économiepublique

Économie publique/Public economics

09 | 2001/3
Modélisation économique et réforme des systèmes de santé

Incentive contracts and the compensation of health care providers

Marie Allard, Helmuth Cremer et Maurice Marchand

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/economiepublique/564
DOI : 10.4000/economiepublique.564
ISSN : 1778-7440

Éditeur
IDEP - Institut d'économie publique

Édition imprimée
Date de publication : 15 juillet 2002
ISBN : 2-8041-3635-3
ISSN : 1373-8496

Référence électronique
Marie Allard, Helmuth Cremer et Maurice Marchand, « Incentive contracts and the compensation of health care providers », *Économie publique/Public economics* [En ligne], 09 | 2001/3, mis en ligne le 07 décembre 2005, consulté le 12 septembre 2020. URL : http://journals.openedition.org/economiepublique/564 ; DOI : https://doi.org/10.4000/economiepublique.564

économie publique

public economics

Revue de **l'Institut d'Économie Publique**

Deux numéros par an

n° 9 – 2001/3

IDEP

Editions De Boeck Université
Rue des Minimes 39, B-1000 Bruxelles

Imprimé en Belgique

ISSN 1373-8496
Dépôt légal 2002/0074/241
ISBN 2-8041-3635-3

économie publique sur internet : www.economie-publique.fr

Centre de la Vieille-Charité

2, rue de la Charité – F-13002 Marseille

La revue économie publique bénéficie du soutien du Conseil régional Provence-Alpes-Côte d'Azur

ISSN 1373-8496

Incentive contracts and the compensation of health care providers

Marie Allard
HEC, Montreal

Helmuth Cremer
University of Toulouse (GREMAQ and IDEI)

Maurice Marchand
CORE-IAG, UC Louvain

1 Introduction

Rapid growth of expenditures on health care has made the incentive effects on provider behavior of reimbursement systems a central issue in the literature that deals with the design of compensation schemes. Starting from Ellis and McGuire (1986) it has mostly focused on the implications of fixed and cost-based payment systems on cost reduction effort and quality of care.[1] It has generally reached the conclusion that systems that mix prospective and retrospective reimbursements are the most effective in trading off optimally these two criteria. More recently, as stressed by Newhouse (1996), their effect on the selection of patients by health care providers has become a major concern. This concern has been raised e.g. in the UK following the allocation of budgets to general practitioners for purchasing selected hospital services on behalf of their patients.[2] In parallel, empirical studies have shown that the way physicians are compensated has generally some significant effect on the utilization of medical services.[3] In particular, the way in which the general practitioner's marginal per-patient compensation

[1] See, for instance, Ellis and McGuire (1990), Ma (1994), Ma and McGuire (1997) and Chalkley and Malcomson (1998).

[2] See, for instance, Matsaganis and Glennerster (1994).

[3] See, for instance, McGuire and Pauly (1991), Krasnik et al (1990) and Gruber and Owings (1996).

varies with their workload has proved to significantly affect their behavior with respect to the number of patients they find optimal to treat.[4]

This literature has generally adopted the assumption that physicians are homogeneous in all respects, ignoring that they differ in terms of productivity or preference for leisure. This assumption is not realistic : heterogeneity of physicians can indeed explain the large discrepancies in workload that are observed across physicians. By treating physicians as identical the existing literature has kept aside an important feature of the market for physician services.

Our purpose in this paper is to contribute to remedying this shortcoming by developing a model of physician remuneration in which doctors differ in productivity. We assume that the public agency in charge of setting the remunerations of physicians cannot observe their individual productivities. This generates a principal-agent problem in which the asymmetry between the regulator and the physicians is related to the productivity of the latter. In the model we build in this paper doctors are assumed to compete for patients through the quality of care they provide. They choose the number of patients they treat by maximizing a utility function that trades off the income they earn, the time they keep for leisure, and the quality of care they supply to their patients. In order to focus on the heterogeneity of physicians, patients are assumed to be identical in all respects, including their medical needs. No co-payment is charged to patients.

The public agency that regulates the health care system is concerned with both the quality of care and the level of expenditures. When deciding upon how to relate a doctor's overall compensation to his number of patients, the agency tries to make him choose the number of patients that is optimal given his productivity while insuring that he participates in the scheme. Through the level of compensation it also tries to achieve the patient-to-doctor ratio that trades off optimally the quality of care and total expenditures : the more generous the per patient compensation paid to doctors the lower the ratio.

The model developed in the next three sections abstracts from the possibility of making doctors pay for the cost of the tests and treatments they prescribe on behalf of their patients. In Section 5 we extend the model to account for this possibility by including those services as an explicit input in the doctor's production function. The quality of service provided to patients then depends upon both the doctor's effort and the amount of medical services prescribed. As in the U.K. for general practitioners we make doctors hold budgets for covering either in part or in full the cost of tests and treatments they prescribe to patients.

[4] Using the Quebec experience, Rochaix (1993) studies the extent to which joint price-quality regulation affects physicians' activity rates. In particular, she was interested by their behavioral responses to the implementation of financial controls (fee freeze) and prospective payment systems (ceiling).

Our purpose is to investigate how this cost sharing should interact with the compensation paid to doctors.

The paper is organized as follows. In the next section the features of our basic model are presented in detail. In particular, we explain how doctors compete for patients. Section 3 studies the full information optimum which is achieved when the regulator can observe the individual productivity of doctors. This optimum is used as a benchmark when we study in Section 4 the regulator's optimum with asymmetry of information. In Section 5, we make the quality of care depend upon both the doctor's effort and the medical services provided to patients. This enables us to study how the cost of services prescribed to patients should be shared between doctors and the public agency.

2 The market for medical services

2.1 Physicians and patients

In the market for medical services, the demand side consists of P patients who have identical medical needs. On the supply side, physicians (or health care providers) differ in their productivity, that is the efficiency with which they transform medical inputs into improvement of their patients' health state. This is summarized in a production function, $h(\pi e)$, which specifies a patient's health improvement as a function of the time (or effort), e, devoted by the physician to each of his patients and of the physician's productivity, π. Later we shall introduce an additional input as argument of the health production function, namely the amount of medical services prescribed by the physician on behalf of his patients. We assume that $h(\pi e)$ is increasing and concave in πe with $h(0) = 0$. Neither e nor π are observable by the regulator financing health care expenditures;[5] moreover health improvement $h(\pi e)$ is not verifiable.[6] Each patient must be registered with a physician. The number of patients registered with a physician is denoted by n; this variable is observable (and verifiable) by the regulator.

A physician's utility is given by

$$u(n, T, e) = T + \gamma n h(\pi e) - v(ne), \tag{1}$$

where T is the provider's compensation (or transfer) paid by the regulator, γ is a parameter measuring his concern for the patient's health (improvement), while $v(ne)$ represents disutility of effort. The disutility function is increasing and convex ($v' > 0, v'' > 0$), with $v(0) = 0$. Note

[5] This regulator can be either the Minister of Public Health or the public insurer.

[6] Consequently, the provider's compensation cannot be (directly) based on h.

that T can also be interpreted as the consumption of a (numeraire) composite good.[7] In what follows, we shall focus on the determination of the physicians compensation schemes. Specifically, we shall study how a provider's compensation $T = T(n)$ should be linked to his number of patients.

As mentioned earlier, physicians differ in their productivity, which is assumed to take two values, $\pi_i (i = 1, 2)$, with $\pi_2 > \pi_1$. The total number of physicians, M, is determined by the regulating authority. Whatever their total number there is a proportion p_i of physicians of type $i (i = 1, 2)$, with $p_1 + p_2 = 1$. Throughout the paper, the subscript i is used to distinguish the two types of physicians.

The net benefit of a patient registered with a physician of type i is given by

$$B_i = h(\pi_i e_i) - w(n_i e_i), \tag{2}$$

where w is a waiting cost function that depends upon the overall time (or effort) spent by the physician for treating his patients. It is increasing and convex ($w' > 0, w'' > 0$). This function reflects the disutility that patients incur because they have to wait for an appointment or in the physician's office. The heavier the workload of a physician the longer these waiting times will be on average. Patients observe B_1 and B_2, and choose the physician who provides them with the highest level of net benefits.

2.2 Equilibrium

The equilibrium in the market for medical services is contingent upon the payment scheme $T(n)$ and the total number of physicians, M (both of which being determined by the regulator). We assume that the regulator has to offer each type of physician at least his reservation utility $\bar{V} > 0$, even for $n = 0$.[8] Formally, the equilibrium is then defined as a vector of (marketwide) net benefits and (physician type specific) patient numbers and effort levels, $(\tilde{B}, \tilde{n}_1, \tilde{n}_2, \tilde{e}_1, \tilde{e}_2)$, which satisfies the following two conditions.

1. Each type of physician chooses his effort level and number of patients to maximize his utility. Consequently, $(\tilde{n}_i, \tilde{e}_i)$ $(i = 1, 2)$ is deter-

[7] It is worth mentioning that the above setting can be reinterpreted in terms of doctors differing in their preferences for leisure rather than in their productivity. To see this let us take $t = \pi e$ and interpret t as the time that doctors devote to each patient. Substituting for e in the disutility-of-effort function, the argument of v becomes nt/π, where $1/\pi$ can be interpreted as the preference for leisure.

[8] This rules out a trivial solution where only high-productivity physicians would be active. Though attractive from a short-run perspective, this solution would be problematic when training decisions are accounted for (in particular when potential physicians are uncertain about their post-training productivity).

mined by

$$\max_{n_i, e_i} u(n_i, T(n_i), e_i) = T(n_i) + \gamma n_i h(\pi_i e_i) - v(n_i e_i), \tag{3}$$

subject to

$$h(\pi_i e_i) - w(n_i e_i) = \tilde{B} \tag{4}$$

2. All patients must be registered with a health care provider :

$$M \sum_{i=1}^{2} p_i \, \tilde{n}_i = P \tag{5}$$

This definition is in the spirit of a competitive equilibrium. Patients being identical, an equilibrium requires that their net benefits be equalized across physician types. This marketwide net benefit level $\tilde{B}$, though endogenously determined at equilibrium, is taken as given by any physician (condition (4)), who determines his demand for patients (as well as his effort level) accordingly.[9] Finally, (5) is the market clearing condition stating that the total demand for patients equals their supply (or number of patients P).

2.3 Regulating authority

The regulator's objective function is evaluated at the market equilibrium induced by its policy (specifying $T(n)$ and M). It takes into account the net benefits to patients and the cost of public funds needed to compensate physicians. With λ denoting the per unit cost of public funds ($\lambda > 1$), the regulator's objective is given by : [10]

$$W = P\tilde{B} - \lambda M \sum_{i=1}^{2} p_i T(\tilde{n}_i) \tag{6}$$

where $\tilde{B}$ and $\tilde{n_i}$ are defined by (3)–(5), that is the conditions determining the equilibrium in the health care market.

The regulator's problem stated this way is quite intricate. To make it tractable, we shall reformulate it and consider the equivalent mechanism design problem. For this purpose, we must first take a closer look at the physician's problem and introduce some additional notation.

[9] This is similar to the "utility taking" assumption in the urban economics (and local public goods) literature.

[10] The regulator's objective function does not take into account the utility of physicians. This is assumed for simplicity and does not change the nature of our results. As we shall see, the first-best solution implies that informational rents are zero. The second-best solution implies a rent equal to zero for low-productivity physicians and a positive rent for high productivity physicians. This positive rent would be higher if physicians' utilities were included in social welfare.

2.4 Behavior and preferences of physicians : a closer look

First of all, it is convenient to eliminate the (unobservable) effort level from the physician's problem. To do this, we introduce the "effort requirement function", $e_i = E_i(n, B)$, which is defined by :

$$h(\pi_i E_i(n,B)) - w(nE_i(n,B)) = B, \qquad i = 1,2 \tag{7}$$

This function specifies the level of effort a physician of type i has to provide for supplying net benefits B to n patients. It has the following properties :

$$\frac{\partial E_i}{\partial B} = [\pi_i h'(\pi_i e_i) - nw'(ne_i)]^{-1}, \tag{8a}$$

$$\frac{\partial E_i}{\partial n} = [\pi_i h'(\pi_i e_i) - nw'(ne_i)]^{-1} e_i w'(ne_i) \tag{8b}$$

It seems natural to assume that an increase in the net benefit required by patients will demand more effort to the physician ($\partial E_i/\partial B > 0$). This is equivalent to assuming :

H1 : $\pi_i h'(\pi_i e_i) - nw'(ne_i) > 0, \qquad i = 1,2,$[11]

which in turn implies that $\partial E_i/\partial n > 0$ and also that $E_1(n,B) > E_2(n,B)$.[12]

Next, we can substitute $E_i(n,B)$ into the direct utility function (1) to obtain the following derived utility function :

$$V_i(T,n,B) = T + \gamma n h(\pi_i E_i(n,B)) - v(nE_i(n,B)), \quad i = 1,2 \tag{9}$$

This utility function is a crucial ingredient of the reformulated problem considered below. In particular, it is used to define indifference curves for a given B in the (n,T) - space. In the appendix we show that under the following hypotheses :

H2 : $v'(ne_i) - \gamma \pi_i h'(\pi_i e_i) > 0, \quad i = 1,2,$

H3 : $e_i v'(ne_i) - \gamma h(\pi_i e_i) > 0, \quad i = 1,2,$

these indifference curves are increasing and convex and they satisfy at any point (n,T) the single-crossing property :

$$MRS^2_{T,n} < MRS^1_{T,n}$$

where

$$MRS^i_{T,n} \equiv -\frac{\partial V_i/\partial n}{\partial V_i/\partial T}$$

[11] Since, by assumption, w' is strictly positive for all e, and h' is decreasing with e_i one could also have $[\pi_i h'(\pi_i e_i) - nw'(ne_i) < 0]$ as e_i tends to infinity. This would mean that the net benefit of effort (required for a small increase in B or in n) is negative. This will never be the case if physicians choose a utility-maximizing effort level for any γ that is not too large.

[12] Redefining E as a function of π, with $E(\pi_i, n_i, B) \equiv E_i(n_i, B)$ differentiating (7) and making use of *H1* yields $\partial E/\partial \pi = -(\pi h' - nw')^{-1} e h' < 0$.

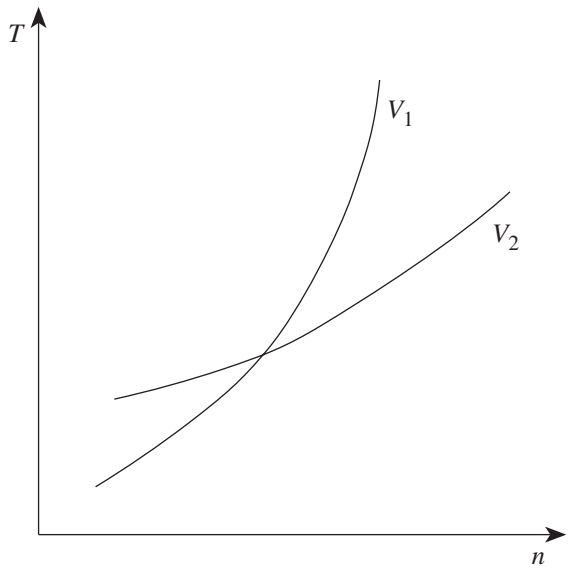

Figure 1 : *Relative slopes of indifference curves*

is the marginal rate of substitution between T and n for physicians of type i. These properties are illustrated in Figure 1, where an indifference curve for each type of physician is represented, the steepest one pertaining to the least productive physician.

The above hypotheses mean that physicians need a higher compensation respectively for providing more effort (for a given n) and for treating more patients (for a given e). They are both satisfied if γ is not too high. Note also that since the physician's utility function is quasi-linear in T, the indifference curves pertaining to each type of physician are for a given B vertically parallel to each other in the (n, T) space.

3 Design of the optimal compensation scheme and the full information optimum

Using the revelation principle, the regulator's problem stated in section 2.3 is equivalent to a mechanism design problem where the regulator searches for the two pairs (n_1, T_1) and (n_2, T_2) that he intends for the first and second types of physicians respectively. In solving this problem, the regulator must make sure that physicians of type i actually prefer the pair (n_i, T_i) designed for them to the other pair and that they are compensated enough to participate. This is accounted for by the incentive-compatibility and individual-rationality constraints of the following problem :

$$\max_{T_i,n_i,M,B} PB - \lambda \qquad M\sum_{i=1}^{2} p_i T_i \qquad (10a)$$

subject to

$$(MC) \qquad M\sum_{i=1}^{2} p_i n_i = P \qquad (10b)$$

$$(IR1) \qquad V_1(T_1, n_1, B) \geqslant \overline{V}$$

$$(IR2) \qquad V_2(T_2, n_2, B) \geqslant \overline{V} \qquad (10c)$$

$$(IC1) \qquad V_1(T_1, n_1, B) \geqslant V_1(T_2, n_2, B) \qquad (10d)$$

$$(IC2) \qquad V_2(T_2, n_2, B) \geqslant V_2(T_1, n_1, B) \qquad (10e)$$

where $\overline{V}$ stands for the physicians' reservation utility, i.e. the level of utility they can reach in the most preferred alternative occupation. Substituting M from the first constraint into the objective function, the Lagrangean of this problem can be written as :

$$\begin{aligned} L = PB - \lambda P \left[\sum_{i=1}^{2} p_i n_i\right]^{-1} \sum_{i=1}^{2} p_i T_i \\ + \sum_{i=1}^{2} \phi_i V_i(T_i, n_i, B) + \mu_1[V_1(T_1, n_1, B) \\ - V_1(T_2, n_2, B)] + \mu_2[V_2(T_2, n_2, B) - V_2(T_1, n_1, B)] \end{aligned} \qquad (11)$$

where the ϕ's and μ's are the dual variables of the individual-rationality and incentive-compatibility constraints respectively. These dual variables satisfy the usual complementarity conditions that they are equal to 0 if the corresponding constraints are not binding. It is useful to first characterize the full information optimum that we shall use as a benchmark. It can be done by deleting the IC constraints from the above problem (i.e. setting $\mu_i = 0$). It is then straightforward to obtain the following conditions for an optimal allocation :

$$MRS^i_{T,n} = \frac{\sum_{i=1}^{2} p_i T_i}{\sum_{i=1}^{2} p_i n_i} \equiv \alpha, \qquad i = 1, 2, \qquad (12)$$

$$P = \lambda M \sum_{i=1}^{2} p_i MRS^i_{T,B} \qquad (12)$$

where $MRS^i_{T,B} = (-\partial V_i/\partial B)/(\partial V_i/\partial T)$ is the additional compensation that a physician of type i requires in response to a unit increase in the net benefit B. In condition (13), α stands for the average cost of treating a patient; according to this condition, minimizing the overall cost of treating the P patients imposes that the additional compensation that either type of physician requires to treat one further patient be equated to that average cost. As to condition (12), it means that the net

benefit B must be pushed to the level where at the margin it is equated with the ratio M/P times the cost of the additional compensation that physicians require on average for an additional unit of B. With full information, condition (12) can be decentralized by means of two different linear payment schemes (one for each physician type) :

$$T_i(n) = A_i + \alpha n, \qquad i = 1, 2 \tag{14}$$

where A_i is set equal to $T_i - \alpha n_i$ with (n_i, T_i) referring to the full-information optimum. This is represented in Figure 2. Note that both pairs (n_i, T_i), $i = 1, 2$, are located on the indifference curves corresponding to $V^i = \overline{V}$.

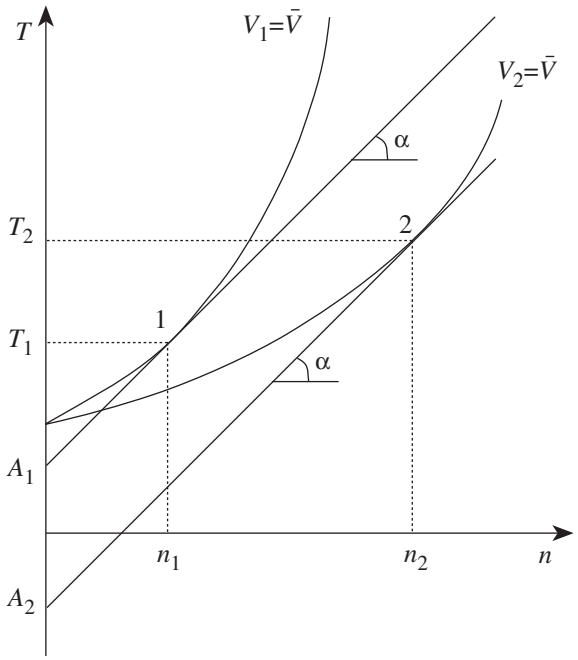

Figure 2 : *Full-information optimum*

4 Second-best solution with asymmetric information

The full-information optimum is not incentive compatible. This can easily be understood from Figure 2. For any given n, the level of effort required to reach a given net benefit B is larger for low-productivity physicians than for high-productivity ones. Consequently, the former require a higher compensation than the latter to reach utility $\overline{V}$ and type 2 physicians would choose the pair (n_1, T_1) and therefore mimic

type 1 physicians. This implies that only the second IC-constraint will be binding, which in turn implies that the second IR-constraint is always satisfied (this results from $V_2(T_1, n_1) > V_1(T_1, n_1)$). In other words, we have $\mu_1 = 0$ and $\phi_2 = 0$.

We can now derive the first-order conditions for an optimum of the regulator's problem with asymmetric information as specified in the Lagrangean. These conditions are given with respect to B, n_1, n_2, T_1 and T_2 :

$$P + \phi_1 \frac{\partial V_1}{\partial B} + \mu_2 \left(\frac{\partial V_2}{\partial B} - \frac{\partial \hat{V}_2}{\partial B} \right) = 0, \tag{15a}$$

$$\lambda P \left(\sum_{i=1}^{2} p_i n_i \right)^{-2} p_1 \sum_{i=1}^{2} p_i T_i + \phi_1 \frac{\partial V_1}{\partial n_1} - \mu_2 \frac{\partial \hat{V}_2}{\partial n_1} = 0, \tag{15b}$$

$$\lambda P \left(\sum_{i=1}^{2} p_i n_i \right)^{-2} p_2 \sum_{i=1}^{2} p_i T_i + \mu_2 \frac{\partial V_2}{\partial n_2} = 0, \tag{15c}$$

$$- \lambda P \left(\sum_{i=1}^{2} p_i n_i \right)^{-1} p_1 + \phi_1 \frac{\partial V_1}{\partial T_1} - \mu_2 \frac{\partial \hat{V}_2}{\partial T_1} = 0, \tag{15d}$$

and

$$- \lambda P \left(\sum_{i=1}^{2} p_i n_i \right)^{-1} p_2 + \mu_2 \frac{\partial V_2}{\partial T_2} = 0 \tag{15e}$$

where $\hat{V}_2 = V_2(n_1, T_1, B)$ refers to the utility of a high-productivity physician mimicking a low-productivity one.

Using the observation that $\partial V_2 / \partial T_1 = 1$, conditions (15c) and (15e) yield :

$$MRS^2_{T,n} = \frac{\sum_{i=1}^{2} p_i T_i}{\sum_{i=1}^{2} p_i n_i} \equiv \alpha \tag{16}$$

Therefore, compared with (12), the same result as with full information is obtained, but it only applies to high-productivity physicians. This is the standard outcome that there is no distortion at the top. Combining conditions (15b) and (15d) and using $\partial V_1 / \partial T_1 = \partial \hat{V}_2 / \partial T_1$, we obtain :

$$p_2 (MRS^1_{T,n} - \widehat{MRS}^2_{T,n}) = p_1 (\alpha - MRS^1_{T,n}) \tag{17}$$

where $\widehat{MRS}^2_{T,n}$ is the high-productivity physicians' marginal rate of substitution between T and n taken at point (n_1, T_1). The single-crossing property implies that $MRS^1_{T,n} > \widehat{MRS}^2_{T,n}$, and thus we infer the following inequality from (17) :

$$MRS^1_{T,n} < \alpha, \tag{18}$$

which means that the number of patients treated by each low-productivity physician is reduced relative to the full information optimum (say by Δn_1). This causes some inefficiency : the corresponding reduction in his compensation (ΔT_1) is lower than the additional cost required to ensure that the patients who are no longer on his list will be treated by other physicians ($\Delta T_1 < \alpha \Delta n_1$). However, this inefficiency is desirable since it reduces the informational rent of high-productivity physicians, and gets them closer to their reservation utility $\bar{V}$. The rationale for this efficiency loss is illustrated in Figure 3. If the full-information conditions $MRS^i_{T,n} = \alpha$ were satisfied for both types of physicians while fulfilling the second IC-constraint, points 1 and 2 would be chosen, and the informational rent left to high-productivity physicians would amount to $V_2 - \overline{V}$. By moving point 1 to 1′ this rent is reduced by ΔT_2 though at the expense of some efficiency loss due to $n'_1 < n_1$ and measured by $\alpha \Delta n_1 - \Delta T_1$.

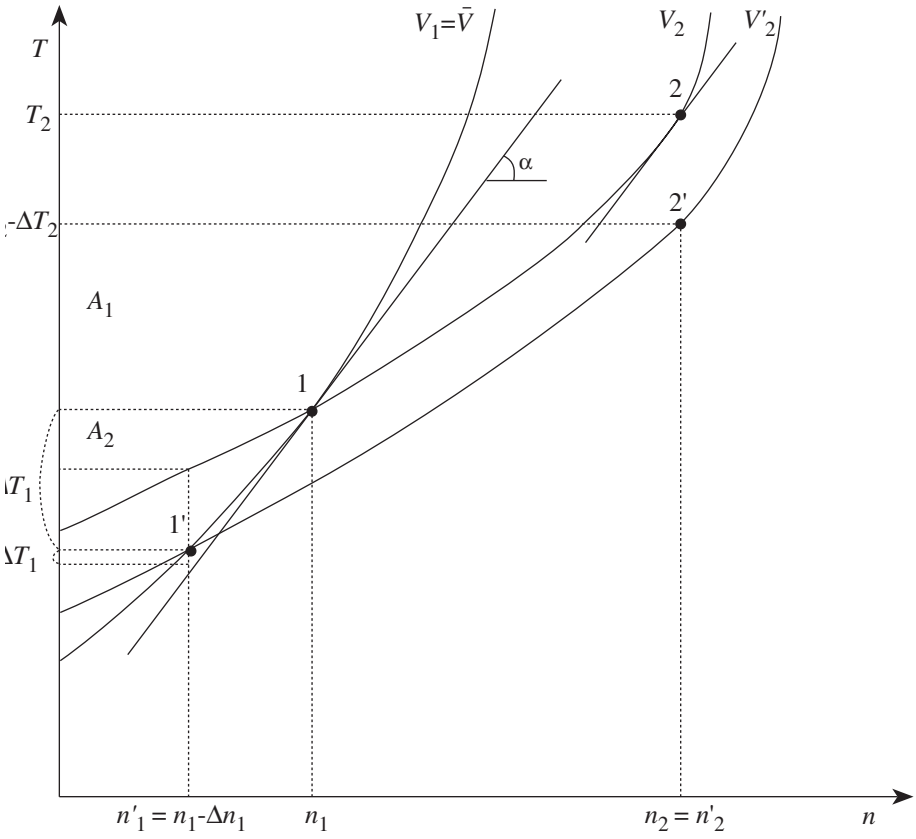

Figure 3 : *The second-best optimum*

The meaning of condition (17) should now be obvious. The reduction in n_1 should be such that the efficiency loss (right-hand side) is

at the margin equated to what is gained from reducing the informational rent (left-hand side). These losses and gains are weighted by the relevant proportion of physicians.

Conditions (16) and (17) can be implemented by any payment scheme $T(n)$ which, in the (n,T) space, goes through the two optimal points $1'$ and $2'$ and is, for any other value of n, strictly below the two relevant indifference curves. The comments for conditions (16) and (17) apply for any given value of B. What remains to be explained is how this value is optimally chosen. Using (15d) and (15e), condition (15a) yields :

$$P = \lambda M \left[\sum_{i=1}^{2} p_i MRS^i_{T,B} + p_2 (MRS^1_{T,B} - \widehat{MRS}^2_{T,B}) \right] \tag{19}$$

where $\widehat{MRS}^2_{T,B}$ is the $MRS_{T,B}$ of high-productivity physicians taken at (n_1, T_1) i.e. the $MRS_{T,B}$ of type 2 physicians mimicking type 1 ones. This condition can be given the same interpretation as condition (12), except that there is here an additional term that pushes up the cost of a unit increase of B. This term can be interpreted in the following manner. Consider a unit rise of B; the additional compensations, ΔT_1 and ΔT_2, that are required to keep physicians at their initial utility levels cause the incentive-compatibility constraint of high-productivity physicians to be violated. Therefore the compensation of these physicians must be further increased by an amount equal to $MRS^1_{T,B} - \widehat{MRS}^2_{T,B}$.

These observations are illustrated in Figure 4, where two pairs of indifference curves are drawn in the (n,T) space. The pair of plain curves are those reached at the optimum for some value of B, with (n_1, T_1) and (n_2, T_2) being the optimal contracts. Suppose that we increase the required net benefit to patient B while keeping the $n_i's$ and the physicians' utilities at their initial levels. Following this move the two indifference curves rise to the dotted ones. Since the increase of B requires the low-productivity physicians to raise their effort by a larger amount than the high-productivity physicians mimicking them, the dotted curve of the former is above that of the latter at the vertical of n_1. As a consequence, the incentive-compatibility constraint of high-productivity physicians is no longer satisfied, which requires a further increase in their compensation, equal at the margin, to $MRS^1_{T,B} - \widehat{MRS}^2_{T,B}$, and which appears in (19).

5 Extension : more general "production technology"

Let us now assume that the improvement in a patient's health state does not only depend on the physician's effort, but also on the amount

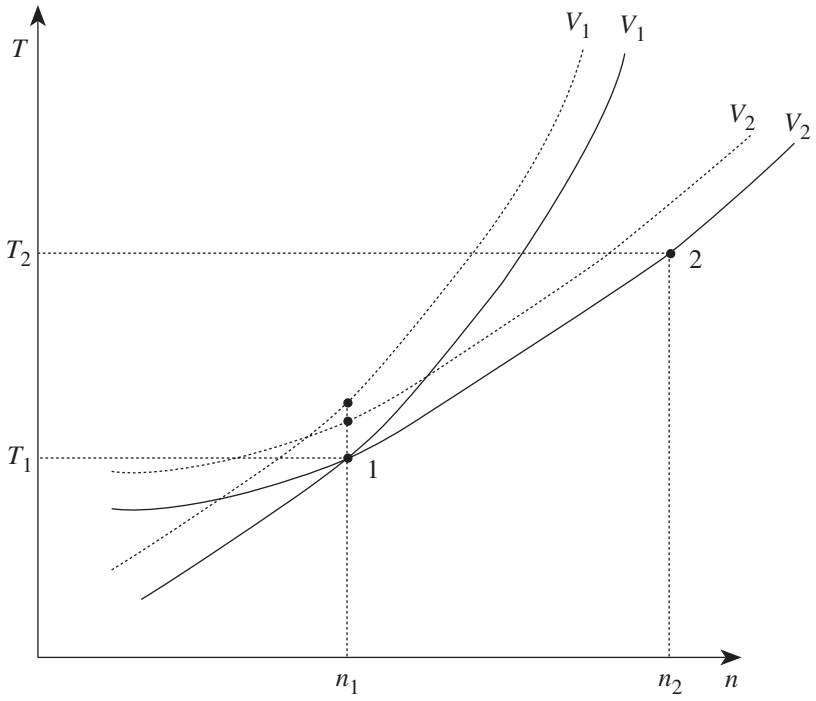

Figure 4 : *The second-best optimum*

of medical services prescribed by the physician. These can be seen as drugs, laboratory tests, physiotherapy, etc. Total spending on these medical services is observable by the regulator and is denoted by y. The production function of physician i ($i = 1, 2$) can then be rewritten $h(\pi_i e_i, y_i)$. We assume that it is increasing in y ($\partial h/\partial y > 0$). Similarly, the effort requirement function of physician i (which continues to specify the level of effort a physician has to provide for supplying net benefits B to n patients) is redefined as $E_i(n, B, y)$. Assuming that physicians' effort and medical services y are substitutes in production, an increase in the level of prescriptions will mean less effort for the provider ($\partial E_i/\partial y < 0$).[13]

The following assumption is made concerning the payment of these medical services : their cost is supported by the physician, but it can be compensated through T. The physician's utility function (1) is then defined accordingly. Substituting $E_i(n, B, y)$ into the utility func-

[13] To see this observe that

$$\frac{\partial E_i}{\partial y} = \frac{-\partial h/\partial y_i}{\pi_i \partial h(\pi_i e_i, y_i)/\partial(\pi_i e_i) - n_i w'(n_i e_i)}$$

is negative because h is increasing in y, while the denominator is negative under $H1$.

tion yields the following derived utility function :

$$V_i(T, n, B, y) = T - ny + \gamma nh(\pi_i E_i(n, B, y), y) - v(nE_i(n, B, y)), \qquad i = 1, 2$$

The physician's compensation function now depends upon both the number of patients registered and the per-patient cost of medical services : $T(n_i, y_i)$. Once the compensation $T(n_i, y_i)$ has been set by the regulator, each type of physician chooses y (and n) to maximize his utility. This is achieved when

$$MRS^i_{T,y} \equiv \frac{-\partial V_i/\partial y}{\partial V_i/\partial T} = \frac{\partial T}{\partial y} \qquad i = 1, 2,$$

where $MRS^i_{T,y}$ is the marginal rate of substitution between T and y for a physician of type i, and $\partial T/\partial y$ is the marginal compensation for y. If $\partial T/\partial y = 0$, the provider bears the full cost of y (at the margin); if, on the other hand, $\partial T/\partial y = n$, he receives full compensation.

Except for the addition of y as argument in functions T_i and V_i, the mechanism design problem is formulated in the same way as earlier in (10a). Proceeding as before we first characterize the full information optimum. The conditions on n_i ($i = 1, 2$) and B that have been obtained earlier carry over without change. It is straightforward to show that at the optimum :

$$MRS^i_{T,y} = \frac{\partial T_i}{\partial y_i} = 0, \qquad i = 1, 2 \tag{20}$$

This implies that physicians of both types should, at the margin, bears the full cost of the medical services they prescribe. It can be implemented by including in T_i a lump-sum component equal to the optimal value of y_i.

Turning to the regulator's problem with asymmetric information, it is straightforward to show that the optimality condition on n_i ($i = 1, 2$) and B obtained in section 3 remains unchanged. The optimality condition on y_2 yields :

$$MRS^2_{T,y} = \frac{\partial T_2}{\partial y_2} = 0, \tag{21}$$

which coïncides with the optimality condition that has just been obtained in the full information case. This is no longer true for the condition on y_1 :

$$MRS^1_{T,y} = \frac{\partial T_1}{\partial y_1} = \frac{p_2}{p_1}(\widehat{MRS}^2_{T,y} - MRS^1_{T,y}), \tag{22}$$

which can be given the following interpretation. Suppose that the expression in parentheses on the right-hand side is positive. This means that as y_1 increases, physicians of type 1 require, for staying on the same indifference curves, an additional compensation that is lower than that required by physicians of type 2 who mimic them. If both are

compensated by the amount required by the former, the latter are made worse off by the change, which relaxes the incentive-compatibility constraint of the latter. This encourages the regulator to make type 1 physicians choose a value of y_1 that is larger than in the full information optimum, which is obtained by subsidizing y_1 at the margin ($\partial T_1/\partial y_1 > 0$).

However, the sign of the expression in parentheses in (22) is ambiguous, and so we cannot assert whether y_1 should be encouraged ($\partial T_1/\partial y_1 > 0$) or discouraged ($\partial T_1/\partial y_1 < 0$).

6 Concluding remarks

The paper has studied the design of compensation schemes for health care providers which relate the providers' earnings to the number of registered patients. The following results have emerged. First, in a full information setting, the marginal compensation for a patient must be the same for all providers and equal to the marketwide average cost of serving a patient. Second, under incomplete information the number of patients registered with the low productivity provider is distorted downward to reduce informational rents of the high productivity provider. Third, the marginal compensation for patients which implements the optimal (second-best) allocation is lower for low productivity providers than for high productivity providers. Consequently, it increases with the number of patients. Fourth, the level of benefits per patients tends to be lower under incomplete information than under complete information.

Finally, we have considered an extension of the model where patients' health improvement also depends on prescribed medical services. We have shown that under full information, providers must bear the full *marginal* cost of the prescribed services. Under incomplete information, the marginal "tax" on the low productivity provider for prescriptions may be lower or higher than the first best level; consequently, it may exceed the full cost of the prescriptions.

Our model clearly only represents a step towards a fully fledged incentive based theory of health care providers' compensation schemes. It admittedly has a number of limitations. The homogeneity of patients is one of them. The re-examination of our model with heterogenous patients is a natural extension. It gives rise to the issue of patient selection by providers and its impact on the power of the optimal incentive scheme. Other extensions include function that of a "regulator" with a different objective (e.g. a private, profit maximizing insurance company) possibly in competition with the public system considered in our setting. These issues are left for future research.

APPENDIX

Lemma 1 : *Let $V_i(T, n, B)$ denote the derived utility function of physician i(i = 1, 2). The indifference curves for given B in the (n, T) space (associated with V_i) are characterized by the following properties :*

i) they are increasing;

ii) they are convex;

iii) they satisfy at any point (n, T) the single-crossing property.

Proof :

i) The slope of the indifference curves is obtained by totally differentiating V_i and making use of $dV_i = dB = 0$. This yields

$$\frac{\partial T}{\partial n} = [nv'(ne_i) - \gamma n\pi_i h'(\pi_i e_i)]\frac{\partial E_i}{\partial n}(n, B) + [e_i v'(ne_i) - \gamma h(\pi_i e_i)] \qquad (A.1)$$

where $\partial E_i/\partial n = [\pi_i h'(\pi_i e_i) - nw'(ne_i)]^{-1} e_i w'(ne_i)$ is positive if w is increasing and $H1$ is satisfied. Property i) ($\partial T/\partial n > 0$) then follows from $H2, H3$ and the fact that $\partial E_i/\partial n$ is positive.

ii) Differentiating $(A.1)$ with respect to n, one obtains[14]

$$\begin{aligned}\frac{\partial^2 T}{\partial n^2} = {} & (v' - \gamma\pi h')\frac{\partial E}{\partial n} + ev''\frac{\partial(nE)}{\partial n} + (w' + new'')\frac{\partial(nE)}{\partial n}\left[\frac{v' - \gamma\pi h'}{\pi h' - nw'}\right] \\ & + new''\left[\frac{v''\frac{\partial(nE)}{\partial n} - \gamma\pi^2 h''\frac{\partial E}{\partial n}}{\pi h' - nw'}\right] \\ & - new''(v' - \gamma\pi h')\left[\frac{\pi^2 h''\frac{\partial E}{\partial n} - w' - nw''\frac{\partial(nE)}{\partial n}}{(\pi h' - nw')^2}\right]\end{aligned}$$

where $\partial(nE)/\partial n = -e\pi h'[\pi h' - nw']^{-1}$ is positive if h is increasing and $H1$ is satisfied. Property ii) ($\partial^2 T/\partial n^2 > 0$) follows from the fact that $\partial E/\partial n$ and $\partial(nE)/\partial n$ are positive, from $H1$ and $H2$, and from assumptions on v (increasing and convex), w (increasing and convex) and h (increasing and concave).

iii) One has to show that $\partial T_1/\partial n > \partial T_2/\partial n$. Suppose $\partial T/\partial n$ is a well-defined function of π, say $(\partial T/\partial n)(\pi_i) \equiv \partial T_i/\partial n$. Differentiating with respect to π leads to

$$\begin{aligned}\frac{\partial}{\partial\pi}\left(\frac{\partial T}{\partial n}\right) = {} & (v' + nv'')\frac{\partial E}{\partial\pi} - \frac{\partial(\pi E)}{\partial\pi} + n(w' + ew'')\frac{\partial E}{\partial\pi}\left[\frac{v' - \gamma\pi h'}{\pi h' - nw'}\right] \\ & + new'\frac{\left[nv''\frac{\partial E}{\partial\pi} - \gamma h' - \gamma\pi h''\frac{\partial(\pi E)}{\partial\pi}\right]}{(\pi h' - nw')} \\ & - new'(v' - \gamma\pi h')\left[\frac{h' + \pi h''\frac{\partial(\pi E)}{\partial\pi} - n^2 w''\frac{\partial E}{\partial\pi}}{(\pi h' - nw')^2}\right]\end{aligned}$$

[14] For the rest of the proof, we omit for simplicity the subscript i and the arguments of functions v, h, w, E_i and of their derivatives. For instance we simply write v' instead of $v'(ne_i)$.

where $\partial E/\partial \pi = -eh'[\pi h' - nw']^{-1}$ and $\partial(\pi E)/\partial \pi = -new'[\pi h' - nw']^{-1}$ are both negative if $H1$ is satisfied and h and w are increasing. Hence the single-crossing property follows immediately from the negativity of $(\partial/\partial\pi)(\partial T/\partial n)$ which, in turn, is implied by $H1, H2$ and the assumptions put on v, w and h.

Bibliographie

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

Ellis, R.P. and T.G. McGuire, 1986, "Provider behavior under prospective reimbursement : Cost sharing and supply", *Journal of Health Economics*, 5, 129–151.

Ellis, R.P. and T.G. McGuire, 1990, "Optimal payment systems for health services", *Journal of Health Economics*, 9(4), 375–396.

Gruber, J. and M. Owings, 1996, "Physician financial incentives and caeserean section delivery", *Rand Journal of Economics*, 27, 99–123.

Krasnik, A. et al., 1990, "Changing remuneration systems : Effects on activity in general practice", *British Medical Journal*, 300, 1698–1701.

Ma, C.-t A., 1994, "Health care payment systems : Cost and quality incentives", *Journal of Economics and Management Strategy*, 3(1), 93–112.

Ma, C. A. and T.G. McGuire, 1997, "Optimal health insurance and provider payment", *The American Economic Review*, 87(4), 685–704.

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

McGuire, T.G. and M. Pauly, 1991, "Physician response to fee changes with multiple players", *Journal of Health Economics*, 10, 385–410.

Newhouse, J.P., 1996, "Reimbursing health plans and health providers : selection versus efficiency in production", *Journal of Economic Literature*, 34, 1236–1263.

Rochaix, L., 1993, "Financial incentives for physicians : The Quebec experience", *Health Economics*, 2, 163–176.

Résumé

Dans cet article, nous utilisons un modèle d'agence pour étudier comment, dans un contexte d'asymétrie d'informations, la rémunération d'un médecin devrait être reliée au nombre de patients traités. Les médecins n'ont pas tous la même productivité; les patients qui ont des besoins homogènes peuvent choisir leur médecin, de sorte qu'à l'équilibre, tous les médecins doivent offrir le même niveau de bénéfices nets (amélioration nette de l'état de santé). Le régulateur qui détermine le schéma de rémunération se préoccupe à la fois de la qualité des soins offerts et du niveau des dépenses encourues. Nous montrons que la solution optimale de second rang donne un schéma de rémunération dans lequel la rémunération marginale par patient augmente avec le nombre de patients. Dans une généralisation du modèle, l'amélioration de l'état de santé du patient peut aussi dépendre des services prescrits par le médecin; nous examinons comment le coût de ces prescriptions devrait être pris en compte dans le schéma de rémunération.

Abstract

In this paper we use a principal-agent model to study how the compensation paid to a physician should be related to the number of his patients. Health care providers are heterogeneous in their productivity; the homogenous patients are mobile so that their level of net benefits must be equalized across providers. The regulating agency is concerned with both the quality of care and the level of expenditures. We show that the second-best (incomplete information) solution implies that the marginal compensation for a patient increases with the number of patients. In an extension, we also account for the possibility that the benefits (health improvement) provided to a patient depend on prescribed services and study how these should enter the compensation scheme.

Mots-clés

Économie de la santé, rémunération des médecins, mécanisme incitatif, sélection adverse.

Keywords

Health economics, physicians'remuneration, incentive mechanism, adverse selection.

Classification JEL : D82, I11, I19