économiepublique

Économie publique/Public economics

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

Mécanismes de rémunération et incitations des médecins

Carine Franc

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/economiepublique/553
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
Carine Franc, « Mécanismes de rémunération et incitations des médecins », *Économie publique/Public economics* [En ligne], 09 | 2001/3, mis en ligne le 07 décembre 2005, consulté le 30 avril 2019. URL : http://journals.openedition.org/economiepublique/553

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

Mécanismes de rémunération et incitations des médecins*

Carine Franc**
CREGAS-INSERM

1 Introduction

La rémunération des médecins est au coeur de nombreuses réformes[1]. Dans cet article, nous essayons de comprendre pourquoi il est si difficile pour un gouvernement, d'instaurer un système efficace de paiement des producteurs de soins et plus précisément des prescripteurs de soins. Nous étudions l'impact du système de rémunération sur le comportement du médecin en termes d'arbitrage entre qualité, illustrée dans ce modèle par la qualité du diagnostic, et maîtrise des dépenses de soins. Si le gouvernement tente de réguler l'offre de soins via le mécanisme de rémunération des médecins, c'est que dans le domaine de la santé, l'assureur public ou privé est confronté au problème d'aléa moral *ex post*. Celui-ci se traduit, en cas de maladie, par la décision *ex post* du coût du traitement.

Certains auteurs supposent que la dépense de soins résulte d'un comportement de demande. Breuil-Genier et al. (1997) montrent que l'assurance complémentaire augmente la probabilité d'avoir connu un

* Je remercie pour leurs remarques constructives, H. Cremer, I. Dubec et P.Y. Geoffard ainsi que les participants au 5^e colloque d'économie publique appliquée (1999), au congrès de l'European Economic Association (1999) et les participants au séminaire assurance santé (DELTA, 1999).

** CREGAS-INSERM U537-CNRS 8052, 80 rue du Général Leclerc, 94276 Le Kremlin-Bicêtre cedex, e-mail : franc@kb.inserm.fr

[1] Les instruments traditionnels de la régulation de l'offre se sont avérés relativement inefficaces. Le système de tarifs (prix administrés), par exemple, combiné avec les règles de rémunération permet au médecin de compenser l'évolution contrôlée des prix par les volumes de soins fournis. En 1993, la « maîtrise médicalisée » des soins de santé prévoit, pour la première fois en France, une action de contrôle de l'offre de soins par le gouvernement (Références Médicales Opposables, Objectif Quantifié National).

épisode de soins. D'autres études empiriques récentes tentent alors d'estimer l'ampleur du risque moral *ex post* par le calcul de l'élasticité de la demande de soins, évaluant ainsi les bénéfices potentiels d'une politique de couverture partielle. Chiappori, Durand et Geoffard (1997) notent ainsi que la présence d'assurance modifie significativement le coût des traitements pour les assurés. Les auteurs remarquent toutefois qu'une consommation accrue de traitements peut être la conséquence d'incitations engendrées par un contrat particulier d'assurance (risque moral) ou, au contraire, du comportement d'un individu qui, connaissant ses besoins, achèterait un contrat particulier (antisélection). Il est alors important de mesurer l'ampleur du problème de risque moral. Les auteurs montrent que l'élasticité de la demande de soins (à savoir le nombre de visites en cabinet d'un généraliste) est très faible voire nulle pour des petites variations de prix.

Au contraire, d'autres auteurs, comme Ma (1994), partent de l'hypothèse selon laquelle la consommation de soins est le résultat d'une décision du prescripteur (indépendamment de la qualité, par exemple). Il s'agit alors de trouver une alternative aux politiques d'assurance partielle et de réguler l'offre de soins par l'intermédiaire des systèmes de paiement des producteurs de soins. De nombreux articles tentent déjà d'apporter des clarifications sur ce point mais plus particulièrement dans le contexte de l'hôpital. Newhouse (1996) propose notamment un survol de la littérature théorique mise en parallèle avec les réformes engagées aux États-Unis. Il présente les différents systèmes de rémunération des hôpitaux, le paiement à l'acte et le système du paiement prospectif. Il détaille à travers les résultats de différents articles[2] les diverses méthodes possibles pour calculer la valeur du forfait anticipé dans le cadre des « Diagnosis Related Groups »[3]. Le cas de la médecine ambulatoire est abordé par Ma et Mac Guire (1997). Ils étudient les problèmes d'agence entre assureurs, assurés et médecins en supposant que la quantité consommée n'est pas observable. Les assureurs se basent sur les reports de traitements établis par le médecin qui, suivant le mécanisme de rémunération, peut ne pas être incité à reporter la quantité effective. En présence de tickets modérateurs et d'un paiement par capitation avec pénalité financière en cas de dépassement, les deux agents sont incités à sous-évaluer la quantité. Les auteurs montrent que, lorsque le médecin se comporte honnêtement (c'est à dire qu'il génère le bénéfice minimum des soins), la contrainte de bon report est relâchée et la combinaison optimale est atteinte.

Dans les précédents travaux, le système de capitation est associé au « Prospective Payment System » qui n'est versé que lorsque le patient

[2] En particulier, Pauly (1980) étudie les propriétés d'optimalité d'un système de paiement à l'acte dans une économie en information parfaite. Pope (1990), Goodall (1990) et Keeler (1990) étudient en fonction de la source de l'hétérogénéité, la méthode de calcul des paiements forfaitaires anticipés (PPS).

[3] Les DRGs sont, par analogie en France, les Groupes Homogènes de Malades (GHM) définis dans le cadre du Programme de Médicalisation du Système d'Information (PMSI).

est déclaré malade et entre à l'hôpital ou consulte le médecin. Pourtant, dans le cadre de la médecine ambulatoire, le système de capitation prévoit le versement du transfert quel que soit l'état de santé du patient.

Le système de capitation (utilisé au Royaume-Uni, aux Pays-Bas ou encore en Suède) fournit un revenu par patient inscrit, basé sur des allocations forfaitaires versées par les organismes payeurs. Chaque patient choisit librement son médecin traitant en sachant qu'en changer est difficile[4]. Ce système encourage la prise en charge globale du patient ainsi que la prévention (à condition que la durée du contrat entre le médecin et le patient soit suffisamment longue). L'avantage évident de ce système est, d'une part, le contrôle important des dépenses et, d'autre part, le fait qu'il s'adapte très bien au financement par l'allocation d'une enveloppe budgétaire décidée par les pouvoirs publics. Néanmoins, cette rémunération n'incite pas à fournir des soins de grande qualité et les files d'attente dans les hôpitaux en sont notamment un exemple. Un autre inconvénient important de ce système est la sélection des patients au moment de l'établissement des listes : le généraliste préfère accepter sur sa liste l'inscription d'individus jeunes ne souffrant pas de longue maladie.

À l'inverse, le système de paiement à l'acte (en France et en Allemagne) procure au médecin un revenu égal à la multiplication du nombre d'actes effectués par le prix de chaque prestation (fixé par négociation entre les organismes payeurs et les professionnels de santé). Le « fee-for-service system » incite naturellement à la sur-production puisque le revenu dépend directement du nombre d'actes réalisés. Sans un abonnement auprès d'un médecin, le patient choisit librement son médecin et a la possibilité de consulter plusieurs fois pour une même maladie. Ce système a des avantages en termes d'amélioration de la qualité des soins, dûs notamment à la concurrence qui naît entre les médecins.

Les études empiriques récentes montrent dans quelle mesure l'un ou l'autre des deux systèmes est insatisfaisant. D'une part, Besley, Hall et Preston (1999) observent à partir de données britanniques que près de 14% de la population est tellement insatisfaite de la durée des listes d'attente que ces individus renoncent à l'assurance publique et achètent un contrat privé d'assurance maladie. D'autre part, Majnoni d'Intignano (1998) conclut, à l'issue d'une comparaison entre l'efficacité du système français et celle de quatre autres pays européens (Allemagne, Danemark, Royaume-Uni et Suède), que le système français est coûteux, efficace pour la médecine curative mais très en retard pour la prévention. Elle préconise le développement de nouvelles formes de rémunération, comme un abonnement auprès d'un médecin ou une

[4] Le coût individuel de changement de médecin n'est pas seulement lié aux règles établies par les institutions. Selon Gravelle et Masiero (2000), dans un système où la qualité est un bien privé, il est aussi la conséquence de la perte d'informations acquises dans la relation patient-médecin.

certaine forme de capitation[5] (de manière à favoriser les activités de prévention).

Dans cet article, nous montrons pourquoi il est impossible pour un gouvernement d'instaurer un paiement des médecins qui permette à la fois de favoriser la qualité et de réduire les dépenses. Nous nous plaçons pourtant dans un cadre simple éludant les problèmes de sélection des patients et offrant une certaine flexibilité des systèmes telle qu'un système mixte peut être instauré. Sous certaines conditions, nous montrons que l'optimum de Pareto est accessible. Cela suppose qu'il n'existe pas de demande induite[6]. Lorsque cette hypothèse est relâchée, le gouvernement doit choisir de favoriser une amélioration de la qualité ou la maîtrise des dépenses en fonction du poids relatif dans l'économie de l'un ou de l'autre des objectifs.

Le modèle considère un consommateur représentatif[7] qui peut souffrir d'une maladie bénigne ou grave. Dans ce cas, le patient nécessite un traitement adapté. Conformément au système de capitation, le patient s'est d'ores et déjà inscrit sur la liste de son médecin qui reçoit un transfert forfaitaire du gouvernement. Une fois le diagnostic établi, le médecin décide non seulement de la quantité globale de soins administrée au patient mais aussi de la répartition des soins entre les actes auto-prescrits (fournis par le médecin lui-même) et ceux prescrits mais fournis par des intervenants extérieurs. Cette distinction entre producteurs permet de comprendre le choix quantitatif du médecin en termes d'auto-prescription. Le diagnostic peut cependant être erroné : il détecte alors une maladie bénigne au lieu d'une maladie grave et le patient requiert un traitement complémentaire. La probabilité de commettre une erreur de diagnostic est tout d'abord supposée exogène, puis endogène dépendant d'un effort consenti par le médecin (cet effort peut être interprété comme la durée de la consultation). Nous étudions deux cas : le premier considère que le médecin prescrit la quantité minimale nécessaire à la guérison du patient; dans le second cas, le médecin use de son pouvoir discrétionnaire et prescrit des actes superflus non nécessaires au traitement (une quantité de soins que n'aurait pas consommée le patient s'il avait eu la même information que le médecin, une demande induite). Pour réduire les dépenses, le gouvernement va inciter le médecin à réduire la fréquence d'erreur. Nous montrons que la contrainte de participation des médecins, la possibilité d'améliorer la qualité et un objectif de maîtrise des dépenses peuvent justifier la mise

[5] Mougeot (1999) est plus sceptique quant à la mise en place d'un système d'abonnement associé à la capitation en raison des nombreux obstacles comme l'attachement à la liberté de choix du médecin.

[6] Cette terminologie, introduite par Evans (1974), caractérise le pouvoir discrétionnaire du médecin sur la demande des patients. Dormont et Delattre (1999) vérifient empiriquement qu'il existe des phénomènes d'induction de la demande de soins pour les médecins de secteur 1.

[7] Keeler (1990) et Ma (1994) justifient le paiement à l'acte par le problème de la sélection des patients. Nous supposons ici que l'ensemble de la population considérée appartient au même GHM.

en place d'un système de rémunération mixte. Toutefois, le gouvernement ne peut pas atteindre simultanément ses deux objectifs. Lorsque le médecin prescrit des quantités excessives de soins auto-prescrits superflus (forte demande induite), seul un système de rémunération par paiement forfaitaire anticipé permet de réduire les dépenses et dans ce cas, le régulateur renonce à améliorer la qualité. Au contraire, lorsque l'ampleur de la demande induite reste limitée, un système mixte de rémunération permet d'obtenir la qualité optimale mais concède une rente au médecin.

Dans la section 2, nous présentons le modèle. Nous étudions ensuite la politique optimale du gouvernement, lorsque la qualité du diagnostic est exogène. Nous distinguons les résultats en fonction du comportement de sur-prescription du médecin. Dans la section 4, la qualité du diagnostic est endogène. Sans le problème de sur-prescription, l'optimum peut être obtenu par un mécanisme de rémunération mixte. En présence de demande induite, nous discutons enfin du mécanisme de rémunération retenu en fonction du choix du gouvernement favorisant la qualité ou l'efficacité (diminution des coûts).

2 Modèle et hypothèses

L'économie est constituée de quatre types d'agents : un patient potentiel, un docteur qui établit le diagnostic et la prescription, les « autres producteurs » de soins et enfin, le gouvernement qui instaure un système d'assurance maladie et le système de rémunération des médecins.

2.1 Patient représentatif

Le consommateur dispose d'un revenu brut I_b et peut être malade avec une probabilité p. Avec cette même probabilité, il consulte un médecin. Il existe deux types de maladies : une maladie grave qui se déclare avec une probabilité p_s et une autre bénigne contractée avec une probabilité $(p - p_s)$. Les individus ne sont pas informés de la gravité de la maladie mais ils le sont sur la nécessité d'un traitement en raison de leur état.

Les individus ont de l'aversion vis-à-vis du risque. La dépense en cas de maladie dépend de la prescription du médecin et le coût résiduel pour le consommateur est fonction du ticket modérateur éventuel. La fonction d'utilité prend en compte l'aversion pour le risque des agents[8], $u(C)$, $u'(C) > 0$ et $u''(C) < 0$ où C est le niveau de consommation déter-

[8] La fonction d'utilité $u(C)$ n'est qu'une fonction de la consommation. Cette hypothèse est la conséquence d'une autre hypothèse implicite : en cas de maladie, le traitement permet de restaurer l'état de santé initial.

miné par les contraintes budgétaires individuelles.(le prix unitaire du bien composite est égal à 1).

2.2 Médecins et autres producteurs de soins

Il existe deux types de producteurs de soins : un médecin et d'« autres producteurs »[9]. Le médecin établit le diagnostic avec une probabilité $(1-\theta)$ de commettre une erreur dans le cas d'une maladie grave. Il prescrit le traitement et réalise lui-même une partie des soins déléguant l'autre aux autres producteurs. La technologie de production tient compte de divers aspects : d'une part, les services fournis par les autres intervenants sont conditionnels à la prescription du médecin et d'autre part, les soins produits par l'un ou l'autre des producteurs ont des rentabilités différentes.

Pour le diagnostic de la maladie i, le médecin prescrit la quantité q_i de soins de santé qui est définie comme le traitement minimal nécessaire :

$$q_i = X_i(y_i) + y_i \qquad \text{pour } i = b, s$$

où $X_i(y_i)$ est la quantité de soins exécutée par le docteur dont le coût unitaire est normalisé à 1 et y_i est la quantité de soins prescrite par le médecin et fournie par les autres intervenants. En supposant qu'il n'existe pas de coopération entre le médecin et les producteurs, il ne surestime pas cette quantité dont le coût unitaire est aussi normalisé à 1.

Le choix de la répartition des soins est, pour le médecin maximisant son profit espéré, le résultat d'un arbitrage financier. Les deux types de soins sont des substituts non parfaits puisque leur rentabilité est différente : la hausse d'une unité de y_i permet de réduire d'au moins autant la quantité de soins auto-prescrits $X_i(y_i)$. Cette fonction se caractérise par $X_i'(y_i) < 0$; plus le médecin prescrit des soins externes, moins il a la possibilité d'en réaliser lui-même et ce à un taux croissant, $X_i''(y_i) > 0$. Les soins y_i peuvent s'interpréter comme une aide au diagnostic qui évite au médecin de tâtonner ou comme une partie du traitement. Lorsque y_i augmente, la quantité globale de soins consommés est réduite et le coût du traitement est donc moins élevé pour le patient comme pour le médecin.

Lorsqu'une maladie grave n'est pas détectée à la première visite (probabilité $p(1-\theta)$), le patient reçoit un traitement q_b insuffisant qu'il faut compléter par la quantité S délivrée par les autres producteurs[10].

[9] Ces « autres producteurs » de soins sont par exemple, un laboratoire d'analyses, un centre de radiologie, de kinésithérapie, l'hôpital, etc. Ils aident sensiblement au diagnostic et au traitement.

[10] L'individu gravement malade, dont le traitement a été retardé par l'erreur du médecin, reçoit par exemple, de l'hôpital, la quantité S de soins (prescrite ou non par le médecin).

Le pouvoir discrétionnaire des médecins est à l'origine des problèmes de demande induite et donc de sur-prescription[11]. Parmi les actes prescrits, certains sont indispensables au traitement comme la consultation alors que d'autres peuvent être superflus ou inutiles et permettent au médecin, dans le cas d'un paiement à l'acte, d'atteindre son revenu cible[12]. Nous supposons que le médecin ne prescrit jamais un traitement lourd réservé aux maladies graves lorsque le diagnostic révèle une maladie bénigne[13]. Nous distinguons deux cas :

- pas de demande induite : le médecin prescrit la quantité de traitement minimale pour l'état de santé qu'il observe : soit q_i pour $i = b, s$;
- si $i = s$, la quantité de soins nécessaires q_s est déjà très importante, mais lorsque $i = b$, le médecin peut utiliser son avantage informationnel et son pouvoir de prescripteur pour recommander ou imposer une quantité de soins que le patient ne choisirait pas de consommer s'il disposait de la même information que le médecin (demande induite). Il prescrit alors une quantité Q de soins qu'il fournit lui même et dont le but est d'accroître son revenu. Si le patient souffre effectivement d'une maladie bénigne, la consommation de Q occasionne une perte d'utilité pour le patient de $v(Q)$, telle que $v'(Q) > 0$, $v''(Q) > 0$ et $v(0) = 0$. En revanche, lorsque le patient est gravement malade et ne reçoit que q_b, la fourniture de Q ne procure pas de perte d'utilité mais n'améliore pas l'état de santé : la quantité supplémentaire requise reste S.

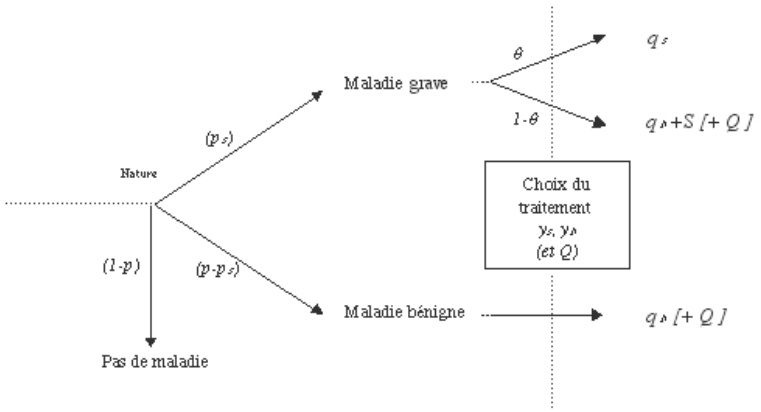

Figure 1 : *Déroulement du jeu.*

[11] Pour éviter les dépenses dues à la sur-prescription (dans ce modèle, la sur-prescription est uniquement la conséquence d'une demande induite), des politiques d'enveloppes globales avec sanctions ont été instaurées voir Hartmann et *al.* (2000) pour une bonne synthèse de ces mécanismes et des premiers constats.

[12] Cette terminologie est introduite par Evans (1974).

[13] Dans le modèle, un seul type d'erreur de diagnostic existe. La dernière hypothèse impose au médecin de choisir le traitement dans la technologie de traitement adaptée à son diagnostic. La maladie grave exige un traitement lourd et cher. Un médecin détectant une maladie bénigne détermine un traitement q_b.

Le patient consomme la totalité de la quantité de soins prescrits q_i, même lorsque celle-ci n'est pas adéquate. La figure résume les hypothèses du modèle.

Quelle que soit la répartition des quantités de soins en fonction des différents producteurs, une maladie grave de type s nécessite toujours plus de soins que le traitement d'une maladie de type b : $X_s(y_s) + y_s > X_b(y_b) + y_b$ ou $q_s > q_b$.

2.3 Gouvernement

L'objectif du gouvernement est de maximiser une fonction de bien-être social de type utilitariste. Les médecins y sont représentés comme l'ensemble des consommateurs et le gouvernement prend en compte la contrainte de participation assurant aux médecins un profit espéré positif ou nul.

Une des priorités est la réduction des dépenses en soins de santé qui permet de réduire le prélèvement fiscal. Le gouvernement instaure un système d'assurance maladie remboursant au taux α les éventuelles dépenses de santé, $(1 - \alpha)$ est le taux du ticket modérateur.

Les ressources du gouvernement sont uniquement fiscales. Si I est le revenu disponible du consommateur représentatif, la recette fiscale est alors $I_b - I$. À travers son rôle d'assureur, le gouvernement supporte au moins une part des dépenses et a donc intérêt à organiser, en fonction de ses objectifs, la rémunération des prescripteurs de soins.

Le gouvernement verse au médecin le forfait K par patient inscrit sur sa liste : K est la partie du revenu du médecin associée au système de capitation et reçue en début de période. En fonction des actes qu'il fournit, le médecin reçoit en plus, le remboursement de la part δ de ses coûts. À sa charge, le médecin conserve la part $(1 - \delta)$ de ses coûts et paie directement aux producteurs extérieurs la part γ du coût des soins de type y prescrits[14]. Si $\gamma < 1$, la part $(1-\gamma)$ des prescriptions y est remboursée au patient par l'assurance sociale (sauf ticket modérateur éventuel).

Quatre outils K, δ, γ, et α sont disponibles pour instaurer les deux mécanismes : une assurance maladie pour les patients et le système de rémunération des médecins.

[14] Le système de capitation à « l'anglaise » (General Practitionners Fundholders) où les médecins généralistes sont des entrepreneurs de soins, est illustré dans le modèle par les paramètres (K, γ), où K est le transfert fourni *ex ante per capita* et γ est la part de l'enveloppe consacrée aux paiements des soins de second recours (prescrits).

3 Information symétrique

L'état de la nature observé *ex post* conditionne les variables i ($i = b, s$). Les consommations individuelles dépendent de l'état de la nature, de l'erreur de diagnostic possible du médecin et éventuellement de la surprescription (demande induite). Ainsi, C_0 représente la consommation lorsque l'individu n'est pas malade, C_b lorsqu'il souffre d'une maladie bénigne, C_s et C_{sS} sont les consommations respectives lorsque le patient est gravement malade et que le médecin réalise ou non le bon diagnostic.

$$\begin{aligned} C_0 &= I, \\ C_b &= I - (1-\alpha)\left[q_b + Q\right], \\ C_s &= I - (1-\alpha) q_s, \\ C_{sS} &= I - (1-\alpha)\left[q_b + Q + S\right]. \end{aligned}$$

Lorsque le médecin ne prescrit pas de soins inutiles $Q = 0$. Cependant, lorsque l'individu consomme $Q \neq 0$ et que sa maladie est bénigne, il subit une perte d'utilité égale à $v(Q)$. L'utilité espérée d'un patient s'écrit donc :

$$\begin{aligned} Eu(C) = (1-p)\, u(C_0) + (p - p_s)\left[u\left(C_b\right) - v(Q)\right] \\ + p_s \theta u\left(C_s\right) + p_s\left(1-\theta\right) u\left(C_{sS}\right). \end{aligned}$$

3.1 Optimum social

Le gouvernement maximise l'espérance d'utilité des consommateurs sous contrainte budgétaire. Il le fait de telle manière que le médecin accepte de soigner un patient quel que soit le diagnostic (contrainte de participation). La contrainte de ressources de l'économie prend en compte la contrainte de rationalité du médecin : le terme de gauche de (1) représente le revenu brut d'un ménage et constitue les recettes disponibles dans l'économie alors que les termes de droite représentent la répartition des richesses entre les différents agents en fonction des états de la nature.

$$\begin{aligned} I_b = &(1-p)C_0 + (p - p_s)C_b + p_s\theta C_s + p_s(1-\theta)C_{sS} \\ &+ p_s\theta X_s(y_s) + (p - p_s\theta)\left[X_b(y_b) + Q\right] \\ &+ p_s\theta y_s + (p - p_s\theta)y_b + p_s(1-\theta)S \end{aligned} \tag{1}$$

Si λ est le multiplicateur de la contrainte de ressources de l'économie, le programme du gouvernement s'écrit alors :

$$\max_{C_0, C_s, C_b, C_{bs}, y_s, y_b} Eu(C_i) - (p - p_s)\, v(Q)$$
$$slc(1)(\lambda)$$

Les conditions de premier ordre permettent d'obtenir l'égalité suivante :

$$u'_C(C_0) = u'_C(C_b) = u'_C(C_s) = u'_C(C_{sS}) = \lambda \tag{2}$$

Les valeurs optimales des niveaux de consommation sont indépendantes de l'état de la nature puisque l'utilité marginale de la consommation est constante. Les propriétés de la fonction u permettent de conclure que la consommation reste inchangée quel que soit l'état de la nature. À l'optimum social, le gouvernement propose une assurance maladie complète $\alpha^* = 1$ indépendamment du coût du traitement[15].

À l'optimum social, la répartition des soins entre les différents producteurs est déterminée par les conditions de premier ordre :

$$X'_s(y_s) = X'_b(y_b) = -1 \tag{3}$$

La quantité optimale de soins y_i est telle que le coût d'une unité supplémentaire fournie par les autres intervenants est égal à son bénéfice marginal en termes de réduction de soins fournis par le médecin lui-même. À l'optimum, la baisse du coût des soins $X_i()$ compense parfaitement le coût d'une unité de traitement supplémentaire y_i (augmenter encore le niveau de y_i n'est alors plus avantageux car la quantité de soins X_i sera réduite de moins que 1). Le niveau socialement optimal de soins délégués est tel que le médecin privilégie la prescription de y_i tant que le bénéfice marginal réalisé par la baisse des $X_i()$ est plus grand que le coût marginal supposé égal à 1. La répartition efficace des soins entre les différents producteurs n'est pas affectée par la présence de demande induite Q.

La condition du premier ordre du Lagrangien de ce programme noté Λ, détermine le niveau optimal de cette quantité de soins Q et confirme parfaitement l'intuition :

$$\frac{\partial \Lambda}{\partial Q} = -(p - p_s)\, v'(Q) - \lambda\,(p - p_s\theta) < 0 \tag{4}$$

Le niveau socialement désirable est bien évidemment $Q^* = 0$. Cette quantité de soins ne contribue pas à l'amélioration de l'état de santé du patient et peut provoquer une perte de satisfaction (lors d'une maladie bénigne). Le premier terme représente cet effet négatif qui croît avec la quantité de soins Q. Le deuxième terme représente le coût supporté par le gouvernement pour financer ce traitement qui n'a pas de contrepartie pour les individus. Le but du gouvernement est donc de limiter au maximum la prescription de traitement inutile $Q^* = 0$.

[15] Ce résultat n'est plus vérifié si le consommateur participe à la détermination du traitement, c'est-à-dire en présence d'aléa moral du côté de la demande (voir Franc (2000)).

3.2 Solution décentralisée

Le gouvernement opte pour les valeurs des outils de régulation (α, δ, γ, K) telles que le médecin soit incité à choisir l'optimum au sens de Pareto (sachant que celui-ci prend en compte la contrainte de participation). Puisque la solution optimale garantit aux consommateurs une couverture complète ($\alpha^* = 1$), le revenu du médecin est entièrement un transfert du gouvernement. En supposant que tous les actes effectués ont le même coût unitaire, le médecin maximise son profit en résolvant le programme suivant :

$$\max_{y_s, y_b, Q} E[\pi] = \{K + (\delta - 1)[p_s \theta X_s(y_s) + (p - p_s\theta)[X_b(y_b) + Q]] - \gamma[p_s\theta y_s + (p - p_s\theta)y_b]\}.$$

Le transfert K, associé au mécanisme de capitation, est reçu *ex ante* et assure la participation du médecin. Le deuxième terme rembourse au titre du paiement à l'acte, la partie δ du coût des actes produits et fournis par le médecin. Le troisième terme est la part de son enveloppe qu'il consacre aux soins prescrits et délégués. Le médecin choisit la répartition des quantités de soins selon :

$$(\delta - 1)X_s'(y_s) = (\delta - 1)X_b'(y_b) = \gamma \tag{5}$$

Quel que soit le type de maladie détectée i, γ est le coût pour le médecin à prescrire une unité de y_i. En prescrivant cette unité et lorsque $\delta > 1$ (l'État lui rembourse plus que son coût), le médecin réalise un manque à gagner représenté par $(\delta - 1)X_i'(y_i)$.

Pour que le médecin choisisse la répartition des soins socialement efficace entre les producteurs, (3) et (5) doivent coïncider. Le gouvernement détermine γ et δ selon la relation :

$$\gamma = 1 - \delta \tag{6}$$

En effet, tant que $\gamma < 1 - \delta$, la part financière du médecin dans le paiement des soins de second recours est moins lourde que le montant qui reste à sa charge sur les soins qu'il fournit, après le paiement à l'acte. Dans ce cas, la totalité des soins, exceptée la consultation, sera confiée aux intervenants extérieurs. Si au contraire $\gamma > 1 - \delta$, le médecin choisit de produire la totalité des soins nécessaires pour bénéficier de toute la marge possible.

Il apparaît clairement que la répartition optimale des soins entre les producteurs est indépendante de Q. En revanche, si la quantité superflue de soins socialement optimale est nulle $Q^* = 0$, elle ne l'est pas pour le médecin :

$$\frac{\partial E[\pi]}{\partial Q} = (\delta - 1)(p - p_s\theta) > 0 \qquad \text{pour tout } \delta > 1 \tag{7}$$

Si $\delta > 1$, le médecin est incité à effectuer trop d'actes par rapport à l'optimum même si la répartition optimale des soins est garantie.

Proposition 1 *En information symétrique, le gouvernement peut empêcher la prescription de soins inutiles au traitement, éludant le problème de demande induite, en proposant le remboursement des coûts tel que :* $\delta \leqslant 1$, *ainsi* $Q = 0$.

Pour s'assurer que le médecin accepte de participer, le gouvernement détermine le montant du forfait de capitation. Sachant que $\gamma = 1 - \delta$, le forfait *per capita* qui garantit la participation du médecin est tel que :

$$E[\pi] = 0 \Leftrightarrow K = (1 - \delta)[p_s\theta q_s + (p - p_s\theta)q_b]$$

Le gouvernement dispose de suffisamment d'instruments pour instaurer l'optimum de Pareto. Il détermine les taux de paiement du médecin de manière à ce qu'aucun des deux types de soins ne soit avantageux pour le médecin. Tant que les coûts unitaires des soins X_i et y_i sont identiques, la relation $\gamma = 1 - \delta$ assure la répartition socialement efficace des soins. Enfin, le gouvernement choisit K tel que le profit espéré du médecin est nul.

Le gouvernement dispose donc de l'ensemble des systèmes de rémunération possibles, variant d'un pur paiement à l'acte à un paiement par capitation[16] :

$\delta = 1$	$\gamma = 0$ $K = 0$	Pur paiement à l'acte
$\delta \in]0, 1[$	$\gamma = 1 - \delta$ $K = (1 - \delta)[p_s\theta q_s + (p - p_s\theta)q_b]$	Système mixte
$\delta = 0$	$\gamma = 1$ $K = [p_s\theta q_s + (p - p_s\theta)q_b]$	Pur système de capitation

Proposition 2 *Si tous les soins ont le même coût unitaire, trois types de rémunération assurent l'efficacité :*

- *le paiement à l'acte* $\delta = 1$, $K = 0$,
- *le paiement per capita* $K = [p_s\theta q_s + (p - p_s\theta)q_b]$, $\delta = 0$,
- *et tout système intermédiaire.*

Nous avons supposé que le coût de production des soins Q est le même que celui des autres soins $X_i()$ et y_i (égal à 1). Seul le cas du système de capitation est insensible à une modification du coût de production des soins Q : comme $\delta = 0$, le médecin n'a aucune incitation à

[16] Pour les médecins, ces schémas de rémunération sont équivalents car il est assez courant de supposer que les médecins sont neutres au risque. Ma et Mac Guire (1997) font, par exemple, l'hypothèse que le médecin n'est jamais malade et qu'il est neutre vis-à-vis du risque financier.

fournir des actes supplémentaires $Q = 0$. En effet, il est peu vraisemblable que des soins qui n'ont aucun effet sur l'état de santé coûtent au médecin le même « travail » et l'hypothèse qui considère que le coût de ces actes est inférieur à 1 semble plus réaliste.

Proposition 3 *Si le coût unitaire des soins de type Q est inférieur au coût unitaire des autres soins X_i() et y_i, seul le système de rémunération per capita garantit l'efficacité* ($Q = 0$) *:* $\delta = 0$, $\gamma = 1$, $K = [p_s\theta q_s + (p - p_s\theta)q_b]$.

Dès que le médecin réalise un profit sur les unités de soins Q, il prescrit $Q = Q_{\max}$.

3.3 Limites

Puisque le gouvernement instaure une assurance complète, il a la charge de payer la totalité des dépenses éventuelles de traitements et donc de rémunérer les différents producteurs de soins. Pour induire le choix optimal de la répartition des soins entre les différents producteurs, il réduit toute possibilité d'opportunité avantageuse. Si $\delta = 1$, alors $\gamma = 0$, les soins auto-prescrits sont totalement remboursés, mais le médecin peut prescrire des soins extérieurs sans que cela ne lui coûte rien. Dans ce cas précis, le paiement par capitation est inutile et $K = 0$. En information symétrique, un système de rémunération efficace est celui qui évite toute demande induite, c'est-à-dire toute prescription de soins inutiles. Lorsque le coût de production des soins Q éventuellement prescrits par le médecin est le même que celui de tout autre soin, tous les systèmes de rémunération variant du pur paiement à l'acte au pur paiement par capitation permettent de décentraliser la solution optimale. En revanche, lorsque le coût de production de Q est moindre (inférieur à 1), le seul système de rémunération qui garantit alors qu'il n'y ait pas de prescription inutile est le système de rémunération *per capita* où le forfait est l'espérance du coût du traitement.

Le modèle ne considère pas de perte d'utilité du patient consécutive à une erreur de diagnostic mais il met l'accent sur le surcoût du traitement dans le cas d'une prescription inadéquate. Si un médecin peut améliorer la qualité du diagnostic en fournissant un effort, il peut alors réduire le coût global espéré des soins. Quel est alors, dans ce nouveau contexte, le système de rémunération efficace ? Le gouvernement peut-il toujours induire l'efficacité parétienne ?

4 Qualité et système de rémunération

Nous supposons maintenant que le médecin peut améliorer la probabilité de diagnostiquer une maladie grave en fournissant un effort e. Cet effort peut, par exemple, représenter la durée de la consultation. En augmentant son investissement dans le diagnostic, le médecin peut réduire la fréquence d'erreur et ainsi : $\theta(e)$ avec $\theta'(e) > 0$ et $\theta''(e) < 0$. Le médecin choisit son niveau d'effort avant le début de la consultation, autrement dit avant de connaître le type de maladie de son patient. Le niveau d'effort e est alors indépendant du type de maladie. Son coût s'exprime en termes de perte de revenu : $\phi(e)$. Les hypothèses sont classiques $\phi'(e) > 0$ et $\phi''(e) > 0$ et $\phi(0) = 0$, le coût augmente avec l'effort à un taux croissant.

Le gouvernement n'observe pas l'effort réalisé par le médecin. Même s'il constate une erreur de diagnostic, il ne lui est pas possible d'en déduire une mauvaise qualité de la consultation. En préservant l'objectif de minimisation des coûts, le gouvernement va inciter le médecin à fournir un effort pour améliorer la qualité du traitement (via un diagnostic plus probablement adapté).

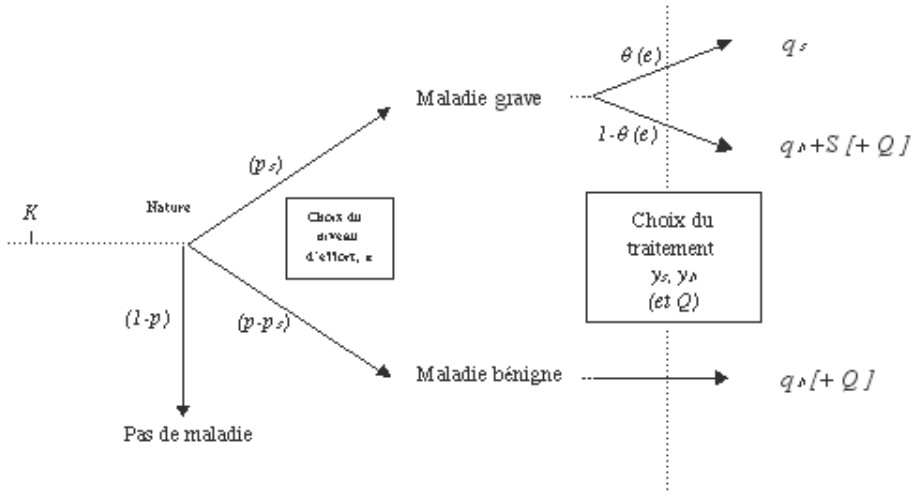

Figure 2 : *Déroulement du jeu avec probabilité endogène.*

4.1 Optimum social

Le gouvernement détermine non seulement la répartition optimale des soins entre les producteurs mais aussi le niveau d'effort optimal en maximisant la fonction de bien-être social représentée par $Eu(C_i) - (p - p_s)v(Q)$ et en disposant des instruments C_i, y_s, y_b, e et Q. La contrainte de ressources prend en compte l'aléa moral et le Lagrangien Δ du pro-

gramme de maximisation du gouvernement s'écrit :

$$\begin{aligned}\Delta =&\{(1-p)u(C_0)+p_s\theta(e)u(C_s)\\&+(p-p_s)\left[u(C_b)-v(Q)\right]+p_s[1-\theta(e)]u(C_{sS})\}\\&+\lambda\{I_b-[(1-p)C_0+(p-p_s)C_b+p_s\theta(e)C_s+p_s[1-\theta(e)]C_{sS}]\\&\quad-[p_s\theta(e)X_s(y_s)+[p-p_s\theta(e)][X_b(y_b)+Q]]\\&\quad-p_s\theta(e)y_s+[p-p_s\theta(e)]\,y_b+p_s\left[1-\theta(e)\right]S-p\phi(e)\}\end{aligned}\tag{8}$$

L'introduction du problème d'aléa moral du côté de l'offre n'affecte pas le comportement du consommateur représentatif. Ainsi, nous retrouvons le résultat similaire à (2) qui montre que l'assurance sociale offre une couverture complète des dépenses de santé :

$$u'_C(C_0)=u'_C(C_b)=u'_C(C_s)=u'_C(C_{sS})=\lambda\Leftrightarrow\alpha^*=1$$

La répartition socialement désirable des soins entre les différents producteurs n'est pas modifiée par la prise en compte du paramètre d'aléa moral. La condition est identique à celle définie dans le cas considérant la qualité exogène (3). Le coût d'une unité de soins y_i compense exactement la baisse de la quantité de soins réalisés par le médecin (coût unitaire égal à 1) $X'_s(y_s)=X'_b(y_b)=-1$.

De façon intuitive, il n'y a aucune raison qu'il y ait un intérêt à laisser le médecin prescrire des actes inutiles et le résultat analytique va dans ce sens :

$$\frac{\partial\Delta}{\partial Q}=-(p-p_s)v'(Q)-\lambda[p-p_s\theta(e)]<0$$

Indépendamment du niveau d'effort consenti par le médecin, toute prescription Q est socialement inefficace et $Q^*=0$. La perte d'utilité consécutive à la consommation de Q n'est pas seule responsable : toute dépense de soins implique un prélèvement fiscal qui a un coût. Le deuxième terme représente le coût social d'une unité de soins Q.

Le niveau d'effort socialement efficace est tel que :

$$\begin{aligned}\frac{\partial\Delta}{\partial e}=&p_s\theta'(e)\left[u(C_s)-u(C_{sS})\right]-\lambda p_s\theta'(e)\left(C_s-C_{sS}\right)\\&-\lambda p_s\theta'(e)\left[X_s(y_s)+y_s-X_b(y_b)-y_b-Q-S\right]-\lambda p\phi'(e)\end{aligned}$$

Comme la couverture d'assurance est complète $C_s=C_{sS}$ et $u(C_s)=u(C_{sS})$. L'expression précédente se simplifie :

$$\begin{aligned}&-p\phi'(e)=p_s\theta'(e)\left[X_s(y_s)+y_s-X_b(y_b)-y_b-Q-S\right],\\&-p\phi'(e)=p_s\theta'(e)[q_s-(q_b+S)-Q]\end{aligned}\tag{9}$$

Le niveau optimal d'effort est déterminé de manière à égaliser le bénéfice marginal et le coût marginal d'une unité supplémentaire d'effort : le bénéfice marginal social retiré d'une unité supplémentaire d'effort consenti se mesure en termes de baisse de la fréquence d'erreur de diagnostic. La conséquence directe de l'amélioration de la qualité est la réduction de la fourniture du traitement complémentaire S. Lorsque $Q = 0$, le coût marginal de l'effort est donc juste égal au gain espéré en termes de compensations distribuées. La différence $q_s - (q_b + S)$ représente le coût du traitement « tardif » d'une maladie grave. En revanche, lorsque le médecin prescrit Q, le bénéfice marginal d'une unité d'effort supplémentaire est moindre.

4.2 Problème d'optimisation du médecin

Le médecin maximise l'espérance de son profit :

$$\max_{y_s, y_b, e, Q} \{K + (\delta - 1)\{p_s\theta(e)X_s(y_s) + [p - p_s\theta(e)]\,[X_b(y_b) + Q]\} - \gamma\{p_s\theta(e)y_s + [p - p_s\theta(e)]y_b\} - p\phi(e)\}$$

L'expression du profit du médecin prend en compte le coût de l'effort, qui n'est réalisé qu'en cas de consultation. Le médecin répartit les quantités de soins entre les différents producteurs de manière à ce que y_s et y_b soient les solutions de :

$$(\delta - 1)X_s'(y_s) = (\delta - 1)X_b'(y_b) = \gamma \tag{10}$$

L'effort est choisi *ex ante*, avant la consultation. La répartition, indépendante du niveau d'effort qu'il décide de fournir, n'est pas modifiée par l'introduction du problème d'aléa moral.

La quantité d'actes superflus est directement dépendante du taux de remboursement des coûts du médecin. Nous obtenons le même résultat que précédemment : si $\delta > 1$, le médecin prescrit le maximum de soins *en sus*.

$$\frac{\partial E[\pi]}{\partial Q} = (\delta - 1)\,[p - p_s\theta(e)] > 0 \qquad \text{lorsque } \delta > 1 \tag{11}$$

Le niveau d'effort optimal est tel que le bénéfice marginal d'une unité supplémentaire soit égal au coût marginal :

$$-p\phi'(e) = -p_s\theta'(e)(\delta - 1)\,[X_s(y_s) - [X_b(y_b) + Q]] + p_s\theta'(e)\gamma(y_s - y_b) \tag{12}$$

Cette expression égalise le coût marginal de l'effort (membre de gauche) et le bénéfice marginal pour le médecin de cette unité d'effort. Prenons le cas où $\delta > 1$. Si le médecin consent à réaliser une unité supplémentaire d'effort, il accroît ses chances de détecter une maladie grave et

donc de fournir une grande quantité de soins $X_s(y_s)$ augmentant son profit espéré. Si $\delta > 1, \gamma < 0$, le médecin est subventionné pour les actes prescrits, il choisit toujours la répartition efficace. Remarquons que la possibilité de prescrire des actes inutiles réduit le bénéfice marginal de l'effort puisque le traitement pour une maladie bénigne lui permet quand même d'exécuter un grand nombre d'actes.

5 Système de rémunération avec aléa moral

Le gouvernement instaure des outils incitatifs qui contraignent le médecin à choisir les valeurs socialement efficaces. La répartition optimale entre les différents producteurs de soins n'est pas modifiée par la présence d'aléa moral. La quantité de soins prescrits est répartie entre le médecin lui-même et les producteurs extérieurs selon la relation $\gamma = 1 - \delta$.

En utilisant cette relation entre δ et γ, (12) peut être simplifiée :

$$-p\phi'(e) = -(\delta - 1)\, p_s \theta'(e)\, [q_s - q_b - Q] \tag{13}$$

Pour que le médecin opte pour le niveau d'effort socialement optimal, (9) et (13) doivent coïncider et le taux de remboursement optimal est alors tel que :

$$\delta = \frac{S}{q_s - q_b - Q}$$

5.1 Sans demande induite, $Q = 0$

Dans le cas où $Q = 0$, le mécanisme de rémunération qui permet au gouvernement d'induire une amélioration de la qualité du diagnostic est un système mixte combinant le système de paiement à l'acte et le paiement *per capita*, tel que $\delta = \frac{S}{q_s - q_b}$.

Proposition 4 *Si $Q = 0$, alors le système de rémunération incitatif est tel que*

$$\delta = \frac{S}{q_s - q_b} \qquad \textit{et} \qquad K = p\phi(e) - (\delta - 1)\{p_s\theta(e)q_s + [p - p_s\theta(e)]q_b\}$$

En supposant que $S + q_b > q_s$, alors $\delta > 1$, le système de rémunération est mixte.

Nous supposons que $S + q_b > q_s$: en cas d'erreur de diagnostic, le coût du traitement complémentaire ajouté à q_b, est bien sûr plus coûteux que le traitement q_s. Le gouvernement a intérêt à favoriser la qualité du diagnostic pour réduire la fréquence d'erreur.

Puisque $S + q_b > q_s$, le taux δ rembourse au médecin plus que le coût unitaire pour chaque acte, $\delta > 1$.

Remarquons tout d'abord que δ est strictement positif et ne dépend pas du niveau d'effort. Ainsi, le médecin reçoit toujours un remboursement d'une partie de ses coûts. Puisque $\delta > 1$, le gouvernement instaure un mécanisme qui compense le médecin d'un montant supérieur à son coût. Il réalise un bénéfice sur chaque acte exécuté. Ce résultat est l'illustration de l'incitation du gouvernement à améliorer la qualité du diagnostic. Bien que δ soit indépendant du niveau d'effort, l'attrait généré par cette marge potentielle sur chaque acte incite le médecin à prescrire une quantité maximale de soins sachant que $Q = 0$. Avec $\delta > 1$, le médecin est incité à détecter une maladie sérieuse pour prescrire $q_s > q_b$. Le médecin améliore alors sa probabilité θ de détecter une maladie grave et donc la qualité de son diagnostic. Comme $\delta > 1$ alors $\gamma < 0$, les traitements fournis par les intervenants extérieurs sont subventionnés. Cette subvention freine le médecin qui pourrait être incité à fournir la totalité des soins, du fait du profit qu'il peut en obtenir.

De part le résultat du taux optimal de paiement à l'acte δ, nous constatons que plus le coût de l'erreur est grand, c'est à dire plus S est grand par rapport à la différence entre q_s et q_b, et plus la marge réalisable par le médecin est importante. En d'autres termes, lorsque le coût d'une erreur est important, le gouvernement incite d'autant plus le médecin à améliorer θ.

Notons aussi que les pouvoirs publics complètent la rémunération du médecin par un transfert $K \neq 0$. Le gouvernement ne peut se contenter d'un pur système de paiement à l'acte pour décentraliser l'allocation optimale. Le transfert *per capita* dépend du niveau d'effort socialement efficace et assure la participation du médecin quelle que soit la maladie du patient. Ce transfert rembourse *ex ante*, le coût de l'effort investi par le médecin durant la consultation. Via ce transfert, le gouvernement récupère le bénéfice espéré qu'il cède au médecin en cas de maladie. Le forfait de capitation K est donc la compensation espérée du coût de l'effort réduite de ce bénéfice espéré. Tant que l'effort du médecin est suffisamment coûteux, le transfert reçu par le médecin lors de l'enregistrement des patients sur les listes est strictement positif[17] :

$$K > 0 \qquad \text{lorsque } p\phi(e) > \left[\frac{S}{q_s - q_b} - 1\right] [p_s\theta(e)q_s + [p - p_s\theta(e)]q_b]$$

Lorsque $Q = 0$, le gouvernement dispose de suffisamment d'instruments pour décentraliser l'allocation optimale et instaure un système mixte, $\delta > 1$ et $K \neq 0$.

[17] Si le coût de l'effort est faible, le transfert est négatif et peut s'interpréter comme un droit d'accès au marché des soins dont le médecin devrait s'acquitter à l'inscription des patients. Lorsque il y a égalité, $p\phi(e) = ([S/(q_s - q_b)] - 1)\,[p_s\theta(e)q_s + [p - p_s\theta(e)]\,q_b]$, le système est un pur paiement à l'acte.

5.2 En présence de demande induite, $Q \neq 0$

Un système de paiement qui laisse au médecin l'opportunité de réaliser une marge sur les soins fournis favorise la qualité mais va à l'encontre des incitations nécessaires à la limitation des dépenses. En maximisant son profit avec $\delta > 1$, le médecin prescrit, en cas de diagnostic d'une maladie bénigne, une quantité maximale de soins superflus $Q = Q_{\max}$. L'effort socialement optimal est réduit par l'éventualité d'une prescription de soins Q (le bénéfice marginal de l'effort (9) est une fonction décroissante de Q). Il est impossible pour le gouvernement, d'instaurer une politique encourageant simultanément une amélioration de la qualité et une limitation des dépenses.

Puisque $Q \neq 0$, le taux de remboursement optimal est une fonction croissante avec Q telle que

$$\delta = \frac{S}{q_s - q_b - Q}$$

Ce résultat paraît contre-intuitif : lorsque δ (supérieur à 1) augmente, le médecin est incité à exécuter plus d'actes. Supposons pour simplifier que le coût unitaire de production de Q est nul[18]. Tant que $\delta > 0$, le médecin prescrit $Q = Q_{\max}$ et dès que $\delta \leqslant 0$ alors $Q = 0$.

• Lorsque $q_s - q_b > Q_{\max}$, soit lorsque la capacité de sur-prescription des médecins est suffisamment faible, inférieure à l'écart $q_s - q_b$, le taux de remboursement δ est positif. Dans ce cas, le système de rémunération met l'accent sur la nécessité d'améliorer la qualité et incite le médecin à fournir l'effort optimal. Si $Q_{\max}$ est suffisamment faible, δ est strictement plus grand que 1.

$$\delta = \frac{S}{q_s - q_b - Q_{\max}} \Leftrightarrow \delta > 0 \qquad \text{et} \qquad S + q_b > q_s - Q_{\max},$$

$$\text{par conséquent } \delta > 1,$$

puisque $S + q_b > q_s$ alors $S + q_b + Q > q_s$. Le système de rémunération favorise l'amélioration de la qualité mais n'empêche pas les médecins de profiter du taux de remboursement et de prescrire $Q > 0$ et $Q = Q_{\max}$.

Nous obtenons le même mécanisme de rémunération que lorsque le médecin ne prescrit que la quantité minimale nécessaire q_i. Le gouvernement ne tient pas compte de l'excès de prescriptions et favorise, dans son choix de politique, une amélioration de la qualité des diagnostics. Même si le transfert K reçu *ex ante* par le médecin, tient compte de la demande induite (K est une fonction décroissante avec $Q_{\max}$), l'allocation obtenue n'est pas la solution de premier rang. La

[18] Si le coût unitaire de production de Q est inférieur à 1 mais non nul, l'arbitrage entre réduction des dépenses et amélioration de la qualité est moins critique, au delà d'un certain seuil, appelé ici $Q_{\max}$, le coût de fourniture pour le médecin devient infini.

consommation de ces soins engendre une désutilité $v(Q)$ pour certains patients.

Proposition 5 *Lorsque la capacité de sur-prescription du médecin est relativement limitée $Q_{\max} < q_s - q_b$, le gouvernement favorise l'amélioration de la qualité. Le système de rémunération incitatif est tel que :*

$$\delta = \frac{S}{q_s - q_b - Q_{\max}} > 1 \text{ et } K = p\phi(e) - (\delta - 1)\{p_s\theta(e)q_s + [p - p_s\theta(e)]\,[q_b + Q_{\max}]\}$$

• À l'inverse si $q_s - q_b - Q_{\max} < 0$, le taux de remboursement des coûts du médecin est négatif ou nul, $\delta \leqslant 0$.

$$\delta = \frac{S}{q_s - q_b - Q_{\max}} \qquad \text{et} \qquad Q_{\max} \to \infty \qquad \text{alors} \qquad \delta \to 0$$

$$\text{Si } \delta = 0, \text{ alors } Q = 0$$

Ce taux de remboursement n'incite plus le médecin à faire le moindre effort. La condition (12) permet d'obtenir ce résultat car si $\delta = 0$ et $\gamma = 1$, le bénéfice marginal de l'effort est donc $p_s\theta'(e)\,[q_s - q_b - Q_{\max}]$ qui est négatif puisque $q_s - q_b - Q_{\max} < 0$. Puisque $Q_{\max}$ est très grand, le médecin n'est pas incité à détecter une maladie grave pour exécuter beaucoup d'actes. Dans ce cas, la politique ne conduit pas à une amélioration de la qualité mais a pour objectif la réduction des dépenses. La probabilité de réaliser un bon diagnostic est inchangée : $\theta(0) = \theta$. Le système de rémunération est le système de capitation. Puisque $Q_{\max}$ est très grand, δ tend vers 0.

Proposition 6 *Lorsque la capacité de sur-prescription du médecin est élevée $Q_{\max} > q_s - q_b$, le gouvernement favorise la réduction des dépenses. Le système de rémunération est le système de capitation tel que $\delta = 0$ et $K = p_s\theta q_s + (p - p_s)\,\theta q_b$.*

Sans outil de régulation supplémentaire, il est impossible pour le gouvernement d'inciter le médecin à adopter le comportement efficace. Les systèmes de rémunération disponibles ne peuvent simultanément inciter le médecin à améliorer la qualité de son diagnostic et le contraindre à ne prescrire que la quantité minimale de soins nécessaires. Dans aucune des deux situations précédemment étudiées l'optimum social n'est atteint. Lorsque la sur-prescription est de faible ampleur, le gouvernement favorise l'amlioration de la qualité et laisse une rente au médecin. Bien que le forfait *per capita* prenne en compte ce bénéfice, le patient subit une perte d'utilité. Lorsque la sur-prescription est très coûteuse, l'objectif principal du gouvernement est de limiter les dépenses, et l'effort du médecin pour améliorer la qualité est nul.

6 Conclusion

Lorsqu'il n'y a pas de demande induite et que la qualité du diagnostic est exogène, nous avons vu que le gouvernement dispose d'un continuum de systèmes de rémunération des médecins variant entre les purs systèmes de paiement à l'acte et de capitation. Tous permettent de décentraliser l'allocation optimale. Lorsque le médecin utilise son avantage informationnel et prescrit des quantités supplémentaires de soins qui, par ailleurs, ne lui coûtent rien, le seul système de rémunération qui permet d'atteindre l'optimalité est le paiement par transfert forfaitaire *per capita*.

Quand la qualité du diagnostic est endogène, il devient impossible pour un même gouvernement d'induire en présence de demande induite, le comportement efficace du médecin. Si les quantités prescrites *en sus* restent suffisamment faibles, le gouvernement peut instaurer un taux incitatif de remboursement des actes. Le système complet de rémunération est alors mixte. Au titre du paiement à l'acte, le médecin est compensé pour ses coûts et peut même réaliser un profit sur les soins fournis. Le forfait reçu, *ex ante*, est d'autant plus faible que la sur-prescription éventuelle est importante. La qualité est améliorée mais le problème de demande induite n'est pas résolu. À l'inverse, si la quantité de soins superflus est excessive le gouvernement s'attache, en priorité, à réduire toute opportunité financière. Le système de rémunération adopté alors est le paiement par capitation qui ne fournit quant à lui aucune incitation pour améliorer la qualité. Les mécanismes de rémunération instaurés dans les deux situations ne conduisent pas à la solution optimale.

Les résultats sont cohérents avec les observations[19] et montrent qu'il est très difficile, voire impossible, pour un gouvernement de mettre en place une rémunération efficace des producteurs de soins, sans disposer d'un grand nombre d'outils. En effet, lorsque le médecin commet une erreur de diagnostic, il suffirait pour restaurer l'efficacité, de le pénaliser[20] en lui faisant supporter une part du coût financier du traitement tardif. Cette mesure nécessite une bonne connaissance par le régulateur de l'état de santé des individus et de l'effort du médecin qui sont par définition, l'un comme l'autre, difficiles à observer. Ici, nous avons supposé que le médecin n'était ni pénalisé en cas d'erreur de diagnostic (ce qui correspond plutôt à la situation française) ni sanctionné par la perte potentielle de patients à la période suivante, lors de la réinscription sur les listes de patients. Il faudrait, pour introduire ce facteur, étudier le modèle sur deux périodes, ce qui reviendrait à

[19] Voir les études empiriques Besley, Hall et Preston (1999) et Majnoni d'Intignano (1998).

[20] Aux États-Unis, de nombreux procès intentés contre des médecins indépendamment du régulateur rétablissent une incitation à améliorer la qualité des soins mais aussi des diagnostics. Toutefois, un excès de cette pratique décourage les médecins qui refusent alors de soigner certains patients.

intégrer des effets de réputation. Gravelle (1999) considère que la concurrence intervient dans la phase d'inscription (différenciation géographique, modèle de Salop), et obtient que la rémunération du médecin devrait être constituée, pour inciter à fournir de la qualité, par un minimum forfaitaire dont une part serait proportionnelle au nombre de patients.

Nous avons également supposé que la sur-prescription ne pouvait être que la conséquence d'une demande induite. Le patient joue pourtant un rôle dans la détermination du traitement. Le modèle ainsi posé élude les problèmes de régulation du côté de la demande et n'envisage donc pas une approche globale de la régulation offre et demande de soins. De même, la prise en compte de la relation entre le médecin prescripteur et les autres producteurs de soins mériterait une étude plus approfondie. Notamment, il est intéressant de modéliser les différents types de soins mais il semble fondamental de tenir compte de la différence des coûts (les soins fournis par les intervenants extérieurs sont en général plus coûteux).

Nous avons supposé, dans ce modèle, que le gouvernement de type utilitariste avait pour objectif la maximisation du bien-être social modélisé par la somme des utilités des consommateurs. Cependant, il semble que cette hypothèse puisse être sujette à discussion. Mougeot (1999) souligne que l'observation des structures et des règles de paiement qui laissent aux offreurs des rentes de situation rappelle les résultats de la théorie de la captation de la régulation.

Nous retrouvons ici, dans le contexte de la médecine ambulatoire, l'arbitrage classique entre la réduction des coûts et l'amélioration de la qualité. Le problème de sélection des patients n'intervient pas puisque tous les consommateurs sont identiques au moment de l'inscription. Même sans ce dernier problème, les résultats montrent que le choix du système de rémunération dénote une orientation politique favorisant l'efficacité ou la qualité. Lorsqu'il existe des phénomènes d'induction de la demande, aucun système ne permet plus d'obtenir l'allocation efficace.

Bibliographie

Besley T., Hall J., Preston I., 1999, "The demand for private health insurance : do waiting lists matter ?", *Journal of Public Economics*, 72, 155-181.

Breuil-Genier P. et al., 1997, « Analyse empirique de la consommation de soins de ville au niveau micro-économique », mimeo.

Chiappori P. A., Durand F., Geoffard P. Y., 1998, "Moral hazard and the demand for physician services : first lessons from a french natural experiment", *European Economic Review*, 42, 499-511.

Dormont B., Delattre E., 2000, « Induction de la demande de soins par les médecins libéraux français. Un test microéconométrique sur données de panel », *Economie et prévisions* n°142.

Evans R.G., 1974, "Supplier induced Demand : Some empirical evidence and implications", Mark Perlman (Ed), The Economics of Health and Medical Care, Proceedings of a conference held by the international Economic association, Tokyo, 162-173.

Franc C., 2000, "Social insurance and redistributive mechanism", mimeo.

Goodall, C., 1990, "A simple Objective method for determining a Percent Standard in mixed Reimbursement Systems", *Journal of Health Economics*, 9, 253-271.

Gravelle H., 1999, "Capitation contracts : access and quality", *Journal of Health Economics*, 18, 315-340.

Gravelle H., Masiero G., 2000, "Quality incentives in a regulated market with imperfect information and switching costs : capitation in general practice", *Journal of Health Economics*, 19, 1067-1088.

Hartmann L., Rochaix-Ranson L., de Kervasdoué J., 2000, « La régulation économique du système de santé » in *Le carnet de santé de la France en 2000,* Fédération de la Mutualité française et éditions Syros.

Keeler E. B., 1990, "What proportion of Hospital Cost Differences is justifiable ?", *Journal of Health Economics*, 9, 359-365.

Ma C.A., 1994, "Health care payment systems : cost and quality incentives", *Journal of Economics & Management Strategy*, Volume 3, Number 1, 93-112.

Ma C.A., McGuire T.G., 1997, "Optimal Health Insurance and Provider Payment", *The American Economic Review*, 4, 685-704.

Newhouse J.P., 1996, "Reimbursing Health Plans and Health Providers : Selection versus Efficiency in Production", *Journal of Economic Literature*, 34, 1236-1263.

Majnoni d'Intignano B., 1998, « La performance qualitative du système de santé français », Complément A in *Régulation du système de santé*, Rapport du Conseil d'Analyse Economique, n°13, La Documentation Française.

Mougeot M., 1999, *La régulation du secteur de la santé*, Rapport du Conseil d'Analyse Economique, n°13, La Documentation Française.

Pauly M. V., 1980, *Doctors and their Workshops*, Chicago, University of Chicago press.

Pope G.C., 1990, "Using Hospital-Specific Costs to Improve the Fairness of Prospective Reimbursement", *Journal of Health Economics*, 9, 237-251.

Résumé

La rémunération des médecins est au coeur de nombreuses tentatives de réformes. Nous montrons dans cet article pourquoi il est si difficile pour un gouvernement (lorsqu'il est l'assureur) d'instaurer un système de paiement optimal. Un patient représentatif peut contracter deux types de maladies (grave ou bénigne). Si la qualité du diagnostic est endogène, le gouvernement ne peut plus instaurer un système qui favorise simultanément la qualité, en incitant le médecin à fournir un effort, et réduise les dépenses, en évitant les prescriptions excessives (demande induite).

Abstract

The design of reimbursement schemes of health care providers is a main issue in most reforms. We show in this paper why it is so difficult for a government (which is also the insurer) to implement an optimal mechanism. The representative consumer can incur two types of illness (serious or benign). If the quality of the diagnosis is depending on an effort of the physician, it is no longer possible for the government to implement a system which, at the same time, favours the quality and prevents the increasing in health expenses (induced demand).

Mots-clés

Capitation, Paiement à l'acte, Aléa moral, Qualité et Demande induite.

Keywords

Capitation system, Fee for service, Moral hazard, Quality and Induced demand.

Classification JEL : D82, I1, H51