CENTRE INTERUNIVERSITAIRE DE RECHERCHE
PLURIDISCIPLINAIRE (CIREP)
STATUT : UNIVERSITE PUBLIQUE
Web : www.cirep.ac.cd
Email : info@cirep.ac.cd

Note du cours de l'Economie de la santé et financement de la santé

OBJECTIFS DU COURS

- ✓ Comprendre les principes généraux en économie de santé
- ✓ Comprendre le fonctionnement du modèle économique en santé
- ✓ Savoir calculer des coûts économiques en santé
- ✓ Faire une analyse économique des coûts en santé
- ✓ Comprendre le mode de financement en santé et son évaluation

INTRODUCTION

Depuis les travaux de Arrow sur « l'économie des soins médicaux » en 1963, l'économie de la santé a connu un développement considérable. Trois grands axes de recherche sont présentés dans cet article : la nature de la demande, l'analyse de l'assurance maladie et les politiques de régulation.

Le premier axe distingue notamment les « besoins » de santé qui traduisent une détérioration de l'état de santé, la transposition de ces besoins en demande de soins qui requiert l'intermédiation — délicate à évaluer — des offreurs de soins, et le lien entre l'état de santé et la consommation de soins — difficile à quantifier en général mais qui peut être appréhendé de manière plus ciblée grâce au développement de l'évaluation médico-économique.

L'assurance maladie a été l'objet d'investigations importantes, notamment l'analyse des asymétries d'informations entre assureurs et assurés, source, d'une part d'anti-sélection, pouvant entraîner l'éviction du marché de l'assurance des individus présentant des niveaux de risque élevés ; d'autre part de risque moral, qui peut conduire à une consommation excessive de biens et services médicaux. L'impact de l'assurance sur les offreurs de soins peut aussi créer un biais technologique qui les amène à développer des innovations dont les bénéfices sociaux sont inférieurs au coût réel.

Enfin, un dernier ensemble de travaux traite de la régulation du système de santé. Une première voie est celle de la responsabilisation financière de la demande, qui conduit à une baisse des dépenses de l'assurance maladie, mais qui suppose que les patients ont la capacité d'évaluer l'opportunité de leurs dépenses. Par ailleurs, concernant la rémunération des professionnels et des établissements de santé, les résultats théoriques montrent la supériorité — en termes d'incitation et de partage des risques — des systèmes de paiement mixtes sur les mécanismes de rembour-sement ex post et les budgets globaux définis ex ante. Enfin, en termes de maîtrise des dépenses, l'intégration des fonctions d'assurance et de production de soins est un mode d'organisation globalement performant, mais il est difficile de faire la part des bénéfices qui relèvent de la rationalisation de la consommation et ceux liés aux stratégies de sélection de la clientèle par les assureurs.

Les dépenses de santé, déjà considérables, sont encore appelées à croître

compte tenu du progrès technique, de la croissance et du vieillissement de la population, de l'incidence des nouvelles pathologies telles que les maladies chroniques ou encore de la valeur accordée à la santé dans les sociétés actuelles. Ce phénomène pose, à l'évidence, la question du financement de ces dépenses avec toujours plus d'acuité et, corrélativement, celle de l'effi- cacité de ces dépenses : tout observateur peut être surpris, par exemple, de la variabilité des pratiques médicales ou de la présence de listes d'attente pour certaines prestations et s'interroger sur leur fondement [2]. Les indicateurs actuels sont relativement frustes pour apprécier la performance globale d'un système de santé au regard de la dépense engagée : l'approche par le résultat (par exemple l'espérance de vie corrigée des incapacités) ne permet pas d'évaluer le rendement marginal des ressources et donc de déterminer, en fonction de leur coût d'opportunité, si le financement est trop élevé, insuf- fisant ou adéquat. En effet, la mesure du produit — la santé — est en elle-même très délicate.

Les voies d'analyse tracées par Arrow constituent les points d'ancrage des études qui se sont développées depuis une quarantaine d'années. On peut les repérer selon trois principales perspectives, qui constituent aussi les lignes directrices de cet article.

À l'origine de la demande soins se trouve la notion de « besoins » de santé définis comme la différence entre l'État de santé constaté et l'État de santé souhaité. Au niveau individuel, cette différence résulte le plus souvent d'une dégradation de l'état de santé relativement imprévisible (c'est ce que l'on appelle le risque épidémiologique ou risque d'être atteint par une maladie), mais aussi d'une dégradation « normale » liée au vieillissement.

La demande de soins, traduction des besoins de santé, est, quant à elle, largement dépendante de l'offre. En effet, le patient, seul, ne dispose pas des connaissances médicales nécessaires pour déterminer ou juger de l'opportunité des soins qui lui sont prodigués en fonction des besoins qu'il exprime. L'utilité et la qualité des soins sont donc difficiles à évaluer par les patients — bien que se développent aujourd'hui des méthodes de révélation des préférences qui tendent à donner un rôle plus actif à ces derniers. Par ailleurs, de nombreux

travaux ont analysé les éventuelles déviations liées à ce pouvoir discrétionnaire des médecins dans le cadre de l'hypothèse de demande induite.

Enfin, il convient aussi de s'interroger sur le lien entre la consommation de soins et l'état de santé qui, au niveau macroéconomique, s'inscrit dans une relation complexe mais qui peut être appréhendé plus facilement dans le cadre de l'évaluation médico-économique des actions de santé.

L'incertitude pesant sur la demande de soins renvoie également au rôle de l'assurance maladie et notamment à ses effets sur le comportement du patient assuré. Il s'agit là de la deuxième perspective qu'offre l'analyse d'Arrow, qui a été depuis largement investie par les économistes et considérablement enrichie par les apports récents de la théorie microéconomique. Cette perspective est traditionnellement dans le cadre de l'analyse de la dynamique du contrat d'assurance : avant/après la signature, avant/après la maladie.

Avant la conclusion du contrat, les dysfonctionnements possibles du marché ramènent au paradoxe d'Akerlof (1971) : en présence de sélection adverse (les individus en mauvaise santé ont tendance à s'assurer plus souvent que ceux qui sont en bonne santé) portant sur les risques des individus, le marché est menacé si les primes sont fixées de façon actuarielle (sur la base des statistiques générales de morbidité, ou le marché est incomplet et inéquitable si les assurances peuvent sélectionner les risques. Ces dysfonctionnements légitimitent à l'intervention de l'État dans l'assurance maladie. Celle-ci peut prendre la forme a minima d'une réglementation mais aussi celle du monopole public dès lors que le système de santé joue un rôle en matière de redistribution des richesses en faveur des bas revenus (qui sont aussi les plus malades).

Après la conclusion du contrat, deux questions sont soulevées : l'assu- rance a-t-elle un impact sur l'occurrence du risque (risque moral ex ante)? Génère-t-elle une surconsommation de soins liés au petit risque (risque moral ex post)?

La dernière partie de cet article est consacrée aux instruments économiques qui peuvent être mobilisés afin de pallier les effets négatifs, en termes d'efficacité et d'équité, de l'incertitude généralisée dans le marché des soins médicaux et de l'assurance maladie. Ces instruments peuvent être regroupés

en deux grandes catégories : ceux qui affectent les patients/assurés via des mécanismes de responsabilisation de la demande et ceux qui agissent sur les incitations économiques auxquelles les offreurs de soins sont confrontés. L'efficacité de ces instruments reste toutefois contingente de la validation d'hypothèses relativement fortes : capacité des patients à évaluer leur consommation de soins pour les premiers ; structure du marché des soins (absence de collusion, de stratégie d'écrémage,…) pour les seconds. L'intégration des fonctions d'assurance et de production au sein d'une même entité est une autre forme de régulation du secteur qui repose sur la combinaison des deux instruments précédents, mais il semble que l'objectif poursuivi par cette intégration — rendre compatible les objectifs de l'ensemble des acteurs — réclame des hypothèses tout aussi lourdes que ces instruments pris isolément.

I. La nature de la demande

C'est Arrow qui déjà attira l'attention sur l'une des spécificités les plus évidente du marché des soins médicaux : le caractère instable, irrégulier et imprévisible de la demande individuelle, de même que l'utilité des soins qui en dérive et le coût d'opportunité de la maladie. Depuis lors, l'analyse de la demande s'est déclinée en plusieurs approches relevant parfois du calcul économique appliqué à la santé : quel est le rôle des besoins de santé et de la demande de santé dans l'expression de la demande de soins ? Quel est le rôle des professionnels de santé dans la traduction de la demande de santé en demande de soins ? Quel est l'impact de la consommation de soins sur l'état de santé ?

1.1. Les besoins de santé

La notion de besoins de santé renvoie à l'écart entre un état de santé constaté et un état de santé souhaité. Cette définition implique que pour évaluer les besoins de santé il faut, d'une part, mesurer l'état de santé et, d'autre part, définir et quantifier un état de santé souhaité. Là encore, la détermination d'un référentiel d'état de santé est difficile : des référentiels absolus définis par des experts risquent d'être déconnectés des besoins ressentis de la population ou tout simplement de refléter les services disponibles ; des

référentiels relatifs définis sur la base de comparaisons territoriales tendent à ériger la moyenne en norme. La mesure de l'état de santé constaté n'est pas aussi exempte de difficultés : les données sur la morbidité restent à ce jour largement insuffisantes et le taux de mortalité (disponible, quant à lui, à une échelle géographique fine) ne donne qu'une image partielle des différences d'état de santé (Lacoste, Salomez, 1999).

Au niveau individuel, l'expression d'un besoin de santé traduit une détérioration — au moins perçue — de l'état de santé. L'incertitude sur la probabilité de connaître une dégradation de l'état de santé et d'avoir recours au système de soins constitue ce que l'on appelle le risque épidémiologique. Arrow soulignait ainsi que le caractère imprévisible des dépenses de santé (et leur ampleur potentielle) à l'échelle individuelle était une spécificité forte de ce poste dans le budget des ménages.

Il est cependant utile de préciser cette particularité de la demande de soins. Le risque épidémiologique regroupe en effet des phénomènes dont le caractère aléatoire n'est pas homogène.

Tout d'abord, sans aller jusqu'à la distinction entre « petits risques » et « gros risques » qui a nourri des débats importants autour du panier de soins, il est d'usage de distinguer entre des risques de faible probabilité mais associés à des coûts importants et des risques de forte probabilité associés à des coûts relativement modérés. Cette distinction s'applique, par exemple, entre le recours à la médecine de ville (forte probabilité mais moyenne et variance des dépenses modérées) et le recours à l'hôpital (probabilité faible mais variance et moyenne élevées). Newhouse (1989) montre ainsi que les caractéristiques individuelles permettent d'expliquer 48 % des disparités entre individus en ce qui concerne les soins primaires mais seulement 8 % des différences d'utilisation des soins hospitaliers [3]. Le risque épidémiologique est donc beaucoup plus important dans ce dernier cas alors que la consommation de biens et services médicaux courants peut être prédite de manière relativement satisfaisante.

Par ailleurs, le caractère aléatoire du risque épidémiologique est aussi lié à l'horizon d'analyse. À court terme, le recours aux soins peut être relativement prévisible du fait du développement des maladies chroniques. Bien que l'appa-

rition de ce type de maladie soit aléatoire, cet état implique des besoins chroniques et des dépenses de santé qui sont généralement peu volatiles dans le temps. Kronick *et al.* (1995) montrent ainsi que plus de la moitié des dépenses de santé des malades chroniques peut être prédite. À plus long terme, les données épidémiologiques rendent aussi la demande individuelle de soins plus prévisible. Ainsi, les maladies liées au vieillissement et leur prévalence sont relativement bien connues (Alzheimer, maladies cardio-vasculaires et troubles musculo-squelettiques). En revanche, le long terme est aussi moins prévisible avec des évolutions possibles, tant du côté épidémiologique que technique, qui peuvent s'inscrire en rupture par rapport aux tendances observées aujourd'hui (incertitude radicale au sens de Knight, 1921).

Grossman, quant à lui, propose une modélisation de la demande de santé dans le cadre de la théorie du capital humain. Il requalifie demande de soins et demande de santé et précise ainsi l'idée de Arrow, à l'état d'ébauche dans son article : « La maladie représente non seulement un risque mais un risque coûteux en soi en dehors du coût des soins » [6]. Si l'état de santé d'un individu (ou son « *capital santé* ») était déjà considéré comme l'une des composantes du capital humain, c'est Grossman qui, le premier, modélise la demande de santé *per se*. Au-delà de la construction théorique, l'originalité de la démarche est de poser les premiers jalons d'une conceptualisation de la demande de soins médicaux comme dérivée de la demande de santé. Les patients ne retirent pas une utilité directe de leur consommation de soins : c'est l'incidence de cette dernière sur la conservation de leur capital santé (et des flux de revenus et de bien-être susceptibles d'en résulter) qui en est la motivation principale. Rochaix (1997) étaye cette distinction entre les deux objets : la santé « *n'a qu'une valeur d'usage à la différence des soins de santé qui ont une valeur d'échange* »

Par la suite, si l'analyse théorique de la demande de santé prolonge le modèle de Grossman tout en l'épurant, le concept de capital santé trouve une traduction tangible dans les mesures élaborées pour estimer l'utilité des états de santé. Ces mesures prennent la forme de questionnaires que les patients doivent renseigner et qui concernent plusieurs dimensions de leur qualité de vie (ou « attributs »). Différentes échelles médicales permettent ainsi d'évaluer la

qualité de vie comme une variable unique définie sur un continuum allant de 0 (mort) à 1 (parfaite santé) et de comparer l'utilité de traitements de nature très différente. L'intérêt de cette approche s'illustre tout particulièrement dans l'usage des QALYs (Quality Adjusted Life Years) qui sont définis comme les années de vie en parfaite santé jugées équivalentes aux années effectivement vécues dans un état de santé donné. Le concept, à l'origine purement empirique, trouve son fondement dans les travaux de Torrance (1976) au Canada. Mais, par la suite, d'indicateurs de santé, les QALYs sont devenus une mesure de l'utilité et ont été utilisés dans le cadre du calcul économique appliqué à la santé (*cf.* encadré 1).

1.2. Des besoins de santé à la demande de soins

La transposition de la demande de santé en demande de soins se traduit, selon Arrow, par une incertitude majeure en ce qu'elle concerne l'utilité de l'achat de soins. L'information sur cette utilité est alors précisément ce qui est acheté aux médecins sous forme de soins qualifiés.

Dans cette optique, les procédures de contrôle et de vérification *a poste- riori* de la qualité des soins sont peu envisageables. L'offreur de soins ne peut s'engager vis-à-vis du patient sur l'efficacité des traitements dès lors que de nombreux facteurs individuels peuvent interférer avec le résultat attendu. Arrow considère donc que les soins médicaux sont des biens de confiance : le seul contrôle de la qualité possible réside alors dans des procédures de certification *a priori* des professionnels de santé.

Rochaix (1989) relève pour sa part que les soins médicaux sont, au moins en partie, des biens d'expérience : les patients sont capables d'avoir un jugement, fût-il subjectif, sur la qualité des soins qui leur ont été prodigués. La diffusion de l'information sur la qualité des médecins et de leurs prescriptions entre les patients exerce alors une pression sur le niveau de qualité offert par les médecins notamment *via* la menace du « vote avec les pieds ». C'est également ce que suggère Tirole (1993) à propos des biens d'expérience : les patients informés, en diffusant cette information sur la qualité, exercent un effet externe positif sur les patients non informés. L'effet de réputation est de nature à inciter les offreurs de soins à améliorer la qualité des soins afin d'attirer la

clientèle.

D'une manière générale, la qualité des soins est un argument présent dans la fonction d'utilité des patients mais son impact sur la demande peut varier considérablement selon les pathologies (les malades chroniques disposant généralement d'informations pertinentes par exemple), selon la structure du marché (l'existence d'offreurs multiples permettant de comparer différentes stratégies de traitement) et selon les caractéristiques individuelles (les patients les plus éduqués étant aussi ceux qui ont un accès plus facile à l'information). Depuis quelques années, on note une volonté de redonner sa place au patient, en améliorant l'information à sa disposition ou en mettant en place des procédures de participation des individus aux choix thérapeutiques qui les concernent. Lebrun (1999) considère que « *ce souci relève d'un double objectif : d'une part [...] accorder au patient la place irremplaçable qui lui revient dans le processus de décision le concernant au premier chef ; d'autre part, dans le contexte de l'allocation optimale des ressources, inventer des méthodologies permettant de procéder à cette allocation optimale* ».

L'évolution de la relation médecin-patient a été marquée par de profonds changements, amorcés dès les années 1960 en Amérique du Nord. Auparavant, le patient déléguait tout son pouvoir décisionnel au médecin, qui considérait en retour être le dépositaire des besoins, désirs et croyances des patients (en sus de la possession du savoir médical). Cette relation, qualifiée de paternaliste, a été aujourd'hui remise en cause, le patient retrouvant sa place en tant que sujet au cœur du processus de décision le concernant. Deux nouveaux modèles de relations entre patients et médecins ont ainsi émergé : le modèle du « patient décideur » (largement théorique) où le médecin transmet toute l'information au patient qui est ensuite seul « décideur » des choix thérapeutiques et le modèle de la décision partagée où ces choix sont effectués par le médecin et son patient ensemble, après avoir échangé mutuellement connaissances médicales et structure des préférences.

Toutefois, le savoir médical est bien souvent complexe et probabiliste. Le transfert de la connaissance du médecin au patient qui permet à ce dernier d'élaborer ses préférences (et ainsi de prendre part activement au processus de décision) est à l'origine de difficultés et de résistances importantes. Il est

donc nécessaire de proposer aux professionnels de santé des outils qui leur permettent de transmettre les informations pertinentes aux patients afin que ces derniers puissent disposer des éléments qui leur permettront de faire des choix ou, tout au moins, de classer différentes options. Dans ce contexte, la révélation des préférences, « vieux » problème économique, est au centre de travaux novateurs en économie de la santé.

L'existence d'un tiers (le médecin) dans la transposition de la demande de santé en demande de soins implique aussi des réflexions sur d'éventuels comportements stratégiques de ce tiers. Le débat sur l'existence d'une demande induite par les professionnels de santé remonte aux travaux d'Evans en 1974. La demande induite renvoie à la capacité (réelle ou supposée) dont dispose un médecin pour générer une demande pour ses propres services. Eu égard à l'avantage informationnel qu'il détient, un médecin est en effet en mesure de recommander, voire d'imposer, une prestation dont l'utilité se situe plutôt du côté de ses objectifs de revenu. L'incertitude médicale rend en effet difficile la séparation entre prestations légitimes (correspondant strictement aux besoins des patients) et prestations induites (visant à augmenter les revenus des professionnels de santé). Outre l'asymétrie d'information entre patients et médecins, deux éléments favorisent ce type de comportement : le mode de rémunération du producteur et la couverture assurantielle des patients qui les rend insensibles aux prix.

Ces effets d'offre concernent tant la médecine ambulatoire que l'hospitalisation. Dans le premier cas, on considère généralement que, dans le cadre du paiement à l'acte, les médecins contribuent à créer le marché, ce qui se traduit par un lien croissant entre la densité médicale et la demande de soins.

Dans le second cas, l'offreur de soins est supposé agir de façon discrétionnaire sur la demande de soins (nombre d'admissions et durée de séjour), ce que semblent confirmer les études empiriques qui mettent en évidence une corrélation entre ces variables et le nombre de lits d'hospitalisation.

L'hypothèse de demande induite suppose donc que le médecin influence les préférences du patient de façon à satisfaire ses fins propres. Toutefois, il n'existe pas de consensus sur la définition de la fonction-objectif du médecin : maximise-t-il son profit ou l'utilité de son patient ? L'idée introduite par Evans

selon laquelle le médecin cherche à atteindre un revenu-cible a été largement controversée, souvent considérée comme bâtie *a posteriori* pour apporter une interprétation à des faits observés empiriquement (en particulier, la discrimination par les prix pratiquée par les médecins américains). C'est aussi ce que montrent Rochaix et Jacobzone (1997), qui rappellent que d'autres interprétations peuvent expliquer le lien entre densité médicale et coût des soins. L'une d'entre elles consiste, en particulier, à supposer que l'accroissement de la densité médicale favorise la différenciation des services vers le haut (diminution des files d'attentes, augmentation des durées de consultation), ce qui pourrait justifier l'augmentation du coût des soins de deux façons : le prix de l'amélioration de la qualité des soins, l'accroissement de demande qui résulte de la variation de qualité des soins. Même si la définition de la demande induite est communément admise parmi les économistes, l'absence de consensus sur un modèle théorique conduit à relativiser la portée de certaines études économétriques concourant à valider l'hypothèse de la demande induite, comme celle de Gaynor et Gertler (1995) qui montre l'influence significative de variables relatives à l'offre de soins — prix, contrats — sur le contenu du recours aux soins du patient. Le foisonnement d'études empiriques peine finalement à démontrer l'existence et l'ampleur du pouvoir discrétionnaire des médecins (McGuire, 2000).

Face à ces difficultés méthodologiques, Phelps, en 1986, propose de centrer les tests non plus sur la densité médicale mais sur la réponse fournie par les praticiens face à un choc exogène affectant leur rémunération (ce choc peut être un gel des tarifs ou une baisse de la fécondité qui affecte le revenu des médecins accoucheurs). Les conclusions de ces études restent toutefois ambiguës : les médecins disposent effectivement d'un pouvoir discrétionnaire qui leur permet de générer une demande pour leurs propres services mais il semble qu'ils n'y ont recours que de manière relativement modérée, en réponse à une modification de leurs conditions d'exercice.

Les travaux réalisés par les économistes de la santé ont donc montré que le lien entre besoins de santé et demande/consommation de soins s'inscrit dans une relation complexe, marquée par l'incertitude médicale, le problème de

la construction et de la révélation des préférences des patients et l'existence d'un tiers. D'autres travaux ont considéré la relation inverse et ont cherché à comprendre l'influence de la consommation de soins sur l'état de santé.

1.3. Dépenses de soins et état de santé

La mesure de l'état de santé appréhendée au niveau global repose sur des indicateurs imparfaits (espérances de vie éventuellement corrigées des incapacités, mortalité infantile, périnatale et maternelle), mais reste un outil d'analyse important des performances des différents systèmes de santé. Les analyses empiriques montrent que si les pays qui consacrent le plus d'argent à la santé sont les plus riches, ce ne sont pas forcément ceux où les indicateurs de santé sont les meilleurs. Cette observation amène à deux remarques : les biens et services médicaux sont, au niveau macroéconomique, des biens supérieurs [7] ; la santé déborde largement le cadre des soins. Cependant, la comparaison des performances des systèmes de santé, exercice auquel se livrent parfois les organisations internationales (OMS, 2000 ; Banque mondiale, 2003) n'est pas sans poser de sérieux problèmes méthodologiques.

Le développement des études sur les déterminants de la santé (notamment au Canada et aux États-Unis dans les années 1980 et 1990) a permis de mettre en évidence la contribution importante des facteurs autres que la consom- mation des soins médicaux tels que par exemple l'environnement physique ou social, les facteurs psychologiques ou le patrimoine génétique (Evans et Stoddart, 1994). Ces facteurs interagissent avec le système de santé ou se combinent au titre d'inputs dans la fonction de production de santé.

L'apport du système de santé dans les gains en espérance de vie et le recul de certaines pathologies (la tuberculose par exemple) est ainsi estimé de 10 à 20 % [8] entre 1950 et 1990. En effet, l'amélioration de l'état de santé d'une population relève, d'une part d'une diminution de la prévalence de la morbidité générale (qui est surtout liée à l'amélioration des conditions de vie et peu au système de soins *per se*) et, d'autre part, d'une meilleure prise en charge des personnes malades (liée quant à elle à l'organisation des soins). Ceci explique que l'on peut, dans le même temps, prétendre que le système de soins n'a pas un impact fort sur l'état de santé d'une population et que le recours aux

soins joue un rôle important pour un individu malade.

Dès lors, comment s'assurer, selon Mougeot (1999), « *d'une part, que le montant des ressources consacrées à la santé est correct (au sens où l'on ne pourrait faire mieux en dépensant moins (plus) pour la santé au profit (au détriment) d'autres services collectifs) et, d'autre part, que ces ressources sont bien utilisées (au sens où l'on ne pourrait améliorer l'état de santé de la population en utilisant autrement les mêmes ressources*) » ? Cette question — qui est celle de *l'efficacité allocative* [9] — renvoie donc à la caractérisation d'une fonction de production de santé, où les soins médicaux représentent une partie des inputs pour lesquels il convient de s'interroger sur leur quantité et leur combinaison.

Une autre question est celle de l'impact du système de santé sur l'état de santé des différents groupes de population ou encore de l'équité horizontale (les états de santé et les modes d'accès aux soins sont-ils indépendants des caractéristiques individuelles ?). Sa mesure repose sur différents indicateurs des états de santé [10], leur comparaison dans l'espace et/ou selon des catégories socio-économiques (revenu, CSP, emploi), de même que leur dynamique. L'appréciation de l'équité horizontale des systèmes de santé renvoie également à un ensemble de mesures de l'accès en termes de recours aux soins. Il s'agit là d'analyser la demande de soins au travers d'études empiriques, descriptives ou probabilistes (sur les données françaises, ce sont principalement les travaux de l'IRDES, de la DREES et de l'INSEE qui renseignent sur les caractéristiques de cette demande), tandis que les différentes dimensions de l'équité dans les systèmes de pays européens font l'objet des travaux du groupe Ecuity (Wagstaff *et al*. 2000).

Des méthodes économiques ont toutefois été développées pour mesurer non pas les performances du système de soins dans son ensemble, mais pour évaluer les résultats d'une action de santé à l'aune des coûts et des bénéfices qui en résultent. Dans le cadre des soins de santé, l'évaluation se heurte à des difficultés importantes ; en particulier, la socialisation des dépenses de santé biaise les mécanismes d'ajustement par les prix et rend donc délicate l'estimation des surplus. La mesure du bénéfice monétaire d'un résultat de santé a donc nécessité l'élaboration d'un certain nombre d'outils qui, bien qu'imparfaits, permettent toutefois d'éclairer les choix des décideurs.

Dans une première étape, des modèles de type capital humain ont été appliqués à l'étude des bénéfices. Cette approche se fonde sur l'actualisation des revenus futurs des individus. Elle est particulièrement utilisée aux États-

Unis dans le cadre des recours juridiques qui supposent une estimation des préjudices subis ou encore dans l'analyse des conséquences de la mortalité évitable. Toutefois, la limite intrinsèque de ce raisonnement est qu'il ne s'applique qu'aux bénéfices liés à une modification de l'activité productive, alors que les résultats d'une action de santé devraient être mesurés par leurs répercussions globales.

Une seconde étape a été marquée par l'application au domaine sanitaire de la notion de consentement à payer (*willingness to pay*). Cette notion, due à Dupuit (1844), évalue la valeur d'un bien au prix maximal qu'un consommateur est prêt à payer pour l'acquérir. Toutefois, dans le domaine de la santé, la mesure du consentement à payer des individus n'est pas un exercice simple ; deux voies ont été retenues : une première fondée sur l'observation des arbitrages monétaires effectués en termes de risques (méthode des préférences révélées) et une seconde basée sur les préférences déclarées par les individus dans le cadre d'un marché fictif (méthode dite d'évaluation contingente).

L'évaluation contingente repose sur le principe de compensation de Hicks-Kaldor qui implique qu'une action de santé sera préférée à la situation initiale si les gagnants à la suite de cette action sont en mesure, au moins théoriquement, de dédommager les perdants tout en demeurant bénéficiaires. Cette approche met donc en regard les bénéfices et les pertes supportés par les différents acteurs. « *La méthode de l'évaluation contingente consiste à proposer à un individu (un groupe d'individus) une situation de marché hypothétique (d'où l'appellation d'évaluation contingente) sur lequel la personne interrogée doit pouvoir se prononcer sur le montant monétaire qu'elle est prête à consentir pour accéder au bien proposé. Ce montant représente le bénéfice monétaire associé à l'accès au marché.* » Allenet et Sailly (1999).

Lorsqu'on applique cette méthode dans le domaine sanitaire, il est important de prendre en compte la présence d'une double incertitude, la première renvoyant à la survenance de la maladie qui va déterminer la demande de soins (le risque épidémiologique) et la seconde étant relative à l'efficacité même de la stratégie (*i.e.* à ce que Mougeot, 1999 nomme l'incertitude médicale).

Une revue de la littérature montre la diversité des actions de santé qui

peuvent être analysées par le biais de l'évaluation contingente : évaluation du coût du handicap, valorisation de l'information apportée par un test diagnostique, évaluation des nouvelles molécules ou stratégies thérapeutiques, etc. Toutefois, les conditions de validité de ce type d'analyse sont de trois ordres et limitent la portée pratique de la méthode :

— le consentement à payer de l'ensemble des individus, dont l'utilité est susceptible d'être affectée par le programme, doit être pris en compte afin de ne pas en sur ou sous-évaluer le bénéfice social ;

— les individus doivent disposer d'information sur leur probabilité d'avoir recours au programme, c'est-à-dire sur l'incertitude exogène ;

— les résultats de ce programme doivent être décrits en termes probabilistes dans le cadre de l'incertitude médicale.

L'enquête est délicate car elle est généralement réalisée par le biais d'un questionnaire, ce qui démultiplie les contraintes méthodologiques : au-delà de la condition de présentation et de compréhension du scénario proposé aux personnes enquêtées à qui l'on doit délivrer une information claire et complète, il s'agit d'éviter les biais d'échantillonnage, les biais relatifs à la construction du questionnaire et les biais relatifs à l'enquêteur et à l'enquêté… La méthode de l'évaluation contingente suscite donc des débats importants concernant sa fiabilité, mais elle reste relativement bien adaptée dans un cadre plus restreint que celui de l'analyse coût-bénéfice. Par exemple, pour mesurer la désutilité liée au handicap ou le coût psychologique de la maladie, les informations relevant du consentement à payer et les différentiels observés devraient être pris en compte dans le processus de décision d'allocation des ressources, en complément à d'autres formes d'évaluation (*cf.* encadré 1).

1. *Le calcul économique appliqué à la santé*

L'évaluation économique des programmes de santé — encore désignée par « calcul économique appliqué à la santé » — fait l'objet d'investigations depuis la fin des années 1970, dans la veine des travaux menés en économie publique (analyses coût-avantage, rationalisation des choix budgétaires,…). Il s'agit d'une boîte à outils d'aide à la décision, mobilisables pour éclairer les choix entre plusieurs programmes concurrents. Ces programmes sont évalués en termes d'efficience, puisqu'il s'agit de rapprocher leurs conséquences et leurs

coûts qui s'étendent généralement sur plusieurs périodes de temps. Toutefois, si l'estimation des coûts en termes monétaires repose sur une méthodologie commune et des outils classiques (actualisation, raisonnement en coût marginal,...), la mesure des conséquences peut procéder de deux logiques différentes.

D'un côté, les conséquences sont estimées en termes non monétaires par le biais d'une Analyse Coût-Efficacité [ACE] (par exemple, des années de vie gagnées) ou d'une Analyse Coût-Utilité [ACU] (par exemple, des QALYs). Dans ce cas, l'objectif est d'allouer des ressources à des programmes concurrents en choisissant le plus efficient. En d'autres termes, la décision est prise sur la base du prix à payer pour atteindre un certain objectif. Cela suppose donc, au moins implicitement, que l'arbitrage quant au montant de l'allocation des ressources a déjà été effectué : il s'agit de juger de l'efficacité productive d'un programme.

De l'autre côté, les conséquences sont évaluées en termes monétaires dans le cadre d'une Analyse Coût-Bénéfice [ACB] : il s'agit alors, comme on l'a vu, d'estimer en termes monétaires les conséquences des actions de santé et de les comparer aux coûts engagés. En attribuant une valeur à une action de santé, cette démarche permet donc de juger du coût d'opportunité d'un programme, éventuellement au vu d'autres programmes sans lien avec la santé. Ce faisant, on adopte une perspective d'efficacité allocative.

D'un point de vue conceptuel comme appliqué, ces différentes méthodes soit expérimentales, soit reposant sur des choix méthodologiques dont le fondement peut toujours être remis en cause (notamment en raison des présupposés théoriques portant sur les préférences individuelles et leur agrégation) — apparaissent de ce fait souvent sujettes à caution (Moatti, 1995 ; De Pouvourville, 1993). De plus, l'expérience de l'État de l'Oregon qui a tenté de déterminer le panier des biens et services remboursables par le biais d'indicateurs de type coût par Qaly s'est avérée désastreuse : les couronnes dentaires étaient prises en charge mais pas les appendicectomies, le traitement de la migraine mais pas celui du Sida...

Toutefois l'apport du calcul économique n'est plus à démontrer, dès lors qu'on en observe l'utilisation concrète pour encadrer les pratiques médicales

(par exemple, par le biais des Références Médicales Opposables [RMO] en France). Il s'est aussi beaucoup développé dans l'industrie pharmaceutique pour orienter la recherche (par le biais de l'évaluation pharmaco-économique) et démontrer auprès du décideur public l'Amélioration du Service Médical Rendu [AMSR] ou le moindre coût des nouvelles spécialités.
Ce bref aperçu permet de jauger toute la complexité de la question de l'efficacité allocative, autrement dit de l'arbitrage en termes d'allocation des ressources ou de ce que peut sous-tendre l'idée de « maîtrise des dépenses de santé ».

2. Assurance maladie et incertitude

L'ignorance du futur, l'incertitude pesant sur l'état de santé, la demande de soins et la qualité de ces soins sont, pour des individus ayant de l'aversion vis- à-vis du risque, sources d'inefficacité dans la mesure où ces individus sont prêts à payer pour diminuer cette incertitude. Le développement des organismes d'assurance répond à ce phénomène mais reste cependant insuffisant pour garantir l'ensemble des risques auxquels les individus peuvent être confrontés. Ceci est particulièrement vrai dans le domaine de la santé où l'assurance ne prend en charge qu'une composante monétaire de ce risque et opère dans un contexte d'asymétries d'information prononcé (Arrow). Les asymétries d'infor- mation ont potentiellement trois conséquences principales : une sélection des risques par les assurances, des phénomènes d'aléa moral du côté des assurés et des biais technologiques du côté des offreurs de soins.

2.1. La sélection des risques

Un des problèmes majeur auquel doit faire face le marché de l'assurance maladie est que les caractéristiques individuelles des acheteurs affectent drasti-quement les coûts de production du bien vendu (en l'occurrence le contrat d'assurance) et que ces caractéristiques ne sont pas observables par l'assureur [12]. Ce phénomène conduit à des dysfonctionnements importants sur le marché de l'assurance, décrits dans l'article de Rothschild et Stiglitz (1976). Ces dysfonctionnements ont trait à la sélection adverse [13] : une personne qui sait avoir une probabilité élevée d'être malade aura plus intérêt à

souscrire un contrat d'assurance généreux qu'une personne qui se sait en bonne santé. La population assurée présentant alors un niveau de risque plus élevé que celui de la population générale, l'assureur va devoir augmenter les primes, ce qui peut entraîner un retrait du marché des risques les plus faibles (parmi les personnes qui avaient choisi de s'assurer au départ). La population assurée voit alors son niveau de risque moyen augmenter, donc les primes demandées par les assureurs vont aussi augmenter, et les risques les plus faibles parmi la population assurée, trouvant le montant des primes trop élevé par rapport à leur propre niveau de risque, vont se désengager, et ainsi de suite. Dans ce cadre, un contrat d'assurance privé est inefficace, voire instable. Dans le modèle de Rotschild et Stiglitz, l'introduction de contrats différenciés permet toutefois d'atteindre un équilibre de second rang offrant une couverture complète aux risques élevés et une couverture partielle pour les risques les plus faibles.

Du côté des assurances, quand les risques individuels ne sont pas observables (et donc que la tarification au risque n'est pas réalisable), les compagnies sont incitées à mettre en place des politiques d'écrémage des risques. Bocognano *et al.* (1998) recensent cinq facteurs à l'origine de la difficulté des assureurs à procéder à une tarification au risque : les contraintes techniques telles que le coût de collecte des informations ou le manque de fiabilité des informations, la taille du groupe assuré qui limite le nombre de facteurs de risques qui peuvent être pris en compte, la multiplication des catégories tarifaires qui entraîne des coûts de gestion très (trop) élevés, la dimension éthique (ou aussi l'effet réputation) qui peut amener l'assureur à ne pas prendre en compte certains facteurs de risque et enfin l'interdiction par l'État de certains critères de tarification. Or, « *à partir du moment où la tarification au risque est imparfaite, l'assureur, qui reçoit un paiement uniforme pour des risques qui ne le sont pas, peut avoir intérêt à mettre en place des stratégies de sélection des risques, dites aussi d'écrémage* ».

Feldman et Dowd (1991), Cutler et Reber (1998) montrent que, dès lors qu'il existe un continuum de risques au sein de la population, ce sont les individus ayant le plus haut niveau de risque qui ont des difficultés à s'assurer du fait de la sélection adverse, contrairement aux prédictions du modèle de Rotschild et Stiglitz [16]. Les

contrats d'assurance ne peuvent généralement pas refuser de prendre en charge les personnes avec un niveau de risque élevé. Toutefois, des stratégies de dissuasion (par exemple restrictions dans l'accès aux professionnels de santé) pour les risques élevés et de promotion pour les bas risques (combinaison de l'offre d'assurance avec des services qui peuvent surtout intéresser les bas risques : adhésion à un club de gym etc.) peuvent être mises en place.

Dans un cadre dynamique, Geoffard (2000) souligne que « *les problèmes liés à l'information peuvent être exacerbés […] lorsque des informations sur le niveau futur du risque sont révélées après la signature du contrat et avant la réali- sation du risque* ». En effet, les individus souhaiteraient pouvoir s'assurer contre la probabilité de devenir un mauvais risque (et donc de voir leurs primes d'assu- rance augmenter), mais un tel mécanisme supposerait la signature d'un engagement bilatéral non renégociable entre assurance et assurés. Or un tel engagement ne peut exister dès lors que l'assurance peut augmenter ses tarifs ou exclure certains assurés et que les assurés peuvent rompre le contrat [17].

Cochrane (1995) propose toutefois une solution qui permet de surmonter cette difficulté *via* un mécanisme de paiements libératoires : chaque individu disposerait d'un compte spécifique crédité ou débité selon l'évolution du risque individuel d'un montant correspondant à la variation de la prime actuarielle à court terme.

2.2. Aléa moral du côté de la demande

L'aléa moral apparaît quand le comportement des assurés est modifié après la signature du contrat et ne peut pas être parfaitement observé par l'assu- rance. Ce comportement peut avoir trait à la prévention primaire (ou *self protection* selon Erlich et Becker, 1972) dont l'objet est de réduire la proba- bilité de sinistre, ou à la prévention secondaire (ou *self insurance*) qui modifie l'étendue du sinistre mais pas sa probabilité d'occurrence.

Dans le domaine de l'assurance maladie, les enjeux en termes de prévention primaire restent circonscrits car l'ensemble des conséquences de la maladie ne peut être garanti, que ce soit en termes de retour à l'état de santé initial ou de compensation de la souffrance. L'assurance maladie qui, dans les faits, peine à distinguer entre recours aux soins curatifs et recours aux soins préventifs (par exemple, le traitement de l'hypertension, qui est un facteur de risque et non une maladie, est pris en charge dans le cadre du risque maladie) aurait même l'effet inverse et

induirait un « aléa moral positif », c'est-à-dire que le recours à la prévention des personnes assurées serait plus élevé. En diminuant le coût de la dépense de prévention primaire pour les assurés, l'assurance constitue bien une incitation à la prévention.

En revanche, l'assurance a un effet moins favorable en ce qui concerne le recours aux soins des individus malades. L'expérimentation de la RAND Corporation et les nombreuses études menées à la suite de modifications des règles de remboursement mettent en évidence l'existence d'un aléa moral *ex post* : si l'assurance a un impact limité sur la prévention primaire (*self protection*), son incidence sur l'étendue du dommage après la survenue de celui-ci (*self insurance*) peut être conséquent (Newhouse *et al.*, 1996).

Les études réalisées d'après l'expérimentation de la RAND soulignent que la consommation des personnes est inversement corrélée avec leur participation financière (*cf.* encadré 2). Ainsi, le taux de surconsommation serait de l'ordre de 25 % pour les personnes bénéficiant de la gratuité des soins. L'expérience québécoise (instauration d'une franchise et d'un ticket modérateur pour les dépenses pharmaceutiques) met aussi en évidence une réduction de la consommation pharmaceutique (au moins à court terme). Cette réduction des dépenses n'est cependant pas homogène et certains types d'assurés y sont plus sensibles : les plus démunis, les gros consommateurs, les personnes souffrant de troubles mentaux.

2.3. Les interactions entre assurance et offreurs de soins

La plupart des économistes de la santé considèrent que, durant les cinquante dernières années, plus de la moitié de la croissance des dépenses de santé est imputable à un effet progrès technologique — introduction et diffusion de nouvelles technologies auprès d'un nombre croissant de patients (Fuchs, 1996).

Si l'assurance modifie le comportement des assurés, la solvabilisation de la demande qu'elle entraîne a aussi un impact sur la structure de l'offre. Certains auteurs parlent d'un aléa moral du côté de l'offre. Cette terminologie n'est pas totalement adaptée dans la mesure où ce ne sont pas les offreurs qui souscrivent un contrat d'assurance, mais il faut tout de même reconnaître que l'assurance maladie, en garantissant la prise en charge des soins médicaux, offre

une forme de « revenus garantis » aux offreurs. Ces derniers font face à une demande peu élastique au prix, ce qui a une incidence forte sur les incitations qu'ils peuvent rencontrer.

Selon Cutler (1997), le progrès technologique dans le domaine de la santé est biaisé du fait de ces incitations, tant quantitativement (montant investi en recherche-développement) que qualitativement (en termes d'orientation des programmes de recherche). Cela signifie, d'une part, que trop de procédures sont adoptées sur la base du bénéfice que peuvent en retirer patients et producteurs sans tenir compte du coût global de ces innovations et, d'autre part, que les patients, comme les professionnels, surestiment l'intérêt des technologies qui permettent d'élargir le champ des traitements disponibles au détriment des technologies qui permettent de limiter les dépenses. Pour reprendre l'analyse de Goddeeris (1984), l'assurance induit un biais technologique dans la mesure où la classification d'un ensemble d'innovations du point de vue de leur profitabilité diffère de celle effectuée en prenant en compte leur valeur sociale nette.

Weisbrod, en 1991, enrichit cette analyse en mettant en évidence les rétroactions entre assurance, recherche-développement et progrès technologique. Le développement des nouvelles technologies entraînant une augmentation des dépenses de santé génère une plus forte demande d'assurance [19] (car la charge financière sera plus lourde en cas de maladie). Cette augmentation de la demande d'assurance passe aussi par l'élargissement de la gamme des prestations pour lesquelles les individus souhaitent une prise en charge. En même temps, l'expansion de l'assurance — hausse concomitante du nombre d'assurés et du nombre de biens et services assurés — offre aux producteurs des incitations plus fortes, *via* la solvabilisation de la demande, pour développer de nouvelles technologies.

Weisbrod et Baumgardner (1991) montrent tout deux que ce biais technologique reste très fortement lié à la nature même du schéma d'assurance. Les assurances traditionnelles fondées sur le remboursement des coûts sont marquées par l'absence de procédures d'évaluation des nouvelles technologies : toute technologie qui améliore la qualité des soins est adoptée indépendamment de son coût. En revanche, les assurances reposant sur des

mécanismes de paiement *a priori* incitent les différents acteurs à utiliser des critères d'évaluation médico-économique avant d'adopter de nouveaux modes de traitement. Sont ainsi privilégiées les technologies qui diminuent les coûts sans nuire exagérément à la qualité des soins.

3. Les politiques de régulation

La section précédente montre que des mécanismes incitatifs doivent être intégrés dans les contrats d'assurance afin d'en préserver l'efficacité. Deux formes principales peuvent être observées : la participation financière des assurés aux dépenses (qui est censée limiter la surconsommation résultant du risque moral) et les schémas de responsabilisation des offreurs (en l'occur- rence les professionnels de santé) qui visent à harmoniser les objectifs de maîtrise des dépenses des assureurs et ceux des fournisseurs de soins. Selon Ellis et McGuire (1993), ce sont les systèmes combinant à la fois participation financière des patients et rémunération mixte des producteurs (c'est-à-dire avec une partie fixe et une partie proportionnelle aux coûts supportés) qui permettent de concilier au mieux efficacité et incitations.

3.1. La responsabilisation de la demande

Comme on l'a vu précédemment, la logique de responsabilisation de la demande repose sur le postulat que la fourniture d'assurance est susceptible d'introduire une incitation à une surconsommation, dès lors que les assurés disposent d'un pouvoir discrétionnaire sur le montant dépensé et que le niveau de remboursement dépend des dépenses engagées par les assurés.

Afin de pallier cet effet, des mécanismes de responsabilisation de la demande peuvent être mis en place par l'assurance maladie, publique ou privée. L'analyse du marché de l'assurance montre ainsi qu'en instituant une participation financière des patients, ces derniers retrouvent les incitations nécessaires à réaliser des arbitrages (selon la logique classique de la disposition à payer). Cette participation financière prend principalement deux formes : une franchise indépendante du montant des dépenses réellement engagées ou un co-paiement (par exemple, le ticket modérateur en France). Ces deux instru-

ments ont des propriétés économiques différentes : la franchise incite les assurés à se prémunir contre le risque d'avoir recours au système de santé alors que le ticket modérateur permet de plus d'inciter les assurés à modérer leurs dépenses le cas échéant (Winter, 1992).

Toutefois, l'opportunité des mécanismes de responsabilisation de la demande au coût des soins reste toute théorique, tant sur le plan de l'efficacité que de l'équité.

Du point de vue de l'efficacité, l'assurance maladie est confrontée au dilemme de Zeckhauser (1970), autrement dit à la recherche du juste arbitrage entre le degré de couverture assurantielle et les incitations individuelles à éviter la surconsommation de soins : le risque individuel de surconsommation serait inversement proportionnel au risque financier supporté par le patient [22]. Comme le rappelle Drèze (1997), « *en théorie le taux de couverture devrait être d'autant plus faible que l'élasticité de la demande de soins par rapport au taux de couverture est plus élevée, et que la tolérance au risque de l'assuré est plus élevé. Ceci conduit à un taux de couverture variable, après une franchise, et avec une couverture à 100 % pour les risques lourds […]. Cette modulation fine est difficile à appliquer en pratique.* ».

Par ailleurs, ce type de mécanisme suppose que les assurés ont la capacité de juger de la « valeur » de leur consommation de soins. Or, comme nous le rappelions dans la première partie, une des caractéristiques du marché des soins est l'asymétrie d'information entre les offreurs de soins et les patients. Ces derniers sont relativement peu informés sur la qualité des biens et services médicaux qu'ils utilisent : comment dès lors procéder à des arbitrages coût/efficacité ?

De nombreuses études empiriques sur l'instauration de mécanismes de co-paiement soulignent cette difficulté. Dans le domaine de la pharmacie par exemple, la participation financière des patients réduit autant leur consommation de médicaments efficaces ou essentiels que celle de médicaments dont l'efficacité est moindre (Tamblyn, 1998 ; Soumerai, 1991). De même, il n'existerait pas d'incidence sur l'utilisation de génériques dont le prix est inférieur du fait de l'augmentation de la participation financière (Leibowitz, 1985).

La question de l'équité est, par ailleurs, posée dès lors que l'on ne dispose pas

d'une mesure normative de la consommation optimale de soins, donc de la surconsommation et donc du coefficient adéquat de partage du coût. Cette question est, par exemple, illustrée par l'expérience de la RAND où, pour les catégories les moins favorisées de la population, le renoncement aux soins est aussi une conséquence — non désirable — de la participation financière ; la présence d'un effet-revenu est également très nette en France comme le montrent les études relatives à l'impact de la Couverture Maladie Universelle Complémentaire sur le recours aux soins (DREES, 2004). La référence à une norme de justice distributive remet également en cause l'efficacité (de second rang) des éventuelles politiques de discrimination au second degré de l'assurance maladie et rend nécessaire l'intervention de l'État dans le marché de l'assurance maladie (Mougeot, 1999).

Les mécanismes de responsabilisation de la demande présentent donc deux inconvénients majeurs : d'une part, ils affectent les personnes les plus démunies pour qui la contrainte financière est plus forte ; d'autre part, ils reposent sur une capacité d'arbitrage des individus entre différents postes de consommation qui reste théorique dans la mesure où les individus ne disposent pas d'informations suffisantes pour réaliser ces arbitrages. Le recours à ces mécanismes suppose *a minima* un encadrement par l'État et renvoie implicitement à la question de son rôle et aux modalités de son intervention au titre de l'efficacité et de l'équité.

3.2. Les modes de rémunération des offreurs de soins médicaux

Tout le problème de la régulation de l'offre de soins procède des caractéristiques du marché qui ont été évoquées jusqu'à présent. Rappelons que l'existence d'assurance a pour conséquence l'absence de réaction des patients aux variations de prix. Il en résulte l'impossibilité d'une décentralisation par les prix, ce qui pose naturellement le problème de la rémunération des offreurs de soins par le financeur. Cette question fait l'objet de nombreux travaux, notamment théoriques, qui comparent les propriétés de différentes règles d'allocation et de paiement dans des contextes variés (demande sensible ou non à la qualité des soins, différenciation horizontale ou verticale, etc.). Ce foisonnement théorique se nourrit des développements actuels de la théorie des incitations (Laffont et Tirole, 1993 ; Mougeot et Naegelen, 1997) et éclaire

ainsi les conséquences des modes de rémunération sur les comportements des offreurs de soins.

Ces travaux privilégient, le plus souvent, l'analyse des situations de risque moral (c'est-à-dire lorsque les offreurs de soins peuvent mettre en place des stratégies de réduction des coûts mais que ces stratégies ne sont pas observables par le financeur). Cette orientation s'inscrit dans la logique des réformes récentes des modes de rémunération des offreurs de soins (tarification à la pathologie pour les hôpitaux, systèmes combinant capitation et paiement à l'acte pour les médecins ou encore politiques d'enveloppe pour les deux secteurs). Ces mesures montrent en effet que le financeur souhaite non pas mettre en place une tarification adaptée à la structure économique des coûts des offreurs de soins, mais plutôt influencer leur comportement en termes d'efficacité productive.

Les différents modèles développés comparent les incitations — à la réduction des coûts ou à l'amélioration de la qualité — auxquelles les offreurs de soins sont soumis dans le cadre des systèmes de paiement rétrospectif (remboursement des coûts constatés) et prospectif (paiement forfaitaire déterminé *ex ante*). Dans le cadre du paiement des établissements de soins (*cf.* encadré 3), on montre généralement que des prix fixes sont préférables à un remboursement des coûts, mais ce résultat est étroitement conditionné par le respect de plusieurs hypothèses dont notamment l'absence de sélection ou d'écrémage de clientèle. Des résultats analogues ont été mis en évidence pour l'offre de soins ambulatoires : les mécanismes de type capitation semblent préférables au paiement à l'acte d'une manière générale (Blomqvist, 1991), mais bien souvent le relâchement de certaines des hypothèses des modèles (notamment la prise en compte de la qualité des soins comme variable) conduisent à préférer les schémas mixtes de paiement (Newhouse 1996).

S'il est aujourd'hui admis que les prix fixes sont avantageux en matière de rémunération des offreurs de soins (au moins dans le cadre d'un schéma mixte de paiement), la détermination de ce prix est une question cruciale [25].

La solution à ce problème a été apportée par Shleifer (1985), qui a théorisé *a posteriori* la pratique du programme social américain *Medicare* : il suffit d'orga-

niser une concurrence par comparaison ou « *yardstick competition* » entre les agents. Shleifer propose un mécanisme de régulation qui tienne compte des caractéristiques des autres entreprises afin d'instaurer une « concurrence fictive » sur la base de prix fixes. Ces prix sont déterminés sur la base des offreurs les plus efficaces, voire sur une disparition du marché. Or, la fixation du prix suppose la connaissance de plusieurs variables (fonction de bien-être social, nature de la demande, niveau de coût observable *ex post*...). Coûts annoncés par l'ensemble des entreprises. En fait, si n firmes sont présentes sur le marché, à chacune de ces firmes correspond un prix fixe, calculé comme la moyenne des coûts annoncés par les autres firmes. À l'équi- libre, ce prix sera le même pour l'ensemble des entreprises soumises à cette régulation et l'effort de réduction des coûts sera optimal pour chaque entre- prise. Ce résultat reste aussi valide si l'on introduit une certaine hétérogénéité entre les établissements. Il suffit que ces différences puissent être appréciées grâce à des caractéristiques observables par la tutelle.

Le modèle de Shleifer illustre l'intérêt des comparaisons de coûts entre firmes identiques mais en situation de monopole spatial (chaque entreprise desservant une zone géographique donnée sur laquelle il n'existe pas de concurrence). Cette structure de marché caractérise de façon assez satisfaisante le marché hospitalier : il existe un grand nombre d'établissements de soins mais la localisation de ces établissements (définie par les schémas régionaux d'organisation sanitaires) rend peu pertinente l'analyse traditionnelle de la concurrence. La concurrence par comparaison, qui sous-tend les mécanismes de paiement à la pathologie présente donc des propriétés théoriques indiscutables. En effet, cet instrument permet d'obvier aux inefficacités liées au risque moral (l'effort de réduction des coûts n'étant pas observable par la tutelle). Toutefois, deux limites doivent être soulignées. D'une part, ce mécanisme ne peut s'appliquer en présence d'une asymétrie d'information sur les caractéristiques exogènes des établissements et, d'autre part, cette procédure peut faire l'objet de manipulations par des ententes.

Par ailleurs, le mécanisme de concurrence par comparaison pose des problèmes évidents de mise en œuvre. Lorsque le mécanisme est basé sur les coûts de production, on peut obtenir l'efficacité productive, mais les offreurs

de soins peuvent choisir le nombre de patients traités (dans la mesure où l'on considère la tarification à la pathologie et non un système de concurrence organisée telle que les « *Managed Care Organizations* »). Le régulateur a cependant la possibilité de substituer à la demande des patients une demande fictive du type enveloppe globale, ce qui constitue une autre forme de concur- rence fictive, par les quantités de soins : le prix varie inversement aux quantités produites.

3.3. Intégration des fonctions d'assurance et de production de soins

L'intégration des fonctions d'assurance et de production de soins est une des caractéristiques traditionnelles des systèmes nationaux de santé comme le NHS britannique. L'État prélève les impôts qui seront utilisés pour financer les dépenses de santé et rémunère directement professionnels et établissements de soins sur la base du salariat, de la capitation ou des budgets globaux. L'un des mérites de cette organisation est que les dépenses de santé sont, de fait, maîtrisées (le budget étant déterminé *ex ante* et non susceptible d'augmen- tation). Cependant, dès lors que le montant global des ressources affectées au système de santé est insuffisant pour répondre à la demande de soins de la population, des mécanismes de rationnement apparaissent avec les listes d'attente [26]. Les principaux effets pervers de l'intégration des fonctions d'assu- rance et de production de soins semblent donc liés au caractère monopolistique de l'offre, alors que ses avantages pourraient être généralisés à d'autres modes d'organisation.

C'est cette hypothèse qui a été testée, notamment aux États-Unis, dans le cadre des *Managed Care Organizations* (MCO) qui combinent mise en concur- rence des assureurs et intégration des fonctions d'assurance et de production de soins. En effet, l'une des tendances de fond observable dans les pays développés aux prises avec l'objectif de maîtrise des dépenses de santé est la mise en œuvre de réformes instituant des mécanismes concurrentiels dans le secteur de la santé, non seulement sous la forme d'une concurrence fictive (ou indirecte, telle que par exemple la concurrence par comparaison), mais aussi sous la forme d'une concurrence organisée (ou directe entre les trois acteurs du système de soins : financeur, offreur, patient).

L'intégration des fonctions d'assurance et de production de soins qui est mise en œuvre au sein des MCO a pour objectif de limiter les comportements opportunistes des patients et des médecins (risque moral et demande induite). En effet, dans un système d'assurance traditionnelle, les intérêts des médecins et des patients sont souvent communs. Ces derniers, qui ne supportent pas la totalité des coûts, auront tendance à surconsommer (aléa moral) et les médecins, payés à l'acte et disposant d'un fort pouvoir discrétionnaire, voient leur revenu augmenter avec leur activité (demande induite). Les deux phénomènes conjugués font que patients et médecins ont une tendance commune à utiliser l'ensemble des biens et services médicaux pouvant améliorer l'état de santé, ignorant pour les premiers et recherchant pour les seconds les dépenses qui y sont associées. L'intégration plus ou moins forte des fonctions d'assurance et de production de soins relève donc d'une logique d'incitation à une plus grande maîtrise des dépenses, ces alliances rendant plus cohérents les objectifs des assureurs et des producteurs (Chambaretaud et Lequet-Slama, 2002).

Le développement récent des MCO répond ainsi aux dysfonctionnements d'un marché où l'information est très dissymétrique. Comme le soulignent Cutler et Zeckhauser (2000), « *les liens naissants entre assureurs et producteurs dans le domaine de la santé [ ] sont une réponse à la difficulté a priori de définir des contrats d'assurance contingents* ». Cette intégration entraîne une modifi- cation importante des modes de rémunération des professionnels de santé, les contrats entre ces derniers et les assureurs visant à transférer une partie des risques de l'assurance vers les producteurs.

Arnould *et al.* (1993) rappellent quelles sont les principales caractéristiques des MCO, à savoir, d'une part des arrangements contractuels avec des offreurs sélectionnés pour fournir un ensemble déterminé de services de soins médicaux aux membres de l'organisation de *managed care*, habituellement à des prix négociés ; d'autre part, des incitations financières significatives afin d'orienter les patients vers les offreurs et les procédures du plan ; par ailleurs, une comptabilité courante mise à disposition des offreurs, relative à leur performance clinique et financière, à travers une assurance qualité formelle et un contrôle d'utilisation (*utilization review*).

L'effet des MCO sur le comportement des professionnels de santé peut être décrit par un modèle de « prix implicites » (Keeler *et al.* 1998, Frank *et al.* 2000). Les MCO, en déterminant un budget fixe pour les différents biens et services médicaux, contraignent les médecins à limiter l'accès à ces soins aux patients qui en ont le plus besoin ou à ceux qui en retireront le plus de bénéfice. Le rationnement des soins (par rapport à une situation où les professionnels de santé sont payés à l'acte) peut-être interprété comme un « prix implicite » qui incite les médecins à prendre en compte le coût de leurs décisions thérapeutiques.

L'impact de ce mode d'organisation sur l'évolution des dépenses de santé n'est pas évident à mesurer, car si le développement des MCO s'est accompagné d'une réduction de la croissance des dépenses aux États-Unis, Bac et Cornilleau (2001) montrent que cette tendance au ralentissement se retrouvait aussi en France, en Allemagne et au Royaume-Uni durant les années 1990.

Cependant, l'existence d'une forte variabilité du taux de pénétration des MCO entre les États permet de tester l'hypothèse d'un effet des MCO sur les dépenses de santé. Cutler et Sheiner (1998) utilisent ainsi des données du ministère de la Santé sur les dépenses de santé des États et des données sur le taux de pénétration des HMO sur la période 1980-1993. Ils montrent ainsi qu'une augmentation de 10 % du nombre d'adhérents à une HMO entraîne une réduction de 5 % du taux de croissance global des dépenses.

À l'origine de cette plus grande maîtrise des dépenses par les MCO, plusieurs explications, non exclusives, ont été avancées. En premier lieu, comme nous le rappelions, ce type d'organisation repose sur la mise en place de mécanismes de paiement des offreurs de soins fortement incitatifs et de référentiels de bonne pratique fondés sur des analyses coût/efficacité. Les MCO encouragent ainsi le recours à la prévention et ont largement contribué au développement du *disease management* qui assure une meilleure coordination de l'ensemble des services de santé dans la prise en charge des patients souffrant de pathologies sévères. En deuxième lieu, leur clientèle présente une structure de risque plus favorable que celle des assurances traditionnelles. Des phénomènes de sélection des risques ont été ainsi mis en évidence par de nombreuses études sur les données américaines.