LES MODES DE RÉMUNÉRATION DES MÉDECINS

LISE ROCHAIX *

La question de la rémunération des médecins n'est, certes, pas nouvelle, ainsi qu'en attestent les nombreux écrits d'auteurs classiques. Mais la réponse limpide offerte, par exemple dans *Le malade imaginaire,* à savoir qu'il est dans l'intérêt du malade de bien rémunérer son médecin, a profondément évolué sous l'influence de deux facteurs : d'une part, la socialisation progressive de la dépense, donc la détermination par un tiers - en France l'Assurance maladie - du mode et du niveau de la rémunération ; d'autre part, le rôle de plus en plus central joué par le médecin prescripteur dans l'affectation de ressources collectives trop rares, eu égard au développement rapide du progrès technique et des besoins. Les contraintes macroéconomiques qui pèsent depuis le début des années 1980 sur les systèmes de santé ont conduit à inscrire cette question en priorité dans les programmes de recherche des économistes de la santé, sans pour autant retrouver la simplicité de la démonstration de Molière.

Une réponse a, tout d'abord, été recherchée en comparant les modes de rémunération des médecins dans différents pays développés, avec les travaux pionniers de Roemer (1962) aux USA, puis de Reinhardt et Sandier (1983) en France. Les analyses comparatives développées depuis ont confirmé l'intuition initiale, à savoir que la diversité des niveaux et des modes de rémunération entre pays industrialisés traduit bien l'absence d'un système parfait qui cumulerait tous les avantages. En parallèle à ces travaux de comparaison, ont été menés deux types de recherches : le premier, dans la lignée des travaux de M. Pauly (1980) vise à établir les propriétés incitatives des systèmes « purs » de rémunération en situation

* IDEP-GREQAM, Université de la Méditerranée.

d'incertitude et d'asymétrie d'information ; le second, plus empirique, mesure la performance des divers modes de rémunération sur un ensemble de critères (efficacité, équité, qualité, accès).

Dès lors que le choix du mode de rémunération des médecins conditionne de manière plus générale la bonne utilisation des soins de santé, il convient de rechercher, au-delà de la « juste » rémunération de la compétence et du temps du médecin, une incitation à la « juste » prescription. Or, cette dernière notion est particulièrement difficile à définir, parce qu'elle évolue dans le temps comme dans l'espace, qu'elle est difficile à mesurer et qu'elle comporte des dimensions éthiques fortes comme celle de l'égal accès des patients au progrès technique.

Cette contribution offre, en premier lieu, une présentation synthétique des principaux résultats obtenus dans les travaux économiques sur les modes de rémunération des médecins. Elle vise surtout à susciter une réflexion constructive sur le mode actuel de rémunération des médecins en France et à préciser les contours et les conditions de réussite de son incontournable réforme. Dans cette perspective, seront présentées les grandes lignes des réformes des pays industrialisés en matière de rémunération des médecins afin d'en tirer quelques enseignements pour la situation française.

PROPRIÉTÉS INCITATIVES DES MODES DE RÉMUNÉRATION : QUELLE APPLICATION EN SANTÉ ?

Les enseignements de la théorie économique en matière de rémunération

L'analyse économique des choix s'offrant à l'entreprise pour définir la rémunération permettant d'obtenir le volume et la qualité de travail requis dans le processus de production conduit à distinguer principalement deux types de rémunération (Elliott, 1991). Le premier est fixe et porte sur la ressource : le temps. Il s'applique lorsque, d'une part, chaque heure de travail fournie est également productive et, d'autre part, l'entreprise peut s'assurer que le nombre total d'heures nécessaires à la production sera effectué. Le deuxième type de rémunération est variable et met directement en relation niveau de rémunération et *output* (rémunération à la pièce). Il conduit à la définition d'un prix uniforme pour chaque unité, la rémunération étant directement liée au nombre total d'unités produites. Pour que ce type de rémunération puisse effectivement augmenter la productivité totale du travail, deux conditions doivent, là encore, être remplies : le niveau de l'*output* doit dépendre du travailleur, et de lui seul ; surtout, il doit être mesurable.

Dans la réalité, aucune de ces conditions n'est vérifiée : la productivité marginale du travail est, en général, décroissante et les coûts de surveillance sont toujours très élevés, de telle sorte que la simple rémunération du nombre d'heures travaillées ne saurait garantir un *output* maximal. Les deux conditions suivantes sont, elles aussi, rarement remplies. Le niveau de production dépend souvent d'autres facteurs que du seul effort et de la compétence du travailleur : des aléas peuvent, en effet, affecter le niveau de production, comme les conditions climatiques, indépendamment du comportement du travailleur ; le travail en équipe est une autre entorse à cette condition. Quant au caractère mesurable de l'output, il suffit ici d'évoquer la notion de qualité, dont on sait bien qu'elle est imparfaitement identifiable à l'issue du processus de production.

Les travaux microéconomiques récents ont tenté de prendre en considération de manière plus explicite ces conditions réelles de fonctionnement des marchés en s'appuyant sur la théorie de la régulation, encore appelée théorie des contrats[1]. Elle a été développée pour analyser les relations entre mandant et mandataire, en situation d'incertitude et d'asymétrie d'information. Elle s'applique à toute situation dans laquelle le principal (le mandant) délègue une action à l'agent (le mandataire), dans le cadre d'un contrat (implicite ou explicite), sachant qu'une incertitude prévaut sur le résultat, indépendamment du comportement de l'agent, et que ce dernier dispose seul d'une information qu'il utilise à son avantage. Ce type d'analyse[2] permet tout d'abord d'identifier les intérêts divergents des deux parties, le principal maximisant en général son profit (bien-être, revenu) alors que l'agent cherche, pour sa part, la production à coût minimal (au niveau de son propre effort et/ou de l'utilisation d'autres *inputs*).

De manière plus importante, cette théorie permet de résoudre ce conflit d'intérêts en précisant les conditions sous lesquelles des modes de rémunération seront « optimaux » au sens où ils permettront de réconcilier au mieux les intérêts divergents des deux parties. Dans le cas, rare, où l'observation du résultat permet de déduire le comportement (la performance) de l'agent, ou encore dans celui, plus rare encore, où les objectifs du principal et de l'agent ne sont pas en conflit (altruisme de l'agent, par exemple), un optimum dit de « premier rang » prévaut, dans lequel l'agent obtiendra la rémunération nécessaire pour effectuer l'action demandée et le principal un niveau de profit maximum. Dans tous les autres cas, seul un optimum de « second rang » est obtenu. Il implique un transfert d'une partie du profit du principal vers l'agent, sous forme d'incitations, afin de s'assurer, non seulement de sa participation au projet mais encore de sa performance maximale. Le coût, pour le principal, de la mise en

œuvre de telles incitations est le prix à payer pour la rente informationnelle dont dispose l'agent.

En situation d'asymétrie d'information, la solution de ce programme double de maximisation (où le principal doit décider, au vu du transfert à effectuer au bénéfice de l'agent, s'il délègue ou pas l'action, et où l'agent doit choisir d'accepter ou non la rémunération proposée), est un contrat de rémunération dit « mixte ». Il implique qu'une part de la rémunération soit indépendante du résultat, pour s'assurer de la participation de l'agent, l'autre partie étant, au contraire, indexée sur l'*output* pour l'inciter à la performance. La rémunération de l'activité prend alors un caractère mixte, avec une partie fixe et une partie variable, l'importance respective de chaque composante variant selon les secteurs. À titre d'illustration, les rémunérations fixes offertes aux commerciaux relèvent de la première logique et visent à s'assurer de leur participation[3], la partie variable, déclinée sous forme d'intéressement aux résultats, relevant de la seconde.

Ces résultats, clairement établis en théorie économique, sont-ils transposables au secteur de la santé ? Pour répondre à cette légitime interrogation, il convient au préalable d'en cerner les particularités.

Sur la spécificité des modes de rémunération des médecins

L'argument selon lequel le secteur de la santé présente des caractéristiques spécifiques telles qu'elles justifient de le soustraire au raisonnement économique n'est aujourd'hui plus recevable. Il est, à présent, reconnu que ce secteur présente, à l'instar d'autres secteurs comme l'éducation, des défaillances de marché (biens publics, externalités, valeur d'option), aucune ne lui étant propre. Mais sa particularité réside dans le cumul de plusieurs défaillances, ce qui rend toute tentative d'application des principes économiques délicate[4]. L'analyse des modes de rémunération des médecins en offre une bonne illustration.

Tout d'abord, la fonction de production de santé comprend comme facteurs, non seulement les soins de santé, mais aussi les conditions socio-économiques, les styles de vie, voire la réaction du patient aux soins. Cette complexité rend malaisée la mesure de la productivité marginale de chaque facteur de production, notamment des soins de santé, sachant de surcroît que les actes des médecins n'en représentent qu'une partie. Par ailleurs, la médecine n'étant pas une science exacte, l'efficacité médicale de certains traitements reste, à ce jour, non prouvée et elle est, par ailleurs, évolutive.

Ces deux premières caractéristiques ont conduit à retenir, de manière dominante, une rémunération hybride en ce sens qu'elle est bien indexée sur la performance individuelle (paiement à la pièce), mais basée sur un résultat intermédiaire qu'est l'acte médical (consultation, visite)

et non final (amélioration de l'état de santé). Dès lors, le lien entre l'acte médical sur lequel est basée la rémunération et le résultat final n'est plus explicite, ce qui permet de s'assurer que le risque de non guérison, lié à cet ensemble d'autres facteurs que le comportement du médecin, ne soit pas uniquement porté par ce dernier[5]. L'une des difficultés majeures rencontrée dans la définition d'une rémunération adéquate tient à l'hétérogénéité du contenu de cet acte médical. Il peut, en effet, soit correspondre à l'établissement d'un diagnostic (avec une éventuelle prescription d'actes d'autres professionnels et de produits pharmaceutiques), soit comprendre à la fois un diagnostic et un traitement posés, cette fois, par le médecin lui-même.

Par-delà le caractère multifactoriel de la fonction de production de santé et l'incertitude médicale, certaines particularités des relations qui se nouent entre agents dans ce secteur rendent délicate la définition d'un mode de rémunération optimal pour les médecins[6] :

- il s'agit, en particulier, de la relation médecin - patient dans laquelle une double action est demandée au médecin en tant qu'agent du patient : d'une part, un diagnostic et, d'autre part, un traitement, sachant que le niveau de sa propre rémunération, établie sur un volume d'actes, dépendra de manière mécanique d'un diagnostic difficilement vérifiable. Se pose ici la problématique chère à Jules Romain, à savoir qu'un bien portant est un malade qui s'ignore[7], idée que R. Evans (1974) transcrira en termes économiques sous le vocable de « demande induite » ;
- l'autre dimension importante concerne la relation médecin-tiers payeur. Ce dernier est, en effet, amené à déléguer au médecin la responsabilité de la prescription d'autres biens et services dont certains peuvent se substituer à son propre temps de travail.

À l'évidence, la réalité est complexe dans ce secteur car les stratégies que les acteurs peuvent adopter, dès lors que l'information n'est plus partagée par tous, sont multiples. Il convient d'en tenir compte autant que faire se peut dans l'application, à la santé, des résultats sur la mixité tirés de l'analyse économique des modes de rémunération. De nombreux travaux ont déjà adapté ces résultats à la rémunération de l'hôpital[8], démontrant ainsi qu'il est possible de concevoir des systèmes mixtes de rémunération et de les mettre en œuvre. De manière surprenante, le secteur de l'ambulatoire a suscité plus tardivement des travaux comparables et a connu de réelles difficultés d'application.

Propriétés des systèmes « purs » de rémunération

Les propriétés des systèmes « purs » de rémunération des médecins ont fait l'objet de nombreuses publications et sont dorénavant présentées dans des ouvrages à visée pédagogique[9]. Elles ne seront donc

évoquées ici que de manière synthétique. Trois systèmes « purs » sont traditionnellement identifiés, le paiement à l'acte, le salaire et la capitation, ce troisième mode impliquant la rémunération forfaitaire, en général pour l'année à venir, de la responsabilité de la santé de la population inscrite sur la liste du médecin[10]. La distinction entre ces trois systèmes de rémunération apparaît de manière plus claire lorsqu'on prend en compte la dimension temporelle. En effet, le paiement à l'acte correspond au remboursement *ex post* d'un coût (le temps passé par le médecin, à compétence donnée) alors que dans les deux autres, il s'agit du paiement prospectif, soit d'un nombre d'heures, soit de la responsabilité du médecin, pour un niveau de demande estimé (en nombre de cas moyen). Si aucun risque n'est présent pour le médecin dans le paiement à l'acte (on peut alors rapprocher cette notion de celle développée en économie industrielle sous le terme de paiements de type *cost plus*), une part importante de risque subsiste dans les deux autres cas (qui s'apparentent plus, en conséquence à un système de type *fixed price*).

Le paiement à l'acte

Il présente un certain nombre d'avantages, souvent définis par les médecins eux-mêmes comme étant une plus grande liberté et l'assurance d'une continuité des soins. Il se traduit aussi concrètement par des niveaux de rémunération globalement plus élevés. L'un des inconvénients du paiement à l'acte est, cependant, de conduire à une multitude de prix, lorsque ceux-ci sont libres, du fait de la grande hétérogénéité entre actes médicaux. Il entraîne par ailleurs une éventuelle surconsommation de services de médecin et/ou d'autres professionnels, identifiée précédemment sous le terme de demande induite. Cette tendance est renforcée par le fait que, côté demande, on ne peut guère s'appuyer sur les préférences des usagers pour établir la valeur des biens et services médicaux. D'une part, ainsi que le notait K. Arrow dès 1963, l'incertitude et l'asymétrie d'information présentes sur le marché des soins médicaux remettent en cause les capacités d'arbitrage des patients ; d'autre part, la solvabilisation de la demande au nom de la solidarité réduit artificiellement le coût du recours aux soins.

Les analyses économiques qui ont tenté de mettre en évidence cette demande induite indiquent, de fait, l'existence d'une certaine marge de manœuvre sur les volumes d'actes prescrits. Les médecins peuvent être amenés à l'utiliser de façon à se prémunir de manière transitoire contre le risque d'une réduction conséquente de leur pouvoir d'achat, soit du fait d'une concurrence plus vive, soit en réaction à des mesures de gel des tarifs médicaux. Ce pouvoir discrétionnaire touche, certes, au volume d'actes prescrits ; il peut aussi se manifester par un biais dans la pratique médicale, conduisant à un recours systématique aux actes les plus

lucratifs de la nomenclature. Cette dernière est, en effet, le résultat des négociations engagées entre les offreurs de soins et le tiers-payeur et ne reflète pas toujours fidèlement la valeur intrinsèque de l'acte posé[11]. Quelle qu'en soit la manifestation, cette induction de la demande par l'offre apparaît aujourd'hui d'envergure plus restreinte qu'initialement estimée[12].

Les systèmes prospectifs

Ils annulent la liaison positive précédente entre nombre d'actes et rémunération et présentent ainsi l'avantage d'une plus grande neutralité à l'égard du jugement médical. La capitation permet de réduire le risque moral[13] dans ses deux dimensions (*ex ante* et *ex post*). En effet, une telle règle conduit le médecin à développer les activités de prévention pour la population dont il a la charge afin de diminuer la probabilité d'occurrence de la maladie (identifié sous le terme de risque moral *ex ante*) et à minimiser le nombre d'actes par cas traité (risque moral *ex post*). Ces incitations sont liées au fait que le médecin est ici créancier résiduel des économies réalisées (en termes financiers et de temps disponible). Le salariat, s'il n'incite guère à la performance, permet, quant à lui, de rémunérer certaines activités professionnelles comme l'enseignement, la recherche, la prévention ou l'administration, activités qui ne peuvent être aisément divisées en services individualisés.

Pour autant, ces modes prospectifs de rémunération ne sont pas exempts d'inconvénients. Se pose, en effet, nécessairement la question de la motivation du médecin, qui sera d'autant moins forte que la répartition de la population entre les différents offreurs de soins s'effectuera de manière administrative (sur une base géographique, par exemple), comme ce fut le cas pour le généraliste dans la Grande-Bretagne d'avant les réformes des années 1990. De plus, dans la mesure où la relation entre soins et état de santé s'inscrit dans le long terme et qu'elle n'est pas univoque, le médecin rémunéré à la capitation risque de réduire la qualité des soins prodigués, ne serait-ce que pour le temps qu'il est disposé à offrir à chaque patient. Les médecins ont, de surcroît, intérêt à pratiquer une forme particulière de sélection des risques en référant de manière abusive vers l'hôpital des cas jugés trop lourds.

Ont ainsi été mis en évidence les deux biais fondamentaux qui caractérisent les principaux modes de rémunération appliqués à la santé, à savoir le risque de surproduction (ou encore l'induction de la demande) dans le cas de paiements variables comme le paiement à l'acte et le manque de motivation (conduisant à une forme de « déflation » de la demande) dans celui des paiements forfaitaires comme la capitation ou le salariat. Chacun des trois systèmes « purs » de rémunération présente donc une série d'inconvénients, mais aussi d'avantages spécifiques et la

combinaison optimale entre ces trois modes de rémunération apparaît d'autant plus légitime qu'elle est sous-tendue par les travaux théoriques présentés précédemment. La recherche d'une telle mixité est manifeste dans les réformes récentes des modes de rémunération des médecins dans les pays industrialisés. Le « dosage » varie cependant selon les pays, en ce qu'il reflète fidèlement les arbitrages politiques et socioculturels de chaque nation.

RÉFORMES DES MODES DE RÉMUNÉRATION : LA RECHERCHE DU « MÉLANGE » IDÉAL

Dans la plupart des pays de l'OCDE[14], y compris la France, une forme de mixité est déjà répandue dans la mesure où une juxtaposition de modes de rémunération prévaut selon le secteur : le salariat domine largement en milieu hospitalier alors qu'en médecine de ville, les deux termes de l'alternative sont le paiement à l'acte ou la capitation. Dès lors qu'un médecin combine, comme c'est souvent le cas, une activité à l'hôpital et en ville[15], sa rémunération s'inscrit déjà dans un régime mixte. Mais la question posée à l'égard des réformes en cours est de savoir s'il est possible, au-delà de cette juxtaposition, d'aller vers une mixité des rémunérations pour un acte et un producteur donnés.

La comparaison des réformes des modes de rémunération entreprises dans divers pays industrialisés a clairement indiqué une convergence sur la recherche de modes de paiement mixtes (Rochaix, 1998b). Cette analyse avait préalablement permis d'identifier les trois principaux objectifs poursuivis par ces réformes : la recherche d'une plus grande productivité, d'une plus grande attention aux préférences des patients et la maîtrise des dépenses de santé. Les deux premiers objectifs se rattachent à la notion d'efficacité microéconomique et recouvrent deux dimensions : l'efficacité productive, qui est définie comme la minimisation des *inputs* pour une production et une qualité donnée (ou de manière symétrique, la maximisation de l'*output*, pour un niveau d'*inputs* donné) ; l'efficacité allocative, qui conduit à s'interroger sur la capacité du système à prendre en compte les préférences des usagers. Le troisième objectif (la maîtrise des dépenses de santé) est, en général, associé à une notion plus large d'efficacité macroéconomique. Techniquement, l'efficacité macroéconomique, pour être assurée, nécessiterait que la règle d'égalisation des utilités marginales entre grandes fonctions collectives de l'État soit respectée. Selon cette règle, chaque euro supplémentaire affecté à chacune de ces fonctions devrait générer le même niveau de bien-être collectif. Malgré les tentatives récentes de mesure de cette efficacité macroéconomique, il est, aujourd'hui encore, bien difficile de comparer la valeur ajoutée, au niveau de l'ensemble de

l'économie, d'un financement supplémentaire qui serait accordé à la santé plutôt qu'à d'autres grandes fonctions collectives de l'État (défense, éducation). On approche donc cette notion d'efficacité macroéconomique par une règle plus simple selon laquelle il serait souhaitable de maintenir constant le pourcentage des dépenses de santé dans le PIB et qui revient à postuler que la part du secteur de la santé dans l'économie est considérée comme satisfaisante et qu'elle doit donc être maintenue telle quelle.

Si les objectifs sont partagés, les voies empruntées diffèrent sensiblement, du fait de situations initiales très hétérogènes :
- les pays dans lesquels les dépenses de santé sont définies *ex ante* (systèmes fiscalisés) et qui recourent principalement au salaire ou à la capitation rencontrent des difficultés relevant essentiellement de la notion d'efficacité allocative. Ils tentent de les réduire par l'introduction de paiements à la performance pour une fraction de l'activité que l'on souhaite encourager ;
- à l'inverse, les pays dans lesquels le paiement à l'acte est le mode de rémunération principal tentent d'atteindre l'objectif d'efficacité macroéconomique en priorité. À cette fin, ils mettent en place des enveloppes globales au niveau collectif, leur non respect se soldant, en général, par des sanctions sur les taux de croissance des honoraires à la période suivante.

Le recours à des instruments différents s'explique donc essentiellement par la nécessité de satisfaire à la fois aux objectifs micro- et macroéconomiques, en partant de situations initiales très différentes.

La recherche de l'efficacité microéconomique

Deux réformes des systèmes prospectifs de rémunération peuvent être envisagées pour améliorer l'efficacité allocative : d'une part, la modulation de la base de calcul de la rémunération ; d'autre part, l'introduction de rémunérations partielles à la performance. Dans le premier cas, un système de tarifs modulés peut être envisagé, à l'instar d'une proposition faite au Québec en 1980[16] et qui revient à faire varier la rémunération forfaitaire en fonction d'un certain nombre de paramètres comme la compétence du médecin, les caractéristiques du patient (âge, morbidité) ou la région de pratique. Dans le deuxième cas, plus courant, il s'agit d'introduire des incitatifs financiers dans les systèmes de rémunération en place[17]. À titre d'illustration, on peut ici évoquer la Grande-Bretagne où plusieurs modes de rémunération sont combinés pour les généralistes. Leur rémunération comprend, outre la capitation et le versement d'indemnités forfaitaires couvrant les frais fixes, une activité rémunérée à l'acte (environ 18 % du total de la rémunération), afin d'encourager, notamment, la prise en charge de certains actes de prévention comme

les vaccinations. De nouveaux incitatifs ont aussi été développés en conférant aux médecins généralistes la responsabilité de budgets pour les soins de second recours des patients inscrits sur leur liste. Les réformes récentes mettent en place une contractualisation de certains programmes, avec une rémunération conditionnée par l'atteinte des objectifs prédéfinis (campagnes de vaccination).

Il est, cependant, difficile de prédire avec précision l'impact de la mise en place de tels incitatifs financiers. L'analyse économique permet, tout au plus, d'estimer la taille des effets de revenu et effets de substitution pour prédire quel sera l'impact global en volume d'heures travaillées d'une augmentation du taux de salaire. Quand il s'agit du paiement à l'acte des médecins, intervient, par-delà le choix du nombre d'heures travaillées, la possibilité pour le médecin de modifier de manière stratégique la combinaison d'actes effectués et/ou de substituer certains actes à son propre temps de travail. De plus, certaines études ont mis en évidence l'absence d'impact de certaines incitations financières (Hughes and Yule, 1992). Ceci tient, entre autres, au caractère très fragmenté du processus de décision en santé, le tarif n'étant, par ailleurs, qu'une des variables explicatives dans le choix d'utilisation d'un service par un médecin.

La recherche de l'efficacité macroéconomique

La recherche de l'efficacité macroéconomique (au sens d'une maîtrise des dépenses de santé) est la préoccupation majeure de systèmes dans lesquels le paiement à l'acte domine. Elle passe inévitablement par l'indexation des taux de croissance des tarifs pour l'année à venir sur la performance, définie comme l'écart entre le taux de croissance négocié *ex ante* et la réalisation. Parmi les mesures de régulation de ce type, il convient, selon Glaser (1993), de distinguer entre enveloppes globales fermées (*expenditure caps*) et ouvertes (*expenditure targets*). Dans le premier cas, une somme fixe est allouée aux soins que l'on cherche à encadrer alors que dans le deuxième, des objectifs sont définis pour l'année en cours et les années à venir. À la régulation stricte imposée par les premières, les enveloppes ouvertes s'apparentent à une régulation plus souple nécessitant la collaboration des offreurs et des payeurs, le rôle du gouvernement se limitant à mettre à disposition une information sur les réalisations.

L'Allemagne fait figure de précurseur en matière d'enveloppes. Les médecins y sont rémunérés à l'acte suivant un barème négocié et sont soumis à une contrainte budgétaire régionale fermée. Les tarifs sont définis en points, la valeur de ce dernier étant flottante et fonction des réalisations au niveau régional[18].

Aux États-Unis, la réforme de Medicare a conduit à l'adoption de

telles enveloppes fermées : en 1989, le Congrès a voté une loi sur le mode de remboursement des médecins. A été mis en place un barème élaboré sur les ressources utilisées (*Resource Based Relative Value Scale* - RBRVS) ainsi qu'une limite aux dépassements. Des objectifs de croissance en volume des dépenses de santé pour les actes posés par les médecins sont définis (*Medicare Volume Performance Standards* - MPVS) et les tarifs sont revus à la baisse lorsque le taux de croissance des dépenses globales dépasse l'objectif (et inversement). Pour ce qui est des soins hors programmes fédéraux, Robinson (1993) a montré l'existence de nombreux mécanismes mis en place pour garantir l'efficacité macro et microéconomique. Ainsi, certaines HMO (*Health Maintenance Organisations*) rémunèrent les médecins à l'acte, mais retiennent un pourcentage (10 à 20 %) de la rémunération jusqu'à la fin de l'année. Ces retenues sont ensuite reversées aux médecins si l'objectif de maîtrise des dépenses a été atteint. Dans d'autres réseaux, une capitation est définie sur un groupe de médecins plutôt qu'individuellement. Chaque médecin du pool est alors payé à l'acte dans la limite des fonds affectés aux services médicaux sur la base de cette capitation. Certaines HMO tentent ainsi d'enrayer la tendance trop fréquente des généralistes à référer les patients qu'ils pourraient traiter eux-mêmes vers les spécialistes en hôpital. À cette fin, est mis en place un système de bonus/malus en fonction du nombre de patients référés.

Dès le milieu des années 1970, le Québec s'est doté d'une régulation macroéconomique par la mise en place d'enveloppes globales pour l'ambulatoire. L'expérience est originale en ce qui concerne les omnipraticiens puisque la Province a retenu un système combinant enveloppe macroéconomique et régulation individuelle. Ainsi, un objectif de croissance des dépenses de soins médicaux est négocié et le respect de la contrainte macroéconomique est assuré par une révision à la baisse de la croissance prévue des tarifs l'année suivante en cas de dépassement. Une contrainte individuelle est emboîtée à cette contrainte macroéconomique par la mise en place d'un plafonnement individuel trimestriel du revenu. Une fois dépassé le plafond, la rémunération de chaque acte est minorée de 75 % jusqu'au début du trimestre suivant. En limitant l'activité très forte de certains omnipraticiens[19], cette contrainte individuelle permet de s'assurer du respect de la contrainte macroéconomique et fait en sorte que le comportement atypique d'un petit nombre de médecins ne conduise pas à une sanction collective. La mise en place conjointe de ces deux types de contraintes a, par la suite, été étendue aux spécialistes.

La comparaison des équilibres retenus par les pays de l'OCDE entre ces divers modes de rémunération est riche en enseignements :
- il apparaît tout d'abord qu'aucun instrument ne permet, à lui seul, d'atteindre les deux objectifs d'efficacité micro et macroéconomique ;

- il existe, par ailleurs, une asymétrie intéressante : il est, en effet, plus difficile de négocier le passage à un mode de paiement mixte à partir d'un système de rémunération majoritairement à l'acte que l'inverse (introduire dans des paiements prospectifs une part limitée de paiement à la performance). Pour les systèmes à dominante paiement à l'acte, la seule solution permettant d'atteindre l'efficacité macroéconomique est alors la mise en place d'enveloppes globales ;
- le choix d'un système de rémunération entraîne avec lui d'autres choix touchant à l'organisation du système de santé. Ainsi en est-il de la capitation qui s'accompagne traditionnellement d'une hiérarchie entre soins de premier et de second recours (soins de spécialistes et soins hospitaliers), le généraliste jouant alors le rôle d'un aiguilleur (*gate-keeper*) dans le système de soins. Dans ces systèmes organisés sur une filière de soins, l'utilisation de paiements directs des usagers est très limitée comparé aux systèmes reposant sur le paiement à l'acte.

Cette comparaison permet donc de rappeler l'interdépendance des différentes composantes du système (modes de rémunération des producteurs, de remboursement et d'accès aux soins des usagers), ainsi que le rôle important de l'histoire dans la détermination des degrés de liberté dont dispose le régulateur en matière de réforme des systèmes de santé.

LES VOIES POSSIBLES DU CHANGEMENT EN FRANCE

Par rapport à d'autres systèmes de paiement à l'acte comme le système allemand ou québécois, l'architecture du système français se caractérise par :
- une liberté d'honoraires pour les médecins de secteur II[20] ;
- une opacité du système d'information sur l'activité des médecins, due essentiellement à une nomenclature trop agrégée, et de surcroît, obsolète ;
- l'absence d'un payeur unique (le patient faisant l'avance de frais et étant ensuite remboursé par les différentes caisses) ;
- une prise en charge intégrale du ticket modérateur, voire même des dépassements dans le cas d'assurances complémentaires très couvrantes ;
- un accès direct aux soins de second recours (par les consultations de spécialistes, les consultations externes et l'urgence en hôpital) auquel tant les patients que les offreurs de soins eux-mêmes semblent être fortement attachés.

Cet ensemble de caractéristiques confère au système un caractère inflationniste marqué. La collusion tacite entre médecin et patient (suscitée par les modes de rémunération et de remboursement existants) conduit à une surutilisation des services de soins. Par ailleurs, la concurrence entre généralistes et spécialistes, favorisée par le libre accès du patient, se solde par un rétrécissement du marché des généralistes et

donc à une intensification de la concurrence entre eux. Des comportements stratégiques se développent, qui visent à fidéliser des patients, comme la multiplication des prescriptions pharmaceutiques et des arrêts de travail, comportements qui ne vont guère dans le sens d'une médecine de qualité. Ainsi, la mise en concurrence par les patients de soins *a priori* de nature complémentaire (et, à ce titre, qualifiés dans de nombreux pays comme respectivement de « premier » et de « second » recours) apparaît plus génératrice de surcoûts que d'efficacité, au vu des spécificités du marché des soins évoquées précédemment. En outre, le coût de cet accès direct est la nécessité de conserver une responsabilisation des usagers en aval, dont les effets régressifs ont été largement documentés[21].

De fait, les réformes structurelles amorcées en France depuis 1990 ont tenté de résorber certains de ces dysfonctionnements. Elles ont impulsé le passage d'un système à guichet ouvert à une logique prospective, en plusieurs étapes : le rapport Santé 2010 (CGP, 1993) repris dans le *Livre blanc sur le système de santé* (1995), puis sa concrétisation dans le cadre du Plan Juppé de 1995 avec la mise en place de l'ONDAM[22]. Cette régulation de type macroéconomique a connu diverses déclinaisons selon le secteur concerné, avec un succès très variable[23].

Pour les cliniques et les laboratoires d'analyse, ont été mises en place dès 1991 des enveloppes dont le respect devait être garanti à la fois par des modulations de tarifs et surtout par des mécanismes de reversements individualisés en cas de dépassement. Pour les médecins, le mécanisme prévu par le Plan Juppé conduisait à conditionner les revalorisations tarifaires de l'année à venir sur le respect de l'objectif ainsi qu'un reversement d'honoraires en cas de dépassement de l'enveloppe. Mais ce mécanisme, pourtant moins dur que celui qui prévaut en Allemagne (où le réajustement concerne l'année en cours et non l'année à venir) n'a pas été appliqué pour les spécialistes lors du dépassement de l'enveloppe en 1997. Les généralistes ont, par contre, pu bénéficier cette année-là d'une augmentation de 9 300 F en moyenne par médecin. L'année suivante, un tunnel de neutralité a été défini, avec une sanction individuelle et des exemptions pour des médecins nouvellement installés. Ce mécanisme complexe ayant fait l'objet d'une opposition très forte de la part de la profession médicale, puis ayant été invalidé par le Conseil d'État en 1998, aucun mécanisme individualisé n'a pu être mis en place. Or une enveloppe fermée sans déclinaison individuelle s'avère injuste pour la majorité des professionnels et ne peut être durablement imposée. En conséquence, l'ONDAM est devenu *de facto* une enveloppe ouverte, régulièrement dépassée.

La Mission de concertation pour la rénovation des soins de ville (Bruhnes, 2001) a, cependant, réaffirmé l'intérêt d'un mécanisme du

type ONDAM et a renoué avec l'introduction de paiements mixtes en recommandant l'usage plus fréquent de paiements forfaitaires dans différentes situations (réseaux de soins, par exemple). Elle a aussi recommandé une véritable délégation à l'assurance maladie qui permettrait à celle-ci de contractualiser de manière individuelle avec les médecins. Mais les revalorisations tarifaires des médecins relèvent, aujourd'hui encore, de décisions ministérielles extérieures à toute contractualisation que les caisses d'assurance maladie (théoriquement en charge de la négociation avec les professionnels de santé) pourraient tenter de mettre en place[24]. Par ailleurs, la mise en place de telles enveloppes, qu'elles soient ouvertes ou fermées, ne permet en rien de résoudre le manque de responsabilisation des usagers. La réduction du risque moral côté usagers implique, alors, dans le cas du système de paiement à l'acte, une participation financière directe des usagers, participation dont les effets « revenu » doivent cependant être tempérés. Clairement, pour être efficace, la maîtrise des dépenses de santé doit passer par une régulation conjointe de l'offre et de la demande.

Enfin, la mise en place de paiements mixtes, même si cette idée devait progresser en France, ne saurait, à elle seule, assurer l'optimalité de la production et de la distribution de soins. Des dispositifs de régulation non financière doivent être mis en place parallèlement, visant à recommander des profils de traitement (comme c'est le cas pour les références médicales opposables en France). À ce stade, l'expérimentation de modes de rémunération mixtes permettant de mesurer l'impact de tels incitatifs, non seulement en termes de quantité et type d'actes effectués, mais aussi et surtout en impact sur la santé des usagers, apparaît indispensable. Elle pourra être effectuée dans le cadre des réseaux qui ont, pour certains, déjà adopté des modes de rémunération forfaitaires.

Les degrés de liberté en matière de réforme des modes de rémunération des médecins apparaissent ainsi étroitement liés à l'histoire et aux équilibres préalables entre ces modes de rémunération. Atteindre les deux objectifs d'efficacité micro et macroéconomique par la mise en place de paiements mixtes en ambulatoire apparaît alors plus aisé dans les systèmes de santé où la rémunération des médecins est à dominante forfaitaire. Pour les pays qui, comme la France, recourent majoritairement au paiement à l'acte en médecine de ville, l'atteinte de ces deux objectifs apparaît plus délicate. La réforme impliquerait, dans ce cas, la mise en place d'enveloppes globales et de mécanismes individuels permettant d'identifier le volume, et surtout la qualité des pratiques. De plus, une réforme du mode de rémunération des médecins qui ne s'accompagne-

rait pas d'une tentative comparable de régulation d'une éventuelle surconsommation de la part des patients serait vouée à l'échec. Enfin, les modes de rémunération ne sauraient, à eux seuls, garantir une utilisation optimale des dépenses de santé. D'autres modalités non financières doivent être mises en place comme l'évaluation par les pairs, la formation et la définition de standards de pratique.

NOTES

1. Cf. Laffont et Tirole, 1993.

2. Nombreuses sont ces applications : entre employeur et employé, entre expert et client...

3. Un commercial auquel on ne proposerait aucun fixe subirait seul le risque d'un mauvais résultat, indépendant de sa volonté et préférerait se mettre à son compte plutôt que d'accepter un tel contrat.

4. Cf. L. Rochaix, 1996.

5. Le seul contre-exemple étant trouvé dans l'ancien temps, en Chine, où le médecin n'était rémunéré que s'il avait guéri son patient.

6. Cf. Rochaix, 1986,1998a.

7. J. Romain, *Knock*, Gallimard, 1924.

8. Cf. Ellis and McGuire, (1986) sur les fondements théoriques de cette application ; Newhouse (1996, 2003) et DREES (2002) pour une revue récente.

9. Cf. Rochaix (1986), Khelifa et Rochaix, (1993), pour de premières synthèses ; Scott and Hall (1995), B. Majnoni d'Intignano et Ph. Ulmann (2001) pour une version synthétique de ces principaux arguments (pp. 329-39) et, plus récemment, Chaix-Coutturier et *al.,* (2003), Gosden et *al.,* (2003).

10. Le médecin qui reçoit la capitation devient responsable de la dispensation d'une gamme plus ou moins étendue de services à une clientèle donnée pendant une certaine période de temps. Le montant du *per capita* peut être ajusté selon différents critères comme les caractéristiques des individus (âge, sexe), les risques épidémiologiques ou le milieu socio-économique, mais il est indépendant de la consommation effective du patient. Ce système est associé à un mécanisme de confinement qui contraint le patient à rester sur la liste d'un médecin pendant un certain temps. La capitation est surtout développée pour les généralistes dans les systèmes construits à l'image du système national de santé britannique (Italie, Espagne, Portugal, Grèce, Danemark).

11. Keeler and Brodie (1993) offrent un exemple de tels biais de pratique à travers l'analyse de ces distorsions de prix en gynécologie aux États-Unis. Ils montrent ainsi que l'important différentiel de tarif en faveur de la césarienne explique en bonne partie le recours plus marqué à ce type d'intervention comparé à d'autres pays. D'autres travaux (Rochaix, 1993) ont indiqué des comportements stratégiques similaires au Québec, dans le choix du niveau de complexité de la consultation à facturer à l'assurance maladie.

12. Les travaux des années 1970-1980 avaient estimé une élasticité unitaire (à savoir qu'une augmentation de 10 %, par exemple, de la densité médicale conduirait à une augmentation comparable des dépenses de santé. Les estimations actuelles montrent plutôt une réaction de l'ordre de 3 % à une telle hausse de la densité médicale ; Cf. Rochaix et Jacobzone, (1997).

13. Le risque moral est défini comme la modification de comportement liée à la présence d'un tiers, en général l'assureur, responsable en dernier ressort. En santé, on distingue le risque moral ex ante (modification de la probabilité d'occurrence de la maladie par moindre effort de prévention de la part du patient, du fait de la présence de l'assurance maladie) et risque moral *ex post* (modification de la taille du dommage, une fois le risque maladie avéré, par surconsommation de soins).

14. Cf. OECD, (1992).

15. À titre d'illustration, en 1997, seuls 47,7 % des omnipraticiens étaient rémunérés exclusivement à l'acte au Québec et plus de 50 % tiraient une partie de leur revenu d'un tarif horaire ou d'honoraires fixes. Ces formes de paiement permettent de rémunérer des activités particulières (Pereira, 2002).

16. Cf. rapport du Comité sur la rémunération des professionnels au Québec, (1980), et Conseil médical du Québec, (1995) et A. P. Contandriopoulos, (1990).

17. Grignon et *al.,* (2004) ont récemment présenté une revue détaillée de tentatives récentes d'introduction de tels mécanismes incitatifs.

18. Cf. Rochaix et *al.,* (2000) pour une présentation détaillée du système de ce point flottant.

19. Cf. L. Rochaix (1993) pour une validation empirique.

20. Créé en 1980, le secteur II comprend les médecins qui ont opté pour la liberté tarifaire, en contrepartie de laquelle ils ont perdu un certain nombre d'avantages toujours consentis aux médecins de secteur I (comme la prise en charge de cotisations d'assurance maladie par la Caisse d'assurance maladie). L'accès en secteur II est aujourd'hui gelé, à quelques exceptions près (anciens chefs de clinique). Les patients de médecins de secteur II payent intégralement le dépassement (la différence entre le tarif fixé par le médecin et le tarif conventionnel), certaines assurances complémentaires assurant, en plus de la prise en charge du ticket modérateur, tout ou partie de ce dépassement.

21. Cf. A. Khelifa et L. Rochaix, CGP, (1993).

22. Objectif national des dépenses d'assurance maladie.

23. Cf. Hartmann L et *al.* (2000).

24. Le dernier exemple en date étant la revalorisation des chirurgiens décidée en août 2004 par l'actuel ministre de la Santé.

BIBLIOGRAPHIE

ARROW J.K., (1963), « Uncertainty and the Welfare Economics of Medical Care », *American Economic Review,* (Dec), 53(6), pp. 941-73.

BRUHNES B, GLORION B., PAUL S. and ROCHAIX L., (2001), « Mission de concertation pour la rénovation des soins de ville », Rapport, juillet, ministère des Affaires sociales, du Travail et de la Solidarité.

CHAIX-COUTURIER C., DURAND-ZALESKI I., JOLLY D. et DURIEUX P. (2000), « Effects of Financial Incentives in Medical Practice: Results from a Systematic Review of the Literature and Methodological Issues », *International Journal for Quality in Health Care,* vol 12, n° 2, pp. 133-142.

COMMISSARIAT GÉNÉRAL DU PLAN, (1993), *Santé 2010.* Rapport du groupe de prospective du système de santé, la Documentation française.

CONSEIL MÉDICAL DU QUÉBEC, (1995), « Avis sur une nouvelle dynamique organisationnelle à implanter : la hiérarchisation des services médicaux », Avis 95-03.

COMITÉ SUR LA RÉMUNÉRATION DES PROFESSIONNELS DE LA SANTÉ AU QUÉBEC, (1980), « Le système des honoraires modulés », rapport.

CONTANDRIOPOULOS, A. P., CHAMPAGNE F., PINEAULT, R. (1990), « Systèmes de soins et modalités de rémunération », *Sociologie du travail,* n° 1/90, pp. 95-115.

DREES, (2002), « La tarification à la pathologie : leçons de l'expérience étrangère », Actes du colloque de Paris, 7 et 8 juin 2001, Dossiers solidarité et santé, Hors Série, *La Documentation française,* juillet.

ELLIOTT R. F., (1991), « Labor Economics: a Comparative Analysis », McGraw Hill, pp. 89-92.

ELLIS R. P. and T. MCGUIRE, (1986), « Provider Behaviour under Prospective Reimbursement », *Journal of Health Economics,* 5(2), 129-151.

GLASER W., (1993), « How Expenditure Caps and Expenditure Targets Really Work », MMFQ, Vol. 71, n° 1, pp. 97 - 128.

GOSDEN T., FORLAND F., KRISTIANSEN I.S., SUTTON M., LEESE B., GIUFFRIDA A., SERGISON M., PEDERSEN L., (2003), « Capitation, salary, fee-for-service and mixed payment: effects on the behaviour of primary care physicians », *Cochrane Review,* the Cochrane Library, Issue 3.

GRIGNON M., PARIS V. et POLTON D. (2002), « L'influence des modes de rémunération des médecins sur l'efficience du système de soins », rapport pour la commission Romanow, Rapport du CREDES n° 35.

HARTMANN L., ROCHAIX L. et J. de KERVASDOUÉ, (2000), « La régulation économique des systèmes de santé », in *Le carnet de santé de la France en 2000,* sous la direction de J. de Kervasdoué, ed. La Mutualité Française, pp. 85-122.

HUGHES D., YULE B., (1992), « The Effect of per-item Fees on the Behaviour of Primary Health Care », *Journal of Health Economics,* 11, 4, pp.413-438.

KEELER E.B., BRODIE M., (1993), « Economic incentives in the Choice between Vaginal Delivery and Caesarean Section », *MMFQ,* Vol. LXXI, n° 3, pp. 365-404.

KHELIFA A. et ROCHAIX L. (1993), Atelier 4 : « Rémunération des producteurs et incitations financières des usagers », Rapport *Santé 2010 : équité et efficacité du système,* Commissariat général du Plan, La documentation française, juin.

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

MAJNONI D'INTIGNANO B. et Ph. ULMANN, (2001), *Économie de la Santé,* Thémis. mai.

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

NEWHOUSE J. P., (2003), « Reimbursing for Health Care Services », *Économie publique,* n° 13, Vol. 2, p. 5-33.

OECD, (1992), « The reform of health care: a comparative analysis of seven OECD countries », *Health policy studies* n° 2.

PAULY M.V., (1980), *Doctors and their Workshops* Chicago Univ. Press.

PEREIRA C. (2002), « La régulation économique de la médecine de ville », Thèse pour le doctorat en sciences économiques, Université de Paris II.

REINHARDT U. et SANDIER S. (1983) « Alternative Methods of Physician Remuneration and their Effects on Physician Activity: an International Comparison », CREDOC.

ROBINSON J.C., (1993), « Payment Mechanisms, Nonprice Incentives, and Organisational Innovation in Health Care », *Inquiry,* (Fall), 30(3), pp. 328-33.

ROCHAIX L., (1993) « Financial Incentives for Physicians: the Quebec experience », *Health Economics,* Vol. II, pp. 163-176.

ROCHAIX L., (1996), « L'analyse du marché des soins médicaux : quelle place pour l'économie de la santé ? » Numéro spécial du 20ème anniversaire, *Revue d'épidémiologie et de santé publique,* nov.

ROCHAIX L., JACOBZONE S., (1997), « L'hypothèse de demande induite : un bilan économique », *Économie et prévision,* n° 129/130, octobre.

ROCHAIX L., (1998a) : «The Physician as Perfect Agent: a Comment », *Social Science and Medicine,* vol. XLVII, n° 3, pp. 355-356.

ROCHAIX L., (1998b), « Performance-tied Payment Systems for Physicians: Evidence form Selected Countries », chapitre d'un ouvrage collectif, *Critical Challenges for Health Care Reform in Europe*, édité par l'OMS Europe chez Open University press.

ROCHAIX L., HARTMANN L. et PEREIRA C., (2000), « Modes alternatifs de rémunération pour la médecine ambulatoire », Rapport de recherche pour la Direction de la Sécurité sociale, ministère des Affaires sociales.

ROEMER M.I. (1962) « On paying the doctor and the implications of different methods », *Journal of Health and Human Behavior*, vol. 3, n° 1, pp. 4-14.

ROMAIN J., (1924), *Knock,* Gallimard.

SCOTT S., HALL J., (1995), « Evaluating the Effects of GP Remuneration : Problems and Prospects », H*ealth Policy,* 31, pp. 183-195.