ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 83/84 – 2006

Reforming Incentive Schemes under Political Constraints: The Physician Agency

Gabrielle Demange*, Pierre Yves Geoffard‡

ABSTRACT. – In many developed countries attempts to reform physicians payment schemes have failed. To analyze some of the difficulties, this paper studies reforms of payment schemes in situations such as the physician agency, where the quality of the good produced is imperfectly observable by the payer. We first study the situation, common in many countries, where physicians face a single scheme. We identify conditions under which no reform can both obtain the consent of a large proportion of physicians (political constraints) and improve patients welfare. We then study whether a menu of contracts, with or without cross subsidies, may solve the difficulties generated by the heterogeneity of producers practice.

Modes de rémunération des médecins et pouvoir de blocage

RÉSUMÉ. – Dans de nombreux pays, les réformes visant à modifier les modes de rémunération des prescripteurs de soins se sont heurtées à leur forte opposition. Ce papier analyse ces difficultés, d'abord dans un cadre de monopole puis de concurrence régulée entre différentes organisations. Quelques facteurs clefs sont mis en évidence : ils sont liés au pouvoir politique des médecins, à l'hétérogénéité de leurs caractéristiques et de leurs comportements.

We thank Marie Allard and Lise Rochaix, two anonymous referees, seminar and conferences participants in Bergen (CEPR Conference on Health Economics), Taipei (Academia Sinica), Leuven (Public Economics seminar), Besançon, Paris (LEI & Journée Jourdan), Venezia (17th Congress of the European Economic Association), and Marseille (2nd European Health Economics Workshop) for their comments and suggestions. Financial support from Fédération Française des Sociétés d'Assurance is gratefully acknowledged. Remaining errors are ours.

* EHESS, PSE, 48 Boulevard Jourdan, Paris 14, 75014, France, and CEPR

† PSE, IEMS (University of Lausanne) and CEPR.

‡ geoffard@pse.ens.fr.

1 Introduction

A major challenge faced by health systems is the regulation of medical practice. Quality of medical care is at least partly unobservable, and the output of medical services cannot be contracted upon. This has justified public interventions that prevent the price of medical care, and especially physician payment schemes, to be determined by market forces. However, regulation has also led to a situation in which physicians have a strong political power, in addition to the rents they may hold thanks to their informational advantage. In many developed countries the last twenty years have witnessed various attempts to reform some medical professions, many of which have failed. The aim of this paper is to investigate how the political power combines with the informational rents held by physicians to impose constraints on payment schemes reforms.

The premise of our analysis is that medical practice reacts to monetary incentives, as documented by growing empirical research. Recent evidence points out that general practitioners increase their working time when the payment scheme provides an incentive to do so; see Delattre and Dormont (2000), Croxson et al. (2002), and McGuire (2000). The same holds true for inpatient care: monetary incentives of physicians (surgeons, obstetricians) condition the rate of several surgical procedures; see Gruber and Owings (1996), Holly *et al.* (1998). Thus, payment schemes affect quality of care, overall welfare, and how this overall welfare is distributed among all agents (patients, doctors).

The history of health systems reforms also stresses that such reforms are difficult to implement if they do not obtain the support of health care professionals, and in particular of physicians. Not surprisingly, this is especially true when the proposed reform aims to modify physician payment schemes which, potentially, may affect physicians income and welfare. Kessel (1958) and especially Havighurst (1978) provide a description of physicians historical reluctance to prepaid group practices. More recently and in a European perspective, Hassenteufel (1997) gives a detailed account of the various ways in which physicians have reacted against supply-side cost sharing[1]. This hostility may be due to risk aversion of physicians. Indeed, supply-side cost sharing transfers some of the health risk to physicians: when they face more severe or more complex cases, they may have to spend more time with their patient; if the payment scheme does not compensate for this additional time, physicians partly bear the risk to face a sicker clientele. The political power of physicians was also a key element in the history of the Medicare program in the US (Corning, 1969). A strong opposition by the American Medical Association to the so-called "socialized medicine" delayed the introduction of the federally funded program; eventually, a generous fee for service was set to obtain the agreement of the AMA to the reform. Recently, Swiss doctors organizations have successfully petitioned against the possibility of selective contracting.

In short, no reform is politically feasible without the support of a large share of physicians. Obtaining this support involves transfers of some of the efficiency

1. For example, in 1913, most german physicians went on strike (against the health insurance funds of that time) to obtain a fee-for-service payment scheme as well as freedom for patients to choose their provider.

gains to those who are threatened to lose their rents. Hence, less gains accrue to the rest of society. This paper investigates the following issue. Under which conditions can a reform produce sufficient efficiency gains so that, once physicians are compensated for their diminished rents, patients are still better off?

Our analysis builds on contract theory. Neither the quality of the service provided by a physician, nor his "talent", nor the health status of his patients are observed. Physicians are paid through contracts, or reward schemes, according to some observable input (say, the number of acts or the total time spent with the patient) that influence the (unobservable) quality they offer. Hence, even though physicians are concerned with the quality of the service, a potential conflict between reducing costs and guaranteeing quality may arise.

In the initial situation (the *status quo*) all physicians face a single payment scheme, assumed to be inefficient. To our knowledge, this paper is the first to address, in a theoretical set up, the problems faced by public policies which aim to reform such payment mechanisms.

We first study "uniform" reforms, which aims to modify this unique scheme. Our analysis derives three elements that may limit the possibility of a reform that would increase patients welfare: heterogeneity of physician practice, important political power of physicians, and low elasticity of physician practice time.

We then investigate a situation in which a single insurance fund introduces several contracts. Cross-subsidies can compensate for differences in cost and quality across contracts, and alleviate the constraints associated with pure competition. In such a case, the introduction of contracts in addition to the status quo one may both keep all physicians at least as well off (hence be politically feasible) and improve patients welfare.

Finally, we turn to more drastic reforms, that introduce some form of managed competition among private insurance companies. Such reforms have been recently undertaken in European countries (the Netherlands, Germany, Switzerland). In some cases, insurance funds often do not have to contract with *all* providers of care, hospitals and physicians[2]. Selective contracting goes along with a variety of contractual arrangements between providers of care and insurance funds, including various payment schemes[3]. In our setup, insurance funds have to compete on both sides, to attract physicians (through payment schemes) as well as patients (through premiums). Double-sided competition in contracts under imperfectly observable quality blends elements of adverse selection *à la* Rothschild-Stiglitz on the physician side, and of product differentiation *à la* Hotelling on the patient side. Since patients can infer some information about quality by observing physicians choice of contract, this double-sided competition imposes strong constraints on the strategies, and we show that no competitive Nash equilibrium exists.

Our analysis extends the literature in two aspects. First, in the contract theory literature, political constraints differ from participation constraints in standard principal-agent models (see, e.g. Laffont and Tirole, 1993). Second, within the health economics literature, our analysis is an attempt to study simultaneously the three nodes of the Arrow (1963) "medical triad." Most analysis to date has focused either on the supply side, by investigating the relationship between insurance funds and

2. In the Netherlands, selective contracting was introduced by the Dekker plan; in Switzerland, it is a key feature of the current revision of the Health Insurance Law.
3. In the realm of Managed Care in the U.S., capitation contracts are standard in IPA or network HMOs, whereas physicians are paid on wage in Group/Staff HMOs.

physicians, or on the demand side, by investigating the risk and incentives faced by patients. However, the way supply-side and demand-side regulation interact is not well understood (Blomqvist, 1991; Ma and McGuire, 1997; McGuire, 2000).

Section 2 introduces the model of physician agency in which quality of service (health improvement) is not contractible, but some input (time) is. Section 3 studies optimal contracts as a benchmark for section 4, in which political constraints are introduced, and deviations from first best efficiency are analyzed. Section 5 studies the situation with many contracts, firstly under cross subsidies and lastly in a competitive setting. Section 6 concludes and section 7 gathers all the proofs.

2 The Market for Care and Health Insurance

This section describes the main features of the market for care considered in this paper. It studies how income and incentives to provide quality are affected by the reward scheme faced by a physician. Physicians preferences over alternative payment schemes, which have implications on reforms, are derived.

2.1 The Physician-Patient Relationship

A patient who suffers from an illness episode meets a physician. The outcome of the service, denoted by l, is referred to as the "quality" of care. We follow a standard assumption in the health economics literature devoted to the physician agency (Blomqvist, 1991; McGuire, 2000) by assuming that quality is a non-contractible input in the health production function.

More precisely, quality of care depends on two main elements. First, given the symptoms expressed by the patient, the physician must set a diagnosis. Second, based upon this diagnosis, the physician prescribes some treatment. Quality of care depends on the accuracy of the diagnosis, and on the adequateness of the treatment to clinical guidelines based upon existing medical evidence; both elements are necessary (i.e. complement) for good care. We assume that the first element depends on the total time t spent by the physician with the patient, during one or more visits, with positive and decreasing returns. Ma and McGuire (1997) argue that the quantity of treatment (which includes the time spent with each patient) is costly to observe by the insurer, and reports may not be truthful. However, simple mechanisms may be implement to circumvent this problem. For instance, in Switzerland, physicians payments depend on the time spent with each patient. At the end of each visit, the physician notifies the length of the visit to the patient (who can easily monitor the truthfulness of the report). Thus, we assume that this input variable t is observable by the insurer at no cost[4].

4. If distinctions can be made on other criteria, our argument is valid for each category of physicians sharing the same criteria: in as much as quality is partially observed through a parameter, contracts should be interpreted as being conditional on each value of this parameter.

Quality is also affected by the health shock the patient has been subject to, denoted by θ. A higher θ indicates a more severe (or more complex) case. Clinical examination may be more difficult, and the physician may be less familiar with some diseases and their treatments. This affects both elements of quality, and therefore l is decreasing with θ. Finally, quality of care also depends on an exogenous characteristic $\beta \in B \equiv \left[\underline{\beta}, \overline{\beta}\right]$ that is specific to each physician. We call β the *talent*. Talent may also affect both elements of quality: more talented physicians may reach a correct diagnosis more rapidly; they may also have a better knowledge of clinical guidelines. In both cases, we assume that quality increases with β. In summary, the production function $l(t,\theta,\beta)$ for quality satisfies:

Positive and decreasing returns:	$l_t > 0, l_{tt} < 0$
More severe patients need more time:	$l_\theta < 0$
Increasing quality with β:	$l_\beta > 0$.

The cross effect of input time t and talent β is less clear. When talent is associated with a better knowledge of medical literature, time and talent are complement, since both are needed to obtain a good quality. When talent is associated with the ability to reach an accurate diagnosis, time and talent may rather be substitute. Hence, we allow both possibilities and will study these two cases; however, we shall assume a Spence-Mirrlees condition which ensures that l_t is monotone in talent. Talent and time are either *substitutes* if for all (t,θ,β), $l_{t\beta} < 0$, or *complements* if for all (t,θ,β), $l_{t\beta} > 0$. This will obviously affect how the total time supplied by each physician depends on his own characteristic β.

Patients value quality with linear preferences: λl is the monetary value for quality. Put differently, λ represents the willingness to pay for a marginal increase in quality l. Information about the physician a patient is matched with, as well as about his own health status, is defined below.

Throughout the paper, we shall assume that a physician knows his own characteristic; for short a β-physician denotes a physician with characteristic β. For each patient, the input decision t is taken by a physician after he observes the health status θ. Accordingly, it is a function of the two parameters (θ,β), and of the reward scheme the physician is facing.

A physician receives a monetary payment for each of his patients. This payment is function of the observable variable t only, which is the total time (or the number of acts) spent by this physician with this patient. Hence, a *reward scheme* is specified by a function of t, $R(t)$. We focus the analysis on *linear* reward schemes, given by

$$R(t) = b + at,$$

where b is a flat payment, and a a fee-for-service rate (restricting to linear schemes is justified later on). A capitation (prospective) contract is associated with a flat scheme $a = 0$ and a constant fixed payment per patient $b > 0$, and a purely retrospective scheme with $b = 0$ and a fee for service rate $a > 0$.

Physicians objective functions are a matter of debate within the health economics literature (McGuire, 2000). However, it is generally assumed that, even though physicians care for their monetary income, they also care for quality. Thus we assume that the objective function of a physician of type β, who spends a time t with a patient with a health status θ is given by:

$$b + at - wt + \alpha l(t;\theta,\beta) \tag{1}$$

Preferences are additive in money. The parameter w is the constant marginal opportunity cost of time. The quality of service also enters the objective function of the physician, with a weight α. The parameter α, which represents the concern for quality, is assumed here to be identical across physicians[5]. This may be interpreted as an ethical norm inducing a concern for the quality of his service (Evans, 1974; Gruber and Owings, 1996). Another interpretation is that patients can partially observe quality (at least *ex post*); in case of a poor service, they may threaten to search for another provider, or share the information with other patients. In both cases, a lower quality diminishes the physician future income (Pauly and Satterthwaite, 1981; Rochaix, 1989; Dranove, 1988)[6]. Notice that this concern for quality should not be interpreted as altruism. A perfectly altruistic physician would value his patient's *total utility*, and thus would also care for the cost of care borne by the patient (either through direct payments or, here, through insurance premiums).

Facing a reward scheme R, the optimal time taken by a β-physician after he observes the health status θ is characterized by the first order condition[7]:

$$R'(t) = a = w = \alpha l_t(t;\theta,\beta), \tag{2}$$

which equates the marginal revenue to the marginal *net* cost, including the concern for quality. We shall denote it by $t^*(a;\theta,\beta)$. Notice that it depends on the fee a, but not on the flat payment b.

Increasing the marginal reward a makes the scheme *more powerful* in the sense that incentives to spend more time and therefore to improve quality are increased[8].

---

5. One could slightly change the interpretation of the talent parameter so as to allow for different levels of concern. We may indeed assume that l writes as $l(t;\theta,\beta) = \beta\tilde{l}(t;\theta)$, in which the $\tilde{l}(t;\theta)$ is the quality of the service.
6. Such an assumption is a common feature in all supplier-induced-demand models, since the induction power must be limited by *some* cost of inducing unnecessary care.
7. Thanks to decreasing returns in time, the objective is concave with respect to t. Hence, the optimal time is unique. Moreover, under standard continuity assumptions, the supremum is reached. Notice that, since quality of care enters his objective, a physician may provide care even when the fee a is lower than the opportunity cost of time w. If the concern for quality α and the marginal effect of time l_t are very large, it could even be the case that physicians would be willing to pay to treat patients. We implicitly assume that α is sufficiently small to rule out such a case.
8. This terminology is to be contrasted with a large branch of the Health Economics literature, which is concerned with cost efficiency issues, especially in hospital care. In such a context a fully prospective payment is a high-powered scheme: letting aside the quality problem and assuming it to be fixed, a prospective payment induces the hospital to minimize its cost, in contrast to a cost-based reimbursement; see, e.g., Newhouse (1996). Our assumptions also lead to the standard feature that labour supply (here, t^*) increases with its reward a, in contrast with the "target income" hypothesis (Fuchs, 1978).

2.2 Health Organizations

We have just described the relationship between a physician with a given characteristic β and a patient with a given health status θ. We now specify the distribution of these characteristics over the population, and how patients and physicians are matched.

The total number of physicians is normalised to 1: the population of physicians is indexed by $j \in J$, uniformly distributed over $J = [0,1]$; the type of physician j is $\beta_j \in B$. The induced probability distribution of types, denoted by F, admits a positive density f over $\left[\underline{\beta}, \overline{\beta}\right]$.

The total number of patients is n, which is also the average number of patients per physician. The type of patient $i \in I = [0,n]$ is denoted by θ_i. It is drawn from an identical probability distribution G with density g, and is unknown by the patient *ex ante*.

Even though health care is a service provided by a physician to a patient, the monetary payments associated with this service are channelled through a health insurance organization *(HO)*. This organization may be a private for-profit or non-profit firm, or a public fund, and may operate in a competitive or a regulated environment. Since patients are *ex ante* identical[9], each one pays a premium independent of his health condition, which is unknown at the time of the subscription. Patients consult their physician whenever they "need", i.e. when they are subject to a health shock θ large enough. Patients and physicians are matched together by the organization to which they subscribe for a given period, say a year. Up to section 5, we consider the situation of a public monopolistic health organization, which offers a unique reward scheme. At the initial situation, the reward scheme in place is denoted by $R^0 = \left(a^0, b^0\right)$ and called the *status-quo*.

All patients subscribe to the *HO*, and all physicians are registered with it. So the *HO* does not play much role apart from choosing the reward scheme R that applies to all physicians and the premium p that is collected from each patient. Monetary transfers are operated without cost.

Patients are randomly matched to physicians: *ex ante*, the case mix, i.e., the distribution of θ, faced by each physician is G, the same for all physicians, and the number of patients per physician (the clientele size) is the same for all physicians, equal to n, the overall ratio of number of patients to number of physicians.

9. Our focus is on the supply side, and the demand side is very sketchy. In particular, we do not consider patients selection (see, e.g., Newhouse, 1996), nor demand side moral hazard issues (for a recent survey on patient demand, see Zweifel and Manning, 2000). This is roughly the situation of a complete insurance contract, consumers paying an overall fixed premium independently of the number of their visits (but the premium possibly depends on their revenue). However, the insurance is not "ideal" in the sense of Arrow (1963): consumers still bear a risk in terms of quality of care and, eventually, in terms of health outcome.

2.3 Physicians and Patients Preferences over Reward Schemes

At the time a reform is contemplated, each physician knows his own characteristic β, and evaluates whether he will lose or gain from the reform. Both patients and physicians are risk neutral.

A β-physician evaluates a reward scheme $R=(a,b)$ according to the expected value of his objective (1). Formally, preferences over reward schemes are represented by the indirect utility function V:

$$V(a,b;\beta)=\left[b+(a-w)T(a;\beta)+\alpha L(a;\beta)\right]n, \tag{3}$$

where T and L are the expected time and quality provided by a β-physician facing a case mix[10] G:

$$T(a;\beta)\equiv E_\theta\left[t^*(a;\tilde{\theta},\beta)\right], \qquad L(a;\beta)\equiv E_\theta\left[l\left(t^*(a;\tilde{\theta},\beta);\tilde{\theta},\beta\right)\right]. \tag{4}$$

Ex ante, before the occurrence of a health shock θ, and having no information on the physician's type with which they are randomly matched β, patients expected utility is given by:

$$E_\beta\left[\lambda L(a;\tilde{\beta})\right]-p.$$

Since both patients and physicians preferences are linear in money and the *HO* operates monetary transfers at no cost, the overall welfare is equal to the sum of patients and physicians utility, and of the *HO* profit. Since monetary transfers cancel out, the overall welfare derived from a linear contract (a,b) depends only on a and may be written as:

$$\overline{W}(a)=nE_\beta\left[W(a,\tilde{\beta})\right]=\int_B W(a,\beta)f(\beta)d\beta \tag{5}$$

where

$$W(a,\beta)=(\lambda+\alpha)L(a,\beta)-wT(a,\beta) \tag{6}$$

is the expected social surplus of a relation between a patient and a β-physician who faces a rate a. Notice that the expected quality L is valued by the patient (weight λ) as well as by the physician (weight α). If physicians were perfectly altruistic, then patients utility should not be counted twice (Jones-Lee, 1991), but in our set up α represents the physician's concern for quality: a better quality improves patients as well as physicians utility, and should therefore enter total welfare in both ways.

---

10. To emphasize uncertainty, we denote a random variable by $\tilde{x}$, and its realization (when observed) by x.

3 Assessing the Need for a Reform

To assess whether a reform is needed, it is useful to characterize a first best allocation of time, without any constraint stemming from reward schemes and the non observability of the physicians characteristics. Still assuming random matching of patients with physicians, the *ex ante* welfare as function of the time spent by physicians $t(\theta,\beta)$, is equal to:

$$n\int_{\beta}\int_{\theta}\left[(\lambda+\alpha)l\left(t(\theta,\beta);\theta,\beta\right)-wt(\theta,\beta)\right]f(\beta)g(\theta)d\beta d\theta.$$

Again transfers across physicians and consumers cancel out. A first best allocation of time maximizes the welfare criterion as given above. The solution is simply obtained by maximising over t, for each (θ,β), the surplus

$$(\lambda+\alpha)l(t;\theta,\beta)-wt.$$

This gives the optimal time t^{FB} as a function of (θ,β), characterized by the first order condition:

$$(\lambda+\alpha)l_t(t;\theta,\beta)-w=0$$

which equates the social marginal value for quality to the marginal cost.

An immediate question is whether the optimal time t^{FB} can be implemented through appropriate payment schemes. Comparing with the time allocated by a physician under a given payment scheme as given by (2), we readily obtain the following result:

PROPOSITION 1. *Any scheme R that satisfies*

$$R'(t)=a^{FB}\equiv w\left(\frac{\lambda}{\alpha+\lambda}\right) \tag{7}$$

leads physicians to choose the optimal time t^{FB}. We call such a scheme first best optimal.

In our model a first best allocation can be obtained through a unique scheme: *there is no reason, on efficiency grounds, to discriminate among the physicians*[11]. Now the main question investigated in this paper is: how to improve upon a (not

11. This is true because there is no heterogeneity in altruism.

first best) status-quo in a situation where physicians are reluctant to be hurt by a reform, are heterogeneous, and schemes cannot be made contingent on their characteristics?

The power of the physicians precisely comes from the possibility to block a reform. Political constraints will impose that a sufficiently large subgroup of physicians benefits from the reform. If individual types were perfectly observed by the *HO*, flat payments could be designed in order to give each β-physician a utility level at least equal to his status quo level. A first best scheme could be implemented, with all physicians supporting the reform. In particular, all efficiency gains could accrue to the consumers. Political constraints together with non observable characteristics distort from first best efficient allocations. By how much?

4 Changing the Status Quo under Political Constraints

We start with the situation where the *HO* is unique and benevolent. It aims to improve patient welfare by changing the status quo contract, and proposes a single scheme R that must satisfy two constraints. Firstly, R is budget balanced:

$$p = E_{\theta,\beta}\left[R\left(t\left(\tilde{\theta},\tilde{\beta}\right)\right)\right] = b + aE_{\beta}\left[T\left(a;\tilde{\beta}\right)\right]. \tag{8}$$

This budget constraint links premiums and payment schemes. Hence, internalizing this constraint, patients utility is given by:

$$U(a,b) = \lambda E_{\beta}\left[L\left(a;\tilde{\beta}\right)\right] - b - aE_{\beta}\left[T\left(a;\tilde{\beta}\right)\right].$$

Secondly, R must be *politically feasible*: a large enough proportion q of the physicians, must accept the new scheme. This leads to the following definition.

DEFINITION 1. *Given the status quo contract R^0, a politically constrained optimum is given by a contract $R = (a,b)$ that maximizes ex ante patient's utility $U(a,b)$ over the politically feasible contracts, i.e. the contracts R that satisfy:*

$$F\left[\beta \in B \middle| V(R;\beta) \geq V\left(R^0;\beta\right)\right] \geq q. \tag{9}$$

Condition (9) states that to be politically feasible, a reform must be preferred to the current situation by a proportion of physicians at least equal to q, each one knowing his type. If unanimity is required, $q = 1$, each physician must be as well off as in the status quo.

To compare political constraints with standard participation constraints, it is natural to assume that the latter are met at the status quo: we have that $V\left(R^0;\beta\right)\geq \underline{v}$ for all β. Therefore, under unanimity, political constraints are more demanding than participation constraints. If unanimity is not required, some schemes may be politically feasible without providing their reservation utility to each physician. Imposing participation constraints in addition to political ones can be easily handled, but does not bring much additional insight.

4.1 Politically Feasible Reforms

To study politically feasible reforms, an analysis of physicians preferences over reward schemes is needed. Which physicians prefer more powerful reward schemes heavily depends on how the optimal time chosen by a physician varies with his type.

PROPOSITION 2.

1. Let a scheme R be given. If talent and time are substitutes (resp. complements) optimal time $t^*\left(R,\theta,\beta\right)$ *decreases (resp. increases) with* β.

2. Let R_1 *be more powerful than* R_2: $a^1 \geq a^2$, *so that physicians spend more time, quality is higher if they face* R_1 *instead of* R_2. *Then if* R_1 *is preferred to* R_2 *by a* β*-physician, it is also preferred by any physician who works more than him, i.e. by any* β'*-physician with* $\beta' < \beta$ *if talent and time are substitutes, or with* $\beta' > \beta$ *if complements.*

In words, property 1 says that in the substitute case, less talented doctors spend more time with their patients than more talented ones, and the opposite in the complement case. Notice that in terms of quality, a change in β has a direct positive effect $\left(l_\beta\right)$ and an indirect effect $\left(l_t t^*_\beta\right)$. Whereas this indirect effect is positive in the complement case, it is negative in the substitute case and may dominate the direct effect.

As for property 2, a scheme R_1 that is more powerful than another one R_2 provides a larger variable reward to time-intensive practices. Therefore, thanks to the envelope theorem, if a physician prefers R_1 to R_2, *a fortiori* any physician with a more intensive practice prefers it as well.

Thanks to these properties, politically feasible linear schemes can be easily described through pivotal characteristics appropriately defined. Let a proposed reform $\left(a,b\right)$ be less powerful than the status-quo $\left(a < a^0\right)$. We know that if the reform is preferred by a physician with type β, it is also preferred by any physician who works less. Hence, the pivotal characteristic is the value β^d such that the proportion of physicians who work *less* than β^d-physician is equal to q. If the reform aims to increase a, then the pivotal characteristic is the value β^u such that physi-

cians who work *more* than the β^u-physician are in proportion q. Formally, we have the following definition:

DEFINITION 2. *The pivotal characteristic*[12] $\beta^c(a,b)$ *for a reform* (a,b) *that aims to decrease (resp. increase) the fee for service,* $a < a^0$, *(resp.* $a > a^0$*) is* β^d *(resp.* β^u*), as given by:*

$$F\left[\beta \middle| T(a;\beta) < T\left(a;\beta^d\right)\right] = q$$

$$\text{and resp.} \quad F\left[\beta \middle| T(a;\beta) > T\left(a;\beta^u\right)\right] = q.$$

In the majority case, $q = 1/2$, the pivotal characteristic is identical whether the proposed reform decreases or increases incentives to work. A majority winner exists if physicians vote on a family of linear contracts $\{(a,b(a))\}$ indexed by a (where $b(a)$ can be supposed to be decreasing in a). This voting equilibrium contract is the one preferred by a β^m-physician, where β^m is the median value of β. The median type β^m is such that more time-intensive physicians (in the substitute case, higher values of β) would favor an increase in a (given the function $b(a)$), whereas less time-intensive physicians would prefer a larger fixed payment.

Given the political power of physicians, however, a much larger support than majority may be needed. For q strictly larger than 1/2, the pivotal characteristics differ whether the proposal is less or more powerful, and we have for any a:

$$T\left(a;\beta^u\right) < T\left(a;\beta^m\right) < T\left(a;\beta^d\right).$$

4.2 Second Best Optimal Reform

The problem of finding a politically feasible reform that makes consumers better off can now be put in a simple form.

A politically constrained optimum is given by a contract $\left(a^p,b^p\right)$ that solves

$$(10) \qquad \begin{cases} \max U(a,b) \\ (a,b) \text{ s.t. } V\left(a,b;\beta^c(a,b)\right) \geq V\left(a^0,b^0;\beta^c(a,b)\right), \end{cases}$$

with $\beta^c(a,b) = \beta^d$ if $a < a^0$, and $\beta^c(a,b) = \beta^u$ if $a > a^0$.

12. The pivotal characteristic is defined in reference to the status quo. Since the status quo is fixed thoughout the paper we have dropped the argument (a^0,b^0). The same remark applies for the compensating variation $b(a, \beta)$ and informational cost $C(a, \beta)$ defined later on.

To analyze the impact of the constraints, it is convenient to write patients utility as the difference between the overall welfare criterion $\bar{W}(a)$, defined by (5), and the aggregate physicians utility. In terms of variation with respect to the status quo we have:

$$n\left(U(a,b)-U\left(a^0,b^0\right)\right)=\left[\bar{W}(a)-\bar{W}\left(a^0\right)\right]-E_\beta\left[V\left(a,b;\tilde{\beta}\right)-V\left(a^0,b^0;\tilde{\beta}\right)\right].$$

Whenever the status-quo differs from a first best scheme $\left(a^0 \neq a^{FB}\right)$, welfare $\bar{W}$ can be increased. To analyze the variation in physicians' utility, let $b(a,\beta)$ be the flat payment that makes a β-physician indifferent between the new scheme $\left(a,b(a,\beta)\right)$ and the status quo $\left(a^0,b^0\right)$. Since the physician's utility is linear in money, it satisfies

$$V(a,b;\beta)-V\left(a^0,b^0;\beta\right)=n\left[b-b(a,\beta)\right]. \tag{11}$$

The flat payment $b(a,\beta)$ is the compensating variation associated with a change in the price of time from a^0 to a. As said previously, the overall efficiency gains would accrue to the consumers while providing each physician with his status quo utility level by giving contingent flat payments $b(a,\beta)$ to each β-physician.

Under non contingent flat payments, changing a requires the flat payment b to be adjusted at $b\left(a,\beta^c\right)$ so as to "buy" the support of the pivotal physician. Therefore, we define the informational cost over all physicians as:

$$C\left(a,\beta^c\right)\equiv E_\beta\left[V\left(a,b;\tilde{\beta}\right)-V\left(a^0,b^0;\tilde{\beta}\right)\right]=n\left(b\left(a,\beta^c\right)-E_\beta\left[b\left(a,\tilde{\beta}\right)\right]\right).$$

If unanimity is required, the informational cost is always positive, since $b \geq b(a,\beta)$ for all β. However, if q is small enough, the informational cost may be negative. For $q=1/2$ for instance, $C(a,\beta)$ is negative if the median value of b is smaller than the mean.

Finally, the change in patients' utility associated to a change in the scheme can be expressed as

$$n\left[U(a,b)-U\left(a^0,b^0\right)\right]=\left[\bar{W}(a)-\bar{W}\left(a^0\right)\right]-C\left(a,\beta^c\right), \tag{12}$$

which is the sum of an efficiency effect, as measured by the variation in welfare, and an informational cost. The following lemma derives the marginal effect of a on these terms.

LEMMA 1. *The marginal effects of changing the fee for service a on welfare W and on informational cost C are given by:*

$$\bar{W}_a(a) = n\left(\frac{\alpha+\lambda}{\alpha}\right)\left(a^{FB}-a\right)E_\beta\left[T_a\left(a,\tilde{\beta}\right)\right], \tag{13}$$

and

$$C_a\left(a,\beta^c\right) = n\left(E_\beta\left[T\left(a,\tilde{\beta}\right)\right]-T\left(a,\beta^c\right)\right) \quad for \quad a \neq a^0 \tag{14}$$

At a given fee a, the marginal welfare gains depend on how far from the first best a is, and on the responsiveness of average physician practice to monetary incentives, $E_\beta\left[T_a\left(a,\tilde{\beta}\right)\right]$.

The marginal informational cost is given by the spread between the average practice time and the one of the pivotal physician. If unanimity is required and $a < a^0$, the pivotal physician, β^d, is the one who spends the largest time; therefore, $T\left(a,\beta^d\right) \geq T\left(a,\beta\right)$ for any β. Thus the marginal cost is negative for $a < a^0$; similarly it is positive for $a > a^0$: this corresponds to the intuition that the further away from the status-quo, the larger the information cost. Also, the more heterogeneity in physician time practice, the more important, in absolute terms, the marginal cost.

A second best optimum trades off marginal welfare gains and marginal informational cost. According to expression (12) and lemma 1, patients' utility increases (resp. decreases) if

$$\left(\frac{\alpha+\lambda}{\alpha}\right)\left(a^{FB}-a\right)+A\left(a,\beta^c\right) \tag{15}$$

is positive (resp. negative) where

$$A(a;\beta) \equiv \frac{T(a;\beta)-E_\beta\left[T\left(a;\tilde{\beta}\right)\right]}{E_\beta\left[T_a\left(a;\tilde{\beta}\right)\right]} \tag{16}$$

and β^c is the pivotal characteristic, β^d or β^u depending on a being lower or greater than a^0. Therefore, the discrepancy between the first best criteria and the second best one is summarized by the quantity $A\left(a;\beta^c\right)$, which is interpreted after the following proposition.

PROPOSITION 3. *Two cases may occur:*

- either the situation is blocked at the status quo, $a^p = a^0$, *which occurs only if*

$$A\left(a^0, \beta^u\right) < \left(\frac{\alpha+\lambda}{\alpha}\right)\left(a^0 - a^{FB}\right) < A\left(a^0, \beta^d\right) \tag{17}$$

- or a^p *satisfies:*

$$a^p = a^{FB} + \left(\frac{\alpha+\lambda}{\alpha}\right) A\left(a^p, \beta^c\right). \tag{18}$$

The key point for understanding why blocking can occur is that the pivotal characteristics differ whether less or more powerful contracts than the status quo are considered except: Technically, expression (15) is discontinuous at a^0, meaning that patients utility presents a kink at the status quo. This is not true however if a simple majority is required, since then both pivotal characteristics coincide with the median one. A reform is politically feasible if it gives the median voter at least his status-quo utility level, so that a second best optimum maximises consumers surplus over the median physician indifference curve[13].

The features that determine whether the situation is blocked and the optimal scheme if it is not blocked are summarized through the function A.

To fix the ideas, assume unanimity is required. Note that without heterogeneity, $A\left(a;\beta^d\right)$ would be null. With heterogeneity, the pivotal physician when a decreases is the one that works the more $\left(T\left(a;\beta\right) \le T\left(a;\beta^d\right)\right)$, hence $A\left(a;\beta^d\right)$ is positive. The quantity $A\left(a;\beta^d\right)$ measures, in terms of fee for service, the impact of the heterogeneity of physicians: $A\left(a;\beta^d\right)$ approximates the raise in a that equalizes the average practice time to that of the pivotal physician $T\left(a;\beta^d\right)$. To see this, let there be a change of the fee for service from a to $a' = a + A\left(a;\beta^d\right)$. Then, assuming that the first order approximation is valid, the effect on the average practice time is given by:

$$\begin{aligned} E_\beta\left[T\left(a';\tilde{\beta}\right)\right] &\simeq E_\beta\left[T\left(a;\tilde{\beta}\right)\right] + A\left(a;\beta^c\right) E_\beta\left[T_a\left(a;\tilde{\beta}\right)\right] \\ &= T\left(a;\beta^c\right). \end{aligned}$$

13. Remark however that the contract chosen by the median voter, say $\left(a^m, b^m\right)$ may well be less efficient than the current scheme: for example if $a^0 > a^{FB}$, the fee for service a^m may be larger than a^0. If this occurs, the patient's welfare is increased at the expense of a larger loss incurred by physicians.

From (17), the larger the positive $A(a;\beta^d)$, and the smaller the negative $A(a;\beta^u)$, the more likely it is that the situation is blocked. In absolute value, it increases with the difference between the practice time of the pivotal physician and the average practice time: this difference increases with heterogeneity of medical practice and with the power of physicians. Also it decreases with its responsiveness: if practice is not very responsive to monetary incentives: very large changes in a would be needed to obtain a desired change in practice time.

Proposition does not characterize optima, and are only local. Finding general conditions under which expression (15) is well behaved are not easy. However, in some cases A is linear in a as in the following example.

Example 1 Let us assume the following form for the quality function:

$$l(t;\theta,\beta) = t^{\gamma} f(\theta,\beta),$$

with $0<\gamma<1$, and $f'_{\theta}<0$, $f'_{\beta}>0$ (complements case). Easy computation gives that, facing a linear scheme $R(t)=b+at$, with $b>0$ and $a<w$, optimal time is given by:

$$t^*(a,\theta,\beta) = \left(\frac{\alpha\gamma f(\theta,\beta)}{w-a}\right)^{\frac{1}{1-\gamma}}.$$

Taking expectation over the case mix distribution gives $T(a,\beta) = K(\beta)(w-a)^{\frac{\gamma}{1-\gamma}}$, for some $K(\beta)$ independent of a. The function A writes as

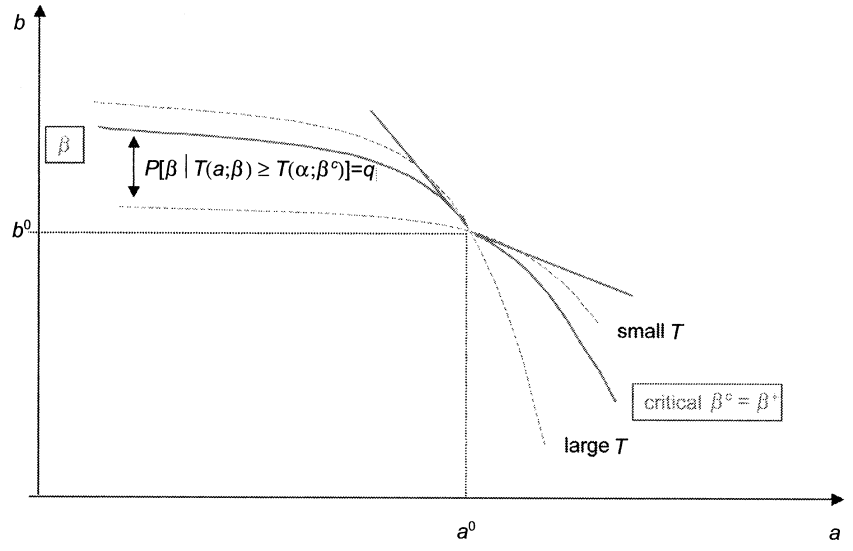
$$(1-\gamma)\frac{K(\beta)-E[K(\beta)]}{E[K(\beta)]}(w-a).$$

The closer the initial fee to marginal cost, the more elastic time practice to monetary incentives (the closer γ to 1), and the more likely the situation is blocked.

To sum sup, our analysis identifies three features as sources for potential deviations from first best efficiency: heterogeneity of medical practice, physicians political power, and elasticity of medical practice with respect to monetary incentives.

These results can be illustrated in the plan (a,b). By linearity in the flat payment, all indifference curves are obtained from each other by vertical translation. Using the envelope theorem, for a physician of type β, the marginal rate of substitution between b (flat payment) and a (fee), V_a/V_b, is equal to the expected time $T(a;\beta)$. Since for any (θ,β), t^* increases with a, indifference curves are concave. In the substitute (resp. complement) case, t^* decreases (increases) with β, and therefore lower (resp. higher) β correspond to steeper indifference curves. The set of politically feasible reforms presents a kink at (a^0,b^0) whenever $q>1/2$.

FIGURE 1



Patients indifference curves over payment schemes are also concave. Patients utility is always decreasing with b, but not with a: when a becomes small enough, the decrease in quality dominates the decrease in the associated payment[14].

If the situation is not blocked at status quo, the second best optimal scheme a^p is such that patients indifference curve is tangent to the pivotal physician indifference curve. For $q > 1/2$, the kink at a^0 in the set of politically feasible reforms may be such that no other feasible contract provides patients with a higher utility level.

5 Menu of Plans

The previous analysis considers a limited type of reform: a new scheme, if accepted, applies to each physician and replaces the status-quo. Since physicians heterogeneity is at the root of the difficulties, a solution could be to offer several distinct schemes within which physicians may choose. A situation with several active schemes may be the result of a competition game among distinct health organizations, or may be proposed by a unique one.

Whatever the situation, we need first specify how patients and physicians allocate themselves across the different proposed plans.

14. Formally, we have that $U_b = -1$, and $U_a = \left(\frac{\alpha+\lambda}{\alpha}\right)\left(a^{FB} - a\right)E_\beta\left[T_a\left(a,\tilde{\beta}\right)\right] - E_\beta\left[T\left(a,\tilde{\beta}\right)\right]$. This last term may be positive when $a < a^{FB}$.

5.1 Equilibrium Allocations

A *plan* specifies a scheme (a,b) to physicians, and a premium p to be paid by patients. Let $\{(a^k,b^k,p^k),k\in K\}$ be the set of proposed plans by health organization(s). Each participant, physician or patient, chooses a single plan (exclusive contract). Taking as given the plans, the resulting allocation of patients and physicians is determined under two basic assumptions: (1) *free mobility* according to which patients and physicians can freely choose within the set of proposed plans and (2) *rational expectations* on the quality proposed by each plan and its clientele.

Before going further, let us mention first a difficulty linked with coordination problems. Given a set of proposed plans, many outcomes are equilibria sustained by well chosen self fulfilling expectations. Suppose for instance that all patients choose the same plan. Then all physicians also pick that plan, which in turn justifies patients' choices. Some expectations solve this coordination problem in a reasonable way, as we explain now.

Let us say that physicians have *homogeneous* expectations if they expect the number of patients per physicians and the distribution of their health shocks θ to be equal to that of the overall population, hence to be identical across plans. This assumption is justified if patients are indifferent between all "active" physicians (as made precise below) and choose among them randomly, as will be true at equilibrium. Given the set of proposed plans $\{(a^k,b^k,p^k),k\in K\}$ the set of plans that can be chosen both by physicians and patients at equilibrium is determined as follows.[15]

Given the set of proposed schemes $\{(a^k,b^k)\}$, each physician determines his choice under homogeneous expectations.[16] Denote by B^k the set of the types of the physicians who choose plan k. Since contracts are exclusive, the sets $\{B^k\}$ form a partition[17] of the set of possible types. A plan is said to be *active* if it is chosen by a non negligible set of physicians and patients. The quality of care provided by an active plan k is given by $L^k = E\left[L(a^k,\tilde{\beta})\middle|\beta\in B^k\right]$.

Patients' choice is determined by the expected quality of care, net of premiums. At the time they choose a plan, although they do not know their health status, they make conjectures about the quality provided by the physicians in each plan. If they correctly assume physicians' choice to be determined as described above, they expect the quality of care provided by physicians in active plan k to be L^k. Now, patients, who are identical ex ante, all derive the same utility from any plan k (i.e., $u^k = \lambda L^k - p^k$), and all choose among those that give them the highest level.

---

15. The selection can be seen as the result of an adaptive process that takes place in real time, or only in agents mind.

16. Note that the value of joining plan k for a β-physician is given by $V(a^k,b^k,\beta)$ where V is defined by equation (3) in section 2.3: $V(a,b;\beta)=(b+(a-w)T(a,\beta)+\alpha L(a,\beta))n$.

17. We consider here distinct contracts, so that the set of physicians who are indifferent between contracts is negligible. If two identical contracts are proposed, as examined in next section, physicians are simply allocated randomly between the *HO*s.

Therefore, plans that do not yield the highest level, if any, are eliminated, and a new allocation of the physicians among the restricted set of plans is determined. Hence, at equilibrium, all patients expect the same utility from any physician of any active plan. Thus, if patients choose randomly among the physicians, physicians homogeneous expectations are correct. The described process leads to a set of active plans that satisfy the following definition:

DEFINITION 3. *A set of distinct plans* $\left\{\left(a^k, b^k, p^k\right), k \in K\right\}$ *are active under homogeneous expectations if*

1. (physicians choice) plan k is chosen by the physicians whose characteristics are in $B^k = \left\{\beta \middle| V\left(a^k, b^k; \beta\right) \geq V\left(a^{k'}, b^{k'}; \beta\right) \forall k'\right\}$ *and* $F\left(B^k\right) > 0$.

2. (patients choice) patients utility level is identical across active HO: $u^k = \lambda E\left[L\left(a^k, \beta\right) \middle| \beta \in B^k\right] - p^k = u$ *for all active k. A clientele of size* $nF\left(B^k\right)$ *subscribes to plan k.*

3. (feasibility) total premiums are larger than total payments to physicians: $\sum_k n\left(p^k - \left(a^k E\left[T\left(a^k, \beta\right) \middle| \beta \in B^k\right] + b^k\right)\right) F\left(B^k\right) \geq 0$.

Equilibrium imposes quite severe constraints. To see this assume that some regulation prohibits any price (premium) discrimination across different plans. This is roughly the situation in the Dutch health insurance system after the Dekker reform: (some) individuals may choose between different insurance funds, each of which selects a list of physicians, but the premium paid by each individual is independent of this choice. In the complement case, *two distinct plans cannot be active without price discrimination*. The intuition is straightforward. Take two contracts with say $a^k < a^{k'}$. Physicians who choose a^k work less than those who choose the more powerful contract $a^{k'}$ (incentive effect). In the complement case, they have lower values of β (selection effect). But since quality increases with time and characteristic, both effects go in the same direction: the average quality in plan k is strictly lower than in k'. So consumers can perfectly rank physicians in terms of expected quality, by observing the contracts they have chosen. If they pay an identical fee for any physician they consult, no consumer would get services from doctors known to provide a lower quality.

5.2 Cross Subsidies

This section considers a public monopoly that introduces several contracts, and may implement cross-subsidies between plans. This means that the overall feasibility constraint is required, but some plans may generate a deficit.

The set of contracts that can be implemented is quite large. To see this, take *any* set of schemes $\left(a^k, b^k\right)$. Physicians self select themselves into a partition $\left\{B^k\right\}$, according to their expected utility level V (as in *1* of definition 3). Exclude con-

tracts that no physician chooses. By charging adequate premiums, exactly those contracts can be active: the partition determines the average quality L^k within each plan k, and the premium is fixed at $p^k = \lambda L^k - u$ for each k, where u is set so as to satisfy feasibility (hence *2* and *3* of definition 3 are also satisfied). Notice that the budget of each single plan is not necessarily balanced.

In our analysis of reform under political constraints, section 4, physicians utility had to be greater, for a sufficiently large proportion of them, than their status quo utility level, obtained under the unique scheme $\left(a^0, b^0\right)$. In line with this analysis, we investigate the case of a monopoly which introduces contracts in addition to the initial status quo contract, and charges premiums according to the implementation strategy just described. To keep notation simple, we investigate the addition of a single contract $\left(a^1, b^1\right)$. We ask whether this initial contract $\left(a^0, b^0\right)$ may be completed with an additional contract $\left(a^1, b^1\right)$, while increasing patients satisfaction. Since by assumption the initial contract is offered, each physician is at least as well off than at the status quo. Accordingly, if patients welfare is also increased, everyone benefits from the introduction of the new contract. This also implies that the two plans will be active if the situation is blocked at the status quo when unanimity is required. An important point to notice is that, as soon as an additional plan is introduced and active, the premium associated to the initial contract $\left(a^0, b^0\right)$ must be modified.

Thanks to linear utilities the sum of the physicians- patients welfare derived from a contract chosen by the set of physicians characteristics in B^1 is $\bar{W}\left(a^1, B^1\right) = n\int_{\beta \in B^1} W\left(a^1, \beta\right) f(\beta) d\beta$ where $W(a, \beta)$, the expected social surplus of a relation between a patient with a β-physician who faces a rate a, is given by equation (6). For each physician β who remains with the status quo plan, the utility level is the same, and the surplus $W\left(a^0, \beta\right)$ is not modified either. Hence, noticing that $\bar{W}\left(a^0, B\right) = \bar{W}\left(a^0, B^0\right) + \bar{W}\left(a^0, B^1\right)$, simple computations give the potential utility gain ΔU due to the introduction of an additional contract as:

$$\Delta U = \bar{W}\left(a^1, B^1\right) - \bar{W}\left(a^0, B^1\right) - n\int_{\beta \in B^1} \left(V\left(a^1, b^1, \beta\right) - V\left(a^0, b^0, \beta\right)\right) f(\beta) d\beta.$$

The first term represents the variation in welfare due to a change in the fee a, which is made over physicians who opt for B^1; if a^1 is closer than a^0 to a^{FB}, it is positive. However, the second term is also positive, since it represents the gain in physicians utility who opt for B^1 or, equivalently, the informational cost. As before, the informational cost may be written as the compensating variation associated with the change of a from a^0 to a^1:

$$(19) \quad \Delta U = \bar{W}\left(a^1, B^1\right) - \bar{W}\left(a^0, B^1\right) - n\int_{\beta \in B^1}\left[b^1 - b\left(a^1, \beta\right)\right] f(\beta) d\beta.$$

The following proposition states that there always exists an additional contract $\left(a^1, b^1\right)$ such that ΔU is positive.

PROPOSITION 4. *Let a^1 be a welfare improving fee for service (say a^1 is between a^{FB} and a^0). The introduction of an additional contract $\left(a^1, b^1\right)$ increases patients welfare, provided that b^1 is not too high.*

As explained above, patients welfare is increased if the informational cost, the second term in equation (19) is smaller than the welfare gains, the first term in equation (19). Consider b^1 the minimum level required to attract some physicians (if $a^1 < a^0$, these physicians are the ones who work the least, say of type $\underline{\beta}$ in the complement case): B^1 is empty. As b^1 increases from this minimal value, B^1 is small. The basic idea is that when B^1 is small, physicians population within B^1 is homogeneous. This implies that the *marginal* informational cost as b^1 increases from this minimal value is equal to zero. Nevertheless, the marginal benefit is strictly positive, since efficiency gains are made over all physicians at the margin (of positive measure $f\left(\underline{\beta}\right)$ in the complement case). If b^1 increases more, the size of the new *HO* and the marginal informational cost increases. At some point, the informational cost may offset the welfare gain if B^1 is too large (which is surely the case if the initial situation is blocked).

Notice that, in the complement case and for $a^1 < a^0$, the average quality in HO^1 is lower than in HO^0, both by an incentive effect on practice time $\left(a^1 < a^0\right)$ and by a selection effect (physicians who opt for HO^1 are those with lower values of β). Since patients utility increases, this means that p^1 must be lower than the premium at the status quo. But the selection effect also implies that the average quality within HO^0 has increased. It may or may not be the case that p^0 is larger than at the status quo.

5.3 Competition

Cross subsidies may be difficult to implement in practice. In a context of physician heterogeneity, competition between several health organizations may be another more natural way of introducing several contracts. This section analyzes a competition game with free entry, in which *HO*s compete for patients and physicians. The specification of such a two-sided competition game is the following one. First health organizations propose plans. Second, given the set of proposed plans, the active plans, the physicians' choice, and the patients utility are determined.

We study Nash equilibria in plans. Let $\left(a^k, b^k, p^k\right)$ $k = 1, .., K$ be proposed, and u be the patients utility level reached at the equilibrium allocation (identical across active plans). A *HO*, say HO^1, takes other plans as given, and contemplates a deviation (a, b, p). In doing so, it considers the new equilibrium allocation of physicians

and patients. There are two possible cases to consider. In the first case, contract (a,b) attracts all physicians whose characteristics are in $B' \subset B$, and patients get exactly the level u, that is the premium p satisfies $p = E\left[\lambda L\left(a,\tilde{\beta}\middle|\tilde{\beta} \in B'\right)\right] - u$. So the profit generated by a contract (a,b,p) is given by

$$\pi(a,b,u,B') = n\int_{B'}\left[\lambda L(a,\beta) - u - aT(a,\beta) - b\right] f(\beta)\,d\beta \tag{20}$$

In the second case, patients get strictly more than u and contract (a,b) is taken by all physicians.[18] Therefore the profit is $\pi(a,b,u+\varepsilon,B)$ for some positive ε, where π is the function defined by (20).

At equilibrium, no single *HO* may increase its profit by deviating from the proposed plans. In principle, both kinds of equilibrium may occur: either all active plans are identical[19] (a pooling equilibrium), or some are distinct (a separating equilibrium). However, the following proposition shows that neither may occur.

PROPOSITION 5. *Under physicians heterogeneity, no equilibrium exists in the competition game.*

That no pooling equilibrium exists follows from the following standard arguments. First if a unique plan (a,b,p) is proposed, possibly by several *HO*s sharing randomly physicians and patients, the expected profit of the plan must be null, thanks to free entry. Hence $\pi(a,b,u,B) = 0$. Second, owing to physicians heterogeneity surely $\pi(a,b,u,B') > 0$ for a well chosen subset B'. Thanks to the single-crossing property of physicians indifference curves (proposition 2), a new plan can be offered that precisely attracts physicians with characteristics in B'. Therefore adjusting the premium so as to give utility level u yields a profitable deviation.

That no separating equilibrium exists is reminiscent of competition in differentiated products à la Hotelling: Health organizations compete for the "center" of the market, which is given by the first best contract. To see this, let us assume several distinct plans be active at an equilibrium. The less powerful one (with the smallest a, say a^1) attracts physicians who work the less. The fact that a *HO*'s strategy is composed with a scheme (a,b) and a premium p allows HO^1 to adjust the fee for service a while attracting the same set of physicians and keeping unmodified the patients utility level. Furthermore, since physicians within HO^1 work less than in any other *HO*, HO^1 may increase a and decrease b (keeping the utility of the marginal physician constant) in such a way that the rent left to all physicians with

18. Such a corner strategy plays an important role in the study of competition among intermediaries. As first shown by Yanelle (1989), an intermediary may benefit from competing hard on one side of the market, here the patients, in order to be in a monopoly position on the other side (see Armstrong (2002) for a recent survey on competition in two sided markets).

19. If all physicians in B' choose contract (a,b) $\pi(a,b,B')$ is the profit of the *HO*. If the contract is proposed by several *HO*s, then each gets a fraction of $\pi(a,b,B)$ (it is not necessary to specify more here).

characteristics in B^1 decreases. If a^1 was smaller than a^{FB}, such a strategy[20] would both improve efficiency and decrease the physicians rent, hence would increase profit. Therefore at a Nash equilibrium, the fee for service of the "less powerful" contract, that is the smallest a, must be larger than a^{FB}. A similar argument shows that the largest a must be smaller than a^{FB}: this gives the contradiction. Health organizations compete for the center of the market, and these centrifugal forces destabilize any separating equilibrium.

REMARK. *Consider a regulated competition, in which each HO must leave patients with a minimum utility level*[21] *denoted by* $\overline{u}$. *Under this regulation, the set of available strategies is smaller since the premium asked by k must satisfy* $p^k \leq \lambda E\left[L\left(a^k,\beta\right)\middle|\beta \in B^k\right] - \overline{u}$. *So one could hope to restore the existence of an equilibrium. However, as the proof makes clear, whatever plans, one HO has a profitable deviation in which the patients' utility levels is left unmodified. Thus an equilibrium does not exist either under this form of regulated competition.*

6 Conclusion

Our analysis shows that political constraints severely restrict the possibility of reforming payment schemes. Indeed, due to imperfectly observable (or not contractible) medical practice, rents have to be left to physicians. When physicians practice is heterogeneous and does not respond much to incentives, the increase in the rent necessary to get a reform supported may outweigh efficiency gains.

We focused our analysis on physicians, but believe our analysis extends to other professions, for which the crucial importance of the service, together with the difficulty to observe the output quality, has justified regulation. Beyond medical professions, examples include train drivers (especially in France) or air traffic controllers. In such areas, past regulation has given professionals a strong political power and a very high status quo position. Our analysis may provide an explanation of the difficulties to reform such professions.

If physician heterogeneity determines the difficulty of payment scheme reforms, this suggests that the introduction of flexibility, in the form of a menu of contracts among which physicians may self-select, could be worth a try, by reducing the cost of information asymmetries. However, the analysis has shown that the introduction of competition may not be an easy solution. The specific ele-

20. The profit over a fixed set of physicians characteristics B' keeping u constant can be decomposed as in the previous section into welfare gains minus a rent to be left to the physicians with characteristics in B^1 (see the proof for more details).

21. This level may be given by the outside option value of not getting insured (and receiving no care) at all: if we denote by $\overline{l}(\theta)$ the "quality of no care", i.e. $\overline{l}(\theta) = l(0,\theta,\beta)$ which is independent of β since $t = 0$, the utility level under no insurance is: $\lambda E\left[\overline{l}(\theta)\right]$. Regulation may also impose a higher patients utility level than their reservation value.

ment that renders competition difficult to implement is that patients can choose between the different plans. Free choice, together with rational expectations, puts strong constraints on the links between the premium and average provided quality not only within a given plan, but also across the various plans. This difficulty directly stems from the fact that three types of "agents" intervene in the system: physicians, patients, and health organizations. Competition gives an important role to health organizations and creates some room for divergence of interests between insurance firms and patients interests. Any solution needs to integrate these conflicting interests, by implementing appropriate cross subsidies between alternative contracts. A monopolistic firm could, in principle, implement such a scheme.

Much has still to be understood in the way regulated competition between health organizations could work in this "medical triad", and provide a way to reduce the cost due to imperfectly observable medical practice. In particular, an interesting question is whether the cross subsidies studied in the last section could be introduced in a competitive setting. Cross subsidies could take the form of "quality-compensation" mechanisms, that would compensate patients for differences in quality across different health organizations. Additional research is needed to study a game in which firms compete in contracts, given such quality compensation mechanisms.

We left aside the important issue of patients selection by physicians or by health organizations. Patients selection should be studied in an extended set up where the physician-patient matching is no longer random but endogenous.

References

ARMSTRONG M. (2002). – "Competition in Two-Sided Markets", mimeo, October 2002.

ARROW K. (1963). – "Uncertainty and the Welfare Economics of Medical Care", *American Economic Review*, 53: pp. 941-973.

BLOMQVIST A. (1991). – "The Doctor as Double Agent: Information Asymmetry, Health Insurance, and Medical Care", *Journal of Health Economics* 10(4): pp. 411-422.

CORNING P. (1969). – *The Evolution of Medicare... From Idea to Law*, Government Printing Office, Washington, D.C.

CROXSON B., PROPPER C. and PERKINS A. (2002). – "Do Physicians Respond to Incentives? The case of British GPFH", *Journal of Public Economics* 79 (2): pp. 375-398.

DELATTRE E. and DORMONT B. (2000). – "Induction de la demande par les médecins libéraux français: Etude microéconométrique sur données de panel", *Economie et Prévision*, 142: pp. 137-161.

DRANOVE D. (1988). – "Demand Inducement and the Physician/Patient Relationship", *Economic Inquiry* 26 (2): pp. 281-288.

EVANS R. (1974). – "Supplier-Induced Demand". In M. Perlman (ed.), *The Economics of Health Care and Medical Care* (London: MacMillan): pp. 162-73.

FUCHS V. (1978). – "The Supply of Surgeons and the Demand for Operations". *Journal of Human Resources*. Vol. XIII, supplement: pp. 35-56.

GRUBER J and OWINGS M. (1996). – "Physician Financial Incentives and Cesarean Section Delivery", *RAND Journal of Economics* 27(1): pp. 99-123.

HASSENTEUFEL P. (1997). – *Les médecins face à l'Etat. Une comparaison européenne,* Paris, Presses de Science Po.

HAVIGHURST C. (1978). – "Professional Restraints on Innovation in Health Care Financing", *Duke Law Journal*: pp. 303-387.

HOLLY A., GARDIOL L., DOMENIGHETTI G. and BISIG B.(1998). – "An Econometric Model of Health Care Utilization and Health Insurance in Switzerland", *European Economic Review*, 42: pp. 513-522

JONES-LEE M. (1991). – "Altruism and the Value of Other People's Safety", *Journal of Risk and Uncertainty* 4: pp. 213-219.

KESSEL R. (1958). – "Price Discrimination in Medicine", *Journal of Law and Economics* 1: pp. 20-53.

LAFFONT J.-J. and TIROLE J. (1993). – *A Theory of Incentives in Regulation and Procurement*, MIT Press.

MA A. and MCGUIRE T. (1997). – "Optimal Health Insurance and Provider Payment", *American Economic Review* 87(4): pp. 685-704.

MCGUIRE T. (2000). – "Physician Agency", ch. 9 in *Handbook of Health Economics*, A. Culyer and J. Newhouse, eds., North-Holland.

NEWHOUSE J. (1996). – "Reimbursing Health Plans and Health Providers: Efficiency in Production versus Selection", *Journal of Economic Literature*, 34(3): pp. 1236-63.

PAULY M. and SATTERTHWAITE M. (1981). – "The Pricing of Primary Care Physician Services: a Test of the Role of Consumer Information", *Bell Journal of Economics* 12: pp. 488-506.

ROCHAIX L. (1989). – "Information Asymmetry and Search in the Market for Physician Services", *Journal of Health Economics* 8: pp. 53-84.

YANELLE M.-O. (1989). – "The Strategic Analysis of Intermediation: Asymmetric Information and the Theory of Financial Markets", *European Economic review*, 3: pp. 294-301.

ZWEIFEL P. and MANNING W.G. (2000). – "Moral Hazard and Consumer Incentives in Health Care", in: J.A. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics* (Amsterdam: Elsevier, Chapter 8).

Appendix: Proofs

Proof of Proposition 2

1. Deriving the first order condition (2): $R'\left(t^*\right) = w - \alpha l_t\left(t^*;\theta,\beta\right)$ with respect to β, straightforwardly gives that if talent and time are substitutes $\left(l_{t\beta} < 0\right)$ optimal time $t^*\left(a,\theta,\beta\right)$ decreases with β.

2. Let R^1 be steeper than R^2 (i.e., $a^1 \geq a^2$). From (2) again, increasing the marginal reward a for all t shifts the marginal benefit of t upwards, hence increases t^*: so $t^*\left(a^1,\theta,\beta\right) \geq t^*\left(a^2,\theta,\beta\right)$.

Denote by $\Delta(\beta)$ the difference in expected utility associated with the two schemes for a β–physician: $\Delta(\beta) = V\left(R_1,\beta\right) - V\left(R_2,\beta\right)$ where V is given by (3):

$$V(R;\beta) = nE_\theta\left[R\left(t^*\left(a;\tilde{\theta},\beta\right)\right) - w\left(t^*\left(a;\tilde{\theta},\beta\right)\right) + \alpha l\left(t^*\left(a;\tilde{\theta},\beta\right);\tilde{\theta},\beta\right)\right].$$

A β-physician prefers R^1 to R^2, if $\Delta(\beta) \geq 0$. The derivative of Δ with respect to β, thanks to the envelope theorem, is given by

$$\Delta'(\beta) = n\alpha E_\theta\left[l_\beta\left(t^*\left(a^1;\tilde{\theta},\beta\right);\tilde{\theta},\beta\right) + l_\beta\left(t^*\left(a^2;\tilde{\theta},\beta\right),\tilde{\theta},\beta\right)\right].$$

If talent and time are substitutes, $l_{t\beta} < 0$, and we know that $t^*\left(a^1,\theta,\beta\right) \geq t^*\left(a^2,\theta,\beta\right)$: Δ' is negative, hence Δ is decreasing. So if a β–physician prefers R^1 to R^2, we have $\Delta(\beta') \geq \Delta(\beta) > 0$ for any $\beta' \leq \beta$. This proves that any physician with a lower characteristic than β, who works more than the β–physician, also prefers R^1 to R^2. The proof is similar in the complement case, with Δ increasing.

Proof of Lemma 1

Differentiating V with respect to a, we have that $V_a(a,b;\beta) = nT(a,\beta)$. This gives:

$$\begin{aligned} W_a/n = U_a + E_\beta\left[V_a\left(a,b;\tilde{\beta}\right)\right] &= E_{\theta,\beta}\left[\lambda l_t t_a^* - t^* - a t_a^*\right] + E_\beta\left[T\left(a,\tilde{\beta}\right)\right] \\ &= E_{\theta,\beta}\left[\left(\lambda l_t - a\right)t_a^* - t^*\right] + E_{\theta,\beta}\left[t^*\right] \\ &= E_{\theta,\beta}\left[\left(\lambda l_t - a\right)t_a^*\right]. \end{aligned}$$

Since $l_t = (w-a)/\alpha$ and $a^{FB} = w\lambda/(\alpha+\lambda)$, we obtain:

$$W_a(a) = n\left(\frac{\alpha+\lambda}{\alpha}\right)\left(a^{FB} - a\right)E_\beta\left[T_a\left(a,\tilde{\beta}\right)\right].$$

As for the informational cost it suffices to use that $V_a(a,b;\beta) = nT(a,\beta)$ for any β.

Proof of Proposition 3

If we denote by μ the Lagrange multiplier associated with the political constraint $V\left(a,b;\beta^c\right) \geq V\left(a^0,b^0;\beta^c\right)$, the first order condition that characterizes an interior solution to problem (10) is defined by:

$$\begin{cases} nU_a + \mu V_a\left(a,b;\beta^c\right) = 0 \\ nU_b + \mu V_b\left(a,b;\beta^c\right) = 0 \end{cases}$$

Since fixed payments are simply payments from patients to physicians, we immediately have that $U_b = -1$ and $V_b = n$, leading to $\mu = 1$. The first condition $nU_a + V_a\left(a,b;\beta^c\right) = 0$ may be written as:

$$nU_a + E_\beta\left[V_a\left(a,b;\tilde{\beta}\right)\right] - E_\beta\left[V_a\left(a,b,\tilde{\beta}\right)\right] + V_a\left(a,b;\beta^c\right) = 0$$
$$W_a(a) - C_a\left(a,\beta^c\right) = 0$$

Substituting the computed values for W_a and C_a gives the result when the optimal value for a is interior. However, C is not differentiate at a^0: its left derivative is equal to $C_a\left(a^0,\beta^d\right)$ and its right derivative to $C_a\left(a^0,\beta^u\right)$. Therefore it may be the case that the marginal cost outweighs the marginal benefit for small changes in either direction. Formally, this happens when $C_a\left(a^0,\beta^d\right) < W_a\left(a^0\right) < C_a\left(a^0,\beta^u\right)$. Hence the result.

Proof of Proposition 5

Given an equilibrium let u be the patients utility level. The "profit" generated in a *HO* by a physician of type β is given by:

$$\pi(a,b,\beta) = n\left[\lambda L(a,\beta) - u - aT(a,\beta) - b\right].$$

(1) We first prove that *no pooling equilibrium exists*.

Let (a,b) be the identical contract proposed by the two *HO*. Since contracts are identical, physicians are indifferent between two health organizations; they are distributed randomly between the different organizations, and the characteristics of the physicians subscribing to each *HO* is identical, equal to the prior distribution of β. Let s^k be the size of HO^k. The profit of HO^k is equal to: $s^k \pi\left(a,b,\left[\underline{\beta},\overline{\beta}\right]\right)$ (note that we do not exclude $s^k = 0$, i.e. a monopoly).

The argument proceeds in two steps.

Step l. The profit of each *HO* is null.

The argument is similar to competition à la Bertrand. Assume that π^1, the profit of HO^1, is strictly positive. If HO^2 deviates by increasing a little bit the fixed payment b and lowering its premium, it attracts all physicians and all patients, and thereby captures almost all HO^1's profit. Hence the case $\pi^1 > 0$ is not compatible with a pooling equilibrium.

Step 2. A *HO* by selecting some physicians can make a positive profit.

Since the profit of each *HO* is null, we have that:

$$\pi\left(a,b,\left[\underline{\beta},\overline{\beta}\right]\right) = \int_{\underline{\beta}}^{\overline{\beta}} \pi(a,b,\beta) f(\beta) d\beta = 0.$$

Since π is not constant with respect β, there is an interval, say $\left[\underline{\beta},\beta'\right]$, such that $\int_{\underline{\beta}}^{\beta'} \pi(a,b,\beta) f(\beta) d\beta$ is not null. Assume it to be positive (otherwise consider $\left[\beta',\overline{\beta}\right]$). This means that the contract (a,b) generates a strictly positive profit if (1) it is only chosen by the physicians with characteristics in $\left[\underline{\beta},\beta'\right]$ and (2) the premium to the patients is fixed accordingly, i.e. $p = \lambda L(a,\beta) - \overline{u}$. We now show that it is possible to modify the contract at the margin so as to attract only these physicians. Let $a' < a$ in the complement case, and $a' < a$ in the substitute case. Choose b' so as to make a β'-physician indifferent between (a,b) and (a',b'). Then all physicians with a lower β strictly prefer (a',b') to (a,b). By choosing a' sufficiently close to a, the profit of the plan is close to $\pi\left(a,b,\left[\underline{\beta},\beta'\right]\right)$, which is strictly positive. So there is a profitable deviation.

Notice that this possibility is a direct consequence of the single-crossing property of physicians indifference curves over contracts (a,b). This completes the first part of the proof: no pooling equilibrium exists.

(2) We now prove that *no separating equilibrium exists*.

Let $\left(a^1,b^1\right)$ and $\left(a^2,b^2\right)$ be two distinct active contracts (possibly within other contracts), with a^1 being the smallest a, and a^2 the largest. If there exists a separating equilibrium, we have $a^1 > a^2$. Physicians who work less (with lower values of T) will join HO^1.

If both are active there exists an interior β' such that physicians with β smaller than β' join HO^1, and those with β larger than β' join HO^2, or the reverse. More

precisely $B^1 = \left[\underline{\beta}, \beta'\right]$ in the complement case and $B^1 = \left[\beta', \bar{\beta}\right]$ in the substitute case.

We consider marginal deviations from $\left(a^1, b^1\right)$ that leave the set of physicians who choose either *HO* unchanged. Such deviations leave the β'-physician indifferent between the two contracts, and keep a fee for service smaller than a^2. Hence, such changes (a,b) from $\left(a^1, b^1\right)$ induce no selection effect, but only "pure" incentive effects: They satisfy the constraints

$$V(a,b,\beta') = V\left(a^2, b^2, \beta'\right), \text{ and } a \leq a^2. \tag{21}$$

Since $V_b = n$ and $V_a = nT(a,\beta)$, such a marginal change satisfies $\frac{\partial b}{\partial a} = -T(a,\beta')$.

Since by construction the level of patients welfare is kept constant, the *HO* profit varies as the surplus over a fixed set of physicians. More precisely the following identity always holds

$$W\left(a, B^1\right) = U + \pi\left(a, b, B^1\right) + V\left(a, b, B^1\right)$$

where U is the utility level of the patients that subscribe to HO^1. By construction, the premium is adjusted so that $U = nF\left(B^1\right)u$ is kept constant. So maximizing profit amounts to maximizing welfare $W\left(aB^1\right)$ under the constraints (21).

Computation and interpretation are similar to those performed in the monopoly section (lemma 1) simply by replacing the whole interval of characteristics B by B^1. This immediately gives the marginal change in profit

$$\left.\frac{\partial \pi}{\partial a}\right|_{\beta' \text{ ct}} = \frac{\partial \pi}{\partial a} - nT(a,\beta')\frac{\partial \pi}{\partial b} =$$

$$n\left(\frac{\alpha+\lambda}{\alpha}\right)\left(a^{FB} - a\right)\int_{B^1} T_a\left(a,\tilde{\beta}\right) f(\beta)\, d\beta + n\int_{B^1}\left[T(a,\beta') - T\left(a,\tilde{\beta}\right)\right] f(\beta)\, d\beta$$

As in the monopoly case, this equation gives the marginal change in profit as the sum of efficiency gains (if a gets closer to a^{FB}) and informational costs (changes in the rent left to physicians) within the *HO*. All physicians in HO^1 work less than the β' physician. So the term $\left[T(a,\beta') - T\left(a,\tilde{\beta}\right)\right]$ is surely positive: increasing a and decreasing b decreases the overall payment to physicians. Since the marginal informational cost is positive, we must have, at a Nash equilibrium, that the marginal efficiency gain is negative, i.e. that $\left(a^{FB} - a^1\right) < 0$.

Consider now HO^2. The argument is reversed. All physicians in HO^2 work more than the β' physician. So the term $\left[T(a,\beta') - T\left(a,\tilde{\beta}\right)\right]$ is surely *negative*: decreasing a^2 and increasing b decreases the overall physicians' welfare. So HO^2

has no profitable deviation only if by decreasing a the surplus is decreased. At a Nash equilibrium, we must have that $a^2 < a^{FB}$.

Starting with $a^1 < a^2$, we showed that π^1 increases with a^1 as long as $a^1 < a^{FB}$, and π^2 decreases with a^2 as long as $a^2 < a^{FB}$. Hence, at a Nash equilibrium, we must have that $a^2 \leq a^{FB} \leq a^1$, which contradicts the starting assumption $a^1 < a^2$.

Proof of Proposition 4

Let a^1 be between the first best and the status quo levels. Given $\left(a^0, b^0\right)$ and $a^1 \in \left[a^{FB}, a^0\right]$, we may parameterize B^1 by β', the type of the physician who is indifferent between B^0 and B^1. To fix the idea take the complement case: physicians with a low value of β work less, therefore $B^1 = \left[\underline{\beta}, \beta'\right]$. If b^1 is too small, then no physician would opt for B^1 and only the initial contract would be active. We show that choosing b^1 higher than the minimal level but small enough does the job. By definition of β', we have that $b^1 = b\left(a^1, \beta'\right)$; hence, from (19), the derivative of ΔU w.r.t. β' in $\left[\underline{\beta}, \overline{\beta}\right]$ is equal to:

$$\frac{\partial \Delta U}{\partial \beta'} = \left[W\left(a^1, \beta'\right) - W\left(a^0, \beta'\right)\right] f\left(\beta'\right) - \frac{\partial b\left(a^1, \beta'\right)}{\partial \beta'} F\left(\beta'\right).$$

The minimal level of b^1 that makes B^1 active corresponds to $\beta' = \underline{\beta}$ For this value, the second term (i.e. the marginal information cost) is equal to zero, and:

$$\left.\frac{\partial \Delta U}{\partial \beta'}\right|_{\beta' = \underline{\beta}} = \left[W\left(a^1, \underline{\beta}\right) - W\left(a^0, \underline{\beta}\right)\right] f\left(\underline{\beta}\right) > 0.$$

The inequality holds since, by assumption, a^1 is closer to the first best than a^0. At the margin, the informational cost to increase β' starting from $\underline{\beta}$ is zero, but the welfare gain is positive. Hence the result.

The proof is similar in the substitute case, except that the new *HO* requires β' to be sufficiently close to $\overline{\beta}$.