

**Microéconométrie de la santé :
remarques sur l'expérience française**

Brigitte Dormont*

N° 99-39

 Thema

Théorie Economique, Modélisation et Applications
UMR 7536 - Centre National de la Recherche Scientifique

Microéconométrie de la santé : remarques sur l'expérience française

Résumé :

Cet article analyse l'expérience française en microéconométrie de la santé à partir de quatre thèmes : le rôle de l'aléa moral dans la consommation de soins, la demande induite, la taille optimale des hôpitaux et l'effet potentiel d'une tarification par pathologie dans le secteur public hospitalier.

Références JEL : C2, I10, I11.

Health econometrics using micro data: some comments on the French experience

Abstract:

The purpose of this article is to provide a view of French experience using micro data in the field of health econometrics. Four issues are addressed: moral hazard in demand for medical care, supply-induced demand, hospital optimal size, potential outcomes of prospective payment systems in public hospitals.

JEL classification : C2, I10, I11.

Microéconométrie de la santé : remarques sur l'expérience française

Brigitte Dormont¹

Communication au congrès de l'AFSE, 23-24 septembre 1999

1 Introduction

Dans un chapitre prévu pour le futur Handbook in Health Economics, Andrew M. Jones souligne le développement, considérable depuis une dizaine d'années, des études d'économétrie de la santé. Toutes les méthodes actuellement disponibles en économétrie ont trouvé leur application dans le domaine de la santé.

Les méthodes de l'économétrie des variables qualitatives, par exemple, avec l'estimation de modèles de choix dichotomiques (décision d'aller ou non chez le médecin, de souscrire ou non une assurance complémentaire...), les modèles de choix multiples (choix d'une d'assurance dans un menu de contrats, choix du traitement à appliquer à un patient...), les modèles avec variable dépendante tronquée (les dépenses de soins). On rencontre aussi des modèles de comptage (pour étudier le nombre de consultations chez le médecin) ou de durée (pour étudier la survie après une hospitalisation pour infarctus, ou la durée écoulée avant de se mettre à fumer...). Les spéci...cations considérées conjuguent souvent un modèle dichotomique qui explicite la décision de participation (aller ou non chez le médecin ; faire partie ou non des fumeurs potentiels) et un modèle décrivant les quantités consommées (dépenses de soins ou nombre de consultations ; quantités de cigarettes fumées). On peut ainsi combiner un modèle probit et l'explication d'une variable dépendante tronquée (modèle tobit généralisé), ou un probit et un modèle de comptage.

En France, on constate aussi un essor rapide des études économétriques appliquées à la santé. Le but de cet article est d'émettre quelques remarques sur l'expérience française.

Dans le domaine de la santé, l'économétrie sur données individuelles paraît tout d'abord se caractériser par des difficultés particulières dues à l'hétérogénéité non observée. On pourra rétorquer que gérer cette hétérogénéité est le lot commun de tous les microéconomètres. Mais en santé, cette tâche devient plus ardue pour deux raisons : la plupart des modèles considérés sont non linéaires ; l'hétérogénéité non observée se conjugue souvent avec des biais de sélection. Un second trait marquant - et plus spéci...quement français - de l'exercice de l'économétrie de la santé tient au contraste existant entre les enjeux considérables des questions posées et le manque de moyens, faute de données pertinentes, d'apporter des réponses décisives.

Dans cet article, on illustre ces caractéristiques en reprenant les travaux réalisés en France sur quatre questions. Les deux premières se rapportent à la

¹Théma, Université Paris X-Nanterre, 200, avenue de la République, 92001 Nanterre Cedex.
E-mail : dormont@u-paris10.fr

médecine ambulatoire et les deux autres concernent le secteur public hospitalier. Mais avant de les aborder, on évoque dans un premier paragraphe les difficultés liées au traitement de l'hétérogénéité non observée.

2 Difficultés liées à l'hétérogénéité non observée

Considérons un modèle linéaire multiple expliquant un phénomène y par un ensemble de k régresseurs X :

$$y_i = \underset{(1;k)}{X_i} \underset{(k;1)}{b} + u_i \quad (1)$$

On veut estimer ce modèle à l'aide d'observations relatives à N individus i ($i = 1; \dots; N$): L'hétérogénéité non observée correspond à des caractéristiques, notées α_i ; des individus i qui expliquent le niveau y_i , mais ne figurent pas, faute d'être observées, dans le vecteur des variables explicatives X_i : La perturbation u_i comprend, entre autres, les caractéristiques individuelles α_i : On peut écrire : $u_i = \alpha_i + \varepsilon_i$:

Par exemple, l'état de santé d'un individu est le déterminant principal de sa demande de soins. Or les indicateurs de morbidité, quand ils existent, constituent généralement des mesures imparfaites de ses besoins. Même dans un modèle où figurent de tels indicateurs, une hétérogénéité non observée en relation avec la consommation de soins de l'individu subsiste dans la perturbation. Pareillement, toutes choses égales par ailleurs, l'activité d'un médecin dépend de son éthique ou de son style de pratique, éléments non mesurables que l'on résume par un effet spécifique α_i . En outre, il est clair que les coûts hospitaliers ne dépendent pas seulement de facteurs comme la taille des établissements, le personnel mobilisé, le type d'affections traitées, etc., mais aussi d'un élément non observable, noté α_h , caractérisant la qualité de la gestion de l'hôpital h .

Dans une spécification simple comme celle du modèle (1), une hétérogénéité non observée n'introduit pas de biais asymptotique dans l'estimation des paramètres, à condition que les effets spécifiques α_i soient non corrélés avec les variables explicatives X_i : Autrement dit, ces dernières doivent être exogènes et vérifier $E(X_i \alpha_i) = E(X_i u_i) = 0$: Si tel n'est pas le cas, il faut recourir à une méthode à variables instrumentales.

Disposer de données individuelles-temporelles permet de gérer l'hétérogénéité non observée. Le modèle s'écrit alors :

$$y_{it} = \underset{(1;k)}{X_{it}} \underset{(k;1)}{b} + \underset{u_{it}}{\alpha_i + \varepsilon_{it}} \quad (2)$$

Dans ce cas, l'hétérogénéité non observée conduit à une structure particulière de la matrice de variance-covariance des perturbations : si les variables explicatives sont exogènes l'estimateur des moindres carrés généralisés est le seul estimateur à la fois convergent et efficace.

Si par contre les effets spécifiques θ_i sont corrélés avec les X_{it} , on peut avoir recours aux méthodes à variables instrumentales mais aussi, plus directement, appliquer une transformation linéaire au modèle permettant d'éliminer θ_i , tout en conservant une spécification linéaire facilement estimable. On obtient alors des estimateurs convergents en appliquant les moindres carrés ordinaires aux écarts aux moyennes individuelles ($y_{it} - \bar{y}_i$) ou aux différences premières ($y_{it} - y_{it-1}$). Ces méthodes présentent toutefois l'inconvénient d'éliminer du modèle toutes les variables purement individuelles, dont on ne peut plus évaluer l'influence sur y : Ainsi, il n'est pas possible d'évaluer l'effet de la taille d'un hôpital sur ses coûts (rendements d'échelle) dès lors que l'on veut prendre en compte l'existence d'un effet θ_h caractéristique de la qualité de sa gestion.

Mais l'essentiel des difficultés liées à l'hétérogénéité non observée résulte de la non-linéarité des modèles, conjuguée à des problèmes de sélection. Pour illustrer notre propos, on considère le problème classique de la difficile distinction entre les phénomènes d'aléa moral et d'anti-sélection dans les comportements de demande de soins. Le modèle généralement considéré a la forme suivante (sur des données en coupe) :

$$\begin{aligned} \frac{1}{2} \quad & A_i^* = X_i \beta + u_i \\ & A_i = 1 \text{ si } A_i^* > 0; \quad A_i = 0 \text{ sinon.} \end{aligned} \quad (3)$$

$$\begin{aligned} \frac{1}{2} \quad & C_i^* = Z_i \gamma + A_i \delta + v_i \\ & C_i = 1 \text{ si } C_i^* > 0; \quad C_i = 0 \text{ sinon.} \end{aligned} \quad (4)$$

Les modèles standards considèrent l'ensemble constitué par les expressions (3) et (4). L'expression (3) figure dans le modèle quand la souscription d'une assurance ($A_i = 1$) repose sur la décision de l'individu. Cette décision dépend des valeurs prises par une variable latente A_i^* : On utilise pour décrire la consommation de soins C_i le modèle non linéaire (4) afin de tenir compte des particularités des données relatives aux dépenses de soins en matière de distribution, notamment de l'importante proportion de zéros². Le modèle (4) est une équation de participation : la probabilité de consommer des soins est fonction des valeurs prises par la variable latente C_i^* ; laquelle dépend des variables Z_i et de la souscription ou non d'une assurance par l'individu. Quand on dispose d'observations sur la valeur des dépenses de soins, on peut utiliser un modèle tobit généralisé : C_i est alors une variable quantitative tronquée en 0, et l'on admet la possibilité d'influences différentes des variables explicatives sur la décision de consommer et sur le niveau de consommation.

La démarche habituellement retenue consiste à estimer ce modèle sur une coupe d'individus, dont certains bénéficient d'une couverture assurantielle plus avantageuse que les autres (par exemple, en France, une assurance complémentaire). L'existence d'un aléa moral est testée en examinant si le paramètre δ est significativement positif.

On sait qu'une telle procédure est particulièrement vulnérable à un biais lié aux phénomènes d'anti-sélection. En effet, les individus qui décident de

²Par exemple, 41 % des individus interrogés par l'Insee lors de l'enquête Santé 1991-92 ne sont pas allés chez le médecin durant la période considérée (Breuil-Genier, 1999).

souscrire une assurance complémentaire se caractérisent peut-être par une morbidité supérieure à celle des autres. Dans ce cas, loin de mesurer l'aléa moral, les estimations mettraient en évidence une causalité inverse : on observerait seulement le fait que les plus exposés à la maladie s'assurent mieux.

Ce biais d'anti-sélection est lié à l'hétérogénéité non observée ϵ_i qui retente l'état de santé de l'individu, mais aussi son aversion au risque, ou encore le poids de la santé dans sa fonction d'utilité. Compte tenu de cette interprétation, il est vraisemblable que cet effet spécifique est partagé à la fois dans les perturbations u_i et v_i des équations (3) et (4) : $u_i = \epsilon_i + \eta_i$ et $v_i = \epsilon_i + \zeta_i$. De façon générale, les perturbations u_i et v_i ne sont donc pas indépendantes. Dans ce cas, l'estimation de (4) par un simple modèle probit conduit à des estimations non convergentes des paramètres β et γ :

Comment résoudre ce problème ? La solution la plus convaincante, mais aussi la plus coûteuse, consiste à déterminer un protocole d'expérience qui permet de s'assurer que A_i est non corrélé avec ϵ_i . C'est le principe de la Health Insurance Experiment menée par la Rand (Manning et al., 1987) : des contrats avec différents taux de couverture ont été répartis de façon aléatoire entre les ménages participant à l'expérience. L'équation (3) est alors éliminée du modèle et A_i , qui ne résulte pas d'une décision individuelle, est bien exogène dans (4).

L'économètre peut aussi profiter des résultats d'une "expérience naturelle" lorsque les contrats d'assurance d'un groupe d'individus sont modifiés de façon exogènes. Mais cette condition ne suffit pas pour évaluer l'aléa moral. Il faut aussi observer les consommations avant et après la date du changement de contrat pour deux échantillons : les individus subissant le changement et un échantillon de contrôle.

En l'absence de données permettant de s'affranchir des effets d'antisélection, ceux-ci doivent être pris en compte dans la méthode retenue pour l'estimation. On doit alors mettre en oeuvre des procédures assez complexes, pas toujours programmées dans les logiciels usuels. Il peut s'agir, dans une approche exclusivement paramétrique, de l'estimation de modèles probit bivariés. On peut aussi recourir à des méthodes semi-paramétriques comme celles développées par Angrist, Imbens et Rubin (1996), et appliquées par McClellan et Newhouse (1997) dans l'analyse coût-efficacité des nouvelles technologies dans le traitement hospitalier des pathologies cardiaques. A moins de disposer de données particulières, on constate ainsi que l'hétérogénéité non observée entraîne rapidement de réelles difficultés méthodologiques pour estimer les modèles non linéaires couramment utilisés en économie de la santé³.

Revenons maintenant au cas de données individuelles-temporelles ($u_{it} = \epsilon_i + \eta_{it}$), en supposant tout d'abord que l'hétérogénéité non observée n'est pas corrélée avec les variables explicatives. Si le modèle est non linéaire, on ne peut pas le transformer simplement pour éliminer l'effet ϵ_i . L'écriture de la vraisemblance doit alors prendre en compte la double dimension des données,

³ Précisons que si le modèle décrivant la consommation était linéaire, on pourrait facilement obtenir une estimation convergente de γ à l'aide d'une méthode en deux étapes.

et notamment intégrer la distribution de θ_i ; pour conduire à une estimation convergente. En...n, si les effets spécifiques θ_i sont corrélés avec les variables explicatives, le problème se corse : toutes les méthodes de type maximum (ou pseudo-maximum) de vraisemblance sont non convergentes, et il convient de faire appel à d'autres méthodes, de type moments généralisés.

3 Quatre questions de microéconométrie de la santé

3.1 Quel est le rôle de l'assurance dans la demande de soins ?

Sur données macro-économiques françaises, L'Horty et al. (1997) évaluent à 6 % la contribution de l'évolution de la couverture sociale à la croissance des dépenses de santé entre 1970 et 1995. Non négligeable, cette contribution est toutefois d'un ordre de grandeur inférieur à celui de l'effet-revenu (41 %) ou à l'influence du progrès technique médical (26 %). En France, le taux de couverture dépend principalement du fait d'avoir ou non souscrit une assurance complémentaire. 84,3 % des Français étaient couverts par une assurance complémentaire maladie en 1996, mais seulement 60,6 % des chômeurs.

L'application du concept d'aléa moral en assurance maladie signifie qu'être assuré favorise les comportements pathogènes et conduit à une sur-consommation une fois la maladie déclarée. Cette conception conduit à préconiser des mécanismes de responsabilisation de type ticket modérateur incompressible (qui pourrait varier avec le revenu (Henriet et Rochet, 1999)). Un tel mécanisme n'est pas à l'ordre du jour en France : les assurances complémentaires couvrent le ticket modérateur et la loi sur la couverture maladie universelle (CMU) va étendre ce système à presque toute la population.

S'interroger sur l'étendue des phénomènes d'aléa moral n'est toutefois pas sans intérêt. Tout d'abord, il faut pouvoir quantifier les effets de la CMU sur l'évolution ultérieure des dépenses. Ensuite, il est important, au moment où l'on cherche à responsabiliser les médecins, d'évaluer la part du risque moral situé du côté de la demande.

On sait en effet que le niveau des dépenses dérive des interactions médecin-patient, le praticien jouant le double rôle d'opérateur de soins et d'agent (demandeur) pour le patient. Ellis et McGuire (1990) distinguent l'aléa moral du côté de l'opérateur et du côté de la demande. Ils montrent qu'un système de santé qui maximise le bien-être des consommateurs doit privilégier les instruments de politique du côté de l'opérateur, avec une bonne couverture assurantielle pour les patients, et un mode de tarification largement prospectif pour les fournisseurs de soins. Newhouse (1996) conteste ce point de vue en défendant le principe d'une responsabilisation bilatérale. Selon lui, l'aléa moral ne joue pas au même niveau selon qu'il intervient du côté de l'opérateur ou du côté de la demande : une bonne couverture par l'assurance inciterait le consommateur à consulter, mais influencerait peu le niveau des dépenses, principalement déterminées par le médecin.

Ce débat débouche ainsi sur une question essentiellement empirique.

La référence en matière d'évaluation de l'influence de l'aléa moral dans la consommation de soins est constituée par les études réalisées à partir des résultats de la Health Insurance Experiment (Manning et al. 1987). Le protocole de cette expérience a permis de mesurer les effets de l'aléa moral sans que les estimations soient biaisées par les phénomènes d'anti-sélection. A la différence des expériences "naturelles", l'étude de la Rand a aussi permis de tester les effets de taux de couverture très variés (allant de 5 à 100 %). Les résultats montrent (i) que la consommation de soins croît avec le taux de couverture, (ii) que celui-ci affecte plus le nombre de contacts médecin-patient que la dépense de soins au cours de chaque rencontre.

En France, l'étude pionnière en ce domaine est celle de Caussat et Glaude (1993). Dans notre pays, le seul effet mesurable est celui de la détention d'une assurance complémentaire. A partir des données de l'enquête Santé réalisée en 1980 par l'Insee, les auteurs estiment un modèle constitué par l'expression (3) et un tobit généralisé : l'intérêt d'une telle spécification réside dans le fait qu'elle admet des paramètres différents en matière de décision de recours aux soins et de valeur de la dépense de soins, comme le suggèrent les résultats de la Rand.

Caussat et Glaude mettent en évidence que la probabilité d'adhérer à une mutuelle est bien liée à la morbidité, mesurée par l'indicateur d'invalidité dont ils disposent. Ils estiment ensuite que bénéficier d'une couverture complémentaire augmente la probabilité de consommer de 12 % et le montant moyen des dépenses de 16 %. Compte tenu de ces deux éléments, la consommation moyenne des mutualistes serait supérieure de 30 % à celle des non-mutualistes.

Sur le plan méthodologique, les auteurs mettent en oeuvre un test d'anti-sélection consistant à examiner si les perturbations u_i et v_i sont indépendantes⁴, ce qui permet de juger si la détention d'une complémentaire A_i est bien exogène dans l'équation (4). L'issue de ce test les conduit à ne pas rejeter l'hypothèse que $\text{Cov}(u_i; v_i) = 0$: Leur évaluation de l'aléa moral est donc en principe convergente. Curieusement toutefois, Caussat et Glaude mettent en doute leurs résultats, déclarant qu'ils sont insuffisants pour servir de base à des décisions relatives aux remboursements des soins. Ils invoquent notamment des "variables individuelles inobservables, comme l'aversion pour le risque" pour expliquer leurs résultats relatifs à la sur-consommation des mutualistes. Pourtant le test qu'ils ont mené leur garantit en théorie que l'effet estimé de l'aléa moral est purgé des problèmes liés à l'hétérogénéité individuelle non observée.

Les résultats de l'enquête Santé 1991-92 ont été utilisés par Genier et al. (1997), puis Genier (1998), qui étudie le rôle de l'assurance dans la consommation de soins à l'aide d'un modèle proche de celui de Caussat et Glaude. La mise en oeuvre d'un test d'anti-sélection la conduit à retenir l'hypothèse que les

⁴Pour simplifier les notations, on raisonne ici dans l'hypothèse d'un modèle constitué par les équations (3) et (4), alors que l'article dont il est question utilise l'équation (3) et un modèle tobit généralisé.

perturbations u_i et v_i sont corrélées. Il y a donc des phénomènes d'antisélection, que Genier rattache plutôt à des propensions à consommer différentes, à morbidité donnée, qu'à des différences d'état de santé entre les individus. Compte tenu de l'anti-sélection, l'estimation de l'effet de l'aléa moral sur la consommation de soins est non convergente. L'influence la plus marquée de l'assurance complémentaire apparaît sur la probabilité de connaître un épisode de soins (+16 %) et sur le nombre d'épisodes. En revanche, l'effet de l'assurance sur la longueur des épisodes et les dépenses de soins est moins évident et pas toujours significatif.

Breuil-Genier (1999) construit des épisodes de soins, définis comme l'ensemble des soins relatifs à une maladie et à un patient donné. Le nombre d'épisodes connus par un patient est déterminé par le nombre de maladies qui l'ont affecté. Seul le nombre de recours au médecin par épisode (la longueur de l'épisode) est susceptible d'être influencé par des comportements de nomadisme médical ou de demande induite. Il apparaît que l'assurance est un élément déterminant du nombre d'épisodes, mais pas de leur longueur ou composition.

Ces études fournissent des résultats très intéressants. Le contraste entre l'influence de l'assurance sur le premier recours (à l'initiative du patient) et l'absence d'effet marqué sur le retour chez le médecin et les dépenses de soins, implique une forte présomption d'aléa moral. Malheureusement, les estimations sont en principe biaisées, dans un sens qu'il reste à déterminer, par les phénomènes d'anti-sélection.

Grâce à des données particulières, Chiappori, Durand, Geoard (1998) peuvent profiter d'une "expérience naturelle" et s'approprier des effets d'anti-sélection. Il dispose d'un échantillon d'environ 4 500 salariés observés sur les années 1993 et 1994 et couverts par une assurance complémentaire. Le contrat de base offre une couverture quasi-totale des dépenses médicales mais, en 1994, 3 689 individus subissent l'introduction d'un co-paiement de 10 %. L'anti-sélection est très limitée car les contrats d'assurance ne sont pas choisis par les individus mais souscrits par leurs entreprises. De plus, la double dimension des données permet de prendre en compte l'hétérogénéité non observée dans les procédures d'estimation. Enfin, la présence d'un groupe de contrôle garantit que les changements de comportement testés ne découlent pas d'un autre événement intervenu de manière concomitante aux changements de contrats.

Les estimations ne révèlent pas de changement significatif, sauf pour les visites à domicile, dont la proportion diminue dans le groupe ayant subi l'introduction du co-paiement. En revanche, des phénomènes d'aléa moral ne sont pas mis en évidence pour les consultations. Chiappori, Durand et Geoard interprètent ce résultat en soulignant l'importance des coûts non monétaires (transport, attente) pour les consultations chez le généraliste, dont les honoraires sont largement pris en charge par l'assurance (sécurité sociale+complémentaire) : l'introduction d'un co-paiement de 10 % correspond de fait à un changement négligeable dans le coût total.

Ces résultats sont obtenus en l'absence de biais d'anti-sélection - là réside tout l'intérêt de cette expérience naturelle. Doit-on considérer qu'ils remettent

en cause ceux de Genier, qui met en évidence un aléa moral non négligeable pour les consultations ? Pas vraiment, puisque le co-paiement est d'un ordre de grandeur (10 %) très inférieur à la différence de couverture existant entre les mutualisés et les non-mutualisés (environ 30 %). En outre, la population sur laquelle a porté l'expérience est particulière : des employés du secteur de la Banque et de l'Assurance, non représentatifs de la population française (notamment en ce qui concerne les coûts associés au temps d'attente chez le médecin). On bute sur la limite inhérente à toute expérience naturelle : le protocole n'est pas défini par l'économiste qui veut exploiter les résultats.

Au total, on constate qu'on ne dispose pas en France d'une évaluation récente et fiable de l'étendue des phénomènes d'aléa moral permettant, par exemple, de quantifier les effets potentiels de la CMU sur les dépenses de santé. Cette situation est certes due à des difficultés méthodologiques, mais aussi au manque de données pertinentes. Pour gérer les problèmes liés à l'anti-sélection, il faut recourir à des méthodes assez complexes. Si l'on doute des résultats ainsi obtenus, force est de recourir à des données expérimentales. De fait, il ne serait pas aberrant de mettre en oeuvre des protocoles d'expériences relativement coûteux pour calibrer des réformes dont les conséquences budgétaires sont non négligeables.

3.2 Quelle est l'importance de la demande induite dans les comportements d'offre de soins ?

Les mécanismes d'induction de la demande sont toujours évoqués quand il s'agit d'analyser la progression des dépenses de santé. Grâce aux asymétries d'informations existant entre le médecin et son patient et à la solvabilisation de la demande garantie par l'assurance maladie, les acteurs disposent d'une certaine latitude pour manipuler la demande. Ce comportement est probable en France, où les médecins sont rémunérés à l'acte. Les politiques envisagées pour maîtriser la progression des dépenses de santé reposent d'ailleurs sur l'hypothèse de demande induite. Ainsi en est-il du *numerus clausus* dans les études médicales, de l'incitation des médecins au départ à la retraite et des mécanismes de reversements-sanctions prévus à l'encontre des praticiens.

Les phénomènes de demande induite sont mis en évidence de manière convaincante par les expériences de contrôle des tarifs aux Etats-Unis et au Québec : les médecins ont réagi au gel des honoraires par une augmentation du volume de leurs prestations (Rochaix (1995)). Mais en l'absence d'un choc bien repérable de politique économique, il est difficile d'évaluer l'importance du phénomène. Très abondante, la littérature⁵ consacrée à ce thème conduit à des résultats multiples et controversés.

De fait, une augmentation de la densité peut très bien entraîner une élévation de la quantité de soins consommés en l'absence de comportements d'induction : sur un marché concurrentiel standard, un choc positif sur la courbe d'offre conduit à un nouvel équilibre caractérisé par des prix plus bas et des échanges

⁵Pour une synthèse de ces études, le lecteur intéressé peut se reporter à Rochaix (1995) et Rochaix et Jacobzone (1997).

plus importants. Si les prix sont ...xés (comme dans le secteur 1 de la médecine ambulatoire en France), un choc positif sur l'offre peut aussi conduire à une augmentation de la consommation, dans le cas où préexistait une situation de rationnement de la demande caractérisée par des ...les d'attente et des temps de transport importants.

Une des difficultés majeures des tests de demande induite réside dans le fait que l'on cherche à détecter une manipulation de la demande par les opérateurs, en l'absence d'information sur la forme de la demande de soins. Ces tests sont ainsi exposés à des problèmes d'identification proches de ceux rencontrés pour évaluer l'aléa moral dans la consommation de soins. Il faudrait pouvoir étudier l'effet d'un choc sur l'offre, à comportement de demande constant. Or les perturbations des équations estimées comportent une hétérogénéité non observée en rapport avec la demande adressée au médecin : coûts d'accès, demande de qualité, information sur les compétences, rationnement, etc. Ces éléments sont vraisemblablement corrélés avec la densité médicale, ce qui conduit à des biais dans l'évaluation de la demande induite.

En France, les premières estimations ont été effectuées par Béjean et Gadreau (1992), qui ont testé l'existence de demande induite sur une coupe instantanée de départements. Elles obtiennent une élasticité élevée de la consommation de soins par rapport à la densité médicale, de l'ordre de 0,8. Mais ces estimations ne permettent pas de distinguer, dans l'effet de la densité, entre ce qui est dû aux phénomènes de demande induite et ce qui peut être attribué aux déterminants de la demande de soins non pris en compte par les variables explicatives. Par exemple, un département peut être caractérisé par un niveau de morbidité particulier qui explique à la fois une densité et une consommation médicale plus élevées. Si la morbidité est mal prise en compte par les indicateurs de structure démographique ...gurant parmi les régresseurs, les estimations sont biaisées. Plus récemment, Béjean (1997) améliore l'approche empirique en construisant un modèle à équations simultanées qui tient compte de l'endogénéité des densités médicales départementales, lesquelles dépendent des choix de localisation des médecins. Les estimations obtenues mettent encore en évidence un effet d'induction non négligeable.

Les études microéconométriques de la demande de soins permettent aussi d'évaluer les phénomènes de demande induite. D'après Genier et al. (1997), ces phénomènes sont limités : la densité médicale ne joue un rôle significatif sur la demande de soins que pour les spécialistes. Les résultats obtenus par Breuil-Genier (1999) confirment ce point de vue : les deux-tiers des épisodes de soins ne comportent qu'un unique recours au médecin et les densités médicales ne jouent pas significativement sur la probabilité d'un second recours. Il convient de souligner toutefois que la longueur des épisodes étudiés est tronquée par la durée de l'enquête Santé d'une part (trois mois) et par la nécessité, d'autre part, de se restreindre aux épisodes de soins non censurés à gauche. De plus, les comportements de demande induite peuvent passer par une manipulation du contenu en actes de la rencontre médecin-patient, et non seulement par une incitation à un deuxième recours.

Grâce à l'utilisation d'un panel représentatif des médecins libéraux français, Delattre et Dormont (1999) peuvent mettre en évidence des comportements d'induction dans le secteur 1. La double dimension des données permet à la fois de prendre en compte l'hétérogénéité non observée et de mettre en oeuvre une méthode d'estimation convergente malgré la non-exogénéité de la densité. L'hétérogénéité non observée découle du fait que l'activité du médecin dépend non seulement des variables explicatives du modèle, mais aussi d'un effet individuel θ_i , caractéristique de son éthique, de son style de pratique et des caractéristiques permanentes de sa clientèle : âge, morbidité, aversion au risque, degré de couverture...

Chaque année, la quasi-totalité des médecins sont exposés à une croissance de la densité médicale dans leur département. Les estimations révèlent que les omnipraticiens et les spécialistes du secteur 1 subissent de ce fait un rationnement de leur œuvre : le nombre de leurs rencontres avec leurs patients diminue lorsque la densité augmente. Toutefois, ce rationnement reste très partiel. De plus, les médecins compensent les rationnements subis sur le nombre de consultations par une augmentation du volume de soins fournis au cours de chaque rencontre. Ces différents constats mettent en évidence l'existence de comportements de demande induite indéniables dans le secteur 1.

Dans le secteur 2, les résultats sont compatibles avec une absence de comportement d'induction de la part des médecins. Les omnipraticiens et les spécialistes réagissent à un choc négatif sur la demande de manière conforme aux prédictions théoriques d'un modèle de concurrence monopolistique : par une diminution de leurs tarifs et une augmentation de leur activité, laquelle s'accompagne d'une élévation du contenu en actes de la rencontre médecin-patient.

Au total, les estimations permettent d'évaluer que 1 point d'augmentation de la densité devrait se traduire, au niveau macroéconomique, par une augmentation de l'activité des médecins du secteur 1 de 0,5 point environ. Ces résultats sont obtenus grâce à la richesse de l'information ouverte par un panel de médecins, et notamment à la double dimension des données, laquelle permet de tenir compte d'un effet spécifique lié aux caractéristiques du médecin et de sa clientèle.

3.3 Quelle est la taille optimale des établissements hospitaliers ?

Pour les hôpitaux, la question de la taille optimale concerne l'efficacité productive mais aussi la qualité des soins. Certains actes doivent être pratiqués souvent pour que la sécurité des soins soit garantie⁶. Les ordonnances d'avril 1996 sur la réforme de l'hospitalisation ont ainsi consacré l'idée qu'une accréditation pouvait être retirée en deçà d'un niveau plancher d'activité (Delahaye-Guillocheau et Mettendorff (1997)).

⁶La société française de cardiologie, par exemple, définit un seuil minimum de 200 angioplasties par an.

Concernant l'efficacité productive, les premiers travaux ont étudié le coût moyen d'un séjour en rapport avec la taille (mesurée par le nombre de lits) d'un établissement ou d'un service hospitalier. La référence est l'étude de Feldstein (1967), réalisée sur des hôpitaux publics anglais : ses estimations mettent en évidence une fonction de coût en forme de U, avec un minimum pour une capacité de 903 lits. Plus récemment, différents travaux ont adopté une modélisation en terme de frontière stochastique, où l'on suppose que l'inefficacité se traduit par un aléa de distribution asymétrique, positif ou nul, ...gurant dans la perturbation de la fonction de coût. En appliquant cette méthode à des données relatives à 1 600 hôpitaux américains du programme Medicare, Zuckerman et al. (1994) trouvent un taux moyen d'inefficacité de 13,6 %. Disposer de données de panel permet d'améliorer la méthode : en supposant que l'inefficacité est constante dans le temps, on l'identifie à un effet spécifique à l'hôpital θ_h . Ceci permet d'éviter d'assimiler toute asymétrie dans la distribution de la perturbation à de l'inefficacité. Si celle-ci est corrélée avec les variables explicatives du modèle (ce qui est vraisemblable), on doit adopter une approche en termes d'effets fixes. Dans ce cas, le modèle comporte des constantes hospitalières θ_h : on ne peut plus identifier l'influence de régresseurs invariants dans le temps (comme la taille). Sur un panel de 49 hôpitaux publics espagnols, Wagstaff (1989) montre qu'un tiers de la variation des coûts peut être attribué à de l'inefficacité.

Pour la France, Leleu et Dervaux (1997) utilisent des techniques non paramétriques d'enveloppement des données afin de calculer des scores d'efficacité pour 137 centres hospitaliers publics. En mesurant l'activité par le nombre d'admissions, ils obtiennent un taux d'efficacité moyen de 0,9. Ils tentent ensuite d'expliquer les scores par les caractéristiques des hôpitaux : les résultats obtenus ne permettent pas de conclure à l'existence de rendements d'échelle. Dervaux et al. (1994) estiment une fonction de coût sur les mêmes données en coupe : les résultats révèlent des économies d'envergure, mais pas de rendements d'échelle. De Pourville et al. (1997) estiment des fonctions de coût sur 217 923 séjours effectués en 1993 dans 22 hôpitaux publics de la base de coût PMSI. Ils trouvent que les rendements d'échelle sont croissants, puis décroissants pour le coût variable médical, avec un minimum du coût atteint pour 1 000 ou 500 lits, selon que l'on inclut ou non le plus gros hôpital de la base dans les données. Ces valeurs sont d'un ordre de grandeur comparable aux résultats de Feldstein et des autres travaux anglo-saxons sur le sujet.

Ces études fournissent des éléments de chifrage indispensables, mais leurs résultats doivent être utilisés avec précaution. Tout d'abord, il s'agit d'établissements publics dont rien ne permet de penser qu'ils se situent sur leur frontière de production⁷ : ils subissent un rationnement lié au système du budget global mais ne sont exposés à aucune contrainte de marché. De ce fait, leurs coûts sont influencés par une hétérogénéité non observée θ_h liée à leur niveau d'inefficacité. Sur des données en coupe l'effet de la taille ne peut pas être identifié indépendamment de θ_h : on peut conclure à des rendements d'échelle si l'hôpital le plus grand est aussi le plus efficace pour des raisons qui ne tiennent pas à sa

⁷ C'est bien le sens des scores calculés par Leleu et Dervaux .

taille, mais à la qualité de sa gestion. Comme la taille est observée seulement au niveau de l'établissement, de Pouvoirville et al. (1997) ne peuvent pas exploiter la dimension individuelle de leurs données et sont aussi confrontés à ce problème. Il devrait cependant être facilement résolu : en suivant les hôpitaux dans le temps, ce qui est possible avec les bases PMSI maintenant disponibles.

Une autre limite de ces études tient à la mesure de la production hospitalière. Dans l'idéal, il faudrait pouvoir évaluer l'effet du passage à l'hôpital sur la santé des patients. En pratique, on mesure généralement la production d'un établissement par le nombre de séjours, associé à un indicateur synthétique de la lourdeur des pathologies traitées, le casemix. Pour chaque séjour, les données du PMSI permettent de connaître les diagnostics et les actes (lesquels sont plutôt des inputs). Mais l'information sur le succès des soins dispensés est très médiocre car on observe des séjours et non des individus : on connaît seulement l'occurrence d'un décès au cours du séjour. La qualité du service rendu en termes d'accueil et de prise en charge de la douleur n'est pas observée, ni les réhospitalisations précoces, ni les infections nosocomiales, ni la qualité de vie du malade après le séjour.

Certes, une information aussi détaillée est rarement disponible dans les pays étrangers. Mais il est quand même possible, par exemple pour les hôpitaux américains du programme Medicare, d'accéder à un suivi des patients, de connaître l'occurrence d'une ou plusieurs réhospitalisations et de mesurer des taux de décès pour des horizons variés, allant de une semaine à deux ans. Une mesure correcte de la qualité et de l'efficacité des soins présente un enjeu considérable pour la régulation du secteur public hospitalier. En France, c'est au nom de la confidentialité des informations individuelles que l'on justifie l'anonymisation des données qui rend impossible tout suivi des patients. De ce fait, la tutelle en est réduite à demander une baisse des coûts, sans contrôler la qualité et l'efficacité médicale des soins. Cette situation valide la critique en termes d'"approche purement comptable de la santé" adressée systématiquement par les médecins aux économistes.

3.4 Quelles seraient les conséquences d'une tarification par pathologie ?

En France, le secteur hospitalier public est financé depuis 1983 sur le principe du budget global. Parallèlement, on a instauré un système d'information (PMSI) permettant de classer les séjours hospitaliers en pathologies clairement repérables (groupes homogènes de malades, GHM). Après une phase d'installation particulièrement longue, ce dispositif statistique est maintenant opérationnel pour évaluer (dans une certaine mesure) la production hospitalière et comparer les coûts des établissements. Pour éviter la réduction des services rendus à laquelle peut conduire le budget global, on s'oriente actuellement vers une utilisation du PMSI pour déterminer les budgets hospitaliers sur la base d'un paiement forfaitaire par GHM.

La tarification par pathologie trouve sa justification dans la théorie de la concurrence par comparaison (Shleifer (1985), Mougeot et Naegelen (1997)). Le

modèle de base suppose que les techniques hospitalières sont homogènes et la qualité des soins élevée : les disparités de coûts sont exclusivement dues à l'aléa moral, c'est-à-dire à l'effort (non observé par la tutelle) fourni par l'hôpital pour réduire son coût. La tutelle obtient une minimisation des coûts de traitement de chaque hôpital en déterminant, pour chaque pathologie, un tarif fixe sur la base des performances obtenues par les autres établissements.

A partir de données sur des séjours effectués pour infarctus du myocarde aigu dans des hôpitaux publics de la base de coût PMSI, Dormont et Milcent (1998, 1999) ont tenté d'évaluer l'effet potentiel d'un tel mode de tarification. Différents traitements de l'infarctus du myocarde aigu peuvent être envisagés : la thrombolyse, le cathétérisme, l'angioplastie et le pontage. Parmi ces traitements, seul le pontage joue un rôle dans le classement des séjours en GHM, alors que le cathétérisme et l'angioplastie sont considérés comme des technologies performantes pour la qualité de vie du malade.

La procédure utilisée vise à mesurer la part de la variance des coûts des séjours due à des écarts permanents de coûts moyens entre les hôpitaux. Ceci revient à calculer la part de la variance expliquée par des effets fixes α_h , spécifiques aux hôpitaux, dans des régressions où figurent déjà les caractéristiques individuelles des patients. Si les hôpitaux ont des caractéristiques techniques identiques, on mesure ainsi l'importance du risque moral, tel qu'il est formalisé dans la théorie, et le gain en homogénéisation des coûts qu'on pourrait attendre de la mise en place d'une tarification par pathologie. L'application empirique montre que l'on pourrait obtenir une réduction de la variance de 7 % pour le coût total et de 20 % pour le coût médical.

Ces évaluations sont fortement révisées à la baisse lorsque l'on tient compte d'actes comme le cathétérisme ou l'angioplastie dans l'explication des coûts. La part de la variance expliquée par l'hétérogénéité α_h est alors seulement de 4 ou 8 %. Ceci provient du fait que certains hôpitaux pratiquent plus volontiers que d'autres des cathétérismes ou des angioplasties. Basée sur les GHM actuels, une tarification par pathologie pourrait pénaliser les hôpitaux pratiquant de tels actes ou les inciter à effectuer une sélection des patients. Pour éviter ces inconvénients, on peut envisager un système de tarification mixte, combinant un tarif forfaitaire et un remboursement partiel du coût.

Un travail de simulation permet de calculer les économies qui pourraient être obtenues en appliquant la tarification par pathologie. De façon à limiter la sélection, on définit grâce aux estimations des paiements qui tiennent compte des caractéristiques individuelles des patients (âge, sexe, etc.). En calculant les tarifs sur la base de l'hôpital le plus efficace, on trouve, pour le GHM 178⁸ par exemple, une économie budgétaire de 35 % ou 37 %, selon que l'on inclue ou non les cathétérismes et angioplasties dans la définition du paiement. Ainsi, on ne perd que 2 points d'économie en incorporant les actes techniques dans les tarifs afin de limiter la sélection. Notons que l'évaluation est réalisée à comportements constants. Le jeu constitué par la tarification par pathologie devrait conduire tous les hôpitaux, y compris le plus efficace d'entre eux, à fournir l'effort optimal

⁸ Infarctus du myocarde avec complications.

: le gain estimé, déjà considérable, doit donc être compris comme une borne inférieure.

Cette étude souffre de la même limite que celle mentionnée supra : la mesure de la production hospitalière ne tient pas compte de la qualité et de l'efficacité des soins. En alignant les tarifs sur l'établissement estimé comme le plus efficace, on se réfère peut-être à un hôpital dont les performances en matière de survie des malades sont très médiocres. Il est donc impératif de collecter une information suivie sur les patients si l'on veut mettre en oeuvre une politique pertinente de maîtrise des coûts. Une information qualitative sur les performances des hôpitaux, dont disposent vraisemblablement les autorités de tutelle, est insatisfaisante : elle peut être affectée par des biais de sélection très importants que seule une approche statistique peut traiter convenablement.

4 Références

Angrist, Imbens et Rubin (1996) Identification of causal effects using instrumental variables, *Journal of the American Statistical Association* 91, p. 444-454.

Béjean S. et Gadreau M. (1992) Asymétrie d'information et régulation en médecine ambulatoire. *Revue d'Economie Politique*, 102, p. 208-227

Béjean S. (1997) L'induction de la demande par l'offre en médecine ambulatoire : quelques évidences empiriques issues du contexte français, *Cahiers de sociologie et de démographie médicale*, XXXVII^{ème} année, n°3-4, p. 311-339.

Blanpain N., Pan Ké Shon J.-L. (1997) L'assurance complémentaire maladie : une diffusion encore inégale. *Insee Première* n°523

Breuil-Genier P. (1999) Episodes et modalités de soins : une approche microéconométrique à partir de l'enquête Santé 1991-92, miméo, Direction de la Sécurité Sociale.

Caussat L. et Glaude M. (1993) Dépenses médicales et couverture sociale. *Economie et Statistique* n°265; P. 31-43.

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

Delahaye-Guillocheau et Mettendorff (1997) La régulation de l'hospitalisation publique et privée : vers un système centré sur la réponse aux besoins de la population. *Droit social* N°9 /10, p. 859-864.

Delattre E. et Dormont B. (1999) Induction de la demande de soins par les médecins libéraux français. Etude microéconométrique sur données de panel. Document de travail Théma, n°99-21, *Economie et Prévision* (à paraître).

de Pourville G., Tibi-Levy Y., Spira R. et Moisson J.-C. (1997) : Les économies d'échelle dans le secteur public hospitalier français, in *Economie de la Santé : trajectoire du futur*, Jacobzone (ed.), *Insee Méthodes*, n° 64-65.

Dervaux B., Jacobzone S. et Leleu H. (1994) : "Productive and Economic Efficiency in French Hospitals", miméo, Laboratoire de Recherches Economiques et Sociales.

- Dormont B. et Milcent C. (1998) " Variabilité des coûts hospitaliers et tarification par pathologie ", rapport pour le Service des Statistiques, des Etudes et des systèmes d'Information (SESI), Ministère de l'Emploi et de la Solidarité.
- Dormont B. et Milcent C. (1999) " Coûts hospitaliers et tarification par pathologie ", miméo, Théma.
- Ellis R. P. et McGuire T.G. (1990) Optimal payment systems for health services. *Journal of Health Economics*, 9, 375-396.
- Feldstein M. S. (1967) Economic analysis for health service efficiency: econometric studies of the British National Health Service, North Holland.
- Genier P., Rupprecht F., Harnois J., Khamlich M., Tomasini M., Wilthien F. (1997) Analyse empirique de la consommation de soins de ville au niveau micro-économique. *Cahiers de sociologie et de démographie médicale*, XXXVII^{ème} année, n°3-4, p. 277-310.
- Genier P. (1998) Assurance et recours aux soins : une analyse microéconométrique à partir de l'enquête Santé 1991-1992 de l'Insee. *Revue Economique*, Vol. 49, n°3, p. 809-819.
- Henriet D. et Rochet J.-C. (1999) Régulation et intervention publique dans les systèmes de santé, complément au rapport Mougeot, Conseil d'Analyse Economique, La Documentation Française.
- Leleu et Dervaux (1997) Comparaison des différentes mesures d'efficacité technique : une application aux centres hospitaliers français. *Economie et Prévision*, n° 129-130, p. 101-119.
- L'Horty Y., Quinet A. et Rupprecht F. (1997) Expliquer la croissance des dépenses de santé : le rôle du niveau de vie et du progrès technique. *Economie et Prévision*, n°129-130; p. 257-268.
- Jones A. M. (1998) Health Econometrics in Handbook in Health Economics, J.P. Newhouse et A.J. Culyer eds (à paraître).
- Manning, W.G., Newhouse J.P., Duan N., Keeler E.B., Leibowitz A. and Marquis S. (1987) Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment, *American Economic Review* 77, p. 251-277
- McClellan M. et Newhouse J. P (1997) : The marginal cost-effectiveness of medical technology: A panel instrumental-variable approach, *Journal of Econometrics* 77, p. 39-64.
- Mougeot M. et Naegelen F. (1997) " La réglementation hospitalière : tarification par pathologie ou achat de soins ? ", *Economie et prévision*, n°129-130, p. 207-219.
- Newhouse J. P. (1996) Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection. *Journal of Economic Literature*, XXXIV, p.1236-1263.
- Rochaix L. (1995) La demande induite : pourra-t-on un jour en mesurer l'envergure ? Commission des comptes et des budgets économiques de la Nation. Ministère de l'Economie et des Finances.
- Rochaix L. et Jacobzone S. (1997) L'hypothèse de la demande induite : un bilan économique. *Economie et Prévision*, n°129-130, p. 25-36.
- Shleifer A. (1985) A theory of yardstick competition, *Rand Journal of Economics*, vol. 16, p. 319-327.

Wagsta^a A. (1989) Estimating efficiency in the hospital sector: a comparison of three statistical cost frontiers, *Applied Economics*, 21, p. 659-672.

Zuckerman S. et al.(1994) Measuring hospital efficiency with frontier cost functions, *Journal of Health Economics*, 13, p. 255-280.