DEMANDE MÉDICALE INDUITE PAR L'OFFRE : CHIMÈRE OU RÉALITÉ (*)

par

Peter ZWEIFEL (**)

RÉSUMÉ. — Dans cet article, nous avons exposé les caractéristiques principales d'un modèle économique du comportement médical. Le médecin y dispose de trois instruments pour atteindre ses objectifs : Un seuil critique d'intensité symptomatique (c), un taux de salaire implicite (q) et le temps moyen (t) qu'il dépense pour un seul cas. Les frais décomptés par cas doivent donc correspondre au produit ($q.t$), indépendamment des prix fixés pour les actes médicaux par les caisses-maladie. Le modèle théorique prédit une réduction de l'élément (q) si le taux de participation aux frais (r) est augmenté pour les assurés. Mais il prédit aussi que le médecin passera plus de temps avec un patient, en moyenne. Par conséquent, la réaction du produit observé ($q.t$) n'est pas déterminée. Dans l'estimation, ces deux éléments peuvent être identifiés dans les décomptes d'une grande caisse-maladie suisse.

Une augmentation du taux de participation (r) de 10 à 20 % induirait les médecins à baisser leur taux de salaire implicite de 18 %. En revanche, leurs services changeraient de qualité, devenant plus intensifs quant au temps dépensé. Cet effet l'emporterait parce qu'il serait de 60 %. Bien que le résultat final soit altéré par des erreurs d'observation, on verrait les frais par cas, équivalent à ($q.t$), augmenter plutôt que baisser. Un tel phénomène ferait certainement penser à une « demande induite par le médecin ». Mais les préférences des patients n'ont pas changé, selon hypothèse préalable. C'est plutôt la manière dont les caisses-maladie rémunèrent les médecins qui explique les phénomènes observés.

ABSTRACT. — SUPPLIER INDUCED DEMAND IN A MODEL OF PHYSICIAN BEHAVIOUR. *This paper is devoted to the main characteristics of an economic model of physicians' behaviour. A physician has three sets of tools at his disposal: a critical level of symptoms, (c), an implicit wage rate (q) and an average time input per case (t). The billings per case should correspond to ($q.t$), regardless of the fees set by sickness funds. The model predicts a reduction of the q component for an increase of the rate of the share paid (r) by the patients, it also predicts the physician will spend more time with his patient, on the average. Therefore the reaction of the observed product ($q.t$) is not determined. These components are identified in the accounts of a swiss sickness fund. An increase of r from 10 to 20 % would lead physicians to lower their implicit rate of fees by 18 %. But their services would change being more intensive with regard to the time spent. This effect would be 60 % of the whole. Though the final result is altered by errors of observation, the cost per case equivalent to ($q.t$) would increase rather than the reverse — which would create the impression of a "supplier induced demand". However the patients' preferences have not changed as assumed at the start. It is the way sickness funds pay physician's fees that explain the facts.*

(*) Version révisée d'une conférence faite lors du VII[e] Colloque International de l'Association d'Économétrie Appliquée à Poitiers, 6-8 février 1980. L'auteur tient à remercier F. CARLEVARO, A. GRIFFITHS, M. GROSSMAN, J. P. NEWHOUSE, J. NIEHANS, J. RAMSEY, A. de SAUSSURE, H. SCHMID, L. D. TAYLOR et P. TSCHOPP de leurs critiques et encouragements.

(**) Institut für Empirische Wirtschaftsforschung der Universität Zürich, Forchstrasse 145, 8032 Zürich. Suisse.

SOMMAIRE

Introduction. 40

1. Les éléments d'un modèle économique du comportement médical. . . . 41

2. Les conditions nécessaires pour un optimum. 44

3. L'analyse comparative statique. 48

4. Les données et le modèle statistique. 51

5. Estimation du modèle. 55

Conclusion. 59

Annexe. 60

Bibliographie. 61

INTRODUCTION

Selon l'opinion de nombreux économistes de la santé, la demande de soins médicaux doit être distinguée de la demande des autres biens et services. Selon eux, le médecin a une grande influence sur les besoins de ses patients en soins médicaux, et il exerce cette influence en fonction de ses buts économiques à lui. Ce point de vue s'oppose à l'opinion traditionnelle américaine selon laquelle le médecin ne fait que jouer le rôle d'un agent parfait de son patient. Par conséquent, même la demande de soins hospitaliers a été expliquée en se référant uniquement aux revenus des patients et à des prix relatifs, sans tenir compte du médecin et ses volontés et contraintes.

La première chose que l'on observe dans cette étude est que pour la plupart des patients le contrôle de leur médecin implique des coûts excessifs. Le médecin n'aura donc pas tant de difficulté à poursuivre ses propres objectifs. Ce fait justifie l'examen d'un modèle économique du comportement médical, qui sera présenté dans la première section. Le modèle est d'une part conventionnel, car les préférences des patients ne peuvent pas être modifiées à volonté par le médecin. D'autre part, il est non conventionnel dans le fait que les préférences des patients sont réduites à l'alternative d'aller consulter le médecin ou non, laissant au médecin le choix de traitement et sa durée.

De plus le modèle contient une variable « éthique professionnelle », ceci dans le but d'analyser l'impact d'une variable souvent mentionnée, mais jamais opérationnalisée. Le médecin dispose de trois variables de décision pour atteindre ses objectifs : un niveau critique d'intensité symptomatique (c) qui détermine l'hospitalisation, un taux de salaire implicite (q) et le temps moyen (t) consacré au traitement d'un patient. Dans la section 2 nous discuterons les conditions nécessaires pour un optimum, puis nous énoncerons dans la section 3 les implications comparatives statiques du modèle. Celles-ci peuvent se résumer

comme suit : Quand la demande de premières consultations est réduite, par exemple par un ticket modérateur plus important, le modèle prédit une baisse de (q), mais une augmentation de (t), c'est-à-dire des traitements moins chers par unité de temps mais de plus longue durée. Pour chaque cas traité, les décomptes avec la caisse-maladie reflètent le produit ($q . t$) du taux de salaire implicite et du temps dépensé. Tandis que le prix effectif du service médical baisse, ce produit observé ne changera pas beaucoup et pourra même augmenter à cause de l'augmentation de (t). En interprétant ce phénomène, on peut donc arriver facilement à la conclusion d'une « demande induite par le médecin ». Dans la quatrième section, nous essaierons d'analyser ces deux éléments (q) et (t) dans des décomptes mis à notre disposition par une caisse-maladie suisse. Enfin on trouvera dans la dernière section un résumé et quelques conclusions.

1. LES ÉLÉMENTS D'UN MODÈLE ÉCONOMIQUE DU COMPORTEMENT MÉDICAL

Le modèle que nous allons développer consiste en une fonction à maximiser et de quatre restrictions portant sur le revenu du travail, le nombre de cas traités, l'éthique professionnelle et le temps de loisir. Selon le modèle, le médecin privé maximise l'espérance mathématique de son utilité de la façon suivante :

$$\bar{U} = E[U(Y, I, L)] \tag{1}$$

E, espérance mathématique;
Y, revenu du travail;
I, degré d'éthique professionnelle, *cf.* équation (4) plus bas;
L, temps de loisir.

L'inclusion du revenu (Y) ne nécessite pas de justification. Comme le médecin ne peut augmenter le volume de service qu'en travaillant plus, il doit tenir compte de son temps libre (L) quand il prend sa décision. Cet aspect a été introduit dans la littérature spécialisée par Sloan [24] ([1]). Quant à l'éthique professionnelle, les économistes de la santé se sentent obligés d'en parler ne serait-ce qu'en note de bas de page. Feldstein [8], par exemple, n'ayant pas réussi à identifier de plausibles fonctions d'offre et de demande de soins médicaux avec des données agrégées, suppose que cet échec est dû à l'éthique professionnelle des médecins. Ceux-ci seraient amenés à fixer un prix plus bas que le prix d'équilibre afin de disposer d'une réserve suffisante de cas intéressants sur le plan scientifique. Dans sa critique d'une étude faite par Sloan (1974) sur l'inflation médicale, Huang (1976) affirme que l'éthique professionnelle doit être reconnue comme motif important, toutefois sans la définir de façon précise. En effet, on peut constater qu'il existe au moins une institution qui protège (économiquement) le médecin qui pratique une certaine éthique professionnelle. Ainsi, dans tous les pays occidentaux, il est interdit aux médecins de former une société

(1) Les chiffres entre crochets renvoient à la bibliographie *in fine*.

anonyme médicale. Car les actionnaires d'une société anonyme auraient toujours la possibilité de remplacer un médecin insuffisamment orienté vers un but économique par un concurrent. Il semble donc valoir la peine d'examiner si une telle éthique modifie le comportement des médecins de façon systématique et peut-être même quantifiable.

La fonction d'utilité expliquée, passons aux restrictions. D'abord, nous avons *pour le revenu :*

$$Y = qt\,P, \tag{2}$$

q, prix brut de 1 heure de service;
t, durée moyenne de traitement par patient;
P, nombre de patients.

Notons que (q), le taux de salaire implicite du médecin, est endogène, malgré les contrats tarifaires entre les associations médicales cantonales et les caisses-maladie. Car si le barème ne correspond pas à la valeur optimale (q^*), le médecin peut redéfinir les actes médicaux ou bien accélérer leur exécution. Comme ceux-ci ont des tarifs fixes, il change ainsi son revenu par unité de temps, soit (q) [1]. Or, en faisant cela, le médecin acquiert la réputation d'être cher. En Suisse, ceci a des répercussions sur la demande des premières consultations car les assurés participent aux frais. Le taux de participation, désigné par (r), est de 10 % en Suisse (au-delà d'une note d'honoraire de 300 F); les détails en seront donnés dans la section 3. Ainsi, le prix net par heure de service (qr) est un argument de la fonction qui détermine le nombre des patients traités :

$$\left\{\begin{array}{l} P = \displaystyle\int_0^{c_h} (qr,\ w,\ s).\,F\,(s,\ \bar{s})\,ds, \\ \qquad 0 < h(\cdot) < 1, \\ \qquad\quad 0 \leqslant s \leqslant 1, \\ \qquad\quad h_{(qr)} < 0, \\ h_s > 0, \qquad h_{(qr).\,s} > 0. \end{array}\right. \tag{3}$$

P, nombre de patients traités;
c, niveau symptomatique critique;
h, demande de premières consultations, définie comme probabilité d'aller voir le médecin;
r, taux de participation aux frais (en guise de ticket modérateur);
qr, prix net de 1 heure de service médical;
w, vecteur de variables économiques influençant $h(\cdot)$;
s, niveau symptomatique, gravité du cas;

(1) L'expérience faite par les caisses-maladie allemandes donne une impression de la marge dont disposent les médecins dans leurs décomptes. Comme signe de solidarité avec le reste de la population, touchée par la récession, les associations médicales convinrent d'augmenter le barème tarifaire par points de 2 % seulement en 1975. Après coup, les caisses-maladie furent frappées par une augmentation des frais par cas traité de 11 %, *cf.* Institut Allemand de Recherche Économique (DIW) (1976).

$\bar{s}$, paramètre (par exemple la moyenne) caractérisant la fonction $F(s)$ inconnue;

F, nombre de personnes au niveau symptomatique (s), inconnu du médecin.

A part le prix de service (q), le médecin a donc un deuxième instrument d'exclusion, à savoir, le seuil critique (c). Quand un patient potentiel présente des symptômes dont l'intensité est au-dessus de (c), il est envoyé à d'autres médecins ou à l'hôpital. Avec la variable de décision (c), le médecin privé règle assez directement la demande de services hospitaliers, qui sont financés en grande partie par le public. En somme, la deuxième restriction fixe le nombre de patients ou cas traités par le médecin considéré.

La troisième contrainte donne son sens au terme « éthique professionnelle » et le rend en principe opérationnel :

$$I = \int_0^c [p(s, 0) - p(s, t)] \cdot h(qr, w, s) \cdot F(s, \bar{s})\, ds, \tag{4}$$

p, mortalité d'un patient potentiel, avec $\partial p / \partial t < 0$.

Pour chacun de ses patients potentiels, le médecin estime deux probabilités de mort : l'une, $p(s, 0)$, sans traitement, l'autre, $p(s, t)$, après (t) heures de soins. C'est donc là l'élément qualitatif de l'éthique professionnelle, qui se reflète dans la contribution du médecin à l'amélioration de la santé de ses patients. L'élément quantitatif de l'éthique est favorisé si le médecin fixe (c) à un niveau élevé ou s'il charge un taux de salaire implicite (q) bas ([1]).

Dans la dernière restriction, le temps de loisir est déterminé :

$$L = T - t P, \tag{5}$$

T, temps total disponible;

t, durée moyenne du traitement;

P, nombre de cas traités;

$t P$, temps de travail au cabinet.

A titre de résumé le modèle entier est présenté comme suit :

$$\begin{cases} \bar{U} = E[U(Y, I, L)] = \displaystyle\int_0^c dU[Y(\bar{s}), I(\bar{s}), L(\bar{s})]\, F(s, \bar{s})\, ds, \\ Y = q t P, \\ P = \displaystyle\int_0^c h(qr, w, s) \cdot F(s, \bar{s})\, ds, \\ I = \displaystyle\int_0^c [p(s, 0) - p(s, t)] \cdot h(qr, w, s) \cdot F(s, \bar{s})\, ds, \\ L = T - t P. \end{cases} \tag{6}$$

(1) Cette formulation de l'éthique professionnelle a été inspirée par le travail de SMALLWOOD et SMITH [25]. Mais là, même la survie des individus anonymes, que le médecin a exclus en chargeant un certain prix, contribue à l'objectif éthique. Dans notre formulation, le médecin est plus égoïste, se bornant à améliorer l'état de santé de ceux qu'il a acceptés comme patients.

Variables de décision primaires

c, niveau symptomatique critique;
q, prix brut de 1 heure de service (taux de salaire implicite);
t, durée moyenne de traitement par patient.

Variables de décision déduites

Y, revenu du travail;
I, degré d'éthique professionnelle, définie comme contribution du médecin à l'amélioration de la santé des patients;
L, temps de loisir;
P, nombre de patients traités dans la période de décision.

Variables exogènes

r, taux de participation aux frais, fixé par la caisse-maladie;
w, vecteur de variables économiques influençant la demande de premières consultations;
s, intensité symptomatique, variable aléatoire, $0 \leqq s \leqq 1$;
$\bar{s}$, paramètre inconnu de la distribution $F(s)$, par exemple intensité symptomatique moyenne;
T, total du temps disponible.

Fonctions

h, demande de premières consultations, définie comme probabilité d'aller voir le médecin;
p, mortalité d'un patient potentiel dans la période de décision;
F, nombre des individus ayant un niveau symptomatique (s) dans la population soignée, caractérisée par ($\bar{s}$).

2. LES CONDITIONS NÉCESSAIRES POUR UN OPTIMUM

D'abord, soulignons quelques hypothèses du modèle. Premièrement, il n'y a pas de problème de diagnostic. Le médecin connaît tout de suite la valeur (s) d'un patient potentiel. En réalité, ce n'est peut-être qu'après une série de consultations que le médecin se rend compte que l'intensité symptomatique d'un patient dépasse son seuil critique (c). Mais les prédictions qualitatives du modèle ne devraient pas changer beaucoup si l'hospitalisation est effectuée lors de la n-ième au lieu de la première consultation. Puis, le médecin accepte des patients ne manifestant pas de symptômes ($s = 0$). Mais d'après le patient, (s) doit être non nul. En général, la perception des symptômes peut bien être différente suivant le médecin ou le patient. Cette distinction n'est pas faite dans le modèle, qui ne contient que le symbole commun (s). Troisièmement, seule la demande de

premières consultations est provoquée par les patients. Le premier contact étant établi, le médecin est libre de choisir sa dépense de temps (t), soumise à des contraintes qui proviennent de lui-même exclusivement. Car les patients potentiels sont indifférents aux variations de (t), *cf.* l'équation de $h(.)$ dans (6). Une augmentation de (t) peut être effectuée par deux moyens qui ne s'excluent pas l'un l'autre : ou bien la durée d'une consultation est prolongée ou bien le nombre de consultations est augmenté. Ces deux variations signifient un coût de temps de la part du patient, avec un impact négatif sur la demande [*cf.* Acton (1975, 1976) ([1])]. D'autre part, le temps dépensé par le médecin est fréquemment interprété comme indicateur de la qualité des soins, ce qui est justifié par l'équation (4) du modèle.

Quant aux conditions optimales de premier ordre, abstraction est faite des inégalités ($0<c<1$, $q>0$, $L>0$) parce qu'elles n'ont pas d'importance pour des médecins qui pratiquent continuellement. En supposant qu'il existe une solution en fonction des variables de décision primaires satisfaisant à ces inégalités, nous avons :

$$E[U_c]=E[U_Y\,Y_c]+E[U_I\,I_c]+E[U_L\,L_c]=0 \tag{7}$$

avec $U_Y=\partial U/\partial Y$. Pour une valeur donnée du paramètre ($\bar{s}$), la dérivation par (c) des équations (6) fournit les résultats suivants :

$$Y_c=Y_P\,P_c=qt.h(qr,\,w,\,c).F(c)>0, \tag{8}$$

$$I_c=[p(c,\,0)-p(c,\,t)].h(qr,\,w,\,c).F(c)>0, \tag{9}$$

$$L_c=-t\,P_c=-t.h(qr,\,w,\,c).F(c)<0. \tag{10}$$

Oublions momentanément les problèmes d'incertitude. Alors, nous pouvons diviser les termes de l'équation (7) par les équations (8) à (10) pour obtenir, après simplification :

$$\frac{U_L}{U_Y}=q+\frac{U_I}{U_Y}.\left[\frac{p(c,\,0)-p(c,\,t)}{t}\right]. \tag{11}$$

Le taux marginal de substitution entre le loisir (L) et le revenu du travail (Y) est donc égal au taux de salaire implicite — plus un terme positif. Ce terme varie proportionnellement à l'importance relative de l'éthique professionnelle dans la fonction d'utilité du médecin (U_I/U_Y). Le deuxième facteur, c'est l'amélioration de la santé apportée aux patients de niveau symptomatique critique. Comme il s'avère que les éléments de l'équation (11) ne changent pas de signe quand ($\bar{s}$) varie, nous aboutissons à la:

PROPOSITION 1. — *A cause de son éthique professionnelle, le médecin travaillerait plus qu'un indépendant d'une autre profession pour le même taux de salaire réalisé.*

---

(1) Dans ses études, ACTON ([1], [2]) définit les coûts de temps par la durée du trajet et de l'attente. Un changement de durée de traitement (t) est ni nécessaire ni suffisant pour un changement de ces coûts.

Il est très difficile de vérifier cette proposition parce que le taux de salaire implicite (q) est une variable de décision, fixée par le médecin en fonction des conditions de marché, *cf.* équations (12) à (15). En effet, d'après les statistiques américaines, les médecins travailleraient plus que les indépendants d'une autre profession, y compris les médecins-dentistes. Dans ces statistiques cependant, la différence entre les heures travaillées par semaine est plus prononcée que celle entre les semaines de travail par an, *cf.* Mennemeyer [20]. Ce détail permet de supposer que ce sont les exigences physiques de la pratique de la médecine dentaire qui sont responsables de cette différence plutôt qu'un manque d'éthique professionnelle.

L'éthique professionnelle devrait aussi influencer la manière dont le médecin fixe le prix de son service, c'est-à-dire son taux de salaire implicite. Cela se voit dans la condition d'optimum :

$$E[U_q] = E[U_Y\, Y_q] + \underset{(-)}{E[U_I\, I_q]} + \underset{(+)}{E[U_L\, L_q]} = 0. \tag{12}$$

Comme expressions de référence, nous avons :

$$\left\{ \begin{array}{l} Y_q = t(P + q\,P_q) = t\,P(1 + e_{P,\,q}), \\ \text{avec :} \\ e_{P,\,q} = \partial P / \partial q \,.\, \dfrac{q}{P}. \end{array} \right. \tag{13}$$

$$I_q = r \,.\, \int_0^c [p(s,\,0) - p(s,\,t)] \,.\, h_{(qr)}\, F(s)\, ds < 0, \tag{14}$$

$$L_q = -\,t\,P_q = -\,rt \,.\, \int_0^c h_{(qr)}\, F(s)\, ds > 0. \tag{15}$$

Si l'éthique professionnelle n'a pas d'effet, le deuxième terme de l'équation (12), $E[U_I\, I_q]$ est zéro, ce qui impliquerait un premier terme $E[U_Y\, Y_q]$ négatif. Par conséquent, dans l'équation (13), $e_{P,\,q}$ doit être inférieur à -1. Donc, un médecin visant à un optimum entre son revenu du travail et son temps de loisir devrait fixer le prix de son service à un niveau si élevé que le nombre de cas traités (P) réagirait de façon élastique. Comme le coût de service marginal des assurés est seulement (qr), cela peut signifier un prix brut exorbitant. Ainsi, le médecin contrecarrerait le but déclaré des caisses-maladie, à savoir, de mettre les assurés à l'abri des prix du marché.

Mais les études empiriques montrent que la demande de soins médicaux est satisfaite à des prix où l'élasticité est beaucoup plus proche de zéro que de -1, *cf.* la revue de Feldstein [9]. Une estimation de l'élasticité des premières consultations par rapport au prix net de service est de -0.11, *cf.* Newhouse et Phelps [22]. Il faut avouer qu'il s'agit là d'une réaction moyenne d'assurés à des

variations de prix se référant au bien « soins ambulatoires ». Par contre, $e_{P,\,q}$ est un paramètre spécifique à un médecin particulier. Mais un calcul approximatif montre que le médecin individuel devrait être protégé par des frais de recherche excessifs de la part du patient pour expliquer la différence entre, disons, -0.11 et -1. Supposons donc que la valeur $\hat{e}_{h.\,(qr)} = -0.11$ trouvée par Newhouse et Phelps [22] est l'élasticité « marché ». L'élasticité par rapport au coût total d'une première consultation peut être calculée suivant la formule :

$$e_{\underset{\text{(observée)}}{h.\,qr}} = \frac{qr}{qr+wt} \cdot e_{\underset{\text{le médecin)}}{h,\,(qr+wt)\ \text{(vraie pour}}} \tag{16}$$

où (w) signifie taux de salaire, et (t) temps de recherche de la part du patient plutôt que temps de traitement, pour l'instant; *cf.* Acton [2], (p. 201). Les patients qui envisagent de changer de médecin sont ceux qui sont prêts à supporter le coût total d'une consultation $(qr+wt)$ comme si c'était le coût monétaire seulement. Ce serait donc l'élasticité $e_{h,\,(qr+wt)}$ qui détermine si oui ou non un médecin cher perd un tel patient marginal. Mais sur la base de l'équation (16), on n'obtiendrait une valeur de -1 ou plus bas pour cette élasticité que si la part des coûts monétaires (qr) dans les coûts globaux $(qr+wt)$ était inférieure à 1/9 ou 11 %. Or, le taux de participation aux frais est environ de 85 % dans l'échantillon utilisé par Newhouse et Phelps [22]. Nous aboutissons donc à la :

PROPOSITION 2. — *La manière dont les médecins choisissent leur taux de salaire implicite ne semble pas être compatible avec la maximisation du revenu. Dans l'équation* (13), *on a donc* $Y_q > 0$.

Dans le contexte d'un système d'assurance sociale avec des caisses-maladie, les médecins auraient toujours la possibilité d'exploiter encore mieux les barèmes tarifaires. Mais, peut-être cet état des choses n'est-il point la conséquence d'une éthique professionnelle mais plutôt d'un horizon de décision à plus longue échéance. C'est pour la même raison que des entreprises à but lucratif s'abstiennent parfois d'exploiter une possibilité d'augmenter leur prix de vente pour retenir leur clientèle. A cet égard, les dentistes se trouvent dans une position très comparable aux médecins quant à la protection de la concurrence et quant au taux de salaire implicite (q). Or, utilisant des données « de branche » Maurizi (1965) estime une élasticité prix de la demande de -1.6. Ainsi, les dentistes semblent exploiter beaucoup mieux leurs possibilités de gains supplémentaires que les médecins. Aussi, la prestation gratuite de services n'a jamais joué le même rôle chez les dentistes que chez les médecins. En conclusion, ce pourrait bien être l'éthique professionnelle qui permettrait d'énoncer la proposition 2.

La troisième condition d'optimalité est donnée par :

$$E[U_Y Y_t] + E[U_I I_t] + E[U_L L_t] = 0,$$

avec :

(17) $$Y_t = q\,P > 0,$$

(18) $$I_t = -\int_0^c h(qr,\ w,\ s).F(s).p_t(s,\ t)\,ds > 0,$$

et :

(19) $$L_t = -P < 0.$$

Après substitution et division par (P), nous obtenons, dans le cas certain ($\bar{s}$ constant) une expression similaire à (11) :

(20) $$\frac{U_L}{U_Y} = q + \frac{U_I}{U_Y} \cdot \left[\frac{-\int_0^c h(qr,\ w,\ s).F(s).p_t(s,\ t)\,ds}{P} \right].$$

Encore une fois, le taux marginal de substitution entre revenu et loisir est égal au taux de salaire implicite plus un terme positif. La comparaison des équations (11) et (20) nous donne la condition dérivée d'optimalité :

(21) $$\frac{p(c,\ 0) - p(c,\ t)}{t} = \frac{-\int_0^c h(.).F(s).p_t(.)\,ds}{P}.$$

Si un médecin visait à maintenir cette égalité, cela serait une preuve sûre de son orientation éthique, car l'équation (21) est valable sous condition que U_I / U_Y ne soit pas égal à zéro. Elle montre que le médecin peut augmenter son véritable output (du point de vue professionnel) en élévant son seuil symptomatique critique. Sa contribution à la santé est exprimée par le nominateur de la partie gauche de l'équation (21). Mais cette élévation du seuil symptomatique critique le forcerait à priver les autres patients de son temps, à un taux marginal qui est égal au taux moyen (t) dans la formulation simplifiée du modèle. L'autre variation marginale serait une intensification du traitement, dont l'impact est donné par le nominateur dans la fraction du côté droit de l'équation (21). Mais ceci signifierait de ne pas accepter des patients par manque de temps. C'est la raison pour laquelle (P) apparaît dans le dénominateur. Enfin, comme le côté à droite de l'équation (21) est positif, le côté à gauche doit l'être aussi. Cela signifie la :

PROPOSITION 3. — *Un médecin pratiquant une certaine éthique professionnelle ne fixera jamais son seuil critique (c) à une valeur si élevée qu'il nuirait aux patients marginaux, sauf à cause de l'incertitude.*

3. L'ANALYSE COMPARATIVE STATIQUE

Supposons maintenant que la demande de premières consultations est limitée par un taux de participation aux frais (r) plus élévé. Ceci force le médecin à s'adapter en cherchant de nouvelles valeurs pour les variables de décision

(c, q, t). Pour une valeur constante de $(\bar{s})$, la différentiation totale des conditions nécessaires d'optimum conduit à :

$$\begin{bmatrix} U_{cc} & U_{cq} & U_{ct} \\ U_{cq} & U_{qq} & U_{qt} \\ U_{ct} & U_{qt} & U_{tt} \end{bmatrix} \cdot \begin{bmatrix} dc \\ dq \\ dt \end{bmatrix} = - \begin{bmatrix} U_{cr} \\ U_{qr} \\ U_{tr} \end{bmatrix} . dr. \tag{22}$$

Afin d'obtenir des restrictions qualitatives portant sur dc/dr, par exemple, les signes du vecteur (U_{cr}, U_{qr}, U_{tr}) doivent être connus. L'expression pour U_{cr} sert à illustrer que c'est impossible sans des restrictions fortes :

$$\begin{aligned} U_{cr} = \frac{\partial}{\partial r}(U_Y Y_c + U_I I_c + U_L L_c) \\ = (\underset{-}{U_{YY}} \underset{-}{Y_r} + \underset{0}{U_{YI}} \underset{-}{I_r} + \underset{0}{U_{YL}} \underset{+}{L_r}) \underset{+}{Y_c} + \underset{+}{U_Y} \underset{0}{Y_{cr}} \\ + (\underset{0}{U_{IY}} \underset{-}{Y_r} + \underset{-}{U_{II}} \underset{-}{I_r} + \underset{0}{U_{IL}} \underset{+}{L_r}) \underset{+}{I_c} + \underset{+}{U_I} \underset{0}{I_{cr}} \\ + (\underset{0}{U_{LY}} \underset{-}{Y_r} + \underset{0}{U_{LI}} \underset{+}{I_r} + \underset{-}{U_{LL}} \underset{+}{L_r}) \underset{+}{L_c} + \underset{+}{U_L} \underset{0}{L_{cr}}. \end{aligned} \tag{23}$$

Ces restrictions, déjà introduites dans l'équation (23), peuvent être justifiées comme suit. D'abord, les variables de décision dérivées dépendent toutes du paramètre inconnu $(\bar{s})$ et sont donc aléatoires. Ainsi, le médecin doit faire son choix entre les différentes possibilités, et s'il craint le risque, cela implique par exemple $U_{YY} < 0$, *cf.* Arrow [4] (p. 92-93). Les dérivées mixtes comme U_{YI} sont égales à zéro parce que des restrictions seulement qualitatives n'apporteraient rien (1). Enfin, les restrictions du type $Y_{cr} = 0$ se fondent sur la supposition que les patients à niveau symptomatique critique $(s = c)$ ne réagissent plus à des variations du prix net de service. En conclusion, nous avons :

$$U_{cr} > 0. \tag{24}$$

Pour mettre des signes aux éléments (U_{qr}, U_{tr}) du vecteur, il faut avoir recours à Arrow [4] qui a montré que le coefficient relatif d'aversion du risque s'approche de l'unité comme limite inférieure si par exemple (Y) devient très grand. Cela signifie :

$$\frac{-U_{YY}}{U_Y / Y} > 1, \tag{25}$$

$-U_{YY}/(U_Y/Y)$ coefficient relatif d'aversion du risque par rapport au revenu.

(1) En outre, la séparabilité stricte de la fonction d'utilité, supposée ici, semble être un peu plus acceptable que dans la théorie de la demande du consommateur, *cf.* BROWN et DEATON [6] qui ont analysé les différents concepts de séparabilité. Car on s'imaginerait aisément que le médecin regarde (Y, I, L) comme trois aspects différents de sa vie. Subjectivement séparés, ils sont néanmoins liés par des restrictions objectives comme celles du modèle (6).

En généralisant pour tous les arguments de la fonction d'utilité, nous avons une restriction disant que les dérivées de second ordre doivent toujours respecter les conditions suivantes :

$$(26) \qquad \frac{U_Y}{|U_{YY}|} < Y, \qquad \frac{U_I}{|U_{II}|} < I, \qquad \frac{U_L}{|U_{LL}|} < L.$$

Ces considérations suffisent à donner au vecteur des impacts la structure suivante :

$$(27) \qquad U_{cr} > 0, \qquad U_{qr} < 0, \qquad U_{tr} > 0.$$

Enfin, il s'avère possible de déterminer le signe d'un des éléments non diagonaux de la matrice hessienne qui apparaît dans l'équation (22). Il s'agit de :

$$(28) \qquad U_{ct} < 0.$$

Malgré une analyse détaillée, on n'a pas réussi à contraindre U_{cq} et U_{qt}. L'ambiguïté a été supprimée par la restriction :

$$(29) \qquad U_{cq} = U_{qt} = 0.$$

Ainsi, le système (27) peut être résolu pour (dq), par exemple, avec le résultat :

$$(30) \qquad \frac{dq}{dr} = \frac{-1}{|H|} \begin{vmatrix} U_{cc} & U_{cr} & U_{ct} \\ 0 & U_{qr} & 0 \\ U_{ct} & U_{tr} & U_{tt} \end{vmatrix} = \underset{-}{\frac{-U_{qr}}{|H|}} \underset{-}{\begin{vmatrix} U_{cc} & U_{ct} \\ U_{ct} & U_{tt} \end{vmatrix}} < 0.$$

Jusqu'ici, l'analyse du système (27) a été effectuée pour une valeur fixe du niveau moyen d'intensité symptomatique $(\bar{s})$. Mais l'intégration sur $(\bar{s})$, requise dans (6), ne changerait rien quant aux implications du modèle. Nous formulons donc la :

PROPOSITION 4. — *Un taux de participation aux frais* (r) *plus élevé inciterait le médecin à baisser son taux de salaire implicite* (q).

Donc, même si les médecins s'abstenaient de réaliser une valeur de (q) qui maximiserait le revenu de travail (*cf.* proposition 2), ils réagiraient de manière très normale à des variations de la demande, contrairement à l'hypothèse avancée par Feldstein [8]. Celui-ci pensait que l'éthique professionnelle était responsable de l'impossibilité d'identifier, dans une analyse empirique, soit la fonction d'offre soit celle de demande.

Quant à la dépense moyenne de temps pour un cas, le modèle prédit que :

$$(31) \qquad \frac{dt}{dr} = \frac{-1}{|H|} \begin{vmatrix} U_{cc} & 0 & U_{cr} \\ 0 & U_{qq} & U_{qr} \\ U_{ct} & 0 & U_{tr} \end{vmatrix}$$

$$= \underset{+}{\frac{-U_{tr}}{|H|}} \underset{+}{\begin{vmatrix} U_{cc} & U_{cq} \\ U_{cq} & U_{qq} \end{vmatrix}} \underset{-}{-\frac{U_{cr}}{|H|}} \underset{-}{\begin{vmatrix} U_{cq} & U_{cq} \\ U_{ct} & U_{qt} \end{vmatrix}} > 0$$

et cela parce que U_{cr} est dominé par U_{tr}, et le deuxième mineur est dominé par le premier ([1]). Voici la :

PROPOSITION 5. — *La réduction de la demande des premières consultations à cause d'un taux* (r) *plus élevé inciterait le médecin à sacrifier plus de son temps pour un cas traité* (t).

Les patients qui ne sont pas intimidés par un ticket modérateur plus important profiteront donc d'un mode de traitement plus personnel, moins technique, et même plus effectif, à en croire l'équation (18). Mais les frais décomptés par cas, reflétant le produit ($q.t$), peuvent bien rester stables ou même augmenter, cachant la baisse du prix effectif du service (q). Si l'on se borne à analyser les frais par cas comme indicateur du prix, on est tenté de parler d'une « demande induite » par le médecin. Evans [7], par exemple, a même prétendu qu'il fallait quitter le cadre de la théorie néoclassique en faveur d'une « target income hypothesis » pour expliquer les phénomènes. Pour notre part, nous adhérons à la :

PROPOSITION 6. — *Le fait que les frais décomptés par cas baissent peu ou même s'élèvent en fonction d'une augmentation de* (r) *ne constitue pas d'évidence en faveur de la thèse de la « demande induite ». L'élément prix* (q) *du produit* ($q.t$) *peut bien avoir baissé tandis que l'élément qualité* (t) *aurait augmenté.*

4. LES DONNÉES ET LE MODÈLE STATISTIQUE

Des données ont été mises à disposition par la caisse CMB, une des caisses-maladie les plus importantes en Suisse ([2]). Il s'agit des factures présentées par un millier de médecins pour les années 1976 à 1978 ([3]). Étant une assurance sociale, la caisse CMB est soumise à des réglementations par l'État. En revanche, elle reçoit des subventions allant de 20 à 25 % des recettes totales ([4]). Il y a trois catégories d'assurés. La première, générale (P1) connaît une franchise de 30 F, prélevée chaque trimestre si le traitement continue. Au-delà de cette limite, la participation est égale à la franchise, ou aux 10 %, quel que soit le montant supérieur. Cela signifie que l'heure médicale marginale est gratuite entre les 30 et 300 F. La ligne budgétaire insérée dans le graphique 1 est donc horizontale entre les points M_1 et M_1'.

---

(1) Pour les détails, *cf.* ZWEIFEL [26], chap. 5.

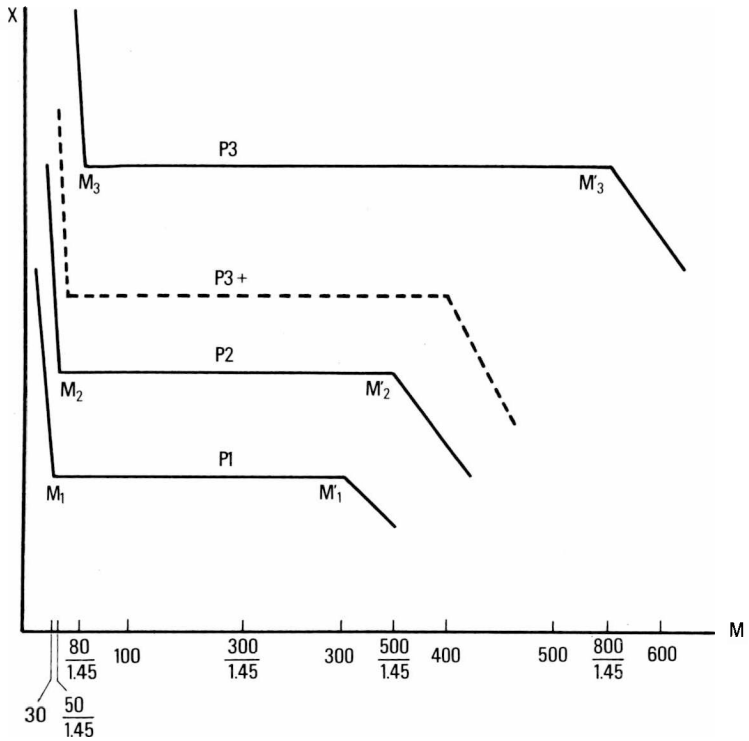
(2) L'auteur voudrait remercier M. H. Schmid, directeur de la Caisse-Maladie pour le Canton de Berne (CMB), d'avoir consenti au transfert des données et d'avoir donné son soutien à l'analyse empirique tout au long de l'étude.

(3) Ces décomptes étant confidentiels, ils n'ont pas pu être combinés avec d'autres sources. Mais pour vérifier les implications relatives à l'éthique professionnelle, on devrait connaître des caractéristiques personnelles du médecin en question. Donc, ces implications ne sont pas soumises à un test statistique.

(4) *Cf.* par exemple le 109e Rapport annuel de l'exercice allant du 1er janvier au 31 décembre 1978, Caisse-maladie pour le Canton de Berne (CMB).

GRAPHIQUE 1

La contrainte budgétaire pour les catégories d'assurés P1, P2 et P3



M, services médicaux ambulatoires, en équivalents monétaires (francs suisses);
X, ensemble de tous les autres biens;
M_1, valeur de (M), applicable à la catégorie P1, où la participation dépasse la franchise;
P3+, contrainte budgétaire pour la catégorie P3 si le médecin met une surcharge de 100 % (au lieu de 45 %).

Les membres de la caisse dont le revenu dépasse un certain seuil sont classifiés dans la catégorie P2. La franchise y monte à 50 F, et les médecins ont le droit de facturer leurs prestations avec une surcharge de 45 %. C'est pour cela que la ligne budgétaire descend encore plus vite entre l'origine et l'équivalent réel de 50 F, soit 35 (= 50/1.45) F. On penserait donc qu'un assuré de la catégorie P2 est exposé à une participation relativement importante. Plus spécifiquement, une population de patients contenant une grande portion de patients du type P2 serait caractérisée par un taux moyen de participation (r) élevé. Mais le graphique montre qu'il existe des montants facturés où l'assuré P1 est déjà sujet à un ticket modérateur de 10 % tandis que l'assuré P2 est toujours libéré de toute participation sauf la franchise. A un patient P1, cela succèderait au fur et à mesure de la probabilité avec laquelle la facture tombe dans l'intervalle [300, 500] F. Le même intervalle est réduit, en termes réels, à [207, 345] F pour

un patient P2 à cause de la surcharge de 45 %. Il est associé à une fréquence de 17 % environ (1).

La troisième catégorie (P3) comprend les patients privés. Comme tels, ils ne profitent plus d'une protection tarifaire. La franchise monte à 80 F, équivalant à des prestations réelles de 55 F, si le médecin se borne à appliquer la surcharge de 45 %. Avec une surcharge de 100 %, ce serait la ligne marquée P3+ qui résulterait. Néanmoins, dans un intervalle correspondant à [207, 551] F réels dans le premier cas et [150, 400] F dans le second, un patient P3 profite d'un taux marginal $r=0$. Malheureusement, il est impossible de chiffrer les probabilités associées avec assez de précision (2). Tout de même, ces considérations peuvent être résumées dans la :

PROPOSITION 7. — *Une portion importante de patients P2 dans la population des patients traités n'indique pas nécessairement un taux de participation (r) élevé en comparaison avec les patients P1. Mais une portion importante de patients P3 signifie une valeur de (r) élevée, relative aux patients P2.*

Ce résultat est important pour l'explication du graphique 2, qui montre les relations entre les variables latentes et observées. A gauche, nous avons les trois quantités exogènes ($\bar{s}$, r, w_y) avec leurs indicateurs. La gravité moyenne des cas présentés au médecin ($\bar{s}$) serait donc représentée par l'âge moyen des patients (AGEP). Les valeurs de cette variable ne sont pas seulement fonction de la quantité latente ($\bar{s}$) mais aussi d'une erreur d'observation, symbolisée par la flèche courte partant de la marge gauche. Nous introduisons REFER % comme deuxième indicateur de ($\bar{s}$) parce qu'un médecin dont la clientèle contient beaucoup de cas référés par ses collègues se trouverait exposé à des cas plutôt difficiles. Enfin, les patients de provenance agricole sont censés être affectés assez gravement avant de voir un médecin parce qu'une telle visite entraîne des coûts d'opportunité importants. Pour cette raison, PAGRI % figure comme troisième indicateur de ($\bar{s}$).

Le taux de participation (r) est une autre quantité exogène inobservée car il est impossible de reconstruire, sur la base des conditions d'assurance, sa valeur applicable *ex ante*. Pire encore, la *proposition* 7 ne nous permet pas même de supposer qu'une population comprenant un grand nombre de patients P2 et P3 soit caractérisée par une valeur de (r) élevée. La classification adoptée par la caisse-maladie fait que P2 % et P3 % sont aussi des indicateurs du revenu (w_Y) prévalant dans la population soignée. Pour cette raison, P2 % et P3 % sont à l'extrémité de deux flèches. Enfin, (w_Y) est aussi reflété par les variables catégoriques SEMIUR et URBN. Comparé aux communes rurales, un médecin pratiquant dans une commune semi-urbaine ou urbaine aura des patients plus riches, en moyenne.

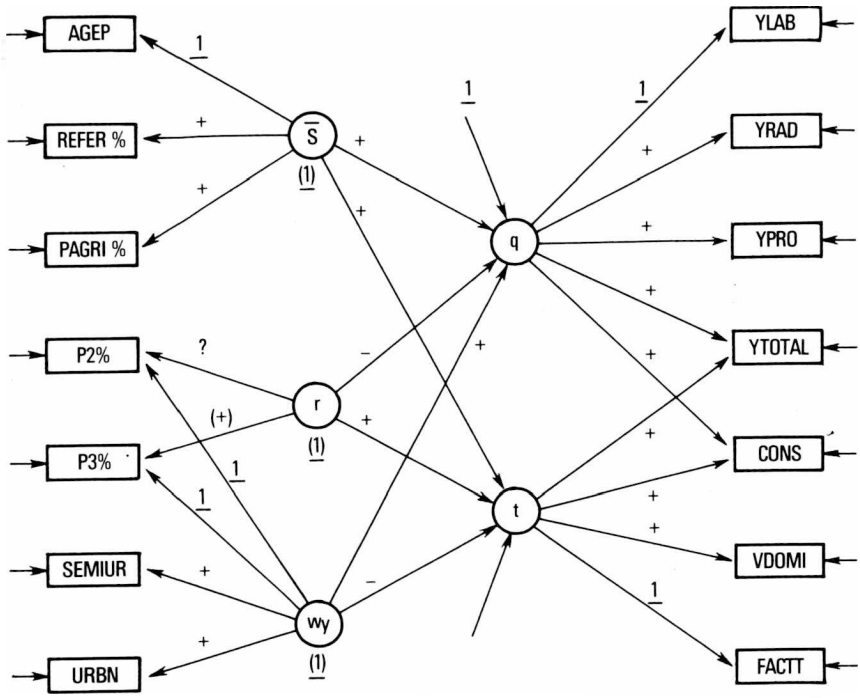
---

(1) *Cf.* ZWEIFEL [26], chap. 8 pour les détails.

(2) La distribution des montants facturés n'est connue que pour l'ensemble des trois catégories. Pour les assurés P3 en particulier, la caisse n'est pas renseignée sur le taux de surcharge appliqué par le médecin.

GRAPHIQUE 2

Le modèle structurel et le modèle de mesure



Remarque. — Les variances des variables exogènes latentes sont standardisées à unité (1); de plus, elles ne sont pas corrélées par supposition.

Au côté droit du graphique 2, nous retrouvons les deux variables de décision, soit le taux de salaire implicite (q) et la durée du traitement (t). Pour des raisons théoriques, expliquées dans la section 1 ci-dessus, le montant décompté par cas (YTOTAL) doit correspondre au produit de ses deux facteurs. Cette double relation est symbolisée par les deux flèches donnant sur YTOTAL. Mais le nombre de consultations par cas (CONS) est, lui aussi, un indicateur ambigu. D'une part, une consultation additionnelle signifie plus de temps accordé à un patient, *ceteris paribus*. D'autre part, un nombre élevé de consultations facturées peut signifier que le médecin considéré a diminué la durée moyenne de la visite pour gagner plus d'argent par unité de temps — c'est-à-dire pour atteindre un taux de salaire (q) plus élevé. Par contre, la fréquence des visites à domicile (VDOMI) est un indicateur pur de (t) parce que ces visites sont toujours mal rémunérées. Aussi, le tarif connaît une position spéciale pour couvrir des consultations de durée exceptionnelle, appelée « facteur temps » (FACTT). Il va sans dire qu'il y aura des abus, des médecins facturant cette position sans passer de temps additionnel avec leur patient. Enfin, nous disposons de trois indicateurs liés uniquement à l'élément prix (q) de YTOTAL. Ni les travaux de laboratoire (YLAB), ni la radiographie (YRAD), ni la

propharmacie à propre compte (YPRO) n'usent beaucoup de temps du médecin. Normalement, ces travaux sont délégués au personnel auxiliaire; ils servent donc à augmenter les recettes par unité de temps, soit le taux de salaire implicite (q).

Le noyau du modèle, ce sont les relations structurelles entre ($\bar{s}$, r, w_Y) et (q, t). Les *propositions* 4 *et* 5 sont à la base des restrictions $e(q, r) < 0$ et $e(t, r) > 0$; (pour les autres, *cf.* Zweifel [26], chap. 5 et 6). Comme les médecins ne réussissent pas toujours à réaliser les valeurs optimales (q^*) et (t^*), des erreurs structurelles sont introduites. Celle influençant (q) est standardisée, sa variance est égale à 1.

5. ESTIMATION DU MODÈLE

Après transformation logarithmique des observations, le modèle du graphique 2 a été estimé par la méthode de la vraisemblance maximale dans le cadre de l'analyse structurelle des covariances, développée par Joereskog [13, 14] (*cf.* l'annexe pour les détails). Or, l'identification du système devient un problème difficile aussitôt que des erreurs d'observation sont admises dans les variables exogènes (Goldberger [11], Malinvaud [18], chap. 10). Elle est pourtant atteinte si les observations de 2 années sont introduites pour chaque indicateur sauf SEMIUR et URBN, à condition naturellement que la qualité des indicateurs soit la même dans les deux périodes. Autrement, le nombre des erreurs d'observation et des coefficients de charge aurait doublé lui aussi. La supposition d'une qualité inaltérée des observations peut être justifiée mieux pour les années 1976 et 1977 que pour 1977 et 1978 parce que la caisse-maladie CMB a changé la structure tarifaire en juillet 1978. Ce changement vise à encourager le médecin à sacrifier plus de son temps au traitement du patient plutôt qu'à avoir recours à des procédures techniques. Cela nous amène à formuler la :

PROPOSITION 8. — *L'hypothèse d'une qualité inchangée des indicateurs relative aux quantités latentes du modèle peut être défendue mieux pour les années* 1976 *et* 1977 *que pour* 1977 *et* 1978.

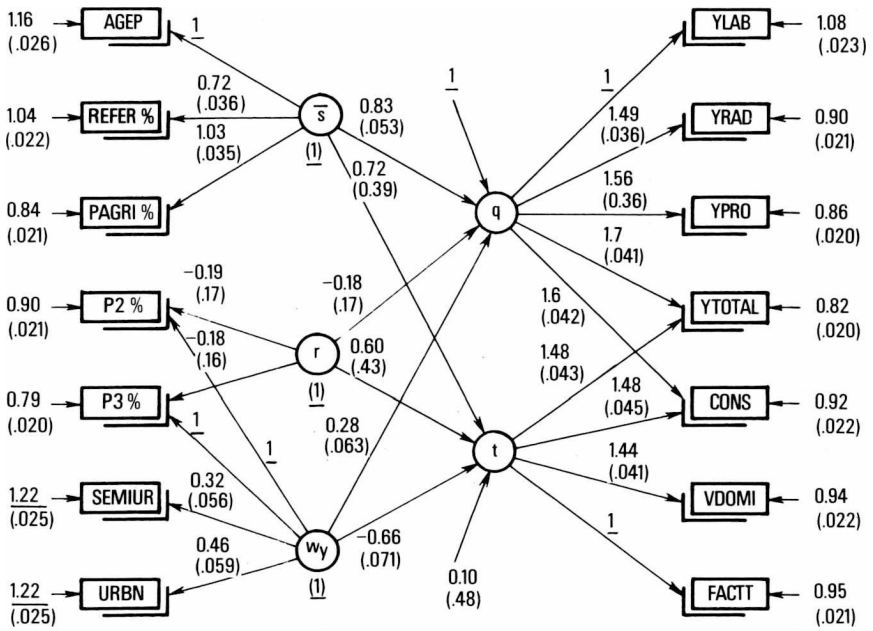
Mais d'abord, voyons si les implications du modèle sont vérifiées. Le graphique 3 contient des résultats assez encourageants. Les six élasticités structurelles correspondent aux prédictions qualitatives formulées dans le cadre du graphique 2. En particulier, $\hat{e}(q, r) = -0.18$ semble une valeur raisonnable; malheureusement, l'écart-type de 0.17 la rend indistinguable de zéro. Du côté exogène, REFER % et PAGRI % sont des indicateurs directs de l'intensité symptomatique moyenne ($\bar{s}$), mais très altérés par des erreurs d'observation. Aussi, l'erreur d'observation contenue dans AGEP aurait elle-même une variance estimée de 1.35 ($=1.16^2$), ce qui dépasse la variance unitaire de l'indicateur standardisé. Quant à P2 % et P3 %, leurs coefficients de charge font penser à une relation inverse avec le taux de participation (r) prévalant dans la population. Les résultats sont moins ambigus du côté endogène. Le logarithme des frais décomptés par cas, YTOTAL, peut en effet être interprété comme la somme d'un élément prix, avec coefficient de charge $\lambda(\text{YTOTAL}, q) = 1.7$

(.041), et un élément intensité avec $\hat{\lambda}$(YTOTAL, t) = 1.48 (.043). Les influences relatives sont très similaires pour CONS. Notre discussion peut être résumée dans la :

PROPOSITION 9. — *Les décomptes cachent un élément taux de salaire* (q) *et un élément temps* (t), *comme prédit par le modèle théorique. Mais les effets liés à des différences de la participation aux frais* (r) *prévalant dans la population n'ont pas été identifiés de façon claire.*

GRAPHIQUE 3

Estimation LISREL, 616 médecins 1976 et 1977



Remarque. — Les écarts-type asymptotiques sont donnés entre parenthèses, et les paramètres soumis à des restrictions sont soulignés. $\chi^2 = 9\,487$, d.l. = 318.

Peut-être que l'hypothèse d'une qualité constante des indicateurs est déjà restrictive sous des conditions normales. Comme YTOTAL est la variable généralement observée et commentée, admettons que l'erreur d'observation puisse avoir une importance changeante dans les 2 années 1976 et 1977. Le graphique 4 montre que ce petit détail a des conséquences considérables. Les six élasticités structurelles correspondent toujours aux exigences du modèle théorique, mais leurs valeurs manquent de stabilité. Comme exemple extrême, $e(t, \bar{s})$ serait alors de 0.033 (.031), comparé à 0.72 (.039) antérieurement.

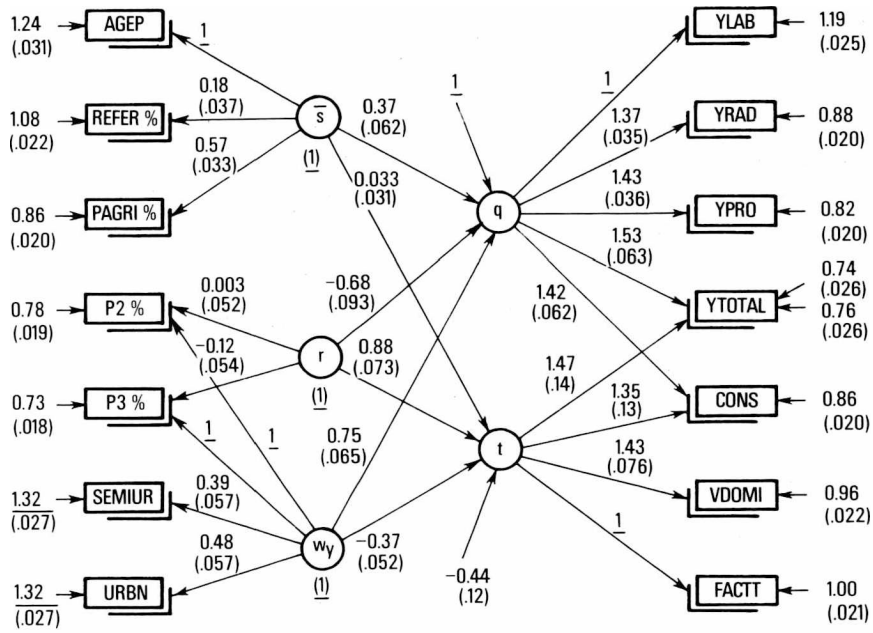
Du côté exogène, les coefficients de charge liés à ($\bar{s}$) ont subi les plus grands changements, la portion de cas référés par tiers (REFER %) ne reflétant qu'à un degré fort diminué des différences en ($\bar{s}$). L'interprétation du côté endogène reste

la même. En particulier, YTOTAL est toujours fonction du taux de salaire implicite (q) aussi bien que du temps dépensé par cas (t) à pondérations à peu près égales. La portion estimée de l'erreur d'observation dans la variance de YTOTAL est 55 % ($=.74^2$) pour 1976 et 58 % ($=.76^2$) pour 1977. Cette expérience nous mène à énoncer la :

PROPOSITION 10. — *Les erreurs de mesure jouent un rôle décisif. Elles sont l'élément ayant la plus grande signification statistique du modèle estimé. De plus, le noyau structurel n'est pas entièrement isolé de leur impact, en dépit de leur reconnaissance explicite.*

GRAPHIQUE 4

Qualité variable de l'indicateur YTOTAL, 616 médecins 1976 et 1977



Remarque. — $\chi^2 = 8\,993$, d.l. $= 317$.

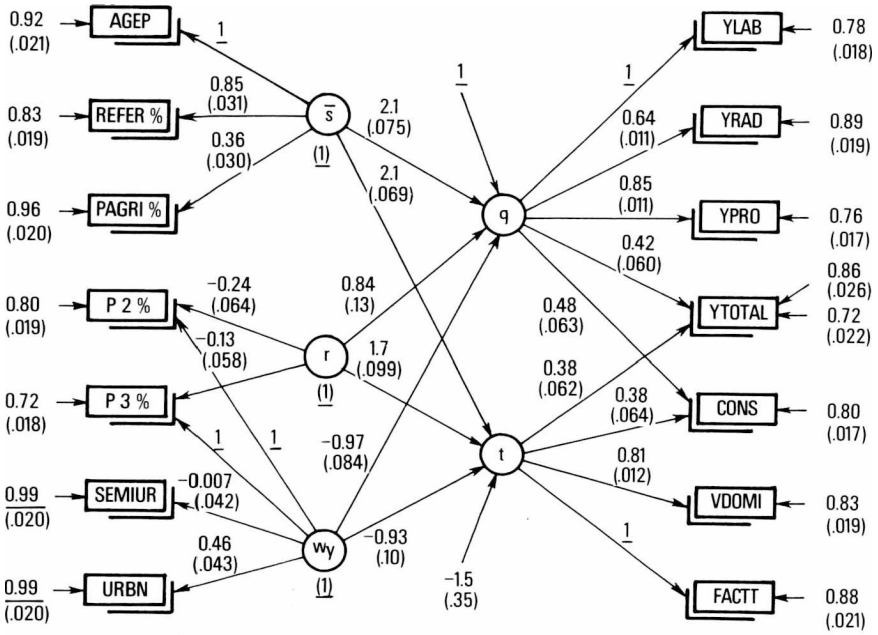
Le même procédé est alors appliqué aux données des années 1977 et 1978, où l'on peut supposer qu'un changement de qualité a eu lieu. Selon le graphique 5, il y aurait en effet une différence marquée entre l'indicateur YTOTAL de l'année 1977 et celui de l'année 1978. Mais ce serait dans la première période que sa qualité est inférieure, la part de la variance attribuable à l'erreur d'observation montant à 74 % ($=0.86^2$). Par contre, cette portion est moins importante dans la deuxième période, avec 52 % ($=0.72^2$).

D'autres résultats ne correspondent pas non plus à nos attentes. Deux élasticités structurelles sont en contradiction avec les implications du modèle, $\hat{e}(q, r) > 0$ et $e(q, w_Y) < 0$. En même temps, les indicateurs YRAD et YPRO

(radiographie et propharmacie) ne reflèteraient plus entièrement les différences dans le taux de salaire implicite (q). En somme, il semble y avoir assez d'instabilité, mais nous n'en avons pas trouvé la raison.

GRAPHIQUE 5

Qualité variable de l'indicateur YTOTAL, 616 médecins 1977 et 1978



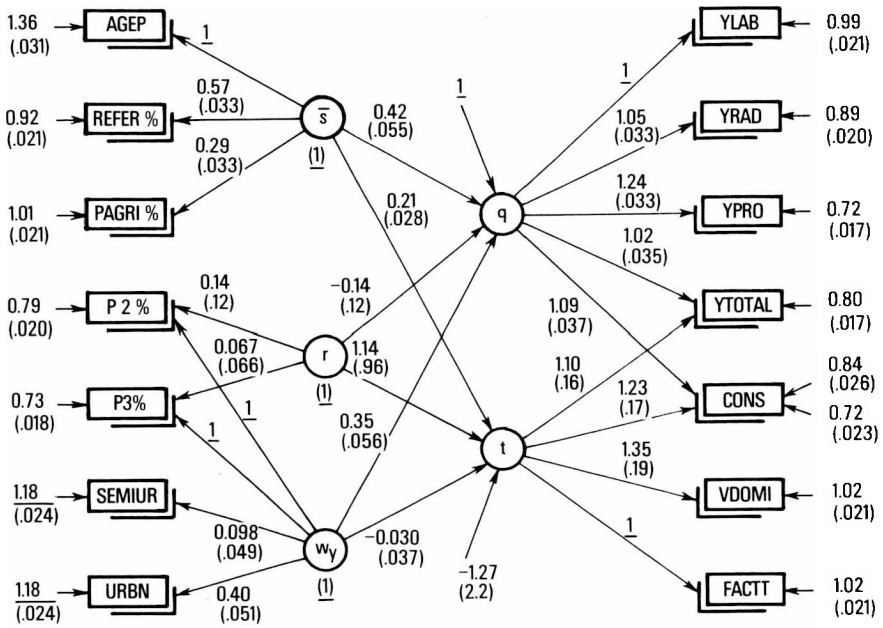
Remarque. — $\chi^2 = 12\,718$, d.l. $= 317$.

Peut-être que YTOTAL n'est pas l'indicateur le plus affecté par le nouveau tarif, entré en vigueur en juillet 1978. Car ce changement ne doit pas nécessairement avoir un impact sur les coefficients des éléments (q) et (t). Et si les coefficients de charge changeaient, cela n'est pas nécessairement reflété dans l'erreur d'observation. En fait, ce sont les consultations (CONS) qui ont été rémunérées mieux en comparaison avec les procédures techniques dans la nouvelle structure tarifaire. Ce changement de structure devrait encourager les médecins qui consacrent beaucoup de temps à leurs patients à se faire reconnaître. Mais les « praticiens à la minute » eux aussi, chercheraient à exploiter cette position tarifaire CONS afin d'en tirer un taux de salaire (q) élevé. Donc, cet indicateur devrait représenter les deux éléments (q) et (t) plus précisément après le changement qu'avant.

Cette hypothèse a été soumise à un test préliminaire dans le graphique 6. Là, c'est alors CONS plutôt que YTOTAL dont l'erreur d'observation est libre de varier entre 1977 et 1978. Les conséquences en sont frappantes. Tout d'un coup, les élasticités structurelles ont les signes attendus avec $\hat{e}(q, r) = -0.14\,(.12)$ dans le voisinage de la valeur $-0.18\,(.17)$, estimée sur la base des données 1976 et 1977. Les autres élasticités ont aussi des valeurs raisonnables; seule $\hat{e}(t, w_Y)$

GRAPHIQUE 6

Qualité variable de l'indicateur CONS, 616 médecins 1977 et 1978



Remarque. — $\chi^2 = 12\,945$, d.l. $= 317$.

manque de stabilité. Quant aux coefficients de charge, ils correspondent assez bien à ceux du graphique 3 confirmant l'impression que YTOTAL est composé de deux facteurs (q) et (t) ayant des importances à peu près égales. Enfin, CONS s'avère être un indicateur plus sûr en 1978 qu'en 1977, la portion de sa variance due à l'erreur d'observation diminuant de 71 % ($=0.84^2$) à 52 % ($=0.72^2$). Ainsi, nous terminons notre discussion avec la :

PROPOSITION 11. — *Si le changement de qualité est reconnu là où il devrait être décisif, les paramètres estimés sont assez stables, et les propositions* 4 *à* 9 *sont vérifiées.*

CONCLUSION

Dans cet article, nous avons exposé les caractéristiques principales d'un modèle économique du comportement médical. Le médecin y dispose de trois instruments pour atteindre ses objectifs : un seuil critique d'intensité symptomatique (c), un taux de salaire implicite (q) et le temps moyen (t) qu'il dépense pour un seul cas. Les frais décomptés par cas doivent donc correspondre au produit ($q \cdot t$), indépendamment des prix fixés pour les actes médicaux par les caisses-maladie. Le modèle théorique prédit une réduction de l'élément (q) si le taux de participation aux frais (r) est augmenté pour les assurés. Mais il prédit

aussi que le médecin passera plus de temps avec un patient, en moyenne. Par conséquent, la réaction du produit observé ($q.t$) n'est pas déterminée. Dans l'estimation, ces deux éléments peuvent être identifiés dans les décomptes d'une grande caisse-maladie suisse. Les résultats du graphique 3 se prêtent au calcul illustratif suivant. Une augmentation du taux de participation (r) de 10 à 20 %, soit de 100 %, induirait les médecins à baisser leur taux de salaire implicite de 18 % [$\hat{e}(q, r) = -0.18$]. En revanche, leurs services changeraient de qualité, devenant plus intensifs quant au temps dépensé. Cet effet l'emporterait parce qu'il serait de 60 % [$\hat{e}(t, r) = 0.60$]. Bien que le résultat final soit altéré par des erreurs d'observation, on verrait les frais par cas, équivalant à ($q.t$), augmenter plutôt que baisser. Un tel phénomène ferait certainement penser à une « demande induite par le médecin ». Mais les préférences des patients n'ont pas changé, selon hypothèse préalable. C'est plutôt la manière dont les caisses-maladie rémunèrent les médecins qui explique les phénomènes observés.

ANNEXE

L'analyse structurelle de covariances par LISREL

Nous nous bornons ici à esquisser les idées principales développées par Joereskog [12, 13]; le programme a été documenté pour la première fois dans Joereskog et van Thillo [14].

Le noyau structurel d'un modèle statistique comme celui présenté dans le graphique 2 consiste en un système d'équations linéaires :

$$\Gamma \mathbf{y} = B \mathbf{x} + \mathbf{v}. \tag{A.1}$$

Le vecteur $\mathbf{y}$ ($m \times 1$) contient les variables endogènes, le vecteur $\mathbf{x}$ ($g \times 1$) les variables exogènes du système. Les m variables aléatoires $\mathbf{v}$ sont les erreurs de la régression classique. Mais ici, non seulement $\mathbf{v}$, mais aussi $\mathbf{y}$ et $\mathbf{x}$ sont latents. Ils produisent des valeurs observées selon les systèmes d'équations :

$$\begin{cases} \mathbf{i} = \boldsymbol{\mu} + K \mathbf{x} + \boldsymbol{\delta}, & \mathbf{i} : \ (g' \times 1) \\ \mathbf{j} = \boldsymbol{\nu} + \Lambda \mathbf{y} + \boldsymbol{\varepsilon}, & \mathbf{j} : \ (m' \times 1). \end{cases} \tag{A.2}$$

Les vecteurs ($\mathbf{i}$, $\mathbf{j}$) sont des indicateurs avec espérances mathématiques ($\boldsymbol{\mu}$, $\boldsymbol{\nu}$), liées aux variables théoriques par les matrices de coefficients de charge (K, Λ) pas nécessairement diagonales. Leurs valeurs observées dépendent aussi des erreurs d'observation ($\boldsymbol{\delta}$, $\boldsymbol{\varepsilon}$), qui, pour être distinguées des erreurs structurelles $\mathbf{v}$, doivent satisfaire à des restrictions. En particulier, on suppose :

$$\begin{gathered} \mathbf{v} \sim N(0, V), \\ \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left[\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} U & 0 \\ 0 & W \end{bmatrix}\right], \\ U, W \text{ diagonaux}, \\ E(\boldsymbol{\delta} \mathbf{v}') = 0, \\ E(\boldsymbol{\varepsilon} \mathbf{v}') = 0. \end{gathered} \tag{A.3}$$

Sous ces conditions, l'analyse statistique peut être basée sur la matrice de covariances créée par les indicateurs. Car nous avons :

$$(A.4)\quad \Sigma_{ii} = E\{(\nu + K\mathbf{x} + \boldsymbol{\delta})(\nu + K\mathbf{x} + \boldsymbol{\delta})' - (\nu + K\nu)(\nu + K\nu)'\}$$
$$= K.\mathrm{Var}(\mathbf{x})K' + \mathrm{Var}(\boldsymbol{\delta}) + 0 \quad \text{à cause de (A.3),}$$

$$(A.5)\quad \Sigma_{ij} = E\{(\quad + K\mathbf{x} + \boldsymbol{\delta})(\mu + \Lambda\mathbf{y} + \boldsymbol{\varepsilon})' - (\nu + K\nu)(\mu + \Lambda\mu)'\}$$
$$= K.E\{(x - K\nu)(y - \Lambda\mu)'\}\Lambda' + 0$$
$$= K.\mathrm{Var}(\mathbf{x})(\Gamma')^{-1}\Lambda' \quad \text{à cause de (A.3),}$$

$$(A.6)\quad \Sigma_{jj} = E\{(\boldsymbol{\mu} + \Lambda\mathbf{y} + \boldsymbol{\varepsilon})(\boldsymbol{\mu} + \Lambda\mathbf{y} + \boldsymbol{\varepsilon})' - (\mu + \Lambda\mu)(\mu + \Lambda\mu)'\}$$
$$= \Lambda.\mathrm{Var}\{\Gamma^{-1}(B\mathbf{x} + \boldsymbol{\nu})\}.\Lambda' + \mathrm{Var}(\boldsymbol{\varepsilon})$$
$$= \Lambda\{\Gamma^{-1}B.\mathrm{Var}(\mathbf{x}).B'(\Gamma')^{-1} + \Gamma^{-1}\mathrm{Var}(\boldsymbol{\nu})(\Gamma')^{-1}\}\Lambda' + \mathrm{Var}(\boldsymbol{\varepsilon}).$$

Les équations (A.4)-(A.6) découvrent les relations bien définies entre les paramètres inconnus du système et la matrice des covariances Σ, qui est représentée de manière cohérente par la matrice des covariances S observée. La matrice $\hat{\Sigma}$, calculée sur la base des paramètres estimés [$\hat{\Gamma}$, $\hat{B}$; $\hat{\ }$, $\hat{\Lambda}$; $\mathrm{Var}(\mathbf{x})$; $\mathrm{Var}(\boldsymbol{\delta})$, $\mathrm{Var}(\boldsymbol{\varepsilon})$] sera plus ou moins différente de S. La tâche du programme LISREL est de minimiser cet écart selon le critère :

$$(A.7)\qquad \ln f = \frac{-1}{2}N\{\ln|\hat{\Sigma}|^{-1}S)\},$$

N, nombre d'observations qui est le logarithme de la fonction de vraisemblance sous les conditions (A.3). En comparaison avec le cas général, le modèle représenté par le graphique 2 est plus simple, avec Γ égal à la matrice d'identité I_2. En revanche, ni K ni Λ ne sont diagonaux parce que P2 % et P3 % reflètent (r) et (w_Y) et YTOTAL est fonction de (q) et (t). Des règles d'identification pour un tel système ne semblent pas encore exister sauf pour des cas spéciaux, *cf.* Geraci [10]. Par contre, une matrice d'information singulière est une indication empirique presque sûre d'un manque d'identification (Silvey [24], p. 81-82).

BIBLIOGRAPHIE

[1] ACTON (J. P.), Nonmonetary Factors in the Demand for Medical Service: Some Empirical Evidence, *Journal of Political Economy*, vol. 83/3, 1975, p. 595-614.

[2] ACTON (J. P.). Demand for Health Care among the Urban Poor, with Special Emphasis on the Role of Time, dans ROSETT (R. N.), éd., *The Role of Health Insurance in the Health Services Sector*, Universities-NBER Conference Series, n° 27, New York, Neale Watson, 1976, p. 165-208.

[3] ARCHIBALD (G. C.), The Qualitative Content of Maximizing Models, *Journal of Political Economy*, vol. 73, n° 1, février 1965, p. 27-36.

[4] ARROW (K. J.), *Essays in the Theory of Risk Bearing*, Amsterdam, North Holland, 1971.

[5] ARROW (K. J.), Uncertainty and the Welfare Economics of Medical Care, *American Economic Review*, décembre 1973, p. 941-973.

[6] BROWN (J. A. C.) et DEATON (A. S.), Survey in Applied Economics: Models of Consumer Behavior, *The Economic Journal*, vol. 82, décembre 1972, p. 1145-1236.

[7] EVANS (R. G.), Supplier-Induced Demand: Some Empirical Evidence and Implications, dans PERLMAN (M) (éd. pour International Economic Association), *The Economics of Health and Medical Care*, London, MacMillan, 1974, p. 162-173.

[8] FELDSTEIN (M. S.), The Rising Price of Physicians' Services, *Review of Economics and Statistics*, vol. 52, n° 2, mai 1970, p. 121-133.

[9] FELDSTEIN (M. S.), Econometric Studies of Health Economics, dans INTRILIGATOR (M. D.) et KENDRICK (D. A.), éds., *Frontiers of Quantitative Economics*, vol. II, Contributions to Economic Analysis, n° 87, Amsterdam, North Holland, 1974, p. 377-434.

[10] GERACI (V. C.), Identif cation of Simultaneous Models with Measurement Error, *Journal of Econometrics*, vol. 4, 1976, p. 263-284.

[11] GOLDBERGER (A. S.), Unobservable Variables in Econometrics, dans ZAREMBKA (P.), éd., *Frontiers in Econometrics*, New York, Academic Press, 1974, p. 193-213.

[12] JOERESKOG (K. G.), A General Method for Estimating a Linear Structural Equation System, dans GOLDBERGER (A. S.) et DUNCAN (O. D.), éds., *Structural Equation Systems in the Social Sciences*, New York, Seminar Press, 1973, p. 85-112.

[13] JOERESKOG (K. G.), Structural Equation Models in the Social Sciences: Specification, Estimation, and Testing, dans KRISHNAIAH (R. P.), éd., *Applications of Statistics*, Amsterdam, North Holland, 1977, p. 265-286.

[14] JOERESKOG (K. G.) et VAN THILLO (M.), LISREL. A General Computer Program for Estimating a Linear Structural Equation System involving Multiple Indicators of Unmeasured Variables, Dept. of Statistics Uppsala, *Research Report* 73-5, 1973.

[15] KESSEL (R. A.), Price Discrimination in Medicine, *Journal of Law and Economics*, vol. 1, 1958, p. 20-53.

[16] LAVE (J.), LAVE (L.) et LEINHARDT (S.), Modeling the Delivery of Medical Services, dans PERLMAN (M.) (éd. pour International Economic Association), *The Economics of Health and Medical Care*, London, MacMillan, 1974, p. 326-351.

[17] LUFT (H. S.), Comment to NEWHOUSE (J. P.) et PHELPS (CH. E.), New Estimates of Price and Income Elasticities of Medical Services, dans ROSETT (R. N.), éd., *The Role of Health Insurance in the Health Services Sector*, Universities-NBER Conference Series, n° 27, New York; Neale Watson), 1976, p. 317-320.

[18] MALINVAUD (E.), *Méthodes Statistiques de l'Économétrie*, Paris, Dunod, 1964.

[19] MASSON (R. T.) et WU (S.), Price Discrimination for Physicians' Services, *The Journal of Human Resources*, vol. IX/1, 1974, p. 64-79.

[20] MENNEMEYER (S. T.), Really Great Returns to Medical Education?, The Journal of Human Resources, vol. 8/1, 1975, p. 75-90.

[21] NEWHOUSE (J.), A Model of Physician Pricing, *Southern Economic Journal*, vol. 37, n° 2, octobre 1970, p. 174-183.

[22] NEWHOUSE (J. P.) et PHELPS (Ch. E.), New Estimates of Price and Income Elasticities of Medical Services, dans ROSETT (R. N.), éd., *The Role of Health Insurance in the Health Services Sector*, Universities-NBER Conference Series, n° 27, New York, Neale Watson, 1976, p. 261-312.

[23] SILVEY (S. D.), *Statistical Inference*, London, Chapman et Hill, 1975.

[24] SLOAN (F. A.), A Microanalysis of Physicians' Hours of Work, dans PERLMAN (M.) (éd. pour International Economic Association), *The Economics of Health and Medical Care*, London, MacMillan, 1974, p. 302-325.

[25] SMALLWOOD (D. E.) et SMITH (K. R.), Optimal Treatment Decisions, Optimal Fee Schedules and the Allocation of Medical Resources, Health Economics Research Center, Madison: *Research and Analytical Report Series*, 2/75, 1975.

[26] ZWEIFEL (P.), *Ein ökonomisches Modell des ärztlichen Verhaltens*, Habilitationsschrift (Un modèle économique du comportement médical, thèse d'agrégation), 1980.