnostic ability are both relatively high (or equivalently, $a > \widehat{a}_3$), both the fundholding and the First-Best strategies are to follow the diagnostic signal. Finally, the fundholding strategies do not coincide with the First Best in the remaining region labelled by (-). In this area, the fundholding GPs (who

are both relatively selfish and have a relatively high diagnostic ability) will wish to always treat the patient with T^L while the First-Best strategy is to follow the diagnostic signal. Thus, such fundholding GPs under-refer their patients to specialty care.

Further notice that under Scenario I, $\widehat{a}_3$ actually reaches a^* when $\beta = 1$. Thus, fully altruistic physicians (i.e., as we define them, GPs who equally weigh their income to their patients' health) will always behave in the First-Best manner, irrespective of their level of diagnostic ability.

[Insert Figure E Here]

Scenario II:

Under Scenario II, $pA < (1-p)B$, which is equivalent to the following condition: $c_{SP}(T^H) - \left[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\right] < (1-p)(n-m)$. That is, the patient's expected utility concern (associated with the gross benefit of being appropriately treated by a specialist when suffering from a relatively severe illness (M^H)) dominates the cost concern, or equivalently, it is worse to make a Type II error than a Type I error. The scenario II condition is more likely to be satisfied when the expected cost difference is small, i.e., when c_{SP} is relatively small and/or the likelihood of a relatively severe illness is high. This condition holds when comparing the fundholding equilibrium strategies to the First Best in Figure F. Accordingly, under scenario II, $\hat{a}_2$ will always figure in the relevant $(\beta, a) - space$ (i.e., there are always some GPs who will wish to always refer). However, comparing the fundholding equilibrium strategies to the First Best when both $\widehat{a}_2$ and $\hat{a}_3$ figure in the relevant $(\beta, a) - space$ gives very similar results to those obtained when comparing the FFS equilibrium strategies to the First Best. As a result, we will focus on the case where $\hat{a}_3$ does not figure in the relevant $(\beta, a) - space$ (i.e., where no physician will wish to always treat).

This particular case corresponds to the situation where the difference between GPs' and

specialists' costs is non-significant. In that case, the First-Best threshold value of the GP's ability a^* tends to be equal to 1. Furthermore, Figure F clearly shows that the fundholding strategies coincide with the First-Best in all but one region, i.e., the area labelled (-), and they completely differ from the FFS strategies.

[Insert Figure F Here]

To summarize, fundholding payment systems often provide disincentive to refer by having GPs pay for their patients' expenses for specialty care. In particular, under Scenario II, where it is worse to make a Type II error than it is to make a Type I error. As such, insufficient referrals should be avoided where under referrals are associated with relatively selfish GPs. However, the opposite would occur in some specific cases, namely when the difference between c_{SP} and c_{GP} becomes non-significant.

5 Comparing across the different payment mechanisms

In this section we turn our attention to comparing the three payment mechanisms examined above. More specifically, we examine for each type of GP (i.e., for each altruism/diagnostic-ability pair (β, a)) which payment system provides the "best" incentives. In order to do so, we present the results from our analysis for each payment mechanism on a single figure under both scenarios. More specifically, in Figures G and H, we rank (using the $\succ$ symbol) the three payment mechanisms for all possible altruism/diagnostic-ability pairs under Scenarios I and II, respectively. Furthermore, we index the payment mechanism with a $*$ symbol whenever its outcome coincides with the First Best.

By examining Figures G and H, one can see quite clearly that no payment mechanism consistently outperforms any other. More specifically, no payment mechanism weakly dominates all others for all potential joint distributions of GP types (i.e., for all joint distributions of β and a). It is nonetheless interesting to note that some general patterns

do emerge.

Under Scenario I, recall that the cost concern associated with wasteful referrals dominates the patient's expected utility concern and that this scenario is more likely to occur when the specialist costs c_{sp} are relatively high and/or the probability of a relatively severe illness (M_H) is low. Therefore, in order to avoid unnecessary referrals, the First-Best strategy for GPs with low diagnostic ability ($a < a^*$) is to systematically treat their patients whereas GPs with a relatively high diagnostic ability ($a > a^*$) should follow their diagnostic signal. Under Scenario I (Figure G), the capitation system is (at least) weakly dominated by both the FFS and the fundholding systems for all potential joint distributions of GP types. This is simply because the capitation system provides incentives for GPs to systematically refer their patients. As a result, we can limit our analysis to comparing the FFS to the fundholding case.

One interesting result is that for GPs with low levels of diagnostic ability (i.e., $a < a^*$), the fundholding system weakly dominates the FFS system irrespective of the GP's level of altruism. Thus, in this case, the fundholding system provides the 'best' incentives for GPs and achieves a First-Best outcome for all GP types. This is so because, under Scenario I where costs associated with wasteful referrals should be avoided, the fundholding system eliminates the incentive to refer patients to specialty care by having the GP pay for the patient's specialty-care expenses; whereas the FFS system provides sufficiently altruistic GPs with financial incentives to either follow their diagnostic signal or systematically refer their patients to specialty care. Thus, the fundholding system can eliminate the perverse incentive to over-refer associated with relatively high altruism in the presence of low diagnostic ability.

When the joint distribution of GPs includes only individuals who have a relatively high diagnostic ability (i.e., $a > a^*$), the FFS and fundholding systems both lead to different

treatment and referral decisions and provide different incentives. More specifically, the FFS system tends to dominate the fundholding system. For instance, for GPs who have intermediate levels of altruism (i.e., $\max\{\widetilde{a}_3, a^*, \widetilde{a}_2\} < a < \widehat{a}_3$), the fundholding system provides GPs with a financial disincentive to refer their patients to specialty care (in order to avoid the costs associated with referrals) even though their diagnostic signal indicates that they should in the First Best; whereas the FFS system provides GPs with the financial incentive to always follow their diagnostic signal (which is the First-Best strategy). Thus, for those GPs, the FFS system stricly dominates the fundholding system.

Finally, it is interesting to note that as the FFS rewards of treatment $R_{GP}(T^L) - c_{GP}(T^L)$ increases, the differences between the financial incentives provided by the FFS and the fundholding systems tend to reduce. Consequently, the differences in the results obtained above, either for GPs with low diagnostic ability ($a < a^*$) or for GPs with high diagnostic ability ($a > a^*$), will also tend to be reduced.

To sum up, under Scenario I, the FFS system generates the First-Best outcome for GPs who are able and have intermediate levels of altruism. Conversely, the fundholding system gives outcomes which are closer to the First Best for all other types of GPs.

[Insert Figure G Here]

Under Scenario II, recall that the patient's expected utility concern dominates the cost concern and that this scenario is more likely to occur when the expected cost difference (between specialists' and GPs' costs) is low and/or the likelihood of a relatively severe illness is high. Therefore, in order to avoid insufficient referrals, the First-Best strategy for GPs with low diagnostic ability ($a < a^*$) is to systematically refer their patients whereas GPs with high diagnostic ability ($a > a^*$) should follow their diagnostic signal.

Under Scenario II, the FFS and the fundholding systems induce very similar behaviors and are therefore treated in the same way in Figure H. Here, no physician payment mech-

anism weakly dominates any other. It is interesting to note, however, that for GPs with relatively low diagnostic ability ($a < a^*$), the capitation payment system weakly dominates both the FFS and the fundholding systems as the capitation strategy coincides with the First Best. This is simply because under Scenario II, GPs with a low diagnostic ability should always refer their patients to specialty care and the capitation payment system provides them with the financial incentives to do so.

For physicians with a relatively high level of diagnostic ability ($a > a^*$), the payment mechanism which leads to the "best" treatment and referral decisions depends on the GPs level of altruism. For relatively low levels of altruism ($a^* < a < \widetilde{a}_3(\hat{a}_3)$), the capitation payment system does "best" as it encourages GPs to systematically refer their patients while both the fundholding and the FFS systems encourage GPs to systematically treat their patients (recall that under Scenario II it is better to systematically refer the patient to specialty care than to systematically treat her with T^L). For GPs who have a relatively higher level of altruism (i.e., $a > \max\{\widetilde{a}_3(\hat{a}_3), a^*, \widetilde{a}_2(\hat{a}_2)\}$), the FFS and the fundholding payment systems dominate the capitation system, as they provide the financial incentives for GPs to follow their diagnostic signal.

Last, FFS tends to induce more (less) referrals than the fundholding system if the FFS margin $R_{GP}(T^L) - c_{GP}(T^L)$ is low(high) enough compared to the expected difference in costs between a treatment by the specialist and the GP $\big(c_{SP}(T^H) - \big[c_{GP}(T^L) + (1-p)\delta c_{SP}(T^H)\big]\big)$. However, as this expected cost difference decreases or even becomes negative (i.e., where very few or even no GP will wish to always treat), the fundholding financial incentives will induce GP's strategies which are similar to those induced by the capitation system. As a result, in this case, the fundholding payment system leads to the First best for most GPs with low diagnostic signal ($a < a^*$) and for all GPs with high diagnostic signal ($a > a^*$).

To sum up, under Scenario II, both FFS and fundholding generate the First-Best be-

havior for GPs who are both able and altruistic enough. On the contrary, the capitation system is closer to the First Best if the majority of GPs are rather selfish and/or unable.

[Insert Figure H Here]

Although no general conclusions can be drawn with respect to which payment mechanism provides the "best" financial incentives for GPs, each one is weakly preferred under particular situations. First, when it is better in expected utility terms to systematically refer patients to specialty care than to systematically treat them (i.e., under Scenario II), the capitation payment system provides the "best" incentives when GPs have relatively low levels of diagnostic ability or have relatively low levels of altruism. When it is better to systematically treat a patient than to systematically refer her (i.e., under Scenario I), the FFS system does "best" when GPs have relatively high levels of diagnostic precision and intermediate levels of altruism. Because such GPs care sufficiently about their patients and their income, they not only wish to rely on their diagnosis for financial reasons but can for altruistic ones. Finally, still under Scenario I, the fundholding system provides the best incentives for GPs when they have low levels of diagnostic ability. This is simply because the fundholding system encourages them to refer less than the other two payment systems. Furthermore, when GPs are very altruistic but have relatively high levels of diagnostic ability, the fundholding system does best as it eliminates the perverse incentives to over-refer when GPs are very altruistic.

6 Conclusion

This paper analyzes and compares the incentive properties of some commonly used payment mechanisms for GPs, namely FFS, capitation and fundholding. It brings important nuances to the incentives already identified in the literature by focusing on gatekeeping GPs, by explicitly recognizing that GPs are heterogeneous in both ability and altruism, and by

introducing dynamics in the health production function.

We conclude that both FFS and fundholding usually result in less referrals to costly specialty care than does capitation. This result contrasts with the cost-containment property that is usually associated with capitation in the literature. The rationale behind our result is that GPs under capitation are better off referring all patients to specialty care for the sake of both economizing on their own treatment expenses and improving their patients' health. However, whenever the cost of having the patient referred to the specialist is low enough relative to the expected costs of GPs care, then fundholding may induce almost as much referrals as capitation, because GPs have little financial incentives to treat their patients directly.

We also show that, although driven by financial incentives of different nature, the GP's strategic outcomes associated with fundholding and those of FFS can be, surprisingly, very much alike. More specifically, only very altruistic yet not very able GPs will refer all their patients to specialty care. GPs who are relatively altruistic and very able will decide to either treat or refer according to their diagnostic ability. Last, the very selfish yet able GPs will directly treat all their patients to either maximize their earnings under FFS or minimize their expected expenses under fundholding.

Finally, whether a regulator should use one or another payment mechanism for GPs depends on (i) his priorities which, in turn, depend on the expected cost difference between GPs care and specialty care, and (ii) the distribution of profiles among GPs. For instance, in the case where specialists costs are relatively high (which is often argued as the raison d'être of gatekeeping), our results suggest the following. If most GPs were very able and relatively altruistic, the FFS system would be the most appropriate payment mechanism. Conversely, if most GPs were less able yet altruistic, the fundholding system would give the most appropriate incentives. However, an empirical evaluation of the distribution of

GPs profiles is definitely necessary before providing reliable policy recommendations.

References

Allard, M., I. Jelovac, and P.-T. Léger (2010). Physicians' self selection of a payment mechanism: Capitation versus fee for service. *GATE LSE Working paper 1024*.

Allard, M., P.-T. Léger, and L. Rochaix (2009). Provider competition under a dynamic setting. *Journal of Economics and Management Strategy 18*, 457–486.

Biglaizer, G. and A. Ma (2007). Moonlighting: public service and private practice. *RAND Journal of Economics 38*, 1113–1133.

Blomqvist, A. (1991). The doctor as double agent: Information asymmetry, health insurance, and medical care. *Journal of Health Economics 10*, 411–422.

Blomqvist, A. and P.-T. Léger (2005). Information asymmetry, insurance and the decision to hospitalize. *Journal of Health Economics 24*, 775–793.

Brekke, K., R. Nuscheler, and O. Straume (2007). Gatekeeping in health care. *Journal of Health Economics 26*, 149–170.

Carey, T. (1995). The outcomes and costs of care for acute low back pain among patients seen by primary care practitioners, chiropractors, and orthopedic surgeons. *New England Journal of Medicine 333*, 913–917.

Chalkley, M. and J. Malcomson (1998). Contracting for healthservices when patient demand does not reflect quality. *Journal of Health Economics 17*, 1–19.

Choné, P. and A. Ma (2011). Optimal health care contracts under physician agency. *Annales d'Économie et de Statistique, forthcoming*.

Deflgaauw, J. (2007). Dedicated doctors: public and private provision of health care

with altruistic physicians. *Tinbergen Institute Discussion Paper 07-010/1*.

Garcia-Marinoso, B. and I. Jelovac (2003). GPs' payment contracts and their referral practice. *Journal of Health Economics 22*, 617–635.

Gonzalez, P. (2010). Gatekeeping versus direct access when patient information matters. *Health Economics 19*, 730–754.

Greenfield, S. (1992). Variations in resource utilization among specialties and systems of care: Results from the medical outcomes study. *Journal of the American Medical Association 267*, 1624–1630.

Grembowski, D., K. Cook, D. Patrick, and A. Roussel (1998). Managed care and physician referral. *Medical Care Research and Review 55*, 3–31.

Jack, W. (2005). Purchasing health care services from providers of unknown altruism. *Journal of Health Economics 24*, 73–93.

Levaggi, R. and L. Rochaix (2007). Exit, choice or loyalty: Patient driven competition in primary care. *Annals of Public and Cooperative Economics 78*, 501–535.

Ma, C. and T. G. McGuire (1997). Optimal health insurance and provider payment. *American Economic Review 87*, 685–704.

Malcomson, J. (2004). Health service gatekeepers. *RAND Journal of Economics 35*, 401–421.

Matsagaris, M. and H. Glennerster (1994). The threat of "cream skimming" in the post-reform nhs. *Journal of Health Economics 13*, 31–60.

7 Appendices

Appendix 1: Proof of Lemma 1

We assume the GPs are paid exactly the cost of treatment (i.e., we assume a perfectly competitive environment where $R_{GP}(T^L) = c_{GP}(T^L)$. We then solve for the patient's expected utility maximizing treatment and referral decisions given a diagnostic signal S^L.

A GP who observes a diagnostic signal S^L (given a diagnostic ability a) can treat the patient with T^L or refer the patient to the specialist. We next calculate the patient's expected utility under these two strategies.[23]

The GP treats the patient with T^L

(a) With probability $\Pr(M^L|S^L)$, the patient suffers from M^L and is treated by the GP with T^L. In this case the patient's utility is:

$$u(h - M^L + T^L) + I - c^L_{GP} + \delta(u(h - M^L + T^L) + I).$$

(b) With probability $\Pr(M^H|S^L)$, the patient suffers from M^H and is treated by the GP with T^L. In this case the patient's expected utility is:

$$u(h - M^H + T^L) + I - c^L_{GP} + \delta(u(h - M^H - L + T^H) + I - c^H_{SP}).$$

Thus, the patient's expected utility is:

$$\begin{aligned} &\frac{pa}{pa + (1-p)(1-a)} \left\{u(h - M^L + T^L) + I - c^L_{GP} + \delta(u(h - M^L + T^L) + I)\right\} + \\ &\frac{(1-p)(1-a)}{pa + (1-p)(1-a)} \{u(h - M^H + T^L) + I - c^L_{GP} + \\ &\delta(u(h - M^H - L + T^H) + I - c^H_{SP})\}. \end{aligned} \tag{A1.1}$$

The GP refers the patient to the specialist:

(a) With probability $\Pr(M^L|S^L)$, the patient suffers from M^L and is treated by the

[23] Including the costs of treatment directly into the maximization problem is equivalent to maximizing the patient's expected utility subject to an actuarially fair insurance constraint.

specialist with T^L. In this case the patient's utility is:

$$u(h - M^L + T^L) + I - c_{SP}^L + \delta(u(h - M^L + T^L) + I).$$

(b) With probability $\Pr(M^H|S^L)$, the patient suffers from M^H and is treated by the specialist with T^H. In this case the patient's utility is:

$$u(h - M^H + T^H) + I - c_{SP}^H + \delta(u(h - M^H + T^H) + I).$$

Thus, the patient's expected utility is:

$$\begin{aligned}&\frac{pa}{pa + (1-p)(1-a)} \left\{u(h - M^L + T^L) + I - c_{SP}^L + \delta(u(h - M^L + T^L) + I)\right\} + \\ &\frac{(1-p)(1-a)}{pa + (1-p)(1-a)}\{u(h - M^H + T^H) + I - c_{GP}^H + \\ &\delta(u(h - M^H + T^H) + I)\}.\end{aligned} \tag{A1.2}$$

Thus, a GP who observes a diagnostic signal S^H should treat the patient with T^H (should refer the patient to the specialist) iff (A1.1) $\geq$ ($<$) (A1.2) or:

$$a \geq (<)\frac{(1-p)B}{pA + (1-p)B} \blacksquare$$

Appendix 2: Proof of Lemma 2

We derive the first-best treatment and referral strategies given a diagnostic signal S^H.

A GP who observes a diagnostic signal S^H (given a diagnostic ability a) can treat the patient with T^L or refer the patient to the specialist. We next calculate the patient's expected utility under these two strategies.

The GP treats the patient with T^L:

(a) With probability $\Pr(M^L|S^H)$, the patient suffers from M^L and is treated by the GP with T^L. In this case the patient's utility is:

$$u(h - M^L + T^L) + I - c^L_{GP} + \delta(u(h - M^L + T^L) + I).$$

(b) With probability $\Pr(M^H|S^H)$, the patient suffers from M^H and is treated by the GP with T^L. In this case the patient's expected utility is:

$$u(h - M^H + T^L) + I - c^L_{GP} + \delta(u(h - M^H - L + T^H) + I - c^H_{SP}).$$

Thus, the patient's expected utility is:

$$\begin{aligned}&\frac{p(1-a)}{p(1-a)+(1-p)a}\left\{u(h - M^L + T^L) + I - c^L_{GP} + \delta(u(h - M^L + T^L) + I)\right\} + \\ &\frac{(1-p)a}{p(1-a)+(1-p)a}\{u(h - M^H + T^L) + I - c^L_{GP} + \\ &\delta(u(h - M^H - L + T^H) + I - c^H_{SP})\}.\end{aligned} \tag{A2.1}$$

The GP refers the patient to the specialist:

(a) With probability $\Pr(M^L|S^H)$, the patient suffers from M^L and is treated by the specialist with T^L. In this case, the patient's utility is:

$$u(h - M^L + T^L) + I - c^L_{SP} + \delta(u(h - M^L + T^L) + I).$$

(b) With probability $\Pr(M^H|S^H)$, the patient suffers from M^H and is treated by the

specialist with T^H. In this case, the patient's utility is:

$$u(h - M^H + T^H) + I - c_{SP}^H + \delta(u(h - M^H + T^H) + I).$$

Thus, the patient's expected utility is:

$$\begin{aligned} &\frac{p(1-a)}{p(1-a)+(1-p)a}\left\{u(h - M^L + T^L) + I - c_{SP}^L + \delta(u(h - M^L + T^L) + I)\right\} + \\ &\frac{(1-p)a}{p(1-a)+(1-p)a}\{u(h - M^H + T^H) + I - c_{SP}^H + \\ &\delta(u(h - M^H + T^H) + I)\}. \end{aligned} \tag{A2.2}$$

Thus, a GP who observes a diagnostic signal S^H should refer the patient to the specialist (should treat the patient with T^L) iff (A2.2) $\geq$ ($<$) (A2.1) or:

$$a \geq (<)\frac{pA}{pA + (1-p)B}\blacksquare$$

Appendix 3: Proof of Proposition 1

Recall from Lemma 1, that a GP who receives a diagnostic signal S^L should refer the patient to specialty care if $a \in \left[\frac{1}{2}, \frac{(1-p)B}{pA+(1-p)B}\right]$ and should treat the patient with T^L when $a \in \left[\frac{(1-p)B}{pA+(1-p)B}, 1\right]$. However, $pA > (1-p)B$ implies that $\frac{(1-p)B}{pA+(1-p)B} < \frac{1}{2}$. Thus, the GP should always treat the patient with T^L when he receives a diagnostic signal S^L (i.e., his diagnostic ability is irrelevant).

Further recall from Lemma 2, that a GP who receives a diagnostic signal S^H should always treat his patient with T^L when $a \in \left[\frac{1}{2}, \frac{pA}{pA+(1-p)B}\right]$ and should refer the patient to specialty care when $a \in \left[\frac{pA}{pA+(1-p)B}, 1\right]$. Thus, the GP's first-best treatment and referral decisions when he receives a diagnostic signal S^H will depend on his diagnostic ability since $pA > (1-p)B$ implies that $\frac{pA}{pA+(1-p)B} > \frac{1}{2}\blacksquare$

Appendix 4: Proof of Proposition 2

Recall from Lemma 2, that a GP who receives a diagnostic signal S^H should treat his patient with T^L when $a \in \left[\frac{1}{2}, \frac{pA}{pA+(1-p)B}\right]$ and should refer the patient to specialty care when $e \in \left[\frac{p_2 A}{p_2 A+p_3 B}, 1\right]$. However, $pA < (1-p)B$ implies that $\frac{pA}{pA+(1-p)B} < \frac{1}{2}$. Thus, the GP should always refer the patient to the specialist when he receives a diagnostic signal S^H (i.e., his diagnostic ability is irrelevant).

Further recall from Lemma 1, that a GP who receives a diagnostic signal S^L should refer the patient to specialty care if $a \in \left[\frac{1}{2}, \frac{(1-p)B}{pA+(1-p)B}\right]$ and should treat the patient with T^L when $a \in \left[\frac{(1-p)B}{pA+(1-p)B}, 1\right]$. Thus, the physician's first-best treatment and referral decisions when he receives a diagnostic signal S^L will depend on his diagnostic ability since $pA < (1-p)B$ implies that $\frac{(1-p)B}{pA+(1-p)B} > \frac{1}{2}$■