DISCUSSION PAPER SERIES

No. 3589

REFORMING INCENTIVE SCHEMES UNDER POLITICAL CONSTRAINTS: THE PHYSICIAN AGENCY

Gabrielle Demange
and Pierre-Yves Geoffard

PUBLIC POLICY

Centre for **E**conomic **P**olicy **R**esearch

www.cepr.org

REFORMING INCENTIVE SCHEMES UNDER POLITICAL CONSTRAINTS: THE PHYSICIAN AGENCY

Gabrielle Demange, DELTA, Paris and CEPR
Pierre-Yves Geoffard, DELTA, Paris, IEMS (Lausanne) and CEPR

CEPR Discussion Paper No. 3589

October 2002

ABSTRACT

Reforming Incentive Schemes Under Political Constraints: The Physician Agency*

The present Paper investigates political constraints that may impair attempts to reform payment schemes of a given profession. When the good produced (e.g. health care) is imperfectly observable by the payer (e.g. health insurance), asymmetries of information limit the possibility to base payments upon outcome (e.g. quality of care), and payment schemes must be based on some verifiable input (e.g. number of acts). The model is applied, but not restricted to, the physician agency. Political constraints are defined as the necessity to obtain the consent of a large proportion of providers to a given reform of their reward schemes. Second-best efficient reforms, which take into account the welfare cost of such political constraints, induce additional spending due to an excessive quality of outcome. More strikingly, no reform, which imposes to shift away from current payment schemes, may be feasible when practice is highly heterogeneous, and the proportion of producers who need to agree to a reform proposal is large. Since heterogeneity of producers' practice is a key issue in terms of reforms' acceptability, we also study whether a menu of contracts may be a way to alleviate the political constraints. In most cases, this requires the introduction of a 'quality compensation' scheme that compensates for quality variations across different competing contracts.

Gabrielle Demange
DELTA
ENS
48 Boulevard Jourdan
75014 Paris
FRANCE
Tel: (33 1) 4313 6327
Fax: (33 1) 4313 6310
Email: demange@java.ens.fr

Pierre-Yves Geoffard
DELTA
ENS
48 Boulevard Jourdan
75014 Paris
FRANCE
Tel: (33 1) 4313 6319
Fax: (33 1) 4313 6310
Email: geoffard@delta.ens.fr

1 Introduction

The last twenty years have witnessed, in many developed countries, several attempts to reform some professions whose payment scheme is tightly regulated. In some cases, these attempts have failed, following strong opposition by members of the profession under consideration for reform. Examples include train drivers (especially in France), air traffic controllers, and most notably medical professions (across European countries). Most often the crucial importance of the service together with the difficulty to observe the quality provided by these professions has justified regulation. Regulation has however also given them a strong political power (that in some cases may be found to be disproportionate in regards to the present situation) and a very high status quo position. The aim of this paper is to investigate how the political power and the status-quo position of a profession combines with the impossibility to contract on outcomes to impose constraints on the reforms of payment schemes. We focus the analysis on physicians, but believe that the model can be applied to some other professions as well.

Most of the proposed reforms in health insurance and health care systems originated by governments, or agencies in charge of the health sector (General Practitioners FundHolders in Great Britain, Dekker plan in Netherlands, Seehofer reform in Germany, Juppé plan in France...). In the U.S., an important organisational change (the development of Managed Care) was mainly an attempt by the insurance sector (both private and public, through Medicaid and Medicare programs) to address the challenge of ever-growing medical expenditures.

Although many goals are assigned to these reforms, one of them is to improve efficiency by including ingredients aimed at reducing the gap between private and social costs, either on the patient side (introduction of deductible and/or co-payment), or on the providers side (supply-side cost sharing). Indeed, since health expenditures are largely covered by public or private insurance in all developed countries, the standard trade-off between risk-sharing and incentives (e.g., Pauly, 1968) apply: at least *some* consumption decisions are taken by individual agents (doctors or patients) who do not face the total marginal cost of these decisions, and may therefore be inefficient from an *ex ante* perspective.

Our objective is to analyse the difficulties that countries have encountered in reforming physicians' incentives. Two kinds of policies have been attempted. A first policy modifies the payment schemes faced by the physicians. In countries where fee-for-service payment schemes prevail (e.g., in

France), reform proposals often include a move towards capitation contracts[1] as a way to moderate expenses; in contrast, in countries where capitation prevail (e.g., in the U.K.), some proposals have suggested to increase the share of payments based upon observed activity, as a way to improve quality of service. Both proposals implicitly assume that physicians adapt their behaviour to the financial incentives they face; they only differ in terms of initial situations.

A second policy, more drastic, introduces some form of managed competition among various institutions (private insurance companies in the Netherlands or in Switzerland, district health authorities or group practices in the UK). These institutions often have the possibility to contract with some providers of care, hospitals and physicians.

Each of these policies has been the object of vivid debates, both within the public opinion of the respective countries, and within the academic community, about the effects of such measures on cost and quality of service, and on equity of the health system. A common feature across countries is the reluctance of health care professionals to accept the introduction of some form of competition, and this reluctance also applies, more generally, to moves towards incentive-based payment schemes[2]. Given the political power of health care professionals, this reluctance has often compromised the implementation of such reforms[3].

Our aim is to analyse these difficulties through a simple model. All participants -the health authority, patients, and physicians- are concerned with the quality of the service. A potential conflict between reducing costs and guaranteeing quality level nevertheless arises in an environment where neither the quality of the service provided by the physician nor the health status of his patients are observed. Whereas the outcome of the physician's act, which is patients health (or health improvements due to care), is not

---

[1] In the realm of Managed Care, capitation contracts are standard in IPA or network HMOs, whereas physicians are paid on wage in Group/Staff HMOs.

[2] 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. 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.

[3] 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 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.

observable, *some* input in the production function is observable. We call this input "time", typically measured by the number of acts Physicians are paid through contracts, or reward schemes, according to the time spent and not according to the (not observable) quality they offer. Assuming that physicians, even if not pure profit maximisers, adapt their practice to payment schemes[4], payment schemes affect the quality of care, and welfare.

We model political constraints by assuming that any reform of payment schemes, to be feasible, must receive the agreement of a sufficiently large proportion of physicians. A reform is contemplated in a given institutional context, in which the initial situation, the status-quo, is given by a unique payment scheme, faced by all physicians. We investigate whether politically feasible reforms may improve consumers welfare.

We first look at a modification of the current payment scheme, which applies to all physicians. We show that any change from the status quo imposes to distribute additional rents to the physicians; the total cost of these rents may outweigh efficiency gains. The situation may even be blocked at the status quo. This case is more likely to occur, the more heterogeneous physicians practice is, and the less physician practice responds to financial incentives.

We then look at the introduction of several health organisations, each one characterised by a payment scheme offered to physicians, and a premium paid by patients/consumers. Both physicians and patients can freely choose among these organisations: physicians self-select themselves across health organisations, which determines the trade off between quality and cost for patients. Furthermore, in order to get the agreement of physicians, the status-quo payment scheme is still offered by one health organisation. We investigate two polar cases: unregulated competition between health organisations, and a regulation that imposes uniform premiums. Typically, in these two cases, no gain of consumers welfare is to be expected from the introduction of competing contracts. Such a reform may be valuable only if adequate cross subsidies are implemented, under the form of a "quality compensation" scheme that compensates for differences in cost and quality across health organisations.

Section 2 introduces the model of physician agency in which quality of service (health improvement) is not contractible, but some input (time)

---

[4]A large part of the empirical literature in health economics has focused on the so-called "supply-induced demand." Even though the conclusions of this literature are ambiguous (Phelps, 1986), it is quite clear that physician's behaviour responds to monetary incentives. See, e.g., Fuchs (1978), Gruber and Owings (1996), Delattre and Dormont (2000) among others; McGuire (2000) also provides many relevant references.

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 analysed. Section 5 studies how the introduction of competing payment schemes, together with an adequate regulation, may soften political constraints and improve patients welfare.

2 The market for care and health insurance

We first describe the main features of our market for care, then study how income and incentives to provide quality are affected by the reward scheme a physician is facing. This allows us to derive physicians preferences over alternative payment schemes, which have implications on the allocation of resources when political constraints matter.

2.1 The market for care

2.1.1 The physician-patient relationship

A patient who suffers from an illness episode meets a physician; the outcome of the service (the "quality") is the patient's health after intervention. We denote it by l. Quality of care is a function of t, the number of acts provided by the physician, with positive and decreasing returns. We interpret this variable as the physician "time" but it may be thought of as any input variable in the quality production function that can be contracted upon[5]. Ma and McGuire (1997) argue that even quantity of treatment is costly to observe by the insurer, and reports may not be truthful. We abstract from this consideration: in this paper, t denotes any variable that enters the production function and that the insurer can observe at no cost.

Quality is also affected by two exogenous parameters. The first one is the magnitude of the health shock the patient has been subject to, denoted by θ. A higher θ indicates a more severe patient, and therefore l is assumed to be decreasing with θ. The second parameter on which quality depends is a characteristic β that is specific to each physician. It measures the ability to reach a diagnosis and a prescription in a given amount of time. We call β the *talent* : the more talented the physician, the less time he needs to offer

---

[5]The assumption that quality itself is a non-contractible input in the health production function is key to the "physician agency" literature, as reviewed, e.g., by McGuire (2000). Also, 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.

a given quality. 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:	$l_\beta > 0.$

Throughout the paper, we shall assume that a physician knows his characteristic; for short a $\beta-$physician denotes a physician with characteristic β.

For each patient, the effort 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 he is facing (this is precised later on). The time supply across physicians depends on how the marginal gain in quality l_t varies with the talent. We shall assume a Spence-Mirrlees condition which ensures that l_t is monotone in talent. More precisely, we shall assume that talent and time are either substitutes or complements, where

β and t are *substitutes* if $l_{t\beta} < 0$, and *complements* if $l_{t\beta} > 0$.

We make the following distributive assumptions: The physicians' types are distributed on an interval $[\underline{\beta}, \overline{\beta}]$, with a cumulative distribution denoted by F, which admits a strictly positive density f. *Ex ante*, all patients are identical and we denote by G the "*case-mix*," i.e. the distribution of θ. Both the total number of doctors and patients are normalised to one ($F(I)$ denotes the number of physicians whose type belongs to interval I).

2.1.2 Physicians preferences

The time spent by a physician costs him $w(t)$, where w' is the marginal opportunity cost of time, and is non-decreasing with t. The quality of service also enters the objective function of the physician, with a weight α. 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 pass the information among other patients. In both cases, a lower quality diminishes the physician future income (Pauly, 1980: Rochaix, 1989; Dranove, 1988)[6]. Finally preferences are additive in money. Therefore, the profit of a physician of type β facing the health status θ, spending t units of time, and

---

[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.

receiving m is given by[7]:

$$m - w(t) + \alpha l(t; \theta, \beta)$$

Here, m is a monetary transfer received from a payer, that we define below as a "health organisation".

2.1.3 Patients preferences

Patients are all *ex ante* identical[8]. They consult a physician whenever they "need", i.e. when they are subject to a health shock θ large enough. Their preferences are additive in quality and money, with a weight λ for quality. So, *ex ante*, before the occurrence of a health shock θ, the expected utility derived from services provided by a physician of type β whose supply time function is $t(., \beta)$ is given by:

$$E_\theta \left[\lambda l(t(\tilde{\theta}, \beta), \tilde{\theta}, \beta) - p \right],$$

where the expectation is taken over the distribution of θ ([9]), and p is the total monetary payment associated with health care. Since we study a situation in which the payer is a health organisation, we will assume that p is a fixed payment that takes the form of a premium, paid ex ante (before the realisation of θ, and not observing physician's type β or time spent t on a specific act of care).

2.2 Health organisation

The monetary payment paid to the physician by the health organisation (the *HO*, which may, or may not, be a private health insurance firm) is function of the observable variable t only. Accordingly, a *reward scheme* is specified by a function of t, $R(t)$, which is assumed to be continuous, differentiable and

---

[7]The parameter α, which represents the concern for quality, is assumed here to be identical across physicians. 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. The incentives analysis goes through (note that concern and time are complements ($l_{t\beta} > 0$), but the welfare analysis has to be modified.

[8]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., Dranove, Ellis and Mac Guire, Ma (1994) and Newhouse (1996)), nor demand side moral hazard issues (for a recent survey on patient demand, see Zweifel and Manning (2000)).

[9]To emphasise uncertainty, we denote a random variable by $\tilde{x}$, and its realisation (when observed) by x.

concave. Actually, we mostly 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. A capitation (prospective) contract is associated with a flat scheme $a = 0$, and a purely retrospective scheme with $b = 0$ and $a > 0$.

Until section 5, we assume that the health organisation is a unique non-for-profit (or public) firm, that operates at no cost: it simply collects premiums p and pays monetary transfers R to the physicians. This yields the zero-profit condition :

$$p - E_{\theta,\beta} R(t(\theta, \beta)) = 0.$$

2.2.1 Physician's optimal time

Facing a reward scheme R, physician's optimal time, which is taken by a $\beta-$physician after he observes both the health status θ, maximises[10] preferences where $m = R(t)$, i.e. it solves:

$$\max_t [R(t) - w(t) + \alpha l(t; \theta, \beta)].$$

Thanks to decreasing returns in time and increasing marginal costs assumptions, the objective is concave with respect to t. Hence, the optimal time is unique, characterised by the first order condition:

$$R'(t^*) = w'(t^*) - \alpha l_t(t^*; \theta, \beta), \qquad (1)$$

which equates the marginal revenue to the marginal *net* cost, including the concern for quality. We shall denote it by $t^*(R; \theta, \beta)$.

Increasing the marginal reward for all t (increasing R') shifts the marginal benefit of t upwards; it makes the scheme *more powerful* in the sense that it increases incentives to spend more time and therefore to improve quality[11].

---

[10] Under standard continuity assumptions, the supremum is reached.

[11] 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 incites the hospital to minimise 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 R', in contrast with the "target income" hypothesis (Fuchs, 1978).

2.2.2 Preferences over reward schemes

Preferences over reward schemes depend on the information that is available at the time they are compared. We analyse the case where a physician compares different reward schemes knowing his own characteristic β. Assuming that physicians are risk-neutral, preferences are therefore described by:

$$V(R;\beta) = E_\theta[R(t^*(R;\tilde{\theta},\beta)) - w(t^*(R;\tilde{\theta},\beta)) + \alpha l(t^*(R;\tilde{\theta},\beta);\tilde{\theta},\beta)] \quad (2)$$

the expectation being taken with respect to the distribution of the health status.

The indirect utility V plays a central role in the study of political support for reforms. Standard participation constraints or, as we shall define below, political constraints, will impose a minimal utility level $V(R;\beta)$ for all, or for a subgroup of, physicians. We call such constraints *agreement constraints*.

3 First best allocation

To assess the impact of informational and agreement constraints, we first analyse first best allocations as a benchmark, and study their implementation through contracts.

In terms of total welfare, since both patients and physicians preferences are linear in money, and the HO is simply a pay-through, monetary transfers cancel out. In contrast with standard regulation theory, we assume that monetary transfers do not induce distorsions costs; therefore, *ex ante* welfare only depends on the time spent by physicians[12], as described by the function $t(\theta,\beta)$, and is equal to:

$$
\begin{aligned}
W &\equiv \underbrace{\left(\int \lambda l - p\right)}_{\text{patients}} + \underbrace{\left(p - \int m\right)}_{\text{HO}} + \underbrace{\left(\int m + \alpha l - w\right)}_{\text{physicians}} \\
&= \int [(\lambda + \alpha)l(t(\theta,\beta),\theta,\beta) - w(t(\theta,\beta))]dG(\theta)dF(\beta). \quad (3)
\end{aligned}
$$

A first best allocation of time and money maximises the welfare criterion W as given by (3). The solution is simply obtained by maximising over t, for each (θ,β), the surplus

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

---

[12]We do not consider the case where patients are affected in a non random way to physicians.

This gives the optimal time t^{FB}, characterised by the first order condition :

$$(\lambda + \alpha)l_t(t; \theta, \beta) - w'(t) = 0$$

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

This optimal time t^{FB} is a function of (θ, β). An immediate question is whether this allocation can be implemented through appropriate payment schemes. Comparing with (1) that determines the time allocated by a physician under a given payment scheme, we readily obtain the following result:

Proposition 1 *Any scheme R that satisfies*

$$R'(t) = w'(t) \left(\frac{\lambda}{\alpha + \lambda} \right) \tag{4}$$

leads physicians to choose the optimal time t^{FB}. We call such a scheme first best optimal. *If the cost of time is linear, optimal schemes are linear, with an identical fee for service $a^{FB} = w(\frac{\lambda}{\alpha+\lambda})$.*

In our model, if there are no agreement constraints, a first best allocation can be obtained through a unique scheme: *there is no reason, on efficiency grounds, to discriminate among the physicians.* If agreement constraints prevail, but individual types are perfectly observed (for example if the group of physicians implements some form of peer monitoring) first best allocations may also be reached: it suffices to design flat payments so as to give to each physician of type β his minimal required utility level, without any welfare cost since monetary transfers are neutral.

This paper considers the situation where the schemes cannot be made contingent on characteristics and agreement constraints prevail, and studies the effects, both in terms of welfare and quality of care, of the political power of physicians, as reflected by agreement constraints.

Remark. Let us mention a case where this power does not prevent efficiency. Assume that a unique group, of which all doctors must be members, negotiates with the health authority, and may accept or reject a given payment scheme[13]. If the group has the power over its members to impose participation, the contracts that are accepted need to satisfy a unique agreement

---

[13]Examples of such professional institutions may indeed be found within the medical profession: in France, *Conseil de l'Ordre* plays such a role. In the U.S., Kessel (1958) recalls that the *American Medical Association* has been constantly involved in regulatory issues such as numerus clausus, etc. In health care, monopolistic professional organisations are mostly in charge of quality certification, and may coexist with several more standard unions, that defend physicians interests.

constraint that bears on a "representative" physician: $E_\beta[V(R;\tilde{\beta})]$ needs to be larger than some given level, for example larger than the average utility level of the current situation. It easily follows from Proposition 1 that[14] *a reform that maximises the consumer surplus, keeping the average expected utility of physicians constant, is first best optimal.* The possibility for the union to act as a representative physician, is, however, a questionable assumption.

4 Political constraints

At the time a reform is contemplated, each physician knows his own characteristic, and it is not clear at all why he should feel concerned with the average expected utility of his colleagues. The political process leading to union decisions may well impose important distortions. Indeed, whether it is channeled through a monopolistic union is not central to that question: very generally, any reform must be accepted by a sufficiently large number of physicians to be adopted.

4.1 Political support for reform

We consider here the situation where a reform modifies the a current contract, also called the status-quo, denoted by R^0, and each physician evaluates whether he will lose or gain from the reform. A reform is defined to be "politically feasible" if it is accepted by a large enough proportion of the physicians:

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.

Formally, political constraints write as:

$$F[\beta, s.t. V(R;\beta) \geq V(R^0;\beta)] \geq q.$$

If unanimity is required, i.e., $q = 1$, each physician must get at least the utility derived from status-quo[15]. The question is whether a reform that simultaneously makes consumers better off and is politically feasible exists.

---

[14]A first best optimal scheme maximises the sum of the consumers' surplus and the expected physicians' utility over all physicians. Since only its slope is determined, it suffices to adjust $R(0)$ so as to satisfy the average agreement constraint on $E_\beta[V(R;\tilde{\beta})]$.

[15]Therefore, if $q = 1$, political constraints amount to participation constraints, for all $\beta, V(R;\beta) \geq \underline{v}(\beta)$, with a type dependent reservation utility level $\underline{v}(\beta) = V(R^0;\beta)$. In general however political and participation constraints differ.

To answer this question, an analysis of physicians preferences over reward schemes is needed.

4.2 Physicians preferences over reward schemes

Which physicians prefer more powerful reward schemes heavily depends on how the optimal time chosen by a physician varies with his type, which in turn depends on whether talent and time are substitutes or complements.

Proposition 2

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

2. *Let R_1 be more powerful (i.e. steeper) than $R_2 : R_1'(t) \geq R_2'(t)$ for any t. Then physicians spend more time, quality is higher if they face R_1 instead of R_2. Moreover, 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. In the complement case, doctors with a strong commitment to quality (or doctors threatened by searching patients or a bad reputation) are those who spend the more time with their patients. As for quality, a change in β has a direct positive effect l_β and an indirect effect $l_t t_\beta^*$. Whereas this indirect effect is positive in the complement case, it is negative in the substitute case and may dominate the direct effect[16].

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

4.3 Voting over linear schemes

¿From now on we restrict attention the analysis to linear schemes (recall that, according to proposition 1, linear schemes implement first best alloca-

---

[16]In the last section, we shall consider a *strong substitute* property, under which the indirect effect dominates the direct one.

tions whenever the opportunity cost of time is linear). The previous propo-sition is very powerful to describe the set of politically feasible reforms.

We first note some simplifications. Faced with a linear scheme $R(t) = at + b$, let $t^*(a; \theta, \beta)$ denote optimal time (which only depends on the fee a) and $V(a, b; \beta)$ a $\beta-$physician's utility level. Note that V is linear in the flat payment and writes

$$V(a, b; \beta) = b + (a - w)T(a, \beta) + \alpha L(a, \beta) \qquad (5)$$

where T and L are the average time and quality provided by a $\beta-$physician facing a case mix G:

$$T(a; \beta) \equiv E_G[t^*(a, \tilde{\theta}, \beta)], L(a, \beta) = E_G[l(t^*(a, \tilde{\theta}, \beta); \tilde{\theta}, \beta)]. \qquad (6)$$

Moreover, using the envelope theorem, for a physician of type β, the margi-nal rate of substitution between b (flat payment) and a (fee) is equal to the expected time. Since for any (θ, β), t^* increases with a, indifference curves are concave.

With linear schemes a reform is politically feasible if it is accepted by a physician of pivotal type that we define now.

Let (a, b) be the proposed reform. If the reform is less powerful than the status-quo, $(a < a_0)$, the *pivotal characteristic* is the characteristic β^- for which the proportion of physicians who work *less* than a β^--physician is equal to q ; similarly, if it is more powerful, it is the characteristic β^+ for which those who work *more* than a β^+-physician are in proportion q.[17] Proposition 2 straightforwardly implies[18] that

> *starting from the current situation given by (a^0, b^0), a scheme (a, b) is politically feasible if and only if it is preferred to the status-quo (a^0, b^0) by a physician of pivotal type.*

In the majority case, $q = 1/2$, the pivotal characteristic is identical whether one wants to decrease or to increase incentives to work, i.e. to decrease or to increase quality. In the more likely situation where q is strictly larger than $1/2$, the pivotal characteristics differ whether the proposal is less or more

---

[17]So in the complement case $F(\beta \leq \beta^-) = F(\beta \geq \beta^+) = q$ and in the substitute case, $F(\beta \geq \beta^-) = F(\beta \leq \beta^+) = q$, where F is the distribution of types.

[18]If the less powerful contract (a, b) is preferred by a physician with type β^-, it is also preferred by any physician who works less: it is accepted by a proportion q of physicians. Conversely, if a β^- physician prefers the status-quo, any physician who works more prefers the status-quo as well: the contract (a, b) is rejected.

powerful. These results may be illustrated in the plan (a, b). By linearity in the flat payment, all indifference curves are obtained from each other by vertical translation. 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 $q > 1/2$.

<div align="center">INSERT FIGURE 1</div>

Remark. It immediately follows from above that, if we consider a majority voting game where physicians vote on a family of linear contracts, which are indexed by a, a majority winner exists which is the contract preferred by a physician with type β^m equal to the median value of β. At this voting equilibrium contract, the median type β^m is such that more time-intensive physicians (in the substitute case, higher values of β) favour an increase in a (given the rate of substitution implied by the budget set), whereas less time-intensive physicians would prefer a larger fixed payment.

4.4 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.

Definition 1 *A (linear) politically constrained optimum is given by a contract (a^p, b^p) that maximises ex ante patient's utility $U = E'_{\theta, \beta}[\lambda l - (at + b)]$, over the politically feasible contracts:*

$$
\begin{cases}
\max U(a, b) \\
(a, b) \ s.t. \ V(a, b; \beta^c(a)) \geq V(a^0, b^0; \beta^c(a)),
\end{cases}
\tag{7}
$$

with $\beta^c(a) = \beta^-$ if $a < a^0$, and $\beta^c(a) = \beta^+$ if $a > a^0$.

Notice that patient's utility fully internalises the non profit constraint on the HO, i.e. how a given scheme affects the premium p. Indeed, the question we ask is: by how much do political constraints distort from first best efficient allocations? To analyse the impact of the constraints, it is convenient to write patients' utility as the difference between the overall welfare criterion W, which is independent on the monetary transfers, here b, and the expected utility of the physicians. In terms of variation with respect to the current situation we have:

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

Whenever the status-quo differs from a first best scheme, i;e. a^0 differs from a^{FB}, the welfare criterium W can be increased. As said previously, the overall efficiency gain can accrue to the consumers while providing each physician with the same utility level he had before the reform if flat payments could depend on β: it suffices to give a $\beta-$ physician the flat payment $b(a, \beta)$ which makes him indifferent between the new scheme $(a, b(a, \beta))$ and the status quo (a^0, b^0). The flat payment $b(a, \beta)$ is nothing but the compensating variation associated with a change in the price of time from a^0 to a. Given that physician's utility is linear in money, we have

$$V(a, b; \beta) - V(a^0, b^0; \beta) = b - b(a, \beta). \qquad (8)$$

Under non observable β however, flat payments should be type independent. Given the desired support as given by q, changing a requires the flat payment to be adjusted so as to "buy" the support of the pivotal physician β^c, that is b is adjusted to the level $b(a, \beta^c)$. So, taking the average over all physicians, one may define the informational cost as:

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

Of course if unanimity is required, $(q = 1)$, the informational cost is positive (since $b - b(a, \beta)$ for all β), but it is no longer the case otherwise. Indeed, when the proportion q is small enough, the informational cost may be negative: it may indeed be *less* costly to obtain support from, say, 40% of all physicians, rather than from a "representative" physician. For $q = 1/2$, $C(a)$ is positive if and only if the median value of b is larger than the mean.

Using this insight, changing the scheme results in

$$U(a, b) - U(a^0, b^0) = [W(a) - W(a^0)] - [b(a, \beta^c) - E_\beta b[(a, \tilde{\beta})]],$$

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

Lemma 1 *The marginal effect of changing the fee for service a on the ex ante welfare W and informational cost are given respectively by:*

$$W_a(a) = \left(\frac{\alpha + \lambda}{\alpha}\right)(a^{FB} - a)E_\beta[T_a(a, \tilde{\beta})], \qquad (9)$$

and

$$C_a(a) = E_\beta[T(a, \tilde{\beta})] - T(a, \beta^c) \text{ for } a \neq a^0 \qquad (10)$$

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, marginal cost is always negative for $a < a^0$: the pivotal physician is the one who spends the largest time, and therefore, $T(a, \beta^c) \geq T(a, \beta)$ for any β. Symmetrically, the marginal cost is positive for $a > a^0$: this corresponds to the intuition that the further away from the status-quo, the largest the information cost. Also, the more heterogeneity between time practice, the more important, in absolute terms, the marginal cost.

A second best optimum trades off marginal benefit and marginal informational cost.

Proposition 3 *Assume that W is concave in a, and C is convex in a (increasing marginal informational cost). Then two cases may occur:*
- either the situation is blocked at the status quo, $a^p = a^0$, which occurs if

$$\frac{E_\beta[T(a^0, \tilde{\beta})] - T(a^0, \beta^-)}{E_\beta[T_a(a^0, \tilde{\beta})]} < \left(\frac{\alpha + \lambda}{\alpha}\right)(a^{FB} - a^0) < \frac{E_\beta[T(a^0, \tilde{\beta})] - T(a^0, \beta^+)}{E_\beta[T_a(a^0, \tilde{\beta})]}$$
(11)

- or a^p is defined by:

$$\left(\frac{\alpha + \lambda}{\alpha}\right)(a^{FB} - a^p) = \frac{E_\beta[T(a^p, \tilde{\beta})] - T(a^p, \beta^c)}{E_\beta[T_a(a^p, \tilde{\beta})]}.$$
(12)

Note that, when a simple majority is required ($q = 1/2$), both pivotal characteristics coincide (with the median one) so that the situation is never blocked (the right and left hand sides of (11) coincide). This is not a surprise from our previous result : politically feasible reforms are those which give to the median voter at least his status-quo level so that a second best optimum maximises consumers' surplus over the median indifference curve. Remark however that the contract chosen by the median voter, say (a^m, b^m) 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 wekfare is increased at the expense of a larger loss incurred by physicians.

Whenever a qualified majority is required, the pivotal physician changes and the situation may be blocked: by construction, $T(a^0, \beta^+) < T(a^0, \beta^-)$ if $q > 1/2$. For example if unanimity is required, whereas the left hand side of (11) represents the spread between the average practice time and the maximal one, the right hand side gives the spread with the minimal one. Therefore, the situation is blocked if the current situation is close enough to the first best payment scheme, the weight of patients in the welfare criterion,

λ, is small, and if the expectation of T_a is small, which is the case if physician practice time is not very responsive to financial incentives.

If the situtation is not blocked, the terms of the trade-off, summarised in equation (12), depend on the elasticity of physician's time practice relative to the heterogeneity of their time. 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.

5   Menu of contracts

Instead of considering a new scheme which would apply to each physician and replace the current scheme, we investigate in this section whether the possibility of several schemes within which physicians may choose facilitates a welfare improving reform. A new scheme is introduced in addition to the existing scheme, as occurs if a new health organisation is set up. If physicians freely choose between health organisations, they should not object against an additional scheme, which simply enlarges their choice. However the analysis should be conducted in an equilibrium set up, taking into account the reaction of both physicians and patients.

5.1   Stable organisations

A health organisation (HO) proposes a contract to physicians, and asks for a premium to patients who subscribes to the organisation. The premium is independent of the patient's health condition, which is unknown at the time of the subscription. Health organisations are indexed by k, with HO^k characterised by (a^k, b^k, p^k) where (a^k, b^k) is the contract offered to physicians, and p^k is the premium.

We are interested in stable sets of contracts under (some form of) rational expectations. Patients and physicians freely choose between health organisations without rationing. Since *ex ante* patients are all identical they must be (*ex ante*) indifferent between any health organisation. So it is natural to assume that patients are distributed randomly to the different organisations. Random matching implies that the distribution of θ for the patients subscribing to any HO is identical, equal to the prior distribution G of health status θ.

Each physician selects one contract, knowing the distribution of the case-mix in each HO, which, under random matching, is identical across all organisations, equal to G. Therefore physicians' preferences over contracts

are still represented by the function $V(a, b; \beta)$, as defined by (5) and (6). Hence the subset of physicians who choose contract k is given by those with characteristics in

$$H^k = \{\beta | V(a^k, b^k; \beta) \geq V(a^j, b^j; \beta) \, \forall j\}. \tag{13}$$

Note that, typically, the set of physicians who are indifferent between two distinct contracts is negligible. So the proportion of physicians who register with HO^k is given by $F(H^k)$. We denote by N^k the number of patients subscribing to HO^k. In line with the assumption that the case mix is identical across HOs, we assume that the number of patients per physician is identical across HOs, and therefore that

$$N^k = F(H^k). \tag{14}$$

Patients do not observe directly the physicians' types. However, when doctors self select by registering to a given organisation, patients are assumed to infer information correctly from physicians choice. Therefore, the utility of a patient subscribing to HO^k is given by:

$$u^k = E[\lambda L(a^k, \tilde{\beta}) | \tilde{\beta} \in H^k] - p^k \tag{15}$$

Finally total spending of HO^k are equal to the total rewards distributed to the physicians :

$$D^k = \int_{\beta \in H^k} [a^k T(a^k, \beta) + b^k] f(\beta) d\beta$$

where it is used again that the distribution of patients in a HO is the prior distribution G, which gives the profit of HO^k:

$$\pi^k = p^k N^k - D^k.$$

Definition 2 *A stable organisation is defined by a set of K different health organisations $\{(a^k, b^k, p^k), k = 0, \ldots, K - 1\}$ such that:*

1. *(physicians choice) H^k is given by $\{\beta | V(a^k, b^k; \beta) \geq V(a^j, b^j; \beta) \, \forall j\}$, and is non empty,*

2. *(patients choice) u^k independent of k where u^k is given by (15), and $N^k = F(H^k)$*

3. *(feasibility) $\sum_k \pi^k \geq 0$.*

This definition applied to the unique contract (a^0, b^0) imposes only random matching, and a premium at least equal to the average expenses, as was considered in the previous section. Otherwise it simply states conditions on the contracts which ensure that all organisations are active under free choice for both patients and physicians, without specifying the form of competition between the health organisations. It leaves open how the premiums p^k are chosen, and it should be clear that we could as well say that several contracts are proposed by the same organisation.

By construction, no physician is worse off at a stable configuration in which the initial contract is still proposed. Two questions are raised:

1. given the initial situation, does there exist a set of additional contracts that would be potentially welfare improving for the consumers ?

2. if additional contracts are potentially welfare improving, can they be implemented ?

5.2 Patients welfare

To answer the first question, we evaluate how total consumers welfare is affected by the introduction of new contracts. Let $\{(a^k, b^k), k = 0, \ldots, K - 1\}$ be a menu of K contracts, and assume random matching. As previously H^k denotes the set of physicians who opt for the contract (a^k, b^k). Under feasibility the sum of the customers utility levels is equal to the total social welfare minus physicians utility:

$$U = \sum_k \left[\int_{\beta \in H^k} W(a^k, \beta) dF(\beta) - \int_{\beta \in H^k} V(a^k, b^k, \beta) dF(\beta) \right],$$

where $W(a, \beta) \equiv (\lambda + \alpha) L(a, \beta) - wT(a, \beta)$ is the expected social surplus of a relation between a patient with a $\beta-$physician. Similarly in the initial situation:

$$\bar{u} = \int_\beta W(a^0, \beta) dF(\beta) - \int_\beta V(a^0, b^0, \beta) dF(\beta).$$

It is convenient to use again the compensating variation $b(a^k, \beta)$ according to which $V(a^k, b^k, \beta) - V(a^0, b^0, \beta) = b^k - b(a^k, \beta)$. Since physicians partition themselves into the sets H^k, we get

$$U - \bar{u} = \sum_{k \neq 0} \left[\int_{\beta \in H^k} [W(a^k, \beta) - W(a^0, \beta)] dF(\beta) - \int_{\beta \in H^k} [b^k - b(a^k; \beta)] dF(\beta) \right]$$

$$(16)$$

As long as the initial contract is still proposed, each physician is at least as well off when additional contracts are introduced: $b^k - b(a^k, \beta)$ must be nonnegative for a β−physician who chooses contract (a^k, b^k), i.e. for β in H^k. It follows from (16) that the average patient's utility can be increased only if the total expected welfare increases. It is however not sufficient since physicians get a positive rent (which is precisely equal to $b^k - b(a^k, \beta)$, for β in H^k). To fix the ideas, take $K = 2$. By (16), $U - \overline{u}$ is equal to the variation of the surplus over the physicians who choose H^1; diminished by the rent to leave to these physicians; let β^c be the characteristic of the physician who is just indifferent between the two contracts: $b^1 = b(a^1, \beta^c)$. If $a^{FB} < a^0$, then $a^1 < a^0$ is needed to increase total welfare, and it increases social surplus for each $\beta : W(a^1, \beta) > W(a^0, \beta)$ holds whatever β. Therefore, increasing the size of H^1 increases welfare gains; this is done however at the cost of a higher rent left to physicians, since increasing the size of H^1 requires a higher value for b^1. The net effect on patients welfare is ambiguous.

5.3 Stable organisations: implementation

This section first shows two results: first, if regulation takes the form of a uniform premium independent of k, no stable organisation exists under general assumptions; second, if unregulated (i.e., "perfect") competition between HO drives profits to zero, then no stable organisation improves welfare with respect to the status quo in the substitute case. Actually the intuition is quite clear. A given menu of contracts determines a partition of the physician set (assuming random matching); for each H^k, total spending D^k as well as average quality $L^k \equiv E[L|H^k]$ are determined. Therefore, patients who register with H^k get a net utility $u^k = \lambda L^k - p^k$. Since at equilibrium, u^k must be equal across HO, differences (if any) in premium must perfectly match differences in quality. A first consequence is that, if the constraints on the premiums are too strong, no organisation except the status-quo is stable. A second consequence is that, as long as the initial contract (a^0, b^0) is still active, the introduction of additional contracts will modify the utility of patients who remain with H^0: since some physicians self-select themselves out of H^0, the average quality over H^0 is modified. If the associated change in premium (which depends on how competition is regulated) does not compensate the changes in quality, then the net utility of patients with H^0 decreases with respect to the initial situation, and no stable organisation is Pareto improving.

5.3.1 No price discrimination

We first study the case in which some regulation prohibits any price (premium) discrimination across different HO. For instance, this is the case 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.

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.

As said above, when premiums are independent of k, the average quality must also be independent of k at a stable organisation. Next proposition shows that this is not possible in general.

Before stating the result, first recall that faced with a scheme, the quality offered by a physician is affected by his characteristic though a direct effect and his behaviour. Whereas both effects are positive if talent and time are complements, it is not true if they are substitute. Talent and time are said to be *strong substitutes* if the marginal rate of substitution l_t/l_β increases with time t, for any value of (t, θ, β). Under strong substitutes, the quality offered by physicians faced all with the same scheme is decreasing with talent β: the decrease in quality induced by the reduction of time practice by more talented physicians outweigh the increasing effect due to their talent[19].

Proposition 4 *In the complement case or in the mild substitute case, or in the strong substitute case, there is no stable organisation for which several contracts are active.*

PROOF. Let two different contracts be chosen. We claim that in the complement case, the average quality provided by the physicians who choose the more powerful contract is strictly higher than the average quality provided by the physicians who choose the other one. The proof is straightforward. Take two active contracts with $a^k < a^j$. In the complement case the physicians who choose contract a^j, have a higher characteristic than those who

---

[19]To see this, note that the derivative of quality with respect to β is equal to the sum of $l_\beta + l_t t_\beta^*$. deriving the first order equation $w - a = l_t$ yields $t_\beta^* = -l_{tt}/l_{t\beta}$. It follows that the sign of $l_\beta + l_t t_\beta^*$ is of the opposite of $l_\beta l_{tt} - l_{t\beta} l_t$, i.e. the sign of the derivative of l_t/l_β. Notice that in the complement case, since l_β increases with t and l_t decreases with t, we have that l_t/l_β decreases with t.

choose a less powerful contract, a^k. Since faced with a given contract, the former work more than the latter, both effects go in the same direction. Formally, for any $\beta^j \in H^j$ and $\beta^k \in H^k$, we know that $\beta^j > \beta^k$. Moreover we have for any θ: $t(a^j, \beta^j, \theta) > t(a^k, \beta^j, \theta) > t(a^k, \beta^k, \theta)$. Since quality increases both with β and time practice, it is higher in H^j than in H^k.

A similar argument works in the strong substitute case: quality can be strictly ordered across HO, with a lower quality associated to a lower fee a. ∎

The proposition states that, if two different contracts at least are chosen, some customer price differentiation is required. The intuition is straightforward: in the complement case, consumers can perfectly rank doctors 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.

In the substitute case, the characteristic and the time spent by the physicians both decrease with the chosen fee for service a. However, since quality increases with the characteristic, quality is not necessarily monotonic in the chosen contract, and a stable organisation may only exist if the two effects perfectly offset each other, i.e. if $E[L|H^k] = E[L|H^0]$, and this is impossible in the strongly substitute case.

In all cases, a set of contracts determines a partition of the set of physicians; this implies that the average quality of service is different in each HO. Since patients are identical at the time they choose a HO, their trade-off between quality and premium is the same. Therefore, the premium must be adapted to compensate exactly the change in quality associated with various contracts.

5.3.2 Unregulated competition

A second case of interest is the situation in which competition between HO is not regulated: they can compete both in price and quality. We assume that, under this unregulated competition between HO, profits π_k are driven to zero[20]. In that situation, each customer pays a premium equal to the average expenses of the HO to which he subscribes:

$$p^k = D^k/N^k = E[a^k T(a^k, \beta) + b^k | \beta \in H^k],$$

and the feasibility condition (3) is surely satisfied.

---

[20]Whether this zero profit condition holds at a Nash equilibrium of a game in which firms compete in contracts is left for future research.

In section 4, we analysed reforms that consisted in deviations from the status-quo contract (a^0, b^0). In line with this analysis, we may wonder whether the introduction of new contracts (a^k, b^k), *in addition* to the existing contract (a^0, b^0), may lead to a feasible welfare improving allocation. The next proposition shows however that, in the substitute case, "fair" premiums do not compensate the changes in quality associated across contracts.

Proposition 5 *Let (a^0, b^0) be the status quo contract, with $a^0 > a^{FB}$. Assume that perfect competition drives profits to zero in each HO, and that H^0 is still active. Then, in the substitute case, the introduction of one or more additional HO^k with $a^k < a^0$ lowers patients welfare.*

PROOF. At a stable organisation, patients net utility is independent of k. Therefore, if it increases, it does so for patients who remain with H^0, which is supposed active. When profits are driven to zero, changes in net utility for patients who remain in H^0 are only attributed to a physician self-selection effect: it is the average, taken over all $\beta \in H^0$, of:

$$\lambda L(a^0, \beta) - a^0 T(a^0, \beta) - b^0,$$

whereas before the introduction, their net utility was the average of the same quantity, but taken over *all* physicians since b^0 is left unchanged by the introduction of additional health organisations. This leads to study the selection effect on $\mu(a^0, \beta) \equiv \lambda L(a^0, \beta) - a^0 T(a^0, \beta)$. In the substitute case, physicians who remain with H^0 are the lowest β. Given that $L(a, \beta) = E_\theta[l(t^*(a, \tilde{\theta}, \beta); \tilde{\theta}, \beta]$, straightforward algebra leads to:

$$\begin{aligned}
\mu_\beta(a^0, \beta) &= \lambda L_\beta - a^0 T_\beta \\
&= \lambda E_\theta[l_\beta]. + \left(\frac{\lambda + \alpha}{\alpha}\right)(a^{FB} - a^0) T_\beta
\end{aligned}$$

The first effect is a direct quality effect, and is positive. The second effect is the indirect effect on the trade-off between quality and cost. In the substitute case, T is decreasing in β and therefore, with $a^{FB} < a^0$, we have that $\mu(a^0, \beta)$ increases with β. Hence, the mean of $\mu(a^0, \beta)$ over H^0 is lower than the mean over all β, and net utility decreases after the introduction of (at least a) competing contract. ∎

The interpretation is simple. If a^0 is larger than a^{FB}, physicians work "too much": patients are not willing to pay for the cost of the time needed to provide the given quality. But the introduction of an additional contract makes things actually worse for patients who subscribe to H^0: physicians

who remain with H^0 are those who have the most intensive practice, and therefore cost more than the average taken over all physicians; under unregulated competition, this higher cost translates into a higher premium. Moreover, in the substitute case, physicians who leave H^0 are those with high β and higher associated quality. Both effects go in the same direction: they would lower the utility of patients who would remain with H^0. Notice that in the complement case, the direct selection effect on quality goes in the opposite direction than the indirect effect[21], and physicians self-selection may indeed improve patients utility.

5.4 Quality compensation schemes

As the previous sections made clear, when patients may choose between several health organisations, imposing additional constraints on a stable configuration such as a uniform premium or zero profit, may prevent any welfare improvement. Therefore, monetary transfers across health organisations are needed for the existence of a welfare improving reform. In some countries with a competitive health insurance sector (e.g., Switzerland), risk compensation scheme are implemented to account for differences in the risk profile of *subscribers*. In our context, monetary transfers should take the form of a risk compensation scheme, aimed at compensating health organisation for differences in the doctors type (and therefore in average quality). We call such a scheme a *quality compensation scheme*.

Formally, if S^k is the (positive or negative) transfer received by HO^k, the premium satisfies $p^k = \frac{1}{N^k}[D^k + S^k]$. The sum of the patient's utility levels is equal to:

$$U = \sum_k N^k u^k = \sum_k \left\{ \int_{\beta \in H^k} [\lambda L - a^k T] - N_k b^k \right\} - \sum_k \pi^k \qquad (17)$$

At a stable organisation the expected utility of a consumer must be equalised across the HO^k, so (17) also gives the utility level of each consumer. A quality compensation scheme is a set of cross-subsidies that makes any patient indifferent between all active HO. Under such a scheme, the utility level of each patient, equalised across the HO, is then equal to the average utility level derived in section 5.2. Hence, adding a new contract leads to a Pareto improving stable organisation (sustained by subsidies) if, and only if, the average net gain $U - \bar{u}$ as given by (16) is positive.

---

[21]Formally, it is not clear that $\mu(a^0, \beta)$ decreases with β.

The following proposition shows that, when the size of the additional organisation HO^1 is small enough, this net gain for patients is always positive. Given a^1, a new contract (a^1, b^1) will attract some physicians only if b^1 is sufficiently high, larger than the minimum value of $b(a^1, \beta)$: in the substitute case, if $a^1 < a^0$, the contract must be preferred by the more talented physician: $b^1 \geq b(a^1, \overline{\beta})$ and in the complement case, $b^1 \geq b(a^1, \underline{\beta})$.

Proposition 6 *Let a^1 be a welfare improving fee for service (say a^1 is between a^{FB} and a^0). If a quality compensation scheme is implemented, the introduction of an additional HO, (a^1, b^1) increases patients welfare, provided that b^1 is not too high.*

PROOF: We show that choosing b^1 higher than the minimal level but small enough does the job. Given (a^0, b^0) and $a^1 \in [a^{FB}, a^0]$, we may parametrise H^1 by β^c, the type of the physician who is indifferent between H^0 and H^1.

In the substitute case, physicians with a high value of β work less. Therefore H^1 is an interval $[\beta^c, \overline{\beta}]$; from (16), the derivative of $U - \bar{u}$ w.r.t. β^c in $]\underline{\beta}, \overline{\beta}[$ is equal to:

$$\frac{\partial(U - \bar{u})}{\partial \beta^c} = -[W(a^1, \beta^c) - W(a^0, \beta^c)]f(\beta^c) - \frac{\partial b(a^1, \beta^c)}{\partial \beta^c}(1 - F(\beta^c)).$$

For $\beta^c = \overline{\beta}$, the second term, i.e. the marginal information cost, is equal to zero, and:

$$\frac{\partial(U - \bar{u})}{\partial \beta^c}\Big|_{\beta^c = \overline{\beta}} = -[W(a^1, \overline{\beta}) - W(a^0, \overline{\beta})]f(\overline{\beta}) < 0.$$

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

The proof is similar in the complement case, except that the new HO to be small enough requires β^c to be sufficiently close to $\underline{\beta}$. ∎

Notice that the proof relies on the fact that, at the margin, the marginal informational cost to decrease β^c from $\overline{\beta}$ (or increase β^c from $\underline{\beta}$) is zero. When the size of the new HO increases, this marginal informational cost also increases, and may offset the welfare gain.

6 Conclusion

The analysis of payment schemes reforms shows that political constraints may impose strong conditions on feasible changes, when the insurer-payer

remains in a monopolistic situation. The conditions are stronger when physician practice is heterogeneous: the rent given to the critical physician may be too high for any reform to be welfare improving.

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 the solution unless it is properly regulated. The specific element that renders competition difficult to implement is that patients can choose between the different plans. Free choice, together with rational expectations, imposes strong constraints on the way a premium is related to the average quality, not only of the given plan, but also of the other plans. This difficulty directly stems from the fact that three types of "agents" intervene in the system: physicians, patients, and insurance plans. Competition gives an important role to such plans and creates some room for divergence of interests between insurance firms and patients' interests, whereas in the monopolistic case, the unique insurance firm was but a representative patient.

Much has still to be understood in the way regulated competition between insurance firms could work in this "medical triad", and provide a way to reduce the cost due to imperfectly observable medical practice. In particular, additional research needs to study Nash equilibria of a game in which insurance firms actually compete in contracts, given a transfer ("quality-compensation") scheme. Finally, if physicians could select patients, the physician-patient matching would no longer be random; this extended setup needs also to be studied.

7 Proofs

PROOF OF PROPOSITION 2.

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

2. Let R_1 be steeper than R_2. From (1) again, increasing the marginal reward for all t (increasing R'), shifts the marginal benefit of t upwards, increases t^*: so $t_1^*(\theta, \beta) \geq t_2^*(\theta, \beta)$.

Let denote by $\Delta(\beta)$ the difference in expected profit associated with the two schemes for a β physician : $\Delta(\beta) = V(R_1, \beta) - V(R_2, \beta)$ where V is given by (2) :

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

A β-physician prefers R_1 to R_2, iff $\Delta(\beta) \geq 0$. The derivative of Δ with respect to β, thanks to the envelope theorem, is given by

$$\Delta'(\beta) = \alpha E_\theta [l_\beta(t_1^*(\tilde{\theta}, \beta); \tilde{\theta}, \beta, R_1) - l_\beta(t_2^*(\tilde{\theta}, \beta), \tilde{\theta}, \beta, R_2)].$$

If talent and time are substitutes, $l_{t,\beta} < 0$, and we know that $t_1^*(\theta, \beta) \geq t_2^*(\theta, \beta)$: Δ' is negative, and Δ is decreasing. So if a β physician prefers R_1 to R_2, $\Delta(\beta) > 0$, any physician with a lower characteristic, who works more, also prefers R_1 to R_2. The proof is similar in the complement case, with Δ increasing. ∎

PROOF OF LEMMA 1:

Differentiating with respect to a, we have that $V_a(a, b; \beta) = T(a, \beta)$, and therefore:

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

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

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

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

PROOF OF PROPOSITION 3:

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

$$
\begin{cases}
U_a + \mu V_a(a, b; \beta^c) = 0 \\
U_b + \mu V_b(a, b; \beta^c) = 0
\end{cases}
$$

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

$$
\begin{aligned}
U_a + E_\beta[V_a(a, b; \tilde{\beta})] &- E_\beta[V_a(a, b; \tilde{\beta})] + V_a(a, b; \beta^c) = 0 \\
W_a(a) &- C_a(a) = 0
\end{aligned}
$$

Substituting the computed values for W_a and C_a gives the result when the optimal value for a is interior.

The concavity assumptions on W and $-C$ imply that the lagrangian defined above is concave, and hence that the second order condition is met. However, since C is not differentiable at a^0, it may be the case that the marginal cost outweighs the marginal benefit for small changes in either direction. Formally, this happens when $C_a(a^0_-) < W_a(a^0) < C_a(a_+)$. Hence the result. ∎

REFERENCES

Arrow, Kenneth (1963), "Uncertainty and the welfare economics of medical care," American Economic Review, 53: 941-973.

Bocognano, Agnès, Agnès Couffinhal, Michel Grignon, Ronan Mahieu, Dominique Polton (1998), *Mise en concurrence des assurances dans le domaine de la santé* CREDES.

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

Delattre, Eric, and Brigitte Dormont (2000), "Induction de la demande par les médecins libéraux français: Etude microéconométrique sur données de panel," Economie et Prévision, 142: 137-161.

Dranove, David (1988), "Demand inducement and the physician/patient relationship," Economic Inquiry 26 (2): 281-288

Gruber, Jonathan, and Maria Owings (1996), "Physician financial incentives and cesarean section delivery," RAND Journal of Economics 27(1): 99-123

Hassenteufel, Patrick (1997), "Les médecins face à l'Etat. Une comparaison europ?enne," Paris, Presses de Science Po.

Hugues, D., and B. Yule (1992), "The effect of per-item fees on the behaviour of general practitioners," Journal of Health Economics 11(4): 413-439

Kessel, R. (1958), "Price discrimination in medicine," Journal of Law and Economics 1: 20-53.

Ma, Albert, and Thomas McGuire (1997), "Optimal health insurance and provider payment," American Economic Review 87(4): 685-704.

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

Newhouse, Joseph (1996), " Reimbursing Health Plans and Health Providers: Efficiency in Production versus Selection," Journal of Economic Lit-

erature, 34(3): 1236-63

Pauly, Mark (1995), "Paying physicians as agents: fee for service, capitation or hybrids," in T. Abbott, ed., Health Care Policy and Regulation, Kluwer

Pauly, Mark and M. Satterthwaite (1981), " The pricing of primary care physician services: a test of the role of consumer information," Bell Journal of Economics 12: 488-506

Propper, Carroll (2002), "Do physicians respond to incentives? The case of British GPFH." Journal of Public Economics, forthcoming.

Reinhard, Uwe (1972), "A production function for physician services," Review of Economics and Statistics: 55-65

Rochaix, Lise (1989), "Information asymmetry and search in the market for physician services," Journal of Health Economics 8: 53-84
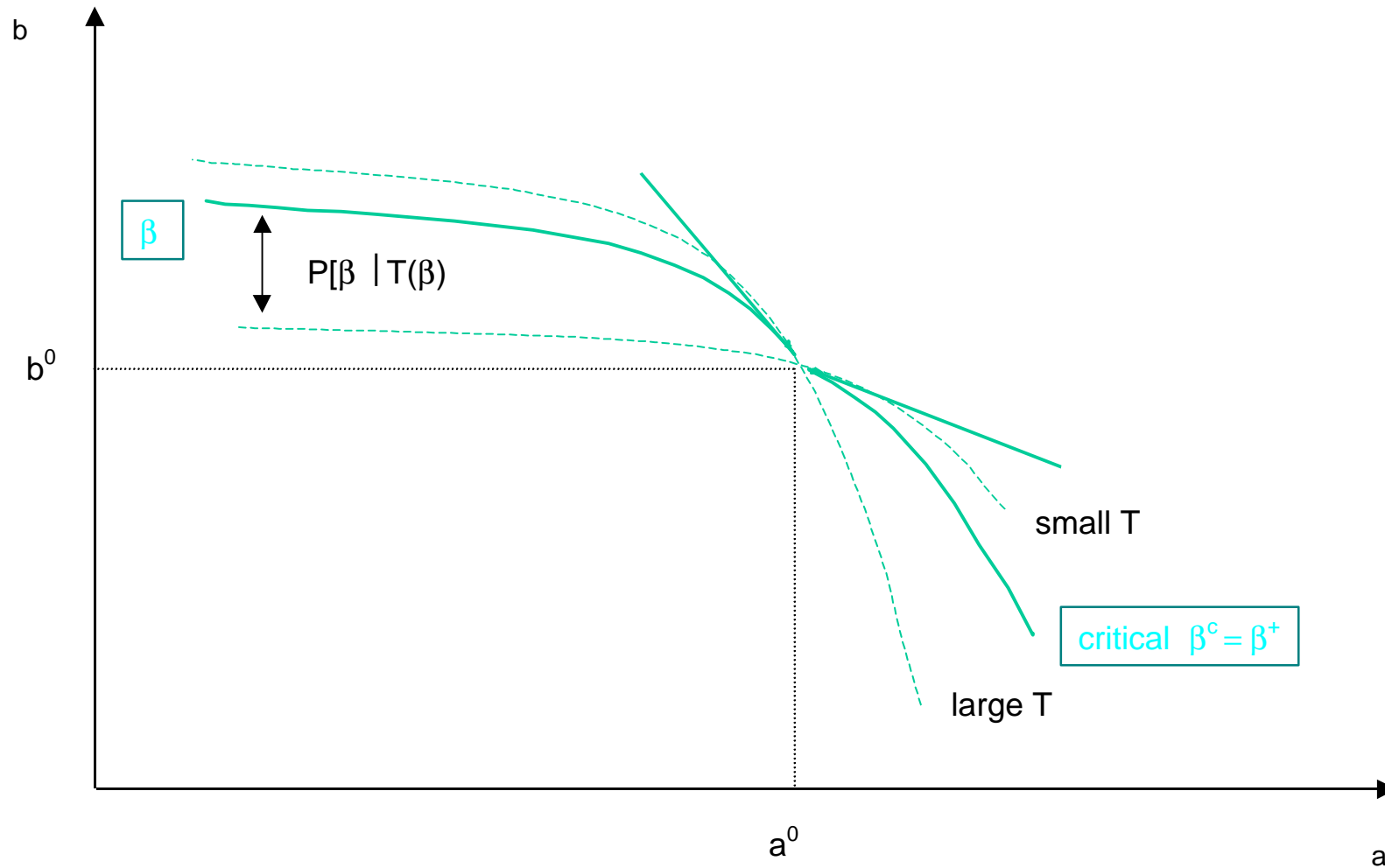
Political support for a reform

Figure 1