GREMAQ
Université des Sciences Sociales
Place Anatole France
31042 Toulouse Cedex

COUT ET EFFICACITE DES DEPENSES DE SANTE

RAPPORT POUR LE COMMISSARIAT GENERAL DU PLAN

Décembre 1998

Ont participé à ce rapport :

Jean-Charles Rochet (Responsable Scientifique)
Helmuth Cremer, Jean-Paul Cresta, Marc Ivaldi, Gwenaël Piaser,
Denis Raynaud (GREMAQ)
Marie Allard et Marcel Dagenais (Université de Montréal),
Agnès Couffinhal (CREDES),
Maurice Marchand (CORE).

Le présent document constitue le rapport scientifique d'une recherche financée par le Commissariat Général du Plan (subvention n° 12-1996).

SOMMAIRE

I. RESUME DES RECHERCHES

II. MESURE ET COMPARAISON DES PERFORMANCES DES SYSTEMES DE SANTE DES PAYS DE L'OCDE (Marcel DAGENAIS et Marc IVALDI)

III. REMUNERATION DES MEDECINS, INCITATIONS ET COÛT DES SOINS DE SANTE (Marie ALLARD, Helmuth CREMER et Maurice MARCHAND)

IV. LE MARCHE DE L'ASSURANCE COMPLEMENTAIRE

IV. 1 La prise en compte des rendements croissants dans la gestion des contrats d'assurance
[Titre anglais : Pooling and Separating Equilibria in Insurance Markets with Adverse Selection and Distribution Costs]
(Marie ALLARD, Jean-Paul CRESTA et Jean-Charles ROCHET)

IV.2 Consultation médicale : l'influence du revenu et de l'assurance complémentaire
(Gwenaël PIASER et Denis RAYNAUD)

V. COUVERTURE MALADIE UNIVERSELLE OU ACCES GRATUIT AUX SOINS
(Agnès COUFFINHAL et Jean-Charles ROCHET)

COUT ET EFFICACITE DES DEPENSES DE SANTE
RAPPORT REALISE PAR LE GREMAQ
POUR LE COMMISSARIAT GENERAL DU PLAN

RESUME

Le but de notre recherche était d'améliorer, à plusieurs niveaux complémentaires, les instruments de mesure déjà disponibles en matière de coût et d'efficacité du système de santé. Nous avons réalisé 5 études :

1) Mesure et comparaison des performances des systèmes de santé des pays de l'OCDE :

Marcel Dagenais et Marc Ivaldi utilisent les méthodes de frontières de production stochastiques pour évaluer l'efficacité des systèmes de santé des pays de l'OCDE. Ils montrent que :

* les dépenses publiques de santé ont un impact important sur l'état de santé moyen dans les pays concernés,

* le Canada et la Suède sont les pays les plus efficaces de l'échantillon, alors que la France et les USA ont un niveau d'efficacité faible, eu égard au montant de leurs dépenses de santé,

* sur la période considérée (1960-1994) l'efficacité du système de santé français a diminué, malgré l'augmentation de ses dépenses.

2) Rémunération des médecins, incitations et coût des soins de santé

Marie Allard, Helmuth Cremer et Maurice Marchand développent un modèle théorique de comportement de médecins qui leur permet d'établir les caractéristiques d'un schéma de

rémunération qui permettrait d'arbitrer au mieux entre la qualité des traitements et la maîtrise des dépenses de santé. Ils montrent en particulier que contrairement à ce que l'on pourrait penser, la rémunération marginale des médecins devrait croître (et non décroître) en fonction du nombre de leurs patients, mais que par contre les médecins devraient supporter une partie du coût de leurs prescriptions pharmaceutiques.

3) La prise en compte des rendements croissants dans la gestion des contrats d'assurance

Marie Allard, Jean-Paul Cresta et Jean-Charles Rochet modifient le modèle de référence du marché d'assurance avec antisélection, dû à Rothschild et Stiglitz. Ils montrent que la prise en compte de rendements croissants dans la gestion des contrats conduit à des prédictions beaucoup plus conformes à la réalité empirique : les contrats mélangeants peuvent apparaître à l'équilibre et ce sont en général les individus les plus risqués qui sont rationnés, à l'opposé de ce qui se passe dans le modèle de Rothschild et Stiglitz.

4) Consultation médicale : l'influence du revenu et de l'assurance complémentaire

Gwenaël Piaser et Denis Raynaud étudient l'influence du revenu et de la couverture complémentaire des ménages sur leur fréquence de consultation médicale à partir de données tirées de l'enquête santé 1991-92 de l'INSEE.

Dans une première partie ils estiment la probabilité de consultation d'un médecin en autorisant des effets non monotones pour le revenu. Les résultats obtenus montrent que le revenu a un effet positif sur la probabilité de consultation pour les individus disposant d'une couverture complémentaire (assurance ou mutuelle), mais cet effet devient négatif à partir d'un certain seuil pour les individus ne disposant pas d'une couverture complémentaire.

Dans une seconde partie ils estiment l'influence de différentes variables sur l'état de santé tel qu'il est perçu par les individus eux-mêmes. Les auteurs montrent que les individus les plus « pauvres » considèrent généralement qu'ils sont en plus mauvaise santé alors que toutes choses égales par ailleurs, les individus les plus « riches » se considèrent en meilleure santé.

5) Couverture maladie universelle ou accès gratuit aux soins

Par un modèle aussi simple que possible, Agnès Couffinhal et Jean-Charles Rochet essaient de capturer les raisons pour lesquelles plus de 40 millions d'Américains renoncent à s'assurer contre la maladie. Ils utilisent ensuite ce modèle pour évaluer les effets qu'aurait eu l'introduction, prévue dans le plan Clinton, d'une couverture maladie universelle aux USA.

I - Résumé des Recherches

La documentation Française : "Coût et efficacité des dépenses de santé : rapport / Université des sciences sociales, GREMAQ ; Commissariat général du Plan ; Jean-Charles Rochet, Helmuth Cremer, Jean-Paul Cresta, \(et al.\)."

I. RESUME DES RECHERCHES

I.1- Mesure et comparaison des performances des systèmes de santé des pays de l'OCDE

L'efficacité du système de santé français est souvent mise en question : la comparaison brute entre le montant des dépenses de santé par habitant et le niveau des indicateurs de santé les plus simples ne semble pas mettre la France dans le peloton de tête des pays de l'OCDE. Or dans toute comparaison de ce type, la difficulté principale consiste à isoler, dans les facteurs explicatifs de ces indicateurs de santé des différents pays de l'OCDE, ce qui provient réellement des caractéristiques du système de santé. Une étude récente de l'OCDE établit en effet que les facteurs exogènes (démographie, niveau de vie, habitudes alimentaires,...) propres à chaque pays expliquent l'essentiel des écarts entre indicateurs de santé des différents pays.

Fort heureusement, cette difficulté peut désormais être surmontée grâce à la conjonction de deux éléments nouveaux :

* les nouvelles méthodes de l'économétrie de la production,

* la disponibilité de données de panel.

Le but de cette étude était d'exploiter ces deux éléments nouveaux pour obtenir une véritable comparaison des performances des systèmes de santé des principaux pays de l'OCDE.

Depuis les travaux de Farell, la question de la mesure de l'efficacité dans les décisions de production est un sujet en perpétuel renouvellement tant du point de vue théorique que de celui des applications en économie de la production. De telles mesures ont un intérêt pour évaluer les conditions de production d'une unité particulière (entreprise ou service), mais aussi pour étudier l'impact de mesures de politique économique sur les décisions de production dans un secteur donné.

Plusieurs approches sont communément utilisées en vue de calculer de telles mesures. Elles ont toutes pour objectif le recouvrement de la frontière des points décrivant les décisions optimales de production. Nous suivons ici une approche paramétrique stochastique. Dans cette approche la notion d'efficacité est introduite via la spécification du terme d'erreur qui est intégré à la fonction de production. Une première composante de ce terme mesure l'oubli de facteurs qui peuvent influer

négativement ou positivement sur le niveau de production. La seconde est supposée négative ou nulle, et mesure l'écart entre la production observée et la frontière de production. On mesure ainsi l'inefficacité de l'unité de production étudiée. Cet écart représente l'omission lors de la spécification de la frontière, de variables qui ont un effet borné sur la production. Le fait que ces variables n'atteignent pas leurs niveaux optimaux est la source de l'inefficacité observée. De telles variables peuvent être l'effort productif des managers, leurs pratiques d'organisations,...

Cette approche connaît un développement important avec la disponibilité de données de panel qui permettent de s'abstraire d'hypothèses trop restrictives concernant notamment la spécification de la seconde composante du terme d'erreur. Ainsi cette composante peut être rendue dépendante d'effets temporels ou de variables exogènes mesurables liées aux sources de l'inefficacité. De plus les données de panel permettent de mieux contrôler les effets des corrélations entre les quantités de facteurs utilisés dans le cycle de production et la composante d'inefficacité, effets qui tiennent au fait que les sources de l'inefficacité jouent un rôle dans l'allocation des ressources.

En utilisant cette approche, nous avons étudié l'efficacité des systèmes de santé des pays de l'OCDE, en testant différents indicateurs de santé (espérance de vie à 60 ans, mortalité infantile, et années de vie potentielle perdues.) pour mesurer la «production». Comme facteurs de production, nous avons considéré le nombre de lits d'hôpitaux occupés pour 1000 habitants, le nombre de médecins, pour 10 000 habitants et les dépenses pharmaceutiques par habitant. Enfin comme facteurs exogènes pouvant expliquer des variations d'efficacité de pays à pays ou d'année à d'année, nous avons introduit la part du financement public dans les dépenses de santé, la durée moyenne d'hospitalisation et la consommation d'alcool et de tabac par habitat.

I.2 Rémunération des médecins, incitations et coût des soins de santé

Les travaux de recherche que nous présentons ici ont été réalisés en collaboration avec Marie Allard (HEC, Montréal, Canada), Helmuth Cremer et Maurice Marchand (CORE, Université Catholique de Louvain, Belgique). Nous commençons par rappeler les principales caractéristiques de notre étude: objectifs, méthodologie, spécificité de l'approche et questions fondamentales. Ensuite, nous donnons une présentation succincte du modèle formel (modèle de base et extension). Enfin, nous passons en revue les principaux résultats.

Notre étude porte sur la détermination du schéma de rémunération approprié pour les médecins dans un contexte d'information asymétrique. Il s'agit là d'un des principaux problèmes que pose aujourd'hui l'organisation du secteur de la santé dans la plupart des pays industrialisés. Nous utilisons un modèle de type principal-agent, adapté aux spécificités du secteur de la santé. En particulier, notre modèle prend en compte les aspects suivants:

* La nature particulière de l'output, qui est reflétée dans la caractérisation de la « technologie de production » que nous considérons.
* L'altruisme ou l'éthique professionnelle dont peuvent faire preuve les médecins. Cet aspect est pris en compte dans la spécification de leur fonction objectif.
* L'existence d'un marché qui donne aux patients la possibilité de choisir leur médecin.
* L'importance des coûts occasionnés par les prescriptions (ou analyses) qui peuvent contribuer à l'amélioration de l'état de santé des patients et qui peuvent être substituts ou compléments d'autres « facteurs de production» tel que le temps que consacre le médecin au patient (ou plus généralement son effort).

Rappelons par ailleurs que nous adoptons un point de vue qui peut être qualifié « d'européen », où le principal consiste en un régulateur public (l'organisme d'assurance sociale, la sécurité sociale etc.) bénévole (dont l'objectif consiste à maximiser le bien-être social, tel qu'il sera défini plus loin).

Les principales questions auxquelles nous avons tenté d'apporter des éléments de réponse sont les suivantes:

* Quelle doit être la forme du schéma de rémunération d'un médecin en fonction, par exemple, du nombre de patients (visites)? En particulier, est-il souhaitable d'avoir un schéma (approximativement) linéaire ou faut-il, au contraire, des schémas convexes ou concaves (c.à.d. ou le paiement par patient augmente ou diminue en fonction du nombre des patients du médecin considéré)?

* Dans quelle mesure convient-il de faire participer le médecin au coût de ses prescriptions? En d'autres termes, faut-il « pénaliser » les médecins qui prescrivent beaucoup?

* Les asymétries d'information combinées à l'utilisation de schémas de paiements incitatifs vont-elles conduire à un nombre de médecins supérieur ou inférieur au nombre optimal (au premier rang)?

Le modèle que nous utilisons dans la première partie de l'article est le suivant. Il y a un certain nombre N (exogène) de patients homogènes. Les médecins, par contre, sont hétérogènes (deux types) et se différencient par leur efficacité (variable de sélection adverse). L'état de santé d'un patient dépend de l'efficacité de son médecin, ainsi que du niveau d'effort du médecin (variable de risque morale). L'utilité d'un patient dépend de son état de santé (effet positif) et du temps d'attente (effet négatif). A l'équilibre du marché tous les médecins (quel que soit leur type) doivent offrir la même utilité aux patients. Observons que le nombre de médecins (actifs) est déterminé de façon endogène à cet équilibre de marché. L'utilité d'un médecin (l'agent) dépend de son revenu (effet positif), de son effort (effet négatif) et de l'état de santé de ses patients (effet positif). L'objectif du régulateur (le principal) est de maximiser l'utilité du patient représentatif (évalué à l'équilibre du marché induit par le schéma de réglementation) en prenant en compte le coût social (coût des fonds publics) lié au financement des soins de santé. Son instrument est le schéma de payement (éventuellement non linéaire) octroyé aux médecins en fonction du nombre de patients.

La solution de ce problème dans un contexte d'information complète constitue un point de référence utile. Elle peut être mise en oeuvre par deux contrats linéaires (un par type de médecin). En information complète le paiement marginal (en fonction du nombre des patients) doit être le même pour les deux types de médecins et identique au « coût moyen d'un patient ». Il s'agit là simplement d'une condition « d'efficacité de la production». Le terme constant est différent selon les types de médecins et en l'occurrence c'est le médecin le plus efficace qui reçoit le paiement le plus faible (on suppose que le niveau d'utilité de réservation est le même pour les deux catégories). En conséquence, cette solution viole les contraintes d'incitation et elle ne peut pas être mise en oeuvre dans un contexte d'information asymétrique.

En information asymétrique, le schéma de rémunération devient plus complexe. Les médecins les plus productifs (type 2) continuent de recevoir un paiement marginal égal au coût moyen d'un patient. Par contre les médecins les moins productifs (type 1) reçoivent un payement marginal inférieur à cette valeur. Ceci conduit (toute autre chose étant égale) à une réduction du nombre de patients par médecins de type 1. Cette différenciation des paiements marginaux est a priori une source d'inefficacité. A bénéfice net (par patient) donné il est moins cher de faire soigner un patient par un médecin peu productif que par un médecin plus productif Cependant, cette distorsion s'avère globalement bénéfique dans la mesure ou elle permet de réduire les « rentes informationnelles » des médecins productifs.

Nous montrons par ailleurs que l'absence d'information complète conduit à une réduction du bénéfice net par patient et par ailleurs tend à conduire à une augmentation du nombre de médecins actifs.

Dans la deuxième partie de notre étude nous considérons une extension du modèle de base. Plus précisément nous introduisons un second « facteur de production» qui est constitué par les prescriptions des médecins (médicaments ou actes de diagnostic). Ce facteur est un substitut à l'effort du médecin: à qualité de soins donné, une médecin qui recourt plus intensément aux prescriptions doit faire moins d'effort (consacrer moins de temps à ses patients).

Nous montrons que la mise en oeuvre de la solution d'information complète requiert que tous les médecins supportent (à la marge) la totalité des coûts entraînés par leurs prescriptions. Si ce n'est pas le cas, « l'efficacité de la production » tend à être violée car les médecins tendent à avoir un recourt excessif à ce second facteur de production (dont ils ne prennent pas en compte la totalité des coûts sociaux).

En information asymétrique, la solution est de nouveau plus complexe. Les médecins productifs doivent supporter (à la marge) la totalité du coût de leurs prescriptions. Cependant, la contribution des médecins moins productifs est différente. Selon les cas ils peuvent être amenés à ne supporter qu'une partie de leurs prescriptions ou encore à faire face à des « pénalités » qui excèdent le coût marginal de leurs prescriptions. Comme pour le payement en fonction du nombre de patients, le schéma de prise en compte de frais de prescriptions conduit donc à une distorsion au niveau des médecins moins productif Par ailleurs, cette distorsion s'avère socialement bénéfique car elle permet de réduire les « rentes informationnelles » des médecins productifs. Cependant, la direction de la distorsion est maintenant ambiguë et l'intuition sous-jacente est plus subtile.

En conclusion, il convient de souligner que notre étude est basée sur un modèle très simple voire caricatural du secteur de la santé. En conséquence, les résultats doivent être interprétés avec précaution et les implications d'ordre politique doivent être considérés de façon très nuancée. Cependant, il nous semble que l'étude conduit à deux conclusions qui peuvent être très pertinentes dans le débat concernant la réforme des systèmes de santé.

Premièrement, les schémas de rémunération des médecins doivent prendre en compte les difficultés soulevées par les phénomènes d'asymétrie d'information. En d'autres termes, des mesures qui ignorent ces considérations peuvent conduire à des résultats contraires au but envisagé. Par

exemple, dans un souci de limiter les dépenses de la santé il est tentant d'envisager un schéma de rémunération « concave » des médecins. La rémunération par patient serait alors une fonction décroissante du nombre de patients. En d'autres termes, au-delà d'un certain seuil, les médecins recevraient une compensation moindre par patient (ou par acte). Notre étude montre qu'un tel schéma peut être tout à fait inapproprié. En l'occurrence, nos résultats suggèrent que le schéma de compensation doive plutôt être convexe et impliquer un paiement marginal plus élevé pour les médecins productifs (qui par ailleurs sont ceux dont le nombre de patients sera le plus élevé).

Deuxièmement, un système dans lequel les médecins ne participent pas du tout au coût impliqué par leurs prescriptions est à coup sûr sous-optimal. Il ne peut qu'entraîner une sur consommation de ses biens (ou service) et cela même si les médecins font preuve d'une certaine éthique professionnelle. Le système de compensation approprié doit reposer sur une « responsabilisation » des médecins fondée sur leur participation aux frais de leurs prescriptions. Le niveau précis de cette participation reste à déterminer et notre modèle est bien trop simple pour trancher cette question. Cependant, l'ambiguïté ne porte pas sur le principe même de la « participation » des médecins: le fait que leur participation marginale doit être positive semble être un résultat robuste. Ce qui est moins clair à ce stade est de savoir si cette participation doit être partielle ou, au contraire, doit excéder le véritable coût marginal de ces prescriptions.

I.3 - Le marché de l'assurance complémentaire

Par comparaison avec d'autres secteurs où les besoins d'assurance sont importants, le secteur de la santé a (en France et dans d'autres pays) la particularité de comporter une part obligatoire de couverture qui correspond à l'assurance maladie prise en charge par la sécurité sociale et financée par des prélèvements liés aux salaires. La forme de cette couverture obligatoire est très rigide et motivée par des arguments d'égalité des citoyens devant la maladie et d'égalité de l'accès aux soins.

Sans se prononcer sur le montant des prélèvements et les problèmes macro-économiques qu'ils soulèvent, il paraît légitime de s'intéresser à la forme optimale que devrait prendre l'intervention publique en terme d'assurance face aux risques de maladie. La couverture uniforme des assurés sociaux a des implications quant au fonctionnement du marché des assurances complémentaires et mutuelles, en particulier dans le cadre de modèles avec antisélection. A budget donné, il est nécessaire de comprendre comment ce marché d'assurance complémentaire pourrait

être modifié par un changement dans le principe de couverture sociale uniforme, une intervention impliquant une forme de discrimination dans l'assurance obligatoire pouvant avoir de meilleures performances tant en termes d'efficacité que de redistribution une fois tenu compte de l'équilibre sur de second marché.

Or, les modèles théoriques employés jusqu'ici pour décrire le fonctionnement de ce marché sont tous dérivés de l'article fondateur de Rothschild et Stiglitz (1976)), qui néglige un aspect fondamental : la présence de rendements croissants dans la gestion des contrats. La première étude de cette partie est consacrée aux conséquences théoriques de ces rendements croissants. La deuxième étude, quant à elle, est une analyse empirique de l'influence du revenu et de la couverture complémentaire sur la fréquence de consultation médicale des ménages français.

I.3.1 La prise en compte des rendements croissants dans la gestion des contrats

[Titre anglais : Pooling and Separating Equilibria in Insurance Markets with Adverse Selection and Marketing Costs]

Les implications du modèle de Rothschild et Stiglitz sont assez peu réalistes, si on les confronte à la réalité du fonctionnement du secteur de l'assurance maladie. En effet, ce modèle prédit (sous certaines conditions qui garantissent l'existence d'un équilibre concurrentiel) l'absence de mutualisation («pooling») entre différents types de risque, et le «rationnement» des «bons» risques, les «mauvais» risques seuls obtenant l'assurance complète, c'est-à-dire le rachat complet du ticket modérateur.

Dans la réalité, le secteur de l'assurance complémentaire fonctionne différemment, notamment en France. La mutualisation des risques y est importante, et quelques mutuelles de taille conséquente se partagent l'essentiel du marché. Par ailleurs, même si les détenteurs d'assurance complémentaire sont en général de plus gros consommateurs de soins (risque moral), ils sont aussi (de façon plus surprenante) en meilleure santé que les autres. L'objectif de cette partie de l'étude est de montrer comment la prise en compte d'un certain type de coûts de transaction permet de rendre compte de ces phénomènes. Plus précisément, nous montrons que si l'offre de chaque type de contrat entraîne pour la mutuelle un coût (fixe) de gestion ou de commercialisation, les caractéristiques de l'équilibre concurrentiel (telles que prédites par le modèle) se rapprochent beaucoup plus de la réalité, et peuvent comporter une mutualisation des risques différents, ainsi qu'un rationnement des risques les plus élevés.

Les principaux résultats obtenus sont les suivants :

* les dépenses publiques de santé ont un impact important sur l'état de santé moyen dans les pays concernés,

* le Canada et la Suède sont les pays les plus efficaces de l'échantillon, alors que la France et les USA ont un niveau d'efficacité faible, en égard au montant de leurs dépenses de santé,

* sur la période considérée (1960-1994) l'efficacité du système de santé français a diminué, malgré l'augmentation de ses dépenses.

I.3.2- Influence du revenu et de la couverture complémentaire sur la fréquence de consultation médicale des ménages français

Gwenaël Piaser et Denis Raynaud étudient l'influence du revenu et de la couverture complémentaire des ménages sur leur fréquence de consultation médicale à partir de données tirées de l'enquête santé 1991-92 de l'INSEE.

Dans une première partie nous estimons la probabilité de consultation d'un médecin en autorisant des effets non monotones pour le revenu. Les résultats obtenus montrent que le revenu a un effet positif sur la probabilité de consultation pour les individus disposant d'une couverture complémentaire (assurance ou mutuelle), mais cet effet devient négatif à partir d'un certain seuil pour les individus ne disposant pas d'une couverture complémentaire.

Dans une seconde partie nous estimons l'influence de différentes variables sur l'état de santé tel qu'il est perçu par les individus eux-mêmes. Nous montrons que les individus les plus « pauvres » considèrent généralement qu'ils sont en plus mauvaise santé alors que toutes choses égales par ailleurs, les individus les plus « riches » se considèrent en meilleure santé.

I.4 - Couverture maladie universelle ou accès gratuit aux soins

Par un modèle aussi simple que possible, Agnès Couffinhal et Jean-Charles Rochet essaient de captiver les raisons pour lesquelles plus de 40 millions d'Américains renoncent à s'assurer contre la maladie. Ils utilisent ensuite ce modèle pour évaluer les effets qu'aurait eu l'introduction, prévue dans le plan Clinton, d'une couverture maladie universelle aux USA.

II - Mesure et comparaison des performances des systèmes de santé des pays de l 'OCDE

1. INTRODUCTION

Les systèmes de santé dans tous les pays développés ont pris une ampleur considérable, s'accompagnant d'une croissance significative des dépenses de santé. Aussi, depuis plusieurs années, l'objet des politiques de santé est de maîtriser l'évolution de ces dépenses. En regardant cette évolution pour les pays de l'OCDE depuis 1960, on s'aperçoit que les dépenses de santé, en proportion du PIB, ont été multipliées par deux en trente ans. Ainsi pour la France, elles étaient de 4,2 % en 1960 et sont d'environ 9 % en 1994. Pour les Etats-Unis, on remarquera que les dépenses en matière de santé sont largement supérieures à celles des autres pays : 14,3 % en 1994. Les soins financés par des fonds publics ont tendance à diminuer dans certain pays comme l'Italie (-12.5 % entre 1960 et 1994), alors que dans d'autres pays ils ont fortement augmenté : + 44,3 % aux Pays-Bas.

De plus, au cours de ces trente dernières années nous avons assisté à une évolution remarquable du secteur de la santé : augmentation du personnel médical ainsi que des moyens dont il dispose, forte augmentation de la consommation médicale, de la couverture sociale (100 % de la population est couverte dans la plupart des pays sauf aux Etats-Unis où seulement 45 % des personnes sont couvertes)... L'avancée du progrès technique et médical a été importante avec la découverte de nouveaux traitements, médicaments, l'utilisation des dépistages...

On peut donc légitimement penser que le secteur de la santé est devenu performant. Pour vérifier cette hypothèse, nous allons construire une mesure d'efficacité des systèmes de santé définie comme étant la capacité à transformer des ressources sanitaires en output de santé.
Cette étude s'effectuera à travers la comparaison des pays de l'OCDE, les systèmes de santé ayant subi une évolution différente dans chacun des pays. Une telle analyse étant difficile à réaliser de part la complexité et la multiplicité des données disponibles, il paraît intéressant de considérer que les principaux indicateurs de santé (espérance de vie, mortalité infantile et années de vie potentielles perdues) peuvent être regroupés pour former un indicateur synthétique de la production. On peut alors calculer l'efficacité pour une telle variable à partir de l'analyse de sa frontière de production.

Farrel (1957) fut le premier à présenter une méthode pour mesurer l'efficacité productive. Une analyse économétrique des frontières de production est proposée en 1977 par Aigner, Lovell et Schmidt. Celle-ci sera enrichie avec l'utilisation des données de panel permettant d'introduire un terme d'efficacité variant dans le temps : Cornwell, Schmidt et Sickles (1990), Battese et Coelli (1992), Lee et Schmidt (1993). Parallèlement, des méthodes d'estimation non-paramétriques se sont développées, notamment la méthode DEA (Data Envelopment Analysis) utilisée par Bosmans et Fecher (1992) pour effectuer une étude sur le secteur de la santé des pays de l'OCDE à partir de données en coupe.

Nous utiliserons ici les données de panel qui permettent de mieux contrôler les effets des corrélations entre les quantités de facteurs utilisés dans le cycle de la production et l'inefficacité.

La section suivante définit le concept d'efficacité et les méthodes de mesure. Des facteurs exogènes seront introduits afin de prendre en compte l'influence du mode de vie et du progrès technique. La section 3 présente les différents indicateurs de santé utilisés comme output et facteurs de production. Plusieurs méthodes d'estimations seront proposées dans la section 4, et la section 5 contient les résultats des mesures d'efficacité.

2. CONCEPT ET MESURE D'EFFICACITE

L'analyse des systèmes de santé des pays de l'OCDE suppose l'existence d'un même critère objectif de comparaison. Aussi, l'utilisation de l'efficacité productive permet d'obtenir un tel critère. Ce concept d'efficacité procède de la notion de frontière de production représentant les productions maximales réalisables par l'entreprise. Une fonction de production peut donc s'interpréter en terme de montant maximal d'output obtenu à partir d'une quantité d'inputs donnée ou encore en terme de quantité minimale de ressources à utiliser pour obtenir un certain niveau d'output. Deux mesures d'efficacité sont à distinguer : l'efficacité technique et l'efficacité allocative.

La figure suivante représente la situation dans laquelle les entreprises (les pays de l'OCDE) utilisent deux inputs, K et L, pour produire leur output. Chaque entreprise est représentée par un point. La fonction de production est représentée par l'isoquant (Q), c'est-à-dire, la frontière de l'ensemble des combinaisons d'input qui peuvent être utilisées pour produire un niveau d'output donné. En d'autres termes, l'isoquant désigne le lieu géométrique associé aux combinaisons d'input les plus efficaces permettant de produire l'output. Une déviation de cet isoquant fournit une mesure de l'efficacité technique. Si l'entreprise localisée en R sur la figure était située au point T sur l'isoquant, elle serait efficace techniquement à 100%. Le ratio OT/OR est donc une mesure de l'efficacité technique.

L'inefficacité allocative survient lorsque la proportion des inputs ne reflète pas la proportion des prix relatifs des inputs et donc le coût total peut être réduit en employant une combinaison d'inputs différente. En d'autres termes, l'efficacité allocative capture l'inefficacité résultant uniquement du mauvais choix de combinaisons techniquement efficaces étant donné le prix des inputs. Sachant que la droite (w) représente le coût associé aux différentes combinaisons d'input pour un niveau donné des prix des facteurs, les niveaux de coût en A et en D sont égaux. Donc lorsqu'une entreprise (ou pays) a un niveau de production situé en T et non en A (qui sont deux allocations techniquement efficaces), une mesure de l'efficacité allocative est donnée par le ratio OD/OT.

L'étude de la performance des systèmes de santé des pays de l'OCDE s'effectuera donc par la mesure de l'efficacité technique (ou productive) de ces systèmes.

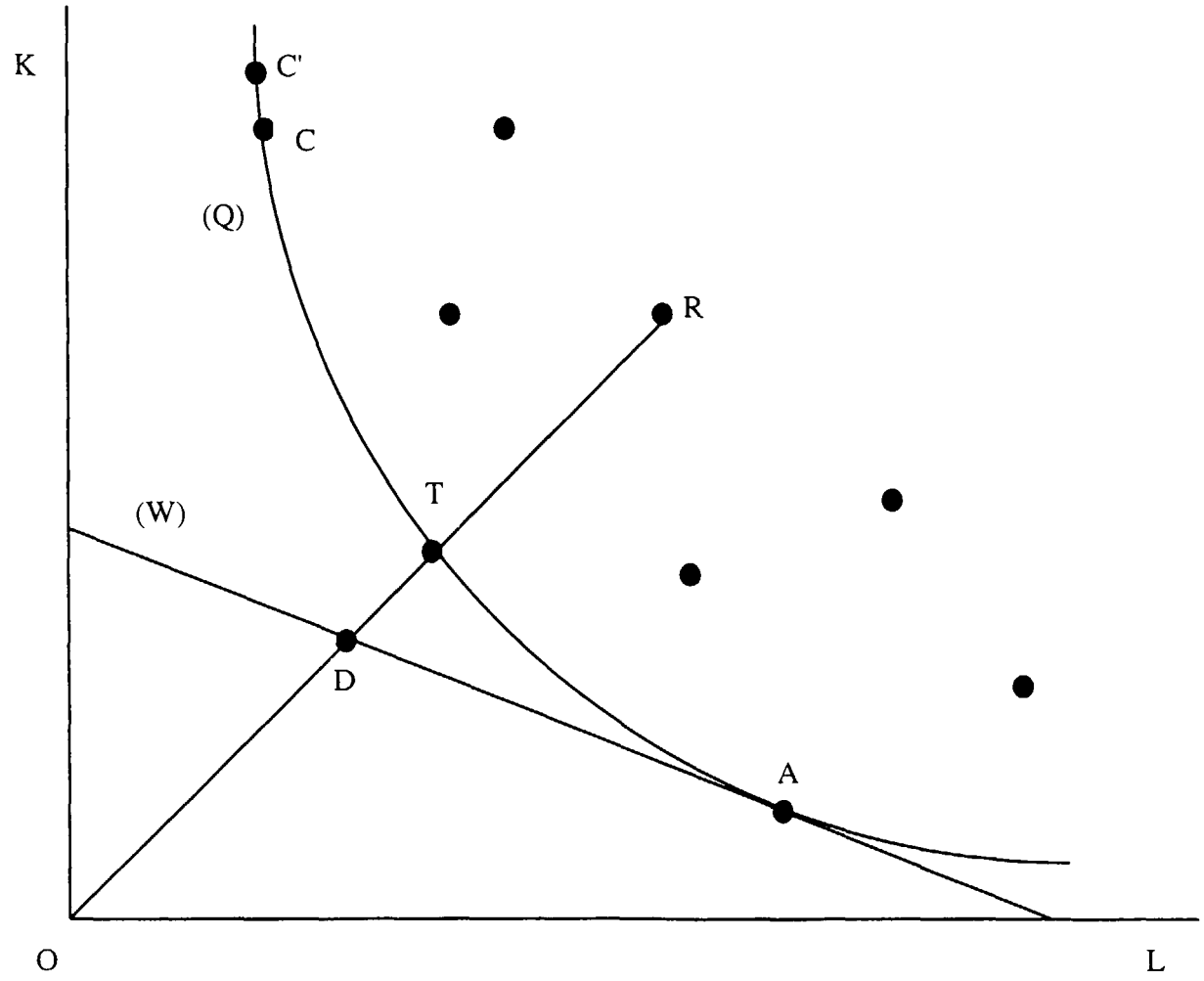


Figure 1 : Efficacité technique et allocative

Le modèle

Une frontière de production stochastique, avec des données de panel, peut être définie par

$$y_{it} = x_{it}\beta + \varepsilon_{it} - u_{it}$$

où y_{it} est la variable dépendante (output), x_{it} est le vecteur des facteurs de production, β est un vecteur de paramètres inconnus.

Le terme d'erreur se décompose en deux parties. La première (ε_{it}) est l'erreur usuelle d'un modèle de régression et représente les variables omises ou non observables ayant un effet illimité sur l'output dans les modèles de frontière de production. La seconde composante (u_{it}) est une variable non négative qui désigne les facteurs inobservables qui ont un effet limité sur l'output (effort, habileté...). Comme ces facteurs n'atteignent pas leur niveau optimal, ils génèrent l'inefficacité.

Nous considérerons, dans le cadre de cet article, le modèle proposé par Schmidt et Sickles (1984) :

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}$$

où l'effet individuel $(\alpha_i = -u_{it})$ apparaît dans la variation de la constante. Les différences dans la constante sont donc interprétées comme étant des niveaux différents d'efficacité invariants dans le temps.

Ces mêmes auteurs calculent l'efficacité à partir de la différence entre l'output observé et l'output estimé :

$$\hat{\varepsilon}_i = \frac{1}{T}\sum_{\tau=1}^{T}(y_{it} - x_{it}\hat{\beta})$$

On obtient alors une mesure de l'efficacité technique :

$$ET_i = \exp(-\hat{\theta}_i) * 100$$

avec $\hat{\theta}_i = \max_j (\hat{\varepsilon}_j) - \hat{\varepsilon}_i$.

Une critique peut être faite sur le modèle présenté ci-dessus. En effet, l'efficacité obtenue avec un tel modèle ne prend pas en compte de variables représentatives de l'environnement des entreprises. Gathon et Pestieau (1992) suggèrent de décomposer les indicateurs traditionnels d'inefficacité en deux parties : l'inefficacité due exclusivement au management et l'inefficacité régulatrice qui est attribuable à l'environnement institutionnel, légal et administratif dans lequel les entreprises opèrent. Ainsi, ils démontrent que l'efficacité technique est affectée par la nature et l'étendue de l'intervention du gouvernement, et qu'elle peut être améliorée en augmentant l'autonomie de l'entreprise. Pour cela ils incluent dans leur modèle des variables correspondant à ces caractéristiques institutionnelles.

Le modèle s'écrit alors :

$$y_{it} = \alpha_i + x_{it}\beta + z_{it}\gamma + \varepsilon_{it}$$

où z_{it} est le vecteur des variables d'environnement et γ un vecteur de paramètres à estimer.

3. MESURE ET DETERMINANTS DE LA PRODUCTION

Les données utilisées sont issues de la base de données sur la santé de l'OCDE (Eco santé OCDE, 1996). Elles concernent les 29 pays membres de l'OCDE pour la période 1960-1994. Du fait de leur tardive adhésion, certains pays seront exclus de l'étude car trop peu de données sont disponibles les concernant (Pologne, Corée, Hongrie et République Tchèque).

Les indicateurs de production que nous avons sélectionnés dans un premier temps sont ceux suggérés par Bosmans et Fecher (1992), à savoir l'espérance de vie à 60 ans, la mortalité infantile et les années de vie potentielles perdues. Pour cette dernière variable, sont pris en compte les décès qui auraient pu être évités si les connaissances médicales avaient été appliquées, si les règles de santé publique connues avaient été en vigueur et si les comportements à risque n'avaient pas l'ampleur qu'on connaît. On constate cependant que ces mesures d'output sont fortement corrélées entre elles. Cet état de fait suggère que l'on puisse générer une mesure synthétique de la production du secteur de la santé en effectuant une analyse en composantes principales et en utilisant la première de ces composante. Cette analyse est réalisée à partir de la matrice de corrélation des variables d'origines afin de donner une importance comparable aux trois indicateurs d'output retenus. Ainsi, pour ces derniers, on remarquera que la somme de leur variance est expliquée à 90.8% par la première composante principale. Par la suite, nous avons modifié l'échelle de cette composante de façon à ce que sa variance soit égale à la somme des variances des variables représentant l'output. Cette mesure (prinf) sera donc celle utilisée finalement comme output dans notre étude.

Les facteurs de production utilisés sont le nombre de médecins pour 10 000 habitants, le nombre de lits occupés pour 1 000 habitants (nombre de lits disponibles multiplié par le taux d'occupation des lits), et les dépenses de produits pharmaceutiques par habitant (consommations intermédiaires). Le fait de prendre en compte le nombre d'infirmiers pour mesurer l'emploi améliore la qualité de l'ajustement global du modèle de régression que d'une façon très marginale. De plus cela provoque une augmentation de la colinéarité des variables explicatives. Aussi seuls seront considérés les médecins.

Nous avons également introduit des variables représentant l'environnement et l'évolution technologique : la part des dépenses publiques dans les dépenses totales de santé, la durée moyenne d'hospitalisation, la consommation de tabac par habitant et enfin la consommation d'alcool par habitant. L'introduction des dépenses publiques permet de tenir compte du mode d'administration du système de santé et en particulier des facilités d'accès aux soins de la population à faible revenu. La

variable "durée d'hospitalisation" a pour objectif de refléter dans quelle mesure les méthodes de production du secteur de la santé incorporent les développements technologiques les plus récents. On présume que ces techniques permettent d'utiliser des thérapies réduisant la durée de séjour dans un hôpital. Finalement, les variables "tabac" et "alcool" sont des variables représentatives du mode de vie qui peuvent augmenter la morbidité et la mortalité des populations.

L'évolution de ces variables pour la période 1960-1994 est donnée en annexe 1.

4. ESTIMATIONS

Il faut noter qu'un nombre important de données manquent pour plusieurs pays. Aussi notre étude se limitera à l'analyse des systèmes de santé de 16 pays de l'OCDE. Le Japon n'a pas été pris en compte, bien que les données soient disponibles pour ce pays. En effet, les valeurs pour la variable "durée de séjour" ne semblent pas comparables à celles des autres pays. Ceci peut s'expliquer par le fait que pour la plupart des pays la durée moyenne de séjour est constituée en majorité des durées d'hospitalisation en hôpitaux généraux ou en court séjour (établissements accueillant des patients dont la durée moyenne de séjour est d'au plus 18 jours). Or, au Japon, la catégorie court séjour étant inexistante, la part des données relatives aux longs séjours est plus importante que dans les autres pays, d'où un allongement de la durée globale d'hospitalisation dans ce pays. Il apparaît donc difficile d'effectuer une comparaison entre le Japon et les autres pays. De plus, le fait d'ajouter le Japon aux données considérées semble diminuer de façon notable la précision obtenue dans l'estimation des coefficients de notre modèle.

La production des systèmes de santé des pays de l'OCDE est estimée à partir d'une fonction Cobb-Douglas :

$$\begin{cases} \log\ y_{it} = \beta\ \log\ x_{it} + \gamma \log z_{it} + \varepsilon_{it} \\ \varepsilon_{it} = \alpha_i + u_{it} \end{cases}$$

avec y'=[prinf]

x'=[médecins lits pharmacie]

z'=[public durée tabac alcool]

On s'attend à ce que les coefficients des variables médecins, lits occupés et pharmacie soient positifs et que ceux des variables durée, tabac et alcool soient négatifs. Plusieurs méthodes d'estimations seront utilisées : effets fixes, effets aléatoires et un test sera effectué pour tester l'exogénéité des trois premières variables citées ci-dessus.

Effets fixes

On constate (tableau 1) que les signes des coefficients estimés correspondent tous aux signes attendus. Les statistiques de Student sont dans l'ensemble relativement élevées sauf pour les variables "lits occupés" et "tabac". On remarquera de plus qu'une augmentation de la part du financement public des soins de santé entraîne une augmentation significative de la production des systèmes de santé. En ce qui concerne les estimations des paramètres représentant l'effet individuel (α_i), il faut rappeler que dans les modèles à effets fixes les estimateurs obtenus sont biaisés si on estime à la fois

chacun des paramètres et une constante générale. Deux solutions sont possibles : éliminer la constante générale ou éliminer le paramètre représentant l'effet individuel pour un pays donné. Dans ce cas, les α_i restants doivent être interprétés comme étant les différences entre le paramètre d'origine d'un pays i et celui du pays de référence qui a été supprimé. Nous avons choisi la France comme pays de référence. On constate, si on adopte un niveau de confiance de 90%, que les coefficients des pays suivants : Danemark, Etats-Unis, Islande, Italie, Nouvelle-Zélande, Royaume-Uni et Suède ne sont pas significativement différents de celui de la France. Donc, en utilisant une mesure d'efficacité globale (pas de distinction entre les facteurs de production usuels et les variables d'environnement), trois groupes peuvent être définis. Le premier correspondrait aux pays pour lesquels cette mesure est supérieure à celle de la France : Canada. Le second rassemblerait les pays ayant une efficacité pratiquement égale à celle de la France (les pays nommés ci-dessus), les pays restant représentant le groupe pour lequel l'efficacité est inférieure.

On pourrait argumenter que dans les fonctions de production, les variables qui correspondent aux facteurs de production, à savoir les variables médecins, lits occupés et pharmacie dans notre cas, pourraient être considérées comme endogènes. Si c'était le cas, il vaudrait mieux alors utiliser une méthode d'estimation de variables instrumentales, afin d'éviter les biais de simultanéité. Dans cette optique, nous avons effectué un test d'exogénéité. Nous avons utilisé comme variables instrumentales associées aux variables médecins, lits occupés et pharmacie ces mêmes variables décalées d'une période. L'hypothèse Ho à tester est que ces facteurs de production peuvent être considérés comme exogènes. La façon la plus simple d'effectuer ce test asymptotique est d'utiliser une procédure de régression augmentée (Davidson et MacKinnon, 1993). Il suffit, dans un premier temps, d'effectuer des régressions avec les trois variables à tester, ces variables apparaissant comme variables dépendantes et les variables indépendantes étant les trois variables instrumentales choisies ainsi que les variables dites d'environnement. Par la suite, les erreurs résiduelles de ces régressions auxiliaires sont ajoutées au modèle d'origine comme variables explicatives additionnelles. Finalement, nous testerons, au moyen d'un test F, l'hypothèse que les coefficients de ces trois erreurs sont simultanément égaux à zéro. Cette hypothèse est équivalente à l'hypothèse Ho mentionnée ci-dessus. La probabilité d'obtenir une valeur de F supérieure à la statistique obtenue étant de 25%, on accepte l'hypothèse d'exogénéité des variables explicatives pour ce modèle à effets fixes.

	Effets fixes		Effets aléatoires	
VARIABLES	**ESTIMATES**	**STUDENTS**	**ESTIMATES**	**STUDENTS**
Intercept	-3.979986	-4.257392	-4.012203	-4.448115
Médecins	0.672609	6.329395	0.543711	5.822535
Lits occupés	0.113958	0.962788	0.234163	2.522843
Pharmacie	0.308490	5.888773	0.342264	7.118056
Public	0.566593	4.654179	0.472555	4.344267
Durée	-0.121845	-1.030372	-0.237047	-2.275119
Tabac	-0.157375	-1.628752	-0.180877	-2.045655
Alcool	-0.330223	-4.608331	-0.236260	-4.312103
	R-Square	0.9685	**R-Square**	0.9272

Tableau 1 : Résultats des estimations.

Effets aléatoires

Les résultats du modèle à effets aléatoires montrent que les statistiques de Student sont significatives pour toutes les variables explicatives.

Une des hypothèses importantes sous-jacentes au modèle aléatoire est que les erreurs résiduelles α_i associées aux différents pays ne sont pas corrélées avec les variables explicatives du modèle. Cependant, dans de nombreux cas il arrive que cette hypothèse soit rejetée (Prob.> m =0.0130 avec notre modèle). Pour éviter que cette corrélation ne rende les estimateurs du modèle non convergents, une approche possible est de supposer que les variables α_i peuvent se décomposer de la façon suivante :

$$\alpha_i = a\left[\overline{x}_i, \overline{z}_i\right] + v_{it}$$

où $\overline{x}_i$ et $\overline{z}_i$ correspondent aux moyennes des x_{it} et des z_{it} pour le pays i sur la période 1960-1994. Dans un premier temps, cette hypothèse revient à introduire tous les éléments de $\overline{x}_i$ et $\overline{z}_i$ dans la régression. Par la suite, on élimine successivement les éléments de ces vecteurs dont les coefficients ne semblent pas significativement différents de zéro, pour ne conserver que ceux dont les coefficients sont significatifs. Dans le tableau suivant, il apparaît que seule la moyenne de la variable tabac (T) a été conservée.

VARIABLES	ESTIMATES	STUDENTS
Intercept	-8.468138	-5.884778
Médecins	0.615428	6.583350
Lits occupés	0.188943	1.893178
Pharmacie	0.303661	6.412842
Public	0.480830	4.421532
Durée	-0.262845	-2.453299
Tabac	-0.173296	-2.000356
Alcool	-0.255609	-4.264915
T	0.604146	3.793794
	R-Square	0.9388

Tableau 2 : Méthode des moyennes

Les résultats laissent apparaître que les variables explicatives sont maintenant exogènes (Prob.> m =0.5932).

5. EFFICACITE DES SYSTEMES DE SANTE

L'efficacité des systèmes de santé peut se mesurer de diverses manières. Tout d'abord, on peut définir l'efficacité comme étant la part de la production non expliquée par l'effet direct des facteurs de production (mesure 1). On peut aussi vouloir définir l'efficacité en neutralisant l'influence des facteurs technologiques ou des facteurs d'environnement qui apparaissent dans le modèle. Ces facteurs peuvent être neutralisés partiellement (mesure 2 : seules les variables alcool et tabac sont considérées en plus des facteurs de production) ou totalement (mesure 3 : toutes les variables du vecteur z sont prises en comptes). Les résultats pour le modèle à effets aléatoires sont représentés par les figures 2, 3 et 4 alors que le modèle à effets fixes est représenté par les figures 5 à 7.

On constate que, pour une même mesure, les ordonnancements obtenus avec les deux modèles ne diffèrent que très peu. Ils diffèrent davantage lorsqu'on passe d'une mesure à l'autre. On remarquera tout d'abord que le Canada et la Suède apparaissent parmi les pays les plus efficaces, contrairement au Portugal et surtout à l'Allemagne dont l'efficacité est toujours inférieure à 50%, et ce quelle que soit la mesure d'efficacité considérée. Si on se réfère aux mesures 2 et 3 qui ont le mérite de neutraliser les effets dus à des facteurs culturels, on constate que la France se situe dans la moitié supérieure. Cependant, elle se situe dans l'autre moitié lorsqu'on ne tient pas compte des effets négatifs de l'alcool et du tabac sur la santé. Le système de santé de la France apparaît d'autant moins efficace que sa part du PIB consacrée aux dépenses de santé est équivalente à celle du Canada. On peut également remarquer, concernant ces deux pays, que la France utilise une quantité plus importante d'inputs que le Canada pour une efficacité moindre. De même, les Etats-Unis ont un niveau d'efficacité relativement faible (mesures 1 et 2) comparé au montant de leurs dépenses en matière de santé.

Les figures 8 et 9 présentent l'évolution, pour les deux modèles, de la "mesure 3" d'efficacité pour chacun des pays et pour deux sous périodes : 1960-1979 et 1980-1994. Ainsi, nous constatons en France une diminution de l'efficacité du système de santé malgré l'augmentation de ses dépenses (en % du PIB) dans ce secteur. D'autres pays ont vu leur efficacité diminuer : Danemark, Islande et Suède par exemple, alors que l'Autriche est devenue plus performante (+10% entre les deux périodes).

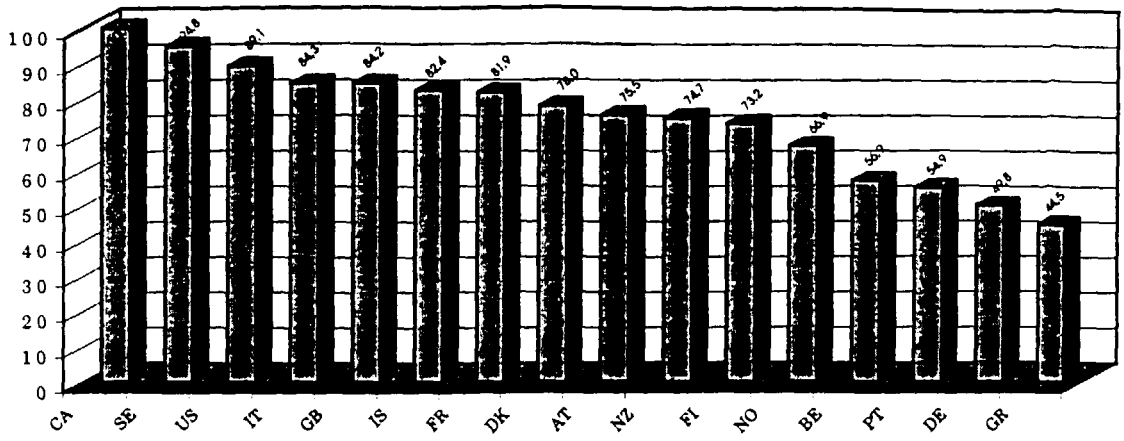


Figure 2 : Efficacité productive et régulatrice (mesure 3).

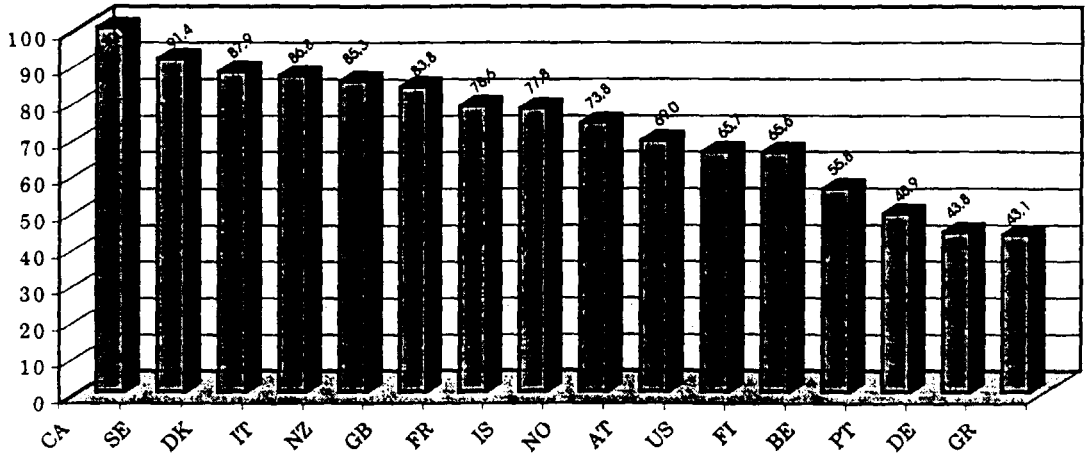


Figure 3 : Mesure 2

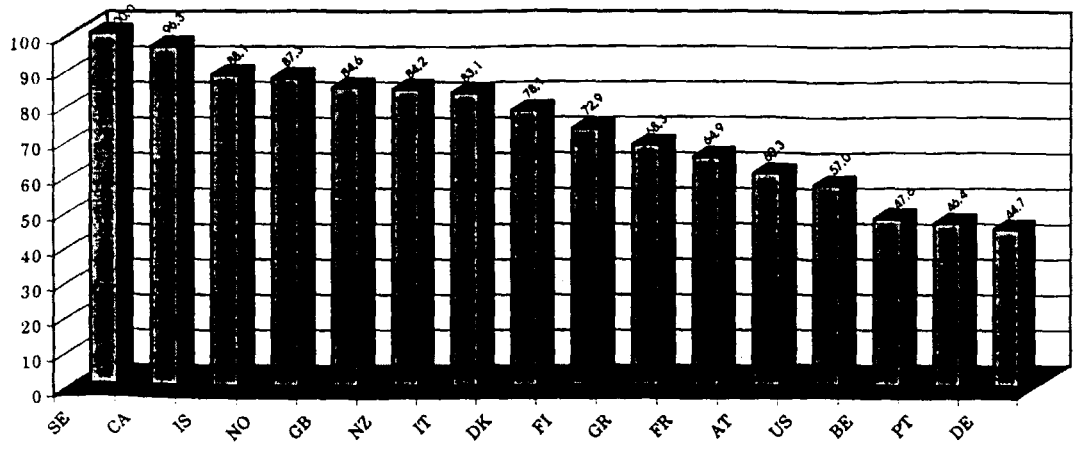


Figure 4 : Efficacité productive (mesure 1).

Modèle à effets fixes

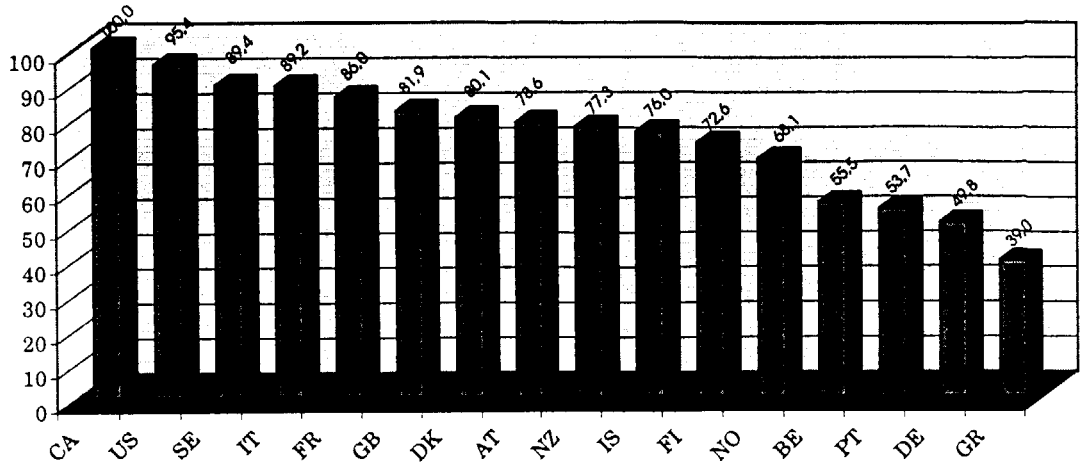


Figure 5 : Efficacité productive et régulatrice (mesure 3).

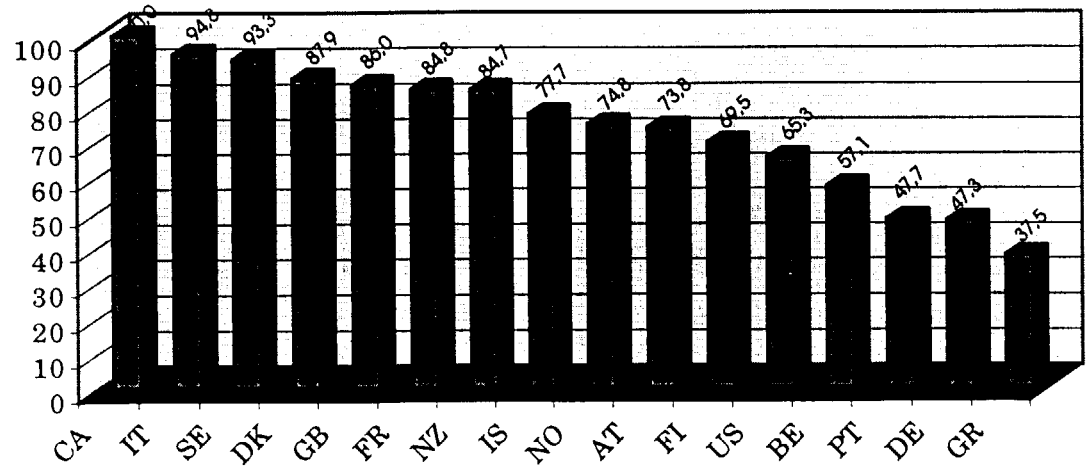


Figure 6 : Mesure 2

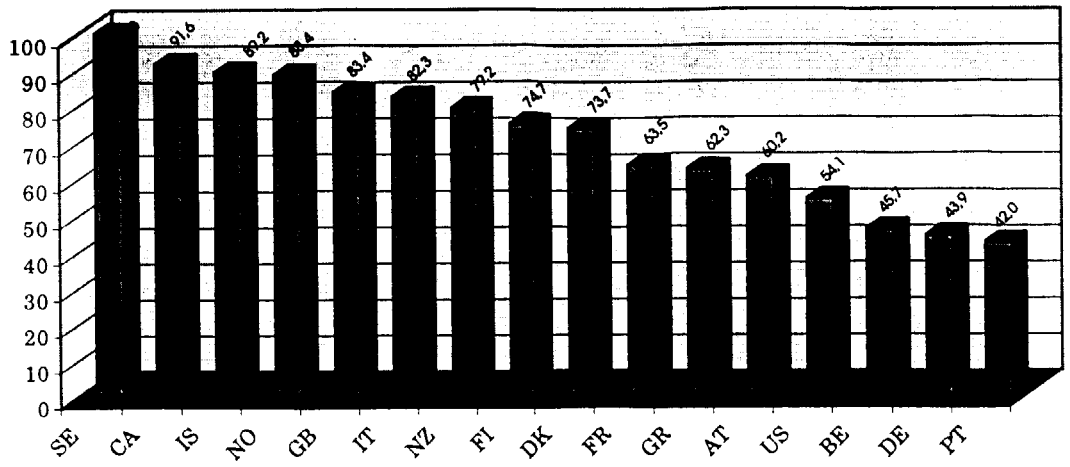


Figure 7 : Efficacité productive (mesure 1).

Figure 8: Evolution de l'efficacité par pays (modèle à effets aléatoires).

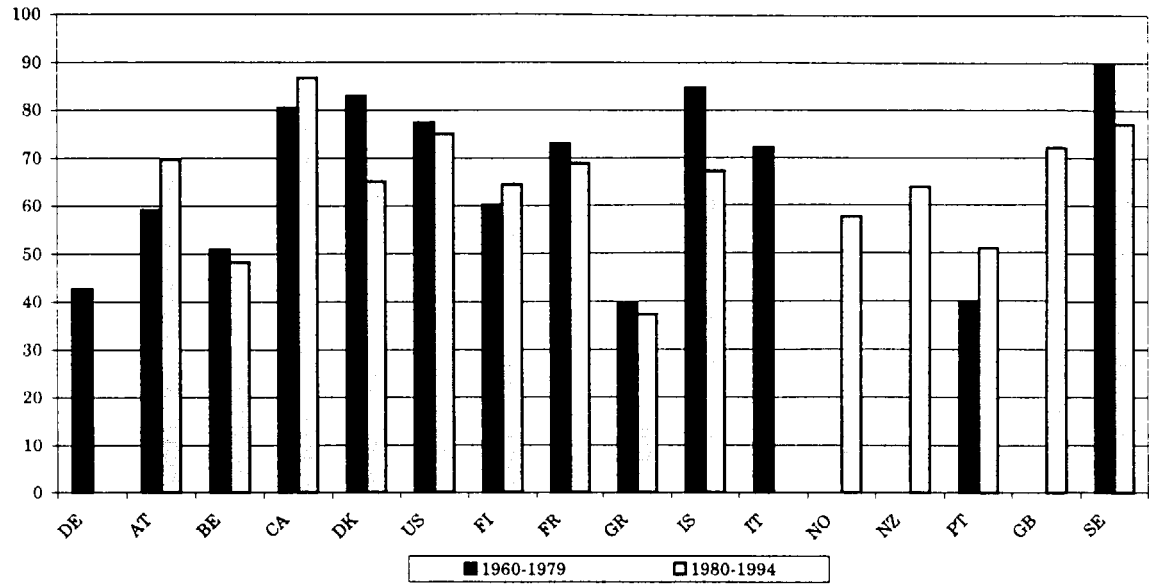
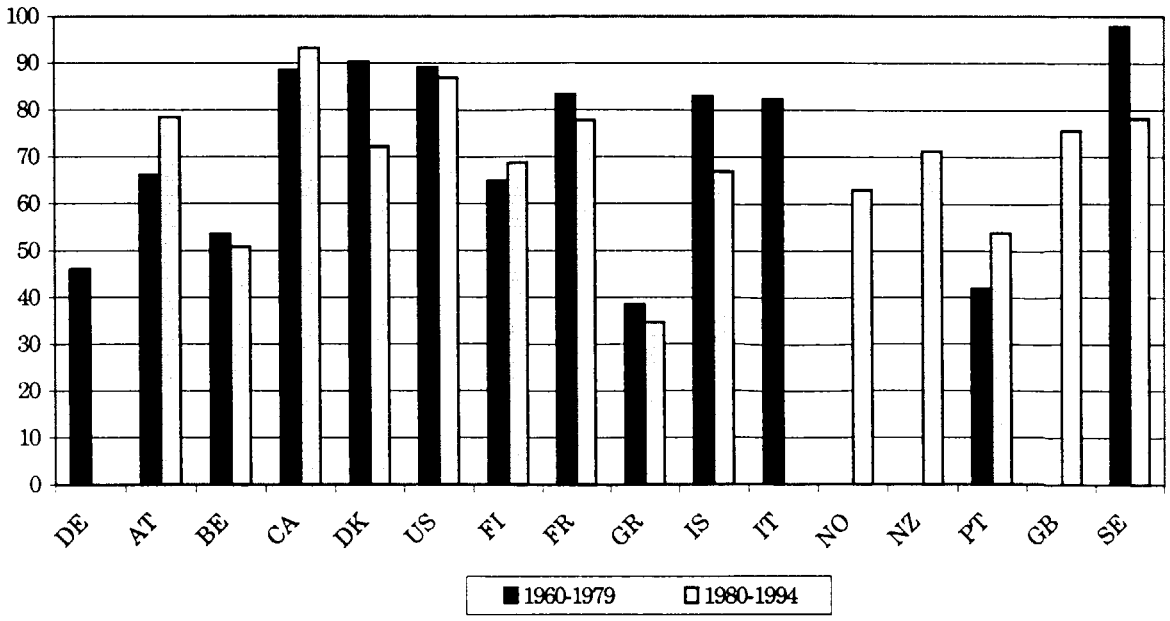


Figure 9 : Evolution de l'efficacité (modèle à effets fixes).



CONCLUSION

Dans cet article nous étudions les systèmes de santé des pays de l'OCDE et les comparons sur la base d'un même critère : l'efficacité technique de ces systèmes. Pour cela, nous avons d'abord estimé leur frontière de production et ensuite établi plusieurs mesures d'efficacité prenant en compte (pour deux d'entre elles) des facteurs institutionnels ou d'environnement.

Les résultats montrent dans un premier temps qu'une augmentation des dépenses publiques de santé améliore l'état de santé des pays. Ensuite, il est apparu que le Canada était le pays le plus efficace au regard des mesures prenant en compte les facteurs d'environnement (il est second pour l'autre mesure). En comparaison, la France a une efficacité relativement plus faible. Pour finir, l'évolution au cours de ces trente dernières années des différentes mesures d'efficacité a montré une tendance à la baisse de l'efficacité des systèmes de santé pour plusieurs pays (dont la France), alors que le Portugal et l'Autriche ont vu leur efficacité augmenter.

ANNEXE 1

		Espérance de vie à 60 ans Années	Vie potent. perdue Total, sauf suicides 100 000 hab	Mortalité infantile Décès /100 nés	Médecins en activité Densité /10 000 hab	Ensemble hôpitaux Lit/1000hab	Tx d'occup. ens. hôpit. % lits offerts
DE	1960-1975	17,2	9793,45	2,47	15,9	11,1	89,5
	1976-1994	19,25	6233,75	0,99	26,3	10,9	84,6
AT	1960-1975	17,1	9445,5	2,76	13,9	10,9	84,7
	1976-1994	19,15	5627,4	1,11	19,3	10,7	82,5
BE	1960-1975	17,25	9492,45	2,28	15,4	8,7	
	1976-1994	19,6	6625,75	1,03	28,7	8,7	85,2
CA	1960-1975	18,75	9225,25	2,13	14,1	6,8	
	1976-1994	21,2	5926,3	0,87	19,8	6,5	83,8
DK	1960-1975	18,7	7361,75	1,6	14,2	8,2	
	1976-1994	19,45	5906,35	0,78	24,8	6,8	82,1
US	1960-1975	18,2	10654,3	2,18	15,3	7,8	81,3
	1976-1994	20,25	7428,45	1,1	21,3	5,3	72,2
FI	1960-1975	16,6	9608,3	1,52	8,5	14	91,9
	1976-1994	18,95	5934	0,64	20,8	13,7	84,4
FR	1960-1975	18,2	8887,95	2,05	12,1	9,9	83,2
	1976-1994	20,65	6295,3	0,86	22,7	10,3	81,3
GR	1960-1975	18,2	9174,5	3,25	15,5	6,1	74,9
	1976-1994	20,6	5815,2	1,38	29,8	5,6	69,1
IS	1960-1975	19,75	7606,9	1,39	13,7	12	94,7
	1976-1994	21,35	4750,8	0,64	24,8	15,5	90,5
IT	1960-1975	18,3	10525,4	3,28	7,3	9,9	79,1
	1976-1994	19,8	6103,15	1,17	12,9	8,4	70
NO	1960-1975	19,05	6992,3	1,36	13,8	15,7	
	1976-1994	20,35	5165,9	0,78	22,6	15,4	85,3
NZ	1960-1975	17,75	9118,85	1,83	10,9	10,8	
	1976-1994	19,45	6923,35	1,09	17,1	9,4	57,3
PT	1960-1975	17,25	16844,8	6,02	9,7	6,2	74,1
	1976-1994	19,2	8738,6	1,8	23,9	5	68,7
GB	1960-1975	17,35	8626,2	1,93		9,5	83,2
	1976-1994	19,25	6172	0,99	14,1	7,1	80,6
SE	1960-1975	19	6313,3	1,26	12,5	15	83,5
	1976-1994	20,6	4672,8	0,64	25,3	13,2	83,8

		Dép. prod.pharm. PPA $/capita	Dép. publ. de santé % dep. tot. de sante	Durée de séjour,tot. Dur. séjour jours	Consommat. d'alcool /Personne Litres	Consommat. de tabac 15ans+ Gr/person.
DE	1960-1975	50	72,9	25,9		
	1976-1994	167,4	78,4	17,9	13,9	2815,5
AT	1960-1975	23,4	66,4	23	13,6	2318,8
	1976-1994	119,9	73,7	15	12,9	2525,3
BE	1960-1975	32,5	76,3		11,1	3486,9
	1976-1994	154,7	84,8	16,3	13	2905,7
CA	1960-1975	23,5	67,2	11,4	8,6	4315,6
	1976-1994	128,9	75,3	13,4	10,5	2875,6
DK	1960-1975	21,1	85,9	18,3	7,9	3178,5
	1976-1994	81,8	84,3	10,3	11,9	2693,6
US	1960-1975	36,2	33,2	16,3	9	4455,6
	1976-1994	161,5	41,7	9,5	10,2	3338,9
FI	1960-1975	18,9	68	25,8	4,9	1614,5
	1976-1994	91,4	78,9	19,6	8,4	1438,2
FR	1960-1975	40,4	69,8	19,5	23,4	2294,2
	1976-1994	184,9	76,7	15,1	18,6	2260,6
GR	1960-1975	19,2	59,9	15,8	1,7	2144,5
	1976-1994	81,4	80,2	11,8	2,3	3201
IS	1960-1975	32,3	82,2	28,8	3,5	2732,9
	1976-1994	148,7	87,3	21,2	4,6	2720,6
IT	1960-1975	20,8	87,6	21	16,7	2065,3
	1976-1994	155,8	79,2	12,6	12,6	2528,3
NO	1960-1975	10,4	86,6	22,1	4,3	1925,4
	1976-1994	78,2	87,6	12,9	5,1	1896,1
NZ	1960-1975		76,6	16,3	9,5	3208,8
	1976-1994	119,9	83,3	11,3	10,8	2307,8
PT	1960-1975		60,2	24,3	13	2219,3
	1976-1994	111,3	60	12,7	11,5	2351
GB	1960-1975	31	86,5	28,3	6,7	2802,5
	1976-1994	99,2	86,6	17	8,9	2293,5
SE	1960-1975	34	81,9	28,5	6,3	2090
	1976-1994	90,9	89,7	19,8	6,5	1945,6

Références

Aigner, D. J., C. A. K. Lovell and P. Schmidt [1977]: Formulation and Estimation of Stochastic Frontier Production Function Models, *Journal of Econometrics*, 6, 21-37.

Battese, G. E. and T. J. Coelli [1988]: Prediction of Firm-Level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data, *Journal of Econometrics*, 38, 387-399.

Cornwell, C. P. Schmidt and R. C. Sickles [1990]: Production Frontiers with Time-Series Variation in Efficiency Levels, *Journal of Econometrics*, 46, 185-200.

Farrell, M. S. [1957]: The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society*, A, 120, 253-281.

Lee, Y. H. and P. Schmidt [1993]: *A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency*, in H. Fried, C. A. K. Lovell and S. Schmidt eds., The Measurement of Productive Efficiency, New York: Oxford University Press.

III - Rémunération des médecins, incitations et coût des soins de santé

(en anglais)

La documentation Française : "Coût et efficacité des dépenses de santé : rapport / Université des sciences sociales, GREMAQ ; Commissariat général du Plan ; Jean-Charles Rochet, Helmuth Cremer, Jean-Paul Cresta, \(et al.\)."

1 The market for medical services

1.1 Physician and patients

In the market for medical services, the demand side consists of patients who have identical medical needs. The supply side consists of physicians (health care providers) who differ in productivity, that is the efficiency with which they transform medical inputs into improvement of their patients' health state. This is summarized in a production function, $h(\pi e)$, which specifies a patient's health improvement as a function of the time (or effort), e, devoted by the physician to each of his patients and of the physician's productivity, π. Later on we shall introduce an additional input as argument of the health production function, namely the amount of medical services prescribed by the physician on behalf of his patients. We assume that $h(\pi e)$ is increasing and concave in πe with $h(0) = 0$. Neither e nor π are observable by the regulator financing health care expenditures; health improvement $h(\pi e)$, on the other hand, is not verifiable.[1] Each patient must be registered with a physician. The number of patients registered with a physician is denoted by n; this variable is observable by the regulator (and verifiable).

A physician's utility is given by

$$u(n, T, e) = T + \gamma n h(\pi e) - v(ne), \tag{1}$$

where T is the transfer from the regulator to the physician (i.e., the providers compensation), γ is a parameter measuring his concern for patient's health (improvement), while $v(ne)$ represents disutility of effort. The disutility function is increasing and convex ($v' > 0, v'' > 0$), with $v(0) = 0$. Note that T can also be interpreted as consumption of a composite (numeraire) good. In what follows, we shall focus on the determination of the physicians payment schemes. Specifically, we shall study how a provider's compensation $T = \boldsymbol{T}(n)$ should be linked to his number of patients.

[1] Consequently, the provider's compensation cannot be (directly) based on h.

Physicians differ only in their productivity, which is assumed to take two values, $\pi_i (i = 1,2)$, with $\pi_2 > \pi_1$. The total number of physicians, M, is determined by the regulating authority. Whatever the total number of physicians there is a proportion p_i of physicians of type $i (i = 1,2)$, with $p_1 + p_2 = 1$. Throughout the paper, the subscript i is used to distinguish the two types of physicians.

There are P identical patients who have to be registered with a physician. The net benefit of a patient who chooses a physician of type i is given by

$$B_i = h(\pi_i e_i) - w(n_i e_i), \tag{2}$$

where w is a waiting cost function that depends upon the overall time (or effort) spent by the physician for treating his patients. It is increasing and convex ($w' > 0, w'' > 0$). This function reflects the disutility that patients incur because they have to wait for an appointment or in the physician office. The heavier the workload of a physician, the longer these waiting times will be on the average. Patients observe B_1 and B_2, and select the physician who provides them with the highest level of net benefits.

1.2 Equilibrium

The equilibrium in the market for medical services is contingent upon the payment scheme $\boldsymbol{T}(n)$ and the total number of physicians, M (both of which being determined by the regulator). It is defined as a vector of (marketwide) net benefits and (type specific) patient numbers and effort levels, $(\tilde{B}, \tilde{n}_1, \tilde{n}_2, \tilde{e}_1, \tilde{e}_2)$, which satisfies the following two conditions.

1. Each type of physician chooses effort level and number of patients to maximize his utility. Consequently, $(\tilde{n}_i, \tilde{e}_i)$ $(i = 1,2)$ is determined by

$$\max_{n_i, e_i} u(n_i, \boldsymbol{T}(n_i), e_i) = \boldsymbol{T}(n_i) + \gamma n_i h(\pi_i e_i) - v(n_i e_i), \tag{3}$$

subject to

$$h(\pi_i e_i) - w(n_i e_i) = \tilde{B}. \tag{4}$$

2. All patients must be registered with a health care provider:

$$M \sum_{i=1}^{2} p_i \, \tilde{n}_i = P. \tag{5}$$

This definition is in the spirit of a competitive equilibrium. Patients being identical, an equilibrium requires that their net benefits are equalized across physician types. This marketwide net benefit level $\tilde{B}$, though endogenously determined at the equilibrium, is taken as given by any single physician (condition (4)), who determines his demand for patients (as well as his effort level) accordingly.[2] Finally, (5) is the market clearing condition stating that the total demand for patients equals supply (number of patients).

1.3 Regulating authority

The regulator's objective function is evaluated at the market equilibrium induced by its policy (specifying $\boldsymbol{T}(n)$ and M). It takes into account the net benefits to patients and the cost of public funds needed to compensate physicians. With λ denoting the per unit cost of public funds ($\lambda > 1$), the regulator's objective is given by:

$$W = P\tilde{B} - \lambda M \sum_{i=1}^{2} p_i \boldsymbol{T}(\tilde{n}_i) \tag{6}$$

where $\tilde{B}$ and $\tilde{n}_i$ are defined by (3)–(5), that is the conditions determining the equilibrium in the health care market.

The regulator's problem stated this way is quite intricate. To make it tractable, we shall reformulate it and consider the equivalent mechanism design problem. For this purpose, we must first take a closer look at the physician's problem and introduce some additional notation.

[2] This is similar to the "utility taking" assumption in the urban economics (and local public goods) literature.

1.4 Behavior and preferences of physicians: a closer look

First of all, it is convenient to eliminate the (unobservable) effort level from the physician's problem. To do this, we introduce the "effort requirement function", $e_i = E_i(n, B)$, which is defined by:

$$h\Big(\pi_i\ E_i(n,B)\Big) - w\Big(n\ E_i(n,B)\Big) = B, \quad i = 1,2. \tag{7}$$

This function specifies the level of effort a physicians has to provide to offer net benefits of B, to n patient. It has the following properties:

$$\frac{\partial E_i}{\partial B} = [\pi_i h_i' \pi_i e_i) - n w'(n e_i)]^{-1}, \tag{8a}$$

$$\frac{\partial E_i}{\partial n} = [\pi_i h'(\pi_i e_i) - n w'(n e_i)]^{-1} e_i w'(n e_i). \tag{8b}$$

It seems natural to assume that an increase in the net benefit required by patients will demand more effort to the physician ($\partial E_i/\partial B > 0$). This is equivalent to assuming:

$$H1: \quad \pi_i h'(\pi_i e_i) - n_i w'(n_i e_i) > O, \quad i = 1,2,$$

which in turn implies that $\partial E_i/\partial n > 0$ and also that $E_1(n,B) > E_2(n,B)$.[3]

Next, we can substitute $E_i(n,B)$ into the direct utility function (1) to obtain the following derived utility function:

$$V_i(T,n,B) = T + \gamma n h\Big(\pi_i E_i(n,B)\Big) - v\Big(n E_i(n,B)\Big), \quad i = 1,2, \tag{9}$$

This utility function is a crucial ingredient of the reformulated problem considered below. In particular, it is used to define indifference curves for given B in the (n,T) space. In the appendix we show that under the following hypotheses:

$$H2: \quad v'(n e_i) - \gamma \pi_i h'(\pi_i e_i) > 0, \quad i = 1,2$$

[3] Redefining E as a function of π, with $E(\pi_i, n_i, B) \equiv E_i(n_i, B)$ differentiating (7) and making use of $H1$ yields $\partial E/\partial \pi = -(\pi h' - n w')^{-1} e h' < 0$.

$$H3: \quad e_i v'(ne_i) - \gamma h(\pi_i e_i) > 0, \quad i = 1,2,$$

these indifference curves are increasing and convex and that they satisfy at any point (n,T) the single-crossing property:

$$MRS^2_{T,n} < MRS^1_{T,n}$$

where

$$MRS^i_{T,n} \equiv -\frac{\partial V_i/\partial n}{\partial V_i/\partial T}$$

is the marginal rate of substitution between T and n for physicians of type i.

The above hypotheses mean that physicians need a higher compensation respectively for providing more effort (for a given n) and for treating more patients (for a given e). They are both satisfied if γ is not too high. Note also that since the physicians' utility function is quasi-linear in T, the indifference curves pertaining to each type of physician are for given B vertically parallel to each other in the (n,T) space.

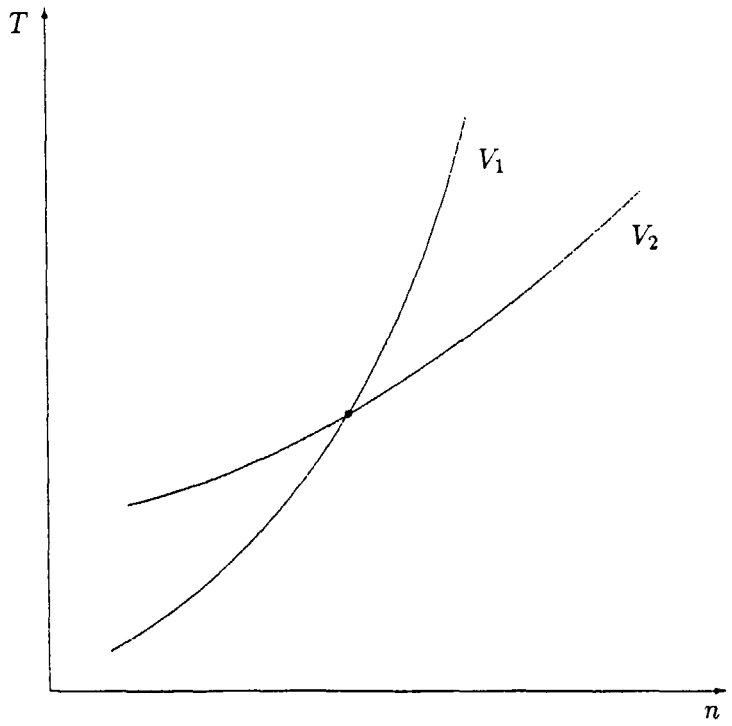
These properties are illustrated on Figure 1, where a typical indifference curve for each type of physician is represented.

2 Design of the optimal payment scheme and the full information optimum

Using the revelation principle, the regulator's problem stated in section 1.3 is equivalent to a mechanism design problem where the regulator searches for the two pairs (n_1,T_1) and (n_2,T_2) that he intends respectively for the first and second types of physicians. In solving this problem, the regulator must make sure that physicians of type i actually prefer the pair (n_i,T_i) designed for them to the other pair. This is accounted for by the incentive compatibility constraints of the following problem:

$$\max_{T_i,n_i,M,B} PB - \lambda \quad M\sum_{i=1}^{2} p_i T_i \qquad (11.1)$$

subject to



$$M \sum_{i=1}^{2} p_i n_i = P \tag{11.2}$$

$$(IR1) \qquad V_1(T_1, n_1, B) \geq \overline{V} \tag{11.3}$$

$$(IR2) \qquad V_2(T_2, n_1, B) \geq \overline{V} \tag{11.4}$$

$$(IC1) \qquad V_1(T_1, n_1, B) \geq V_1(T_2, n_2, B) \tag{11.5}$$

$$(IC2) \qquad V_2(T_2, n_2, B) \geq V_2(T_1, n_1, B) \tag{11.6}$$

where $\overline{V}$ stands for the physicians' reservation utility, i.e. the level of utility they can reach in the most preferred alternative occupation. Substituting M from the first constraint into the objective function, the Lagrangian of this problem can be written as:

It seems natural to assume that an increase in the net benefit required by patients will demand more effort to the physicians' reservation utility, i.e. the level of utility they can reach in the most preferred alternative occupation. Substituting M from the first constraint into the objective function, the Lagrangian of this problem can be written as:

$$\begin{aligned} L = {} & PB - \lambda P\left[\sum_{i=1}^{2} p_i n_i\right]^{-1} \sum_{i=1}^{2} p_i T_i \\ & + \sum_{i=1}^{2} \phi_i \sum_{i=1}^{2} V_i(T_i, n_i, B) + \mu_1[V_1(T_1, n_1, B) \\ & - V_1(T_2, n_2, B)] + \mu_2[V_2(T_2, n_2, B) - V_2(T_1, n_1, B)] \end{aligned} \tag{12}$$

where the ϕ's and μ's are the dual variables of the individual-rationality and incentive-compatibility constraints respectively. These dual variable satisfy the complementarity condition that they are equal to 0 if the corresponding constraints are not binding. It is useful to first characterize the full information optimum that we shall use as a benchmark later on. It can be done by deleting the IC constraints from the above problem (i.e. setting $\mu_i = 0$). It is then straightforward to obtain the following conditions for an

optimal allocations:

$$MRS^i_{Tn} = \frac{\sum_{i=1}^{2} p_i T_i}{\sum_{i=1}^{2} p_i n_i} \equiv \alpha, \quad i = 1,2 \tag{13}$$

$$P = \lambda M \sum_{i=1}^{2} p_i MRS^i_{TB} \tag{14}$$

where $MRS^i_{TB} = -\frac{\partial V_i}{\partial B} / \frac{\partial V_i}{\partial T}$ is the additional compensation that a physician of type i requires in response to a unit increase in the net benefit B. In condition (13), α stands for the average cost of a patient, and minimizing the overall cost of treating the P patients imposes that the additional compensation that either physician requires to treat one further patient be equated to that average cost. As to condition (14), it means that net benefit B must be pushed to the level where at the margin the overall net benefit is equated with the cost of the additional compensation that physicians require for an additional unit of B. With full information, condition (14) can be implemented by means of two different linear payment schemes (one for each physician type):

$$T_i(n) = A_i + \alpha n, \quad i = 1,2 \tag{15}$$

where A_i is set equal to $T_i - \alpha n_i$ with (n_i, T_i) referring to the full-information optimum. This is represented in Figure 2. Note that both pairs $(n_i, T_i), i = 1,2$ are located on the indifference curves corresponding to $V^i = \overline{V}$.

3 Second-best solution with asymmetric information

The full-information optimum just described is not incentive compatible, which is easy to understand. It suffices to note that as represented in Figure 2, for any n the level of effort required to reach a given net benefit B is larger for low-productivity physicians than for high-productivity ones; so, the former require a higher compensation than the latter for attaining utility $\overline{V}$. As shown by Figure 2, the physicians of type 2 would choose the pair (n_1, T_1) and therefore mimic physicians of type 2. This implies that only the second of the IC-constraints will be binding, which in turn implies that the second

of the IT-constraints is always satisfied (this results from $V_2(T_1, n_1) > V_1(T_1, n_1)$). In other words, we have $\mu_1 = 0$ and $\phi_2 = 0$.

We can now derive the first-order conditions for an optimum of the regulator's problem with asymmetric information as it is specified in the Lagrangian expression. These conditions are given with respect to B, N_1, n_2, T_1 and T_2:

$$P + \phi_1 \frac{\partial V_1}{\partial B} + \mu_2 \Big(\frac{\partial V2}{\partial B} - \frac{\partial \hat{V}2}{\partial B} \Big) = 0, \qquad (16.1)$$

$$\lambda P \Big(\sum_{i=1}^{2} p_i n_i \Big)^{-2} p_1 \sum_{i=1}^{2} p_i T_i + \phi_1 \frac{\partial V_1}{\partial n_1} - \mu_2 \frac{\partial \hat{V}_2}{\partial n_1} = 0, \qquad (16.2)$$

$$\lambda P \Big(\sum_{i=1}^{2} p_i n_i \Big)^{-2} p_2 \sum_{i=1}^{2} p_i T_i + \mu_2 \frac{\partial V_2}{\partial n_2} = 0, \qquad (16.3)$$

$$-\lambda P \Big(\sum_{i=1}^{2} p_i n_i \Big)^{-1} p_1 + \phi_1 \frac{\partial V_1}{\partial n_1} - \mu_2 \frac{\partial \hat{V}_2}{\partial T_1} = 0, \qquad (16.4)$$

and

$$-\lambda P \Big(\sum_{i=1}^{2} p_i n_i \Big)^{-1} p_2 + \mu_2 \frac{\partial V_2}{\partial T_2} = 0 \qquad (16.5)$$

where notation $\hat{V}_2 = V_2(n_1, T_1, B)$ refers to the utility of a high-productivity physician mimicking a low-productivity one.

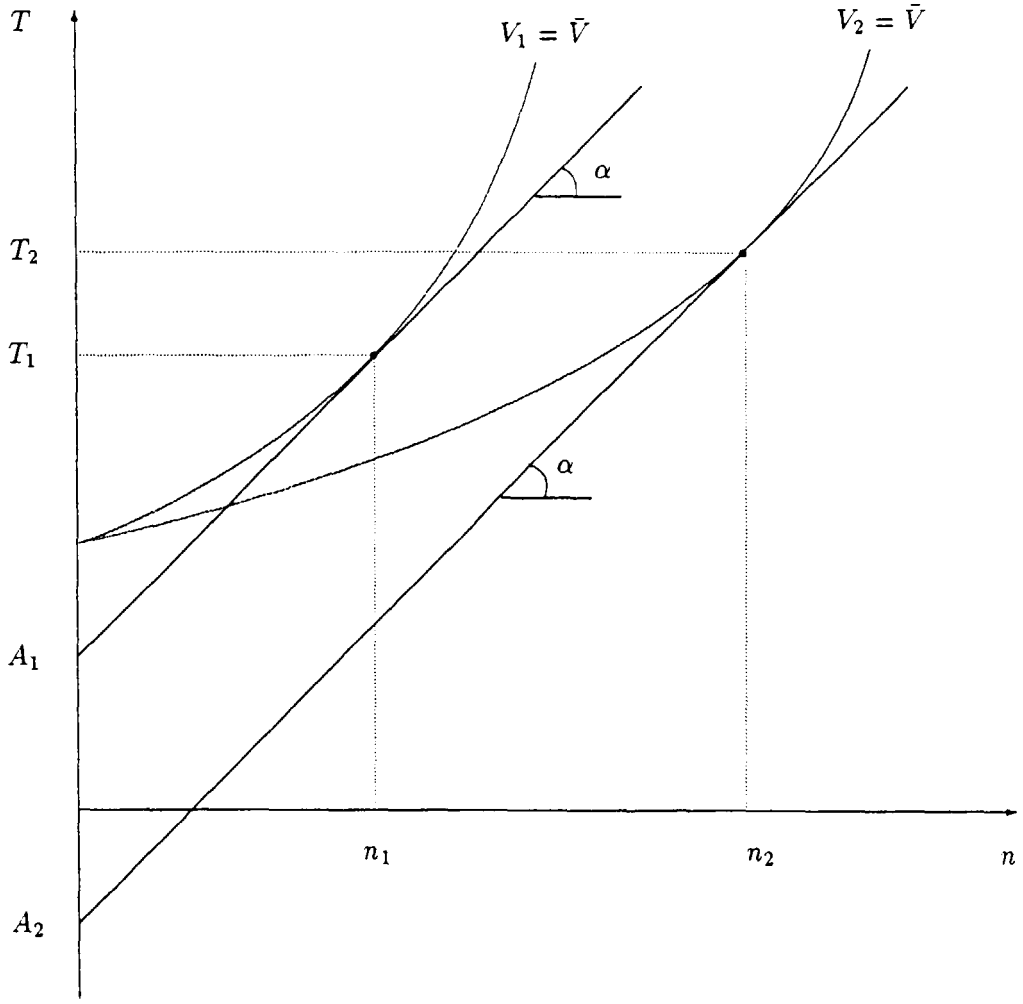
Using the observation that $\frac{\partial V_i}{\partial T_i} = 1 (i = 1, 2)$ and $\frac{\partial \hat{V}_2}{\partial T_1} = 1$, conditions (16.3) and (16.5) yield:

$$MRS^2_{T,n} = \frac{\sum_{i=1}^{n} p_i T_i}{\sum_{i=1}^{n} p_i n_i} \equiv \alpha. \qquad (17)$$

Therefore, the same result as with full information is here obtained (see 13), but it only applies to high-productivity physicians. This is the standard outcome that there is no distortion at the top. Combining conditions (16.2) and (16.4) one gets:

$$p_2 (MRS^1_{T,n} - \hat{MRS}^2_{T,n}) = p_1 (\alpha - MRS^1_{t,n}) \qquad (18)$$

where $\hat{MRS}^1_{T,n}$ is the high-productivity physicians' marginal rate of substitution between T and n taken at point (n_1, T_1). The single-crossing property implies that



$MRS^1_{T,n} > \hat{MRS}^2_{T,n}$, and thus we infer from (18) the following inequality:

$$MRS^1_{T,n} < \alpha, \tag{19}$$

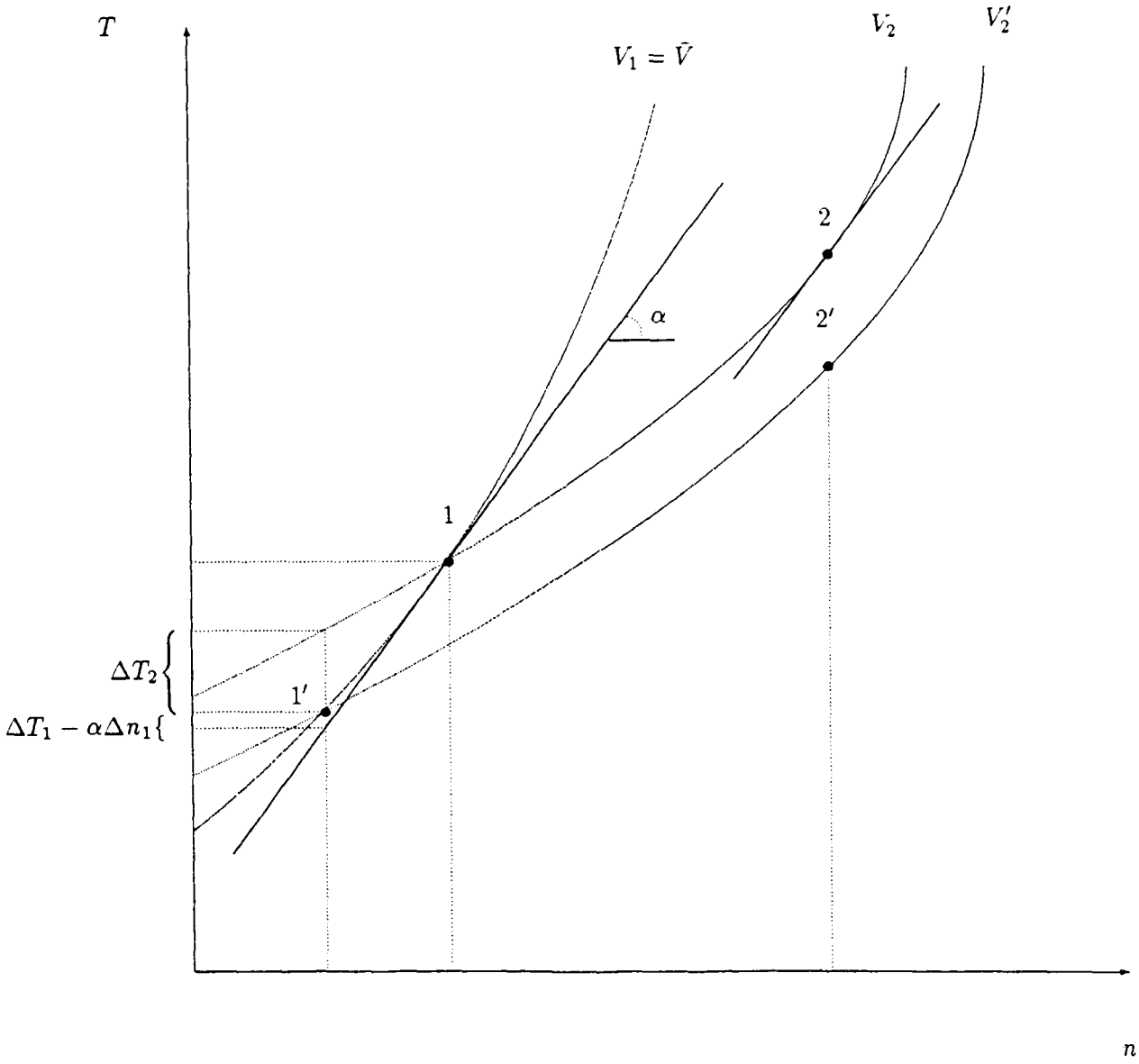
which means that the number of patients treated by each low-productivity physician is reduced relative to the full information optimum (say by $\Delta n_1 < 0$). This causes some inefficiency: the corresponding reduction in his compensation ($\Delta T_1 < 0$) is lower in absolute value than the additional expenses that are incurred for making his deleted patients treated by other physicians ($|\Delta T_1| < \alpha|\Delta n_1|$). However, this inefficiency is desirable since it allows to reduce the informational rent of high-productivity physicians, that is to get closer to their reservation utility by lowering T_2. This rationale is illustrated in Figure 3. If the full-information conditions $MRS^1_{T,n} = \alpha$ were satisfied for both types of physicians, points 1 and 2 would be chosen, and the informational rent left to high-productivity physicians would amount to $V_2 - \overline{V}_2$. By moving point 1 to $1'$ this rent is reduced by ΔT_2 though at the expense of some efficiency loss equal to $|\alpha\Delta n_1 - \Delta T_2|$.

æ

The meaning of condition (18) should now be obvious. The reduction in n_1 should be such that at the margin, the loss of efficiency (right-hand side) is equated to what is gained from reducing the informational rent (left-hand side). These loss and gain are weighted by the relevant proportions of physicians.

Condition (17) and (18) can be implemented by any payment scheme $T(n)$ whose graph in the (n,T)-space goes through the two optimal points (n_1, T_1) and (n_2, T_2) and is, for any other value of n, strictly below the two relevant indifference curves. What has been written above on conditions (17) and (18) applies for any given value of B. It remains to explain how this value is optimally chosen. Using (16.4) and (16.5), condition (16.1) yields:

$$P = \lambda M\left[\sum_{i=1}^{2} p_i MRS^i_{T,B} + p_2(MRS^1_{T,B} - \hat{MRS}^2_{T,B})\right] \tag{20}$$



where $\hat{MRS}^2_{T,B}$ is the $MRS_{T,B}$ for high-productivity physicians taken at (n_1, T_1) (that is the $MRS_{T,B}$ of the mimickers). This condition can be given the same interpretation as condition (14) except that there is here an additional term that pushes up the cost of a unit increase of B. This term can be interpreted in the following manner. Following a unit rise of B the additional compensations, ΔT_2 and ΔT_2, that are required to keep physicians at the same levels of utility violate the incentive-compatibility constraint of high-productivity physicians, and therefore the compensation of these physicians must further be increased by an amount equal to $MRS^1_{T,B} - \hat{MRS}^2_{T,B}$. These observations can be made clear by looking at Figure 1, where at the vertical of n' the dotted indifference curve of low-productivity individuals is above that of high-productivity ones.

4 Extension : more general "production technology"

We now turn to the study of a more general health production function. To do so, we shall first examine how the behavior and preferences of physicians are modified. Then, we shall study how the full information and the second best optimum results are affected by the introduction of this more general production technology.

4.1 The physician's problem revisited

Let us now assume that a patient's health improvement does not only depend on the efficiency units of physician's effort (πe), but also on the amount of medical services prescribed by the physician, y. These prescriptions can be seen as drugs, laboratory tests, physiotherapy, etc; their level is observable by the regulator. The production function of physician $i (i = 1, 2)$ can then be rewritten $h(\pi_i e_i, y_i)$. We assume that it is increasing in $y (\partial h / \partial y > 0)$. Similarly, the effort requirement function of physician i (which continues to specify the level of effort a physician has to provide to offer net benefits of B to n patients) is redefined as $E_i(n, B, y)$. Under Assumption $H1$, an

increase in the level of prescriptions will mean less effort for the provider ($\partial E_i/\partial y < 0$).[4]

The following accounting convention is made concerning the level of prescriptions y : its cost is supported by the provider, but it can be compensated through T. The physician's utility function (1) is then redefined according to this accounting convention and to the extended health production function. Substituting $E_i(n, B, y)$ into the redefined direct utility function yields the following derived utility function:

$$V_i(T, n, B, y) = T - ny + \gamma nh(\pi_i E_i(n, B, y), y_i) - v(nE_i(n, B, y)), \quad i = 1, 2.$$

Once again, this utility function is a crucial element of the problems to be revisited below. Once the "new" payment scheme $\mathbf{T}(n_i, y_i)$ is decided by the regulator, each type of physician chooses y (and n) to maximize his utility. This is achieved when

$$MRS^i_{T,y} \equiv \frac{-\partial V_i/\partial y}{\partial V_i/\partial T} = \frac{\partial \mathbf{T}}{\partial \mathbf{y}} \qquad i = 1, 2,$$

where $MRS^i_{T,y}$ is the marginal rate of substitution between T and y for physician of type i, and $\partial \mathbf{T}/\partial \mathbf{y}$ is the marginal compensation for y.
If $\partial \mathbf{T}/\partial y = 0$, the provider supports the fullcost of y (at the margin); if, on the other hand, $\partial \mathbf{T}/\partial y = n$, he receives full compensation.

4.2 Full information optimum revisited

We shall not rewrite the regulator's problem since its initial formulation as well as its equivalent mechanism design formulation are not modified. However, one should bear in mind that the payment schemes and the utility functions appearing in it have been transformed to take account of y.

As before, the full optimum problem is obtained by deleting the IC constraints from the regulator's problem (i.e. $\mu_i = 0, i = 1, 2$). It is then straightforward to characterize

[4] To see this observe that

$$\frac{\partial E_i}{\partial y} = \frac{-\partial h/\partial y_i}{\pi_i \partial h(\pi_i e_i, y_i)/\partial(\pi_i e_i) - n_i w'(n_i e_i)}$$

is negative because h is increasing in y, while the denominator is negative under $H1$.

the full information optimum. First, the conditions on $n_i (i = 1,2)$ and B are not modified and, consequently, give the same results as before. As to the optimal level of y, the derivative of the Lagrangian (of the full optimum problem) with respect to y yields:

$$MRS^i_{T,y} = 0 \quad i = 1,2.$$

This condition says that, with full information, the marginal costs of y should be equal to their marginal benefits. In other words, the full information optimum requires that both physicians support the full cost of y (at the margin).

4.3 Second-best optimum

We consider the same regulator's problem with asymmetric information as in section 3. We assume that, again, only the $IR1$ and $IC2$ constraints are binding (i.e. $\phi_2 = 0$ and $\mu_1 = 0$). This will allow us to compare (more easily) our results to those obtained above. We derive the first - order conditions of this problem. This is done by differentiating the corresponding Lagrangian function with respect to $B, n_1, n_2, T_1, T_2, y_1$ and y_2. From these derivatives, we first note that the conditions regarding $n^i_1 (i = 1,2)$ and B are not modified. Therefore, they give the same result as above. As to the condition on y_2, one gets :

$$MRS^2_{T,y} = 0$$

which coincide with the full information condition.

APPENDIX

Lemma 1 : *Let $V_i(T,n,B)$ denote the derived utility function of physician $i(i = 1,2)$. The indifference curves for given B in the (n,T) space (associated with V_i) are characterized by the following properties :*

i) they are increasing ;

ii) they are convex ;

iii) they satisfy at any point (n,T) the single-crossing property.

Proof: :

i) The slope of the indifference curves is obtained by totally differentiating V_i and making use of $dV_i = dB = 0$.

This yields

$$\frac{\partial T}{\partial n} = [nv'(ne_i) - \gamma n\pi_i h'(\pi_i e_i)]\frac{\partial E_i}{\partial n}(n,B) + [e_i v'(ne_i) - \gamma h(\pi_i e_i)] \qquad (A.1)$$

where $\partial E_i/\partial n = [\pi_i h'(\pi_i e_i) - nw'(ne_i)]^{-1} e_i w'(ne_i)$ is positive if w is increasing and $H1$ is satisfied. Property i) $(\partial T/\partial n > 0)$ then follows from $H2, H3$ and the fact that $\partial E_i/\partial n$ is positive.

ii) Differentiating $(A.1)$ with respect to n, one obtains [5]

$$\frac{\partial^2 T}{\partial n^2} = (v' - \gamma\pi h')\frac{\partial E}{\partial n} + ev''\frac{d(nE)}{\partial n} +' w' + new'')\frac{\partial(nE)}{\partial n}\left[\frac{v' - \gamma\pi h'}{\pi h' - nw'}\right]$$

$$+ new''\left[\frac{v''\frac{\partial(nE)}{\partial n} - \gamma\pi^2 h''\frac{\partial E}{\partial n}}{\pi h' - nw'}\right]$$

[5] For the rest of the proof, we omit for simplicity the subscript i and the arguments of functions v, h, w, E_i and of their derivatives. For instance we simply write v' instead of $v'(ne_i)$.

$$- new''(v' - \gamma\pi h') \left[\frac{v'' \frac{\partial(nE)}{\partial n} - w' - nw'' \frac{\partial(nE)}{\partial n}}{(\pi h' - nw')^2} \right]$$

where $\partial(nE)/\partial n = -e\pi h'/\pi h' - nw'$ is positive if h is increasing and $H1$ is satisfied. Property ii) ($\partial^2 T/\partial n^2 > 0$ follows from the fact that $\partial E/\partial n$ and $\partial(nE)/\partial n$ are positive, from $H1$ and $H2$, and from assumptions on v (increasing and convex), w (increasing and convex) and h (increasing and concave).

iii) One has to show that $\partial T1/\partial n > \partial T2/\partial n$. Redefining $\partial T/\partial n$ as a function of π, say $\partial T/\partial n(\pi_i) \equiv \partial T_i/\partial n$, and differentiating it with respect to π leads to

$$\frac{\partial}{\partial \pi}\left(\frac{\partial T}{\partial n}\right) = (v' + nv'')\frac{\partial E}{\partial \pi} - \frac{d(\pi E)}{\partial \pi} + n(w' + ew'')\frac{\partial E}{\partial \pi}\left[\frac{v' - \gamma\pi h'}{\pi h' - nw'}\right] \tag{21}$$

$$+ new' \frac{\left[nv'' \frac{\partial E}{\partial \pi} - \gamma h' - \gamma\pi h'' \frac{\partial(\pi E)}{\partial \pi}\right]}{(\pi h' - nw')} \tag{22}$$

$$- new'(v' - \gamma\pi h') \left[\frac{h' + \pi h'' \frac{\partial(\pi E)}{\partial \pi} - n^2 w'' \frac{\partial E}{\partial \pi}}{(\pi h' - nw')^2} \right] \tag{23}$$

where $\partial E/\partial \pi = -eh'/\pi h' - nw'$ and $\partial(\pi E)/\partial \pi = -new'/\pi h' - nw'$ are both negative if $H1$ is satisfied and h and w are increasing.

Hence the single-crossing property follows immediatly from the negativity of $\partial/\partial\pi(\partial T/\partial n)$ which, in turn, is implied by $H1, H2$ and the assumptions put on v, w and h. ■

IV-1 La Prise en compte des Rendements Croissants dans la Gestion de Contrats d 'Assurance

POOLING AND SEPARATING EQUILIBRIA IN INSURANCE MARKETS WITH ADVERSE SELECTION AND DISTRIBUTION COSTS

by

Marie ALLARD*
Jean-Paul CRESTA† and Jean-Charles ROCHET‡

This version September 12, 1997

*Ecole des Hautes Etudes Commerciales (HEC), Montréal, e-mail : Marie.Allard@HEC.CA.
†IDEI, GREMAQ, LEA (Université de Toulouse le Mirail).
‡IDEI, GREMAQ, Université de Toulouse I, e-mail : Rochet@cict.fr.

Abstract

In the Rothschild-Stiglitz (1976) model of a competitive insurance market with adverse selection, pooling equilibria cannot exist. However in practice, pooling contracts are frequent, notably in Health Insurance and Life Insurance. This is due to the fact that distribution costs are non negligible and increase rapidly when more contracts are offered. We modify accordingly the Rothschild-Stiglitz model by introducing such distribution costs. We find that, however small these costs may be, they entail possible existence of pooling equilibria. Moreover, in these pooling equilibria, it is the high-risk individuals who are rationed, in the sense that they would be willing to buy more insurance at the current premium/insurance ratio.

Key Words : Adverse Selection, Insurance Markets, Marketing costs, Pooling Equilibria, Separating Equilibria.
JEL Classification system: G22, D82, D43, I11.

1 Introduction

The seminal paper of Rothschild and Stiglitz (1976) is important for many reasons. It was among the first to recognize that "some of the most important conclusions of economic theory are not robust to considerations of imperfect information" (p. 629). It provides a very simple framework to analyse how competitive markets work while facing adverse selection phenomena. Even though it is highly stylised, the Rothschild-Stiglitz model has, for the last twenty years, given rise to a vast literature and, therefore, is rightly considered as a classic.

Whatever the worth and merits of this article, it predicts that there cannot exist pooling equilibria, i.e. equilibria which entail the same insurance plan for different types of customers. However, pooling contracts are frequent in practice. For example, in the context of medical insurance in the USA, Newhouse (1996) writes: "...a plan generally receives the same premium for all single employees of a given firm who choose that plan, even though some may have a chronic disease that increases their expected cost to the plan and others do not". Also, HMOs are organisations that typically offer the same contract to heterogeneous risks. Another example is group contracts in life insurance.

Neipp and Zeckhauser (1985) (cited in Newhouse (1996)) gave a first possible answer to the question of why one observes such a phenomenon. They introduce the notion of stickiness among the consumers who choose an insurance plan: enrollees tend to maintain in the previously chosen plan. However, as suggested by Newhouse (1996), plans might well be able to overcome such stickiness by offering attractive new options.

Another possible reason for such a pooling, first explored by Newhouse (1996)[1] is the presence of distribution costs. The importance of these costs, as well as the presence of non negligible scale economies, is well documented, notably in health insurance (see for instance Diamond (1992)). It is therefore interesting to study the interaction of such scale economies with adverse selection.

The purpose of this paper is to show that a simple modification to the traditional Rothschild-Stiglitz model can alter its conclusions in a way that seems more consistent with what is observed. The modification we propose is introducing distribution costs of a simple type : we assume that each contract offered carries a fixed set-up cost c (therefore offering a contract is no more costless). This is the only change we consider with respect to the Rothschild-Stiglitz model.

Our main results can be sketched as follows. We obtain a complete characterization

[1] Newhouse assumes for simplicity that only the contracts offered to low-risk individuals incur such costs. We relax this assumption and systematically explore the nature of the competitive equilibrium as a function of the relative importance of distribution costs and adverse selection.

of Nash Equilibria, as a function of c and t, where t represents the proportion of high-risk individuals. Like in the Rothschild-Stiglitz model, we are able to show that there cannot be more than one type of equilibrium[2] and that an equilibrium exists for a large enough value of t. However, contrary to Rothschild-Stiglitz, we are also able to show that for any $c > 0$, the equilibrium may be pooling. Moreover, when a pooling equilibrium exists there is, generally, a range of possible pooling equilibria containing the pooling equilibrium described by Newhouse (1996). Of course, it is hardly surprising that introducing set-up costs leads to the possibility of pooling equilibria. Our punchline is rather that, for **arbitrarily small** values of c, pooling equilibria may exist. Another interesting aspect of our results is that when a pooling equilibrium exists, it is high-risk individuals who are rationed (not low-risk persons like in the Rothschild-Stiglitz (1976) model), while low risks could be forced to buy more insurance than they want at the current price. These results seem to better accord with reality.[3],

The rest of the paper is organized as follows. In section 2 we present the model and the characterization of equilibria. We then characterize the values of parameters (t, c) such that equilibria do exist. This is done for pooling (section 3) and separating equilibria (section 4).

2 The Model

We consider the well known model of Rothschild and Stiglitz (1976) and we slightly modify it by introducing a set-up cost c per contract offered, corresponding to the cost of designing the contract and marketing it.[4]

As in their model, an individual can insure himself against a risk (hereafter called an accident) of damage D by signing an insurance contract $Z = (q, p)$. According to this contract, the insuree has to pay to an insurance company a premium p in return for which he will receive q if an accident occurs. As usual, we require $q \leq D$. The market consists of two kinds of customers: low-risk individuals with accident probability π_L and

[2]We mean that a pooling equilibrium cannot coexist with a separating equilibrium. However there are typically several pooling equilibria.

[3]For instance, Newhouse (1996) states : "it is high-risk, not low-risk individuals who tend to have trouble obtaining the desired amount of insurance" (p. 1242). However, we must be careful that we are reasoning on a group of individuals with the same **observable** characteristics. Therefore, in a pooling contract, high risks cannot be detected a priori. It is only their future accident (or illness) records that will allow to separate them from low risks.

[4]Other management costs (selling costs, proportional to the number of customers who buy the contracts, or settlement costs, proportional to the damage reimbursements paid to insurees), could have been introduced (see for example Newhouse (1996)) without fundamentally altering the conclusions of Rothschild and Stiglitz. The crucial property of our marketing costs is that they are fixed, which introduces increasing returns to scale.

high-risk individuals with accident probability $\pi_H > \pi_L$. Apart from this difference in their accident probabilities, individuals are identical in all respects: they have a gross wealth (or income) W and their preferences are represented by a concave von Neumann-Morgenstern utility function denoted u. The utility that an individual of type i ($i = H, L$) gets from an insurance contract $Z = (q, p)$ is

$$V_i(Z) = \pi_i u(W - p + q - D) + (1 - \pi_i)u(W - p).$$

Insurance companies know the statistical distribution of types and in particular the average probability of accident in the population $\pi_m = t\pi_H + (1 - t)\pi_L$, where t is the fraction of high-risk individuals in the population. Taking account of the fact that offering a contract is no more costless (every contract offered carries a marketing cost c), the profit (per customer) obtained from a contract $Z = (q, p)$ can be:

- $p - \pi_m q - c$ if everybody[5] buys it,
- $p - \pi_L q - \frac{c}{1-t}$ if only low risks buy it,
- $p - \pi_H q - \frac{c}{t}$ if only high risks buy it.

Let Δ_m, Δ_L and Δ_H represent the associated zero profit lines (fair-odds lines). Contrarily to the Rothschild-Stiglitz model, these lines don't intersect for $q = 0$. The intersection of Δ_m and Δ_L will be denoted by $K = (q_K, p_K)$ -see figure 1 below-.

We consider a sequential game in which at each stage a new firm decides to offer a contract or not, given the contracts offered by the firms already in place.

It is easy to see that any equilibrium of this game is characterized by the two properties :

- none of the contracts offered makes losses,
- no entrant can offer a new contract and make a profit.

Such an equilibrium can exhibit pooling (i.e. the same contract Z_m for both groups) or be separating (i.e. two separate contracts: Z_H for high risks and Z_L for low risks). The marketing cost c introduces non convexities which complicate the analysis with respect

[5]Note that contrarily to the Rothschild-Stiglitz model, the number of active firms and the way they share the market become relevant. If several firms were to offer the same contract, any one of them would have interest to undercut its rivals by offering a sligthly lower price, for attracting their clients and thus decreasing their marketing costs per unit. This may introduce inexistence problems for the free entry equilibrium, even without adverse selection. We solve these problems by assuming sequential entry (see our definition of equilibrium below).

to that of Rothschild and Stiglitz. For example, we may have equilibria where some individuals do not buy insurance. For the sake of simplicity, we will temporarily rule out these equilibria by assuming that insurance is compulsory (as it is implicit in the Rothschild-Stiglitz model when the possibility of offering a pooling contract is examined).

Another interesting feature of our model is that low risks can benefit from pooling[6], which was never the case in the standard Rothschild-Stiglitz framework. This happens when

$$\pi_m q + c \leq \pi_L q + \frac{c}{1-t}.$$

This condition is equivalent to

$$q \leq q_K = \frac{c}{\Delta\pi(1-t)}$$

where $\Delta\pi$ is the difference between π_H and $\pi_L(\Delta\pi = \pi_H - \pi_L)$. When q_K exceeds the maximum possible coverage D, everyone always gains from pooling. We will focus here on the more interesting case where $q_K < D$ (see figure 1), which corresponds to the following condition :

$$c < \Delta\pi(1-t)D. \tag{1}$$

It is clear that candidates for equilibrium have to be on the (adequate) zero profit lines (i.e. $Z_m \in \Delta_m$, $Z_H \in \Delta_H$, $Z_L \in \Delta_L$). Moreover they have to be undominated on these zero profit lines. For example, for Z_m to be a pooling equilibrium, it must be that: there is no other pooling contract ($Z'_m \in \Delta_m$) such that $V_H(Z'_m) \geq V_H(Z_m)$ and $V_L(Z'_m) \geq V_L(Z_m)$ (with at least one strict inequality).

Notice that, along Δ_m, V_H increases with q, while V_L has a maximum in $E = (q_E, p_E)$. Therefore the set of undominated pooling contracts on Δ_m is exactly the segment EF where $F = (D, \pi_m D + c)$ is a pooling contract with complete coverage. (see figure 1 below).

Similarly a separating equilibrium has to be undominated on the set of feasible separating (pairs of) contracts defined by:

$$\mathcal{F} = \Big\{\big(Z_H = (q_H, p_H), Z_L = (q_L, p_L)\big) / p_H \geq \pi_H q_H + \frac{c}{t},$$

$$p_L \geq \pi_L q_L + \frac{c}{1-t}, V_H(Z_H) \geq V_H(Z_L), V_L(Z_L) \geq V_L(Z_H)\Big\}.$$

A straightforward adaptation of the reasoning of Rothschild and Stiglitz shows that

[6]High risks always benefit from pooling since scale economies comfort their gain from a smaller marginal premium.

there is a unique undominated pair of contracts (H, L), defined by[7] :

$$\begin{cases} H = \left(D, \pi_H D + \frac{c}{t}\right) \\ L \in \Delta_L \quad \text{and} \quad V_H(H) = V_H(L). \end{cases}$$

The following figure represents all these important benchmarks.

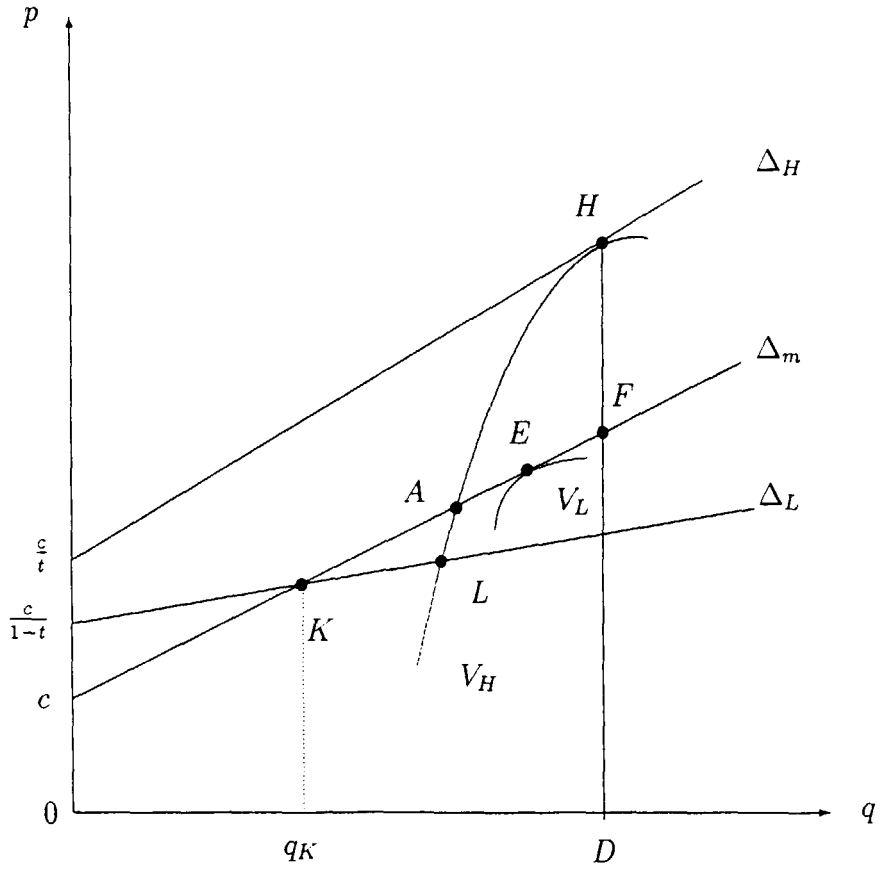


Figure 1

The sets of undominated contracts on the zero profit lines are thus :

- the interval $[E, F]$ for pooling contracts,
- the singleton (H, L) for (pairs of) separating contracts.

We denote by A the intersection of Δ_m with the indifference curve of high risks that goes through H and L.

[7] When condition (1) is not satisfied, one may obtain strange situations where H is below Δ_L and therefore low risks always prefer H to their own contract.

3 Characterization of Equilibria and Pareto Optima

The next proposition gives the characterization of existence of a pooling equilibrium.

Proposition 1 : *Let $Z = (q, p)$ belong to the segment EF, that is, the set of undominated pooling contracts. Z is a pooling equilibrium if and only if:*

$$\begin{cases} V_H(Z) \geq V_H(H) & (1) \\ Z \text{ is below } \Delta_L, \text{ i.e. } p \leq \pi_L q + \frac{c}{1-t} & (2) \end{cases}$$

Proof: We have to prove that the introduction of any new contract Z' would be unprofitable. This is clear if Z' attracts all consumers, since Z is undominated on the zero-profit line Δ_m. Now consider the set of contracts Z' that would attract high risks alone: it corresponds to the dashed region in figure 2.

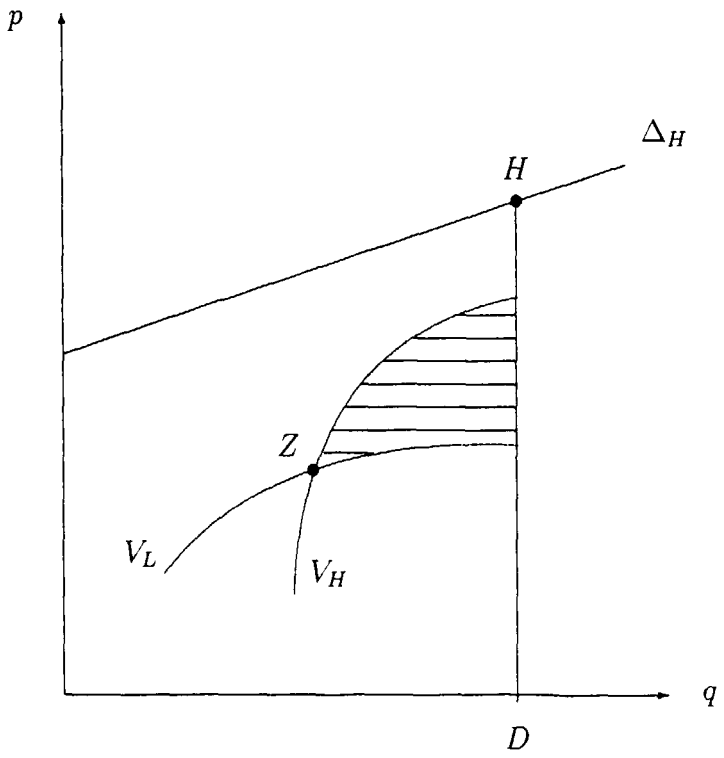


Figure 2

(This corresponds to the case where K' is on the left of A and E)

This region does not contain any profitable contract if and only if it is entirely below Δ_H, a condition which is equivalent to (1).

We then turn to the set of contracts Z' that would attract only low-risk individuals: it corresponds to the dashed region in figure 3. This region does not contain any profitable

contract if and only if condition (2) is satisfied. Notice that in the original model of Rothschild-Stiglitz, Δ_m is entirely above Δ_L, therefore condition (2) is never true.

■

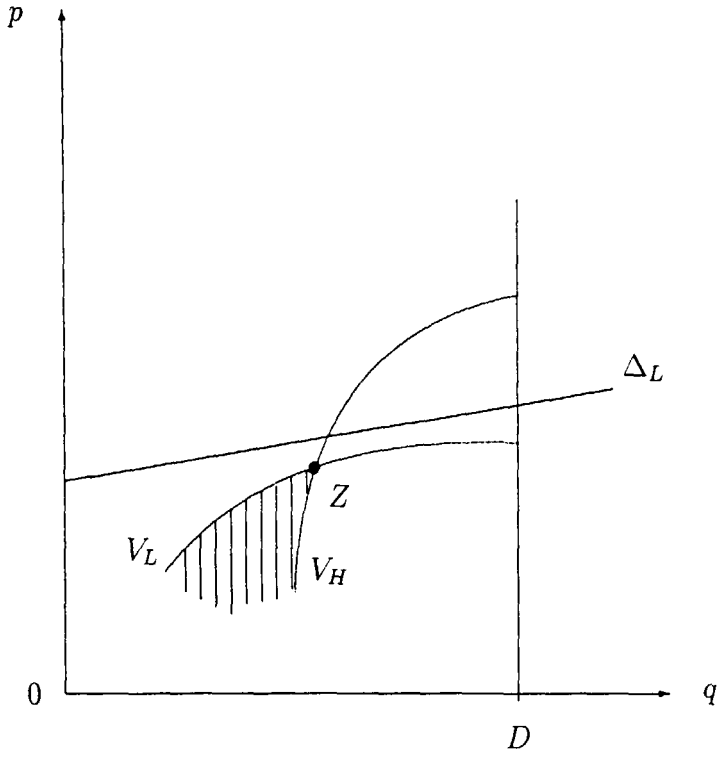


Figure 3

Corollary 1: *The set of pooling equilibria is exactly the intersection of* $[E, F]$ *(the set of undominated pooling contracts) and* $[A, K]$ *(the set of pooling contracts that are not dominated by* (H, L)*.*

Proof: Immediate from proposition 1, given the definitions of A and K.

■

Before characterizing the existence of a separating equilibrium (which can only be (H, L)) we need to remark that it is destabilized by a pooling contract $Z \in \Delta_m$ if and only if $V_H(Z) \geq V_H(H)$ and $V_L(Z) \geq V_L(L)$ (with at least one strict inequality).

Proposition 2 *:* (H, L) *is a separating equilibrium if and only if* $V_L(L) > V_L(E)$*.*

Proof: Let a new contract Z be introduced:

- if Z attracts only high risks it is unprofitable since H is the most preferred contract by high risks on Δ_H;

- if Z attracts only low risks it is also unprofitable since L is the most preferred contract by low risks in $\Delta_L \cap \{Z | V_H(Z) \leq V_H(H)\}$.

Therefore (H, L) is an equilibrium if and only if it is not destabilized by a pooling contract. This implies proposition 2, since E is the most preferred contract on Δ_m by low risks.

Lemma 1 : *If* $V_L(L) > V_L(E)$ *then* $q_K < q_A$.

Proof: By contradiction: suppose $q_A \leq q_K$ then $q_A \leq q_L$ (see figure 4) and $V_L(A) \geq V_L(L)$ (since V_L decreases along the indifference curves of V_H). Moreover by definition of E, $V_L(E) \geq V_L(A)$ therefore we obtain the desired contradiction that $V_L(E) \geq V_L(L)$.

Corollary 2: *Pooling equilibria cannot coexist with separating ones.*

Proof: Immediate from corollary 1 and lemma 1.

■

Characterization of Pareto Optima

As we have seen the set of Pareto optimal contracts (under self selection and non negative profit conditions) is included in the reunion of $\{(H, L)\}$ (the Pareto dominating separating contract) and $[EF]$ (the set of undominated pooling contracts). We are now going to examine the different configurations corresponding to the respective locations of these sets in the utility plane (V_L, V_H).

1st case: $V_L(L) \geq V_L(E)$

This is the case where (H, L) is an equilibrium

The set of Pareto Optima is the reunion of $\{H, L\}$ (the separating equilibrium) and the set $[A, F] \cap [E, F]$ of undominated pooling contracts that high risks prefer to H. However none of these contracts can be an equilibrium since they don't attract low risks.

2nd case: $V_L(L) < V_L(A)$

(H, L) is Pareto dominated and the set of Pareto Optima is $[E, F]$.

3nd case: $V_L(A) \leq V_L(L) < V_L(E)$

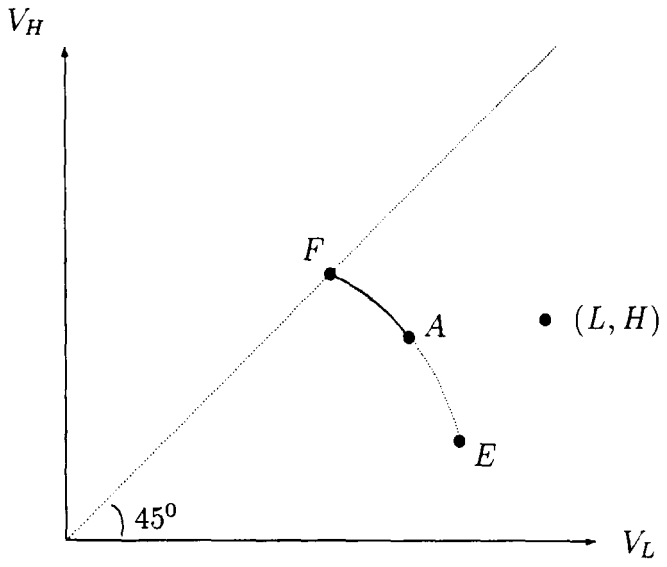


Figure 4a: 1st case

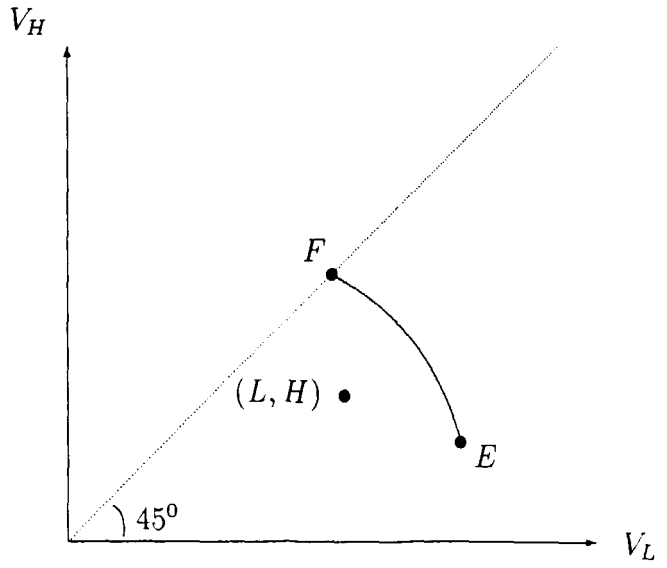


Figure 4b: 2nd case

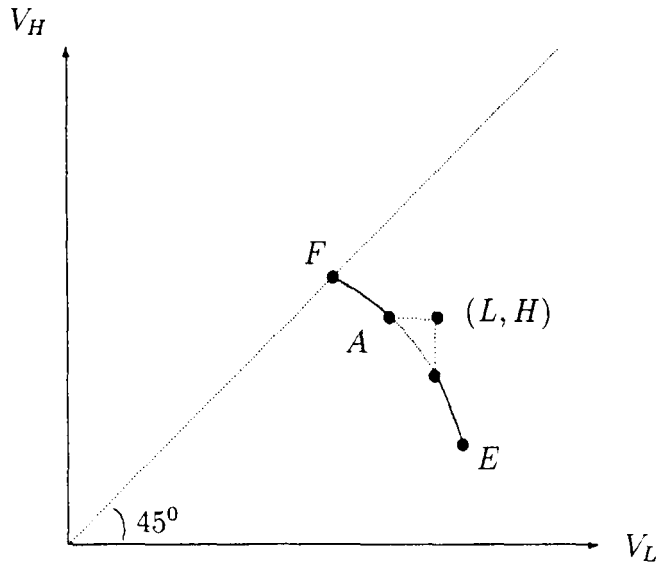


Figure 4c: 3nd case

Here (H, L) is Pareto optimal even though it is not an equilibrium.

In the last two cases, the set of equilibria (which are necessarily pooling) depends on the position of K. However all equilibria are necessarily Pareto Optimal but contrarily to the Rothschild-Stiglitz model, there are cases in which (H, L) is Pareto Optimal but not an equilibrium (case 3 above).

We are now going to characterize the values of parameters (t, c) such that either type of equilibria exists. We start with the pooling case.

4 Pooling equilibria

Lemma 2 : *When $c < \Delta\pi(1-t)D$, a pooling equilibrium exists if and only if*

$$q_A \leq q_K \tag{3}$$

and

$$q_E \leq q_K. \tag{4}$$

Proof : The condition $c < \Delta\pi(1-t)D$ exactly means that $q_K < q_F = D$. Therefore the set of pooling equilibria is just the line $[\max(A, E), K]$. It is non empty if and only if conditions (3) and (4) are satisfied. ■

The next figure illustrates the set of pooling equilibria when $q_A \leq q_E \leq q_K$, i.e. the line $[E, K]$. It is very easy to see (but not illustrated) that if $q_E \leq q_A \leq q_K$, the set of pooling equilibria will then be the line $[A, K]$.

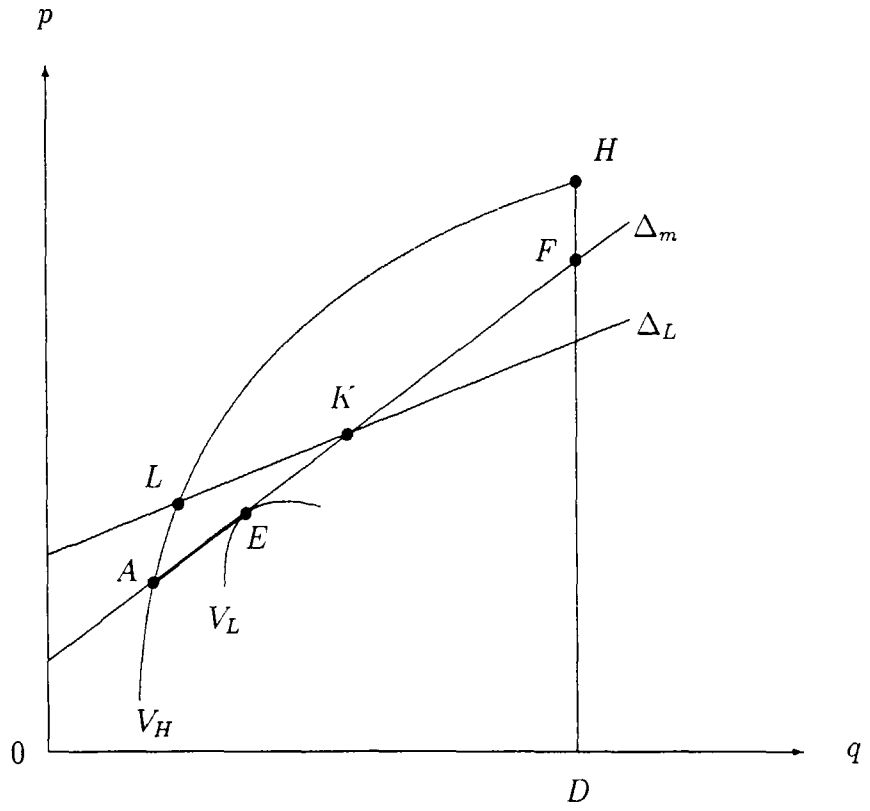


Figure 5: The set of pooling equilibria is $[AE]$.

The next two lemmas characterize the values of parameters (t, c) such that a pooling equilibrium exists. While lemma 3 determines when condition (3) is satisfied, lemma 4 is dedicated to condition (4).

Lemma 3 *: There exists a function $c_1(t)$ such that*

$$q_A \leq q_K \qquad \Leftrightarrow \qquad c \geq c_1(t).$$

Moreover $c_1(0) = c_1(1) = 0$.

Proof : We have $q_A \leq q_K \Leftrightarrow V_H(A) \leq V_H(K)$. But

$$q_K = \frac{c}{\Delta\pi(1-t)} \text{ and } p_K = \frac{\pi_H c}{\Delta\pi(1-t)} = \pi_H q_K,$$

so that :

$$V_H(K) = (1-\pi_H)u(W - \pi_H q_K) + \pi_H u[W - D + (1-\pi_H)q_K].$$

Moreover $V_H(A) = V_H(H) = u(W - \pi_H D - \frac{c}{t})$.

Let the auxiliary function $\gamma(q)$ be defined implicitly by :

$$u[W - \pi_H D - \gamma(q)] = (1 - \pi_H)u(W - \pi_H q) + \pi_H u[W - D + (1 - \pi_H)q].$$

$\gamma(q)$ is the risk premium associated by high risks to a contract $(q, \pi_H q)$. It is then clear that $V_H(A) \leq V_H(K)$ if and only if :

$$\gamma(q_K) \leq \frac{c}{t}, \quad \text{with} \quad q_K = \frac{c}{\Delta\pi(1-t)}$$

(since u is increasing). It is easy to see that γ decreases on $[0, D]$ and that $\gamma(D) = 0$.

For t fixed $\in (0,1)$ and c_1 varying in $(0, D\Delta\pi(1-t))$, let us define

$$\phi(c_1) = \gamma\left[\frac{c_1}{\Delta\pi(1-t)}\right] - \frac{c_1}{t}$$

γ being continuous and decreasing, so is ϕ. Moreover $\phi(0) = \gamma(0) > 0$ and $\phi[D\Delta\pi(1-t)] = -\frac{D\Delta\pi(1-t)}{t} < 0$. Therefore for all $t \in (0,1)$, there is a unique c_1 such that $\phi(c_1) = 0$. Thus we can define a function $c_1(t)$ such that

$$\gamma\left[\frac{c_1(t)}{\Delta\pi(1-t)}\right] = \frac{c_1(t)}{t}. \tag{5}$$

We then have just to remark that γ being decreasing, we have $\gamma\left[\frac{c}{\Delta\pi(1-t)}\right] \leq \frac{c}{t}$ if and only if $c \geq c_1(t)$, which implies the desired property :

$$q_A \leq q_K \Leftrightarrow c \geq c_1(t).$$

Finally, it remains to notice three things :

- by construction, we have

$$0 \leq c_1(t) \leq D\Delta\pi(1-t). \tag{6}$$

- γ is bounded above by $\gamma(0)$ so that (5) implies

$$0 \leq c_1(t) \leq \gamma(0)t. \tag{7}$$

- strictly speaking, $c_1(0)$ and $c_1(1)$ are not defined. However, by continuity, (6) and (7) show that it is natural to set : $c_1(0) = c_1(1) = 0$.

■

Lemma 4 : *There exists a function $c_2(t)$ such that*

$$q_E \leq q_K \qquad \Leftrightarrow \qquad c \geq c_2(t).$$

Proof : Since E is the maximum of V_L on Δ_m, and since K also belongs to Δ_m, we have, by concavity of V_L :

$$q_E \leq q_K \Leftrightarrow -\left(\frac{\partial V_L}{\partial q} \Big/ \frac{\partial V_L}{\partial p}\right)(K) \leq \pi_m. \tag{8}$$

A crucial property is that, when c varies (from 0 to $\Delta\pi D(1-t)$), K remains on the fixed line $p = \pi_H q$ (and q varies from 0 to D). Along this line, let $A(q)$ denote the marginal rate of substitution of low risks, i.e. :

$$A(q) = -\left(\frac{\partial V_L}{\partial q} \Big/ \frac{\partial V_L}{\partial p}\right)(q, \pi_H q).$$

It is easy to see that for $0 \leq q \leq D$, $A(q)$ decreases from $A(0)$ to $A(D) = \pi_L$.

Since $q_K = \frac{c}{\Delta\pi(1-t)}$ and $\pi_m = \pi_L + t\Delta\pi$, condition (9) is equivalent to :

$$H(c,t) \stackrel{def}{=} A\left(\frac{c}{\Delta\pi(1-t)}\right) - \pi_L - t\Delta\pi \leq 0. \tag{9}$$

Denoting by $q^*(\cdot)$ the inverse function of $A(\cdot)$, this is equivalent to :

$$c \geq c_2(t) \stackrel{def}{=} \Delta\pi(1-t)q^*(\pi_L + t\Delta\pi). \tag{10}$$

Notice that when $c = \Delta\pi(1-t)D$ we have

$$H(c,t) = A(D) - \pi_L - t\Delta\pi = -t\Delta\pi \leq 0.$$

Therefore, since H is decreasing in c,

$$c_2(t) \leq \Delta\pi(1-t)D. \tag{11}$$

In fact, $q^*(\pi_m)$ is negative when $\pi_m > A(0)$, in which case (11) is always satisfied. For convenience we will then set $c_2(t) = 0$, which guarantees that $c_2(t) \geq 0$. Condition (11) then implies that $c_2(1) = 0$. Notice also that

$$c_2(0) = \Delta\pi q^*(\pi_L) = \Delta\pi D.$$

Moreover, $A(\cdot)$ being decreasing, so is $q^*(\cdot)$, which implies that $c_2(\cdot)$ is also decreasing.

■

In summary, we have :

Proposition 3 : *A pooling equilibrium exists if and only if*

$$c \geq \max(c_1(t), c_2(t)).$$

Therefore, contrarily to the well-known result of Rothschild-Stiglitz, pooling equilibria may exist. We will see that for any $c > 0$ there exist values of t such that this is true.

5 Separating Equilibria

Lemma 5 : **a)** *If $q_A < q_K$ there is no separating equilibrium.*
b) *As a consequence, if*
$q_A < q_K < q_E$ there is no equilibrium at all (in pure strategies).[8]

Proof : **a)** is immediate from lemma 1 and proposition 1.
b) is an easy consequence of lemma 2.

■

Lemma 6 : *If $q_A > q_K > q_E$ the separating contract (H, L) is an equilibrium.*

We have to establish that (H, L) is not dominated by a pooling contract Z.

Proof : Recall that the set of undominated pooling contracts is the interval $[E, F]$. Given that $q_A > q_K > q_E$, we will consider the two following cases :

a) $Z \in [K, F]$: then $V_L(Z) \leq V_L(K)$ since $q_E < q_K \leq q_Z$ and V_L decreases with q along Δ_m for all $q \geq q_E$. Moreover since $q_A > q_K$ then $q_L > q_K$ (see figure 5), and therefore $V_L(L) > V_L(K)$ (since both L and K belong to Δ_L). As a consequence $V_L(Z) < V_L(L)$, which means that the contract Z cannot dominate (H, L).

b) $Z \in [E, K[$: then $V_H(Z) < V_H(K)$ since $q_Z < q_K$. But we also have $V_H(K) < V_H(A) = V_H(H)$. Consequently $V_H(Z) < V_H(H)$ which implies that Z cannot dominate (H, L).

Hence there is no pooling contract Z that dominates (H, L). By proposition 2, (H, L) is always an equilibrium. This result is illustrated in figure 6.

■

[8] Obviously if we use lemma 2 and lemma 3, lemma 4 can also be stated as follows: If $c > c_1(t)$ there is no separating equilibrium. As a consequence, if $c_1(t) < c < c_2(t)$ there is no equilibrium at all.

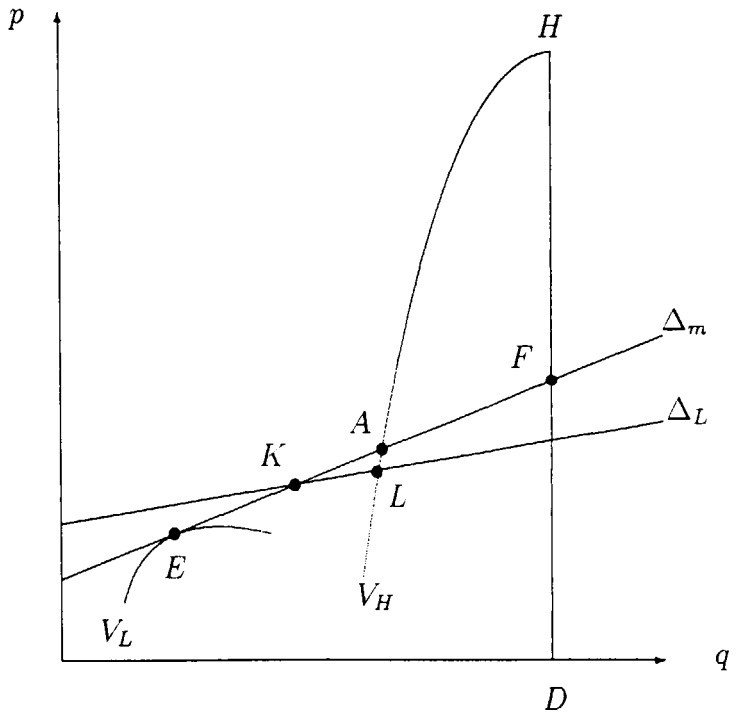


Figure 6 : Existence of a separating equilibrium when $q_A > q_K > q_E$.

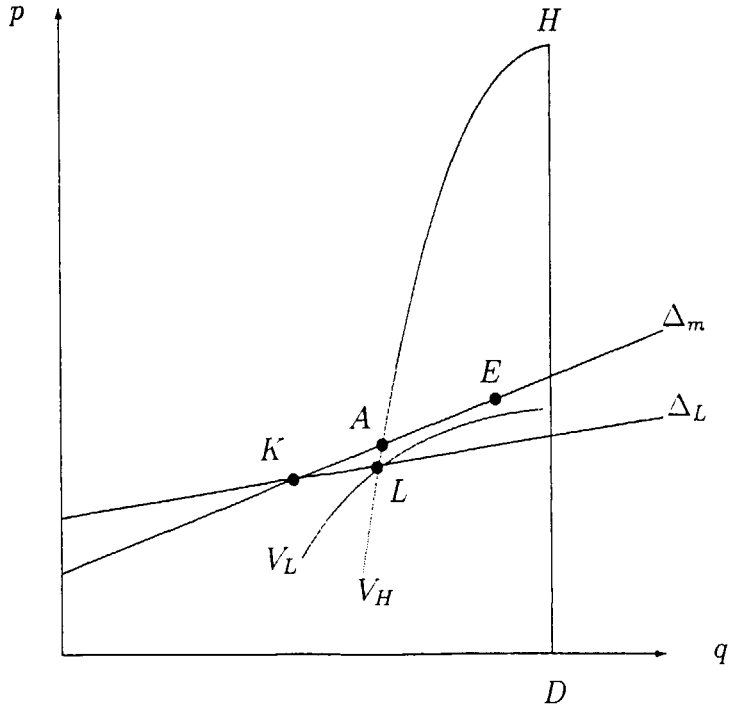


Figure 7: Existence of a separating equilibrium when $V_L(L) > V_L(E)$.

■

This result is illustrated in figure 7. Notice that, by lemma 4, we necessarily have in this case $q_A \geq q_K$. The following lemma clarifies that the converse is not true.

Lemma 7 : *When $q_E > q_K$ and $V_L(L) < V_L(E)$, there is no equilibrium (in pure strategies).*

Proof : We first prove that when $V_L(L) < V_L(E)$, there is no separating equilibrium.

Indeed, L is destabilized by E, which makes no loss whether or not high risks are also attracted. It is then obvious that there is no equilibrium since, by lemma 1, $q_E \leq q_K$ is a necessary condition for the existence of a pooling equilibrium.

■

Notice that when $q_A \geq q_K$ and $q_E \geq q_K$ (like in lemmas 6 and 7), we obtain a direct extension of a result in Rothschild-Stiglitz (1976) : the separating contract (H, L) is an equilibrium if and only if low-risk individuals prefer L to E, and therefore to any (pooling) contract on Δ_m (see figure 7). It remains to characterize the set of parameters (t, c) for which this property is true: like in Rothschild-Stiglitz, it will be satisfied for large enough values of t.

Lemma 8 : *There exists a function $t(c)$ such that*

$$V_L(L) > V_L(E) \qquad \Leftrightarrow \qquad c \leq \max(c_1(t), c_2(t)) \text{ and } t > t(c).$$

Proof : We define an auxiliary function $\psi(c, t)$ as follows :

$$\psi(c, t) = V_L(q_L, p_L) - V_L(q_E, p_E)$$

where

$$V_L(q_E, p_E) = \max_q V_L(q, \pi_m q + c), \tag{12}$$

$p_L = \pi_L q_L(c, t) + \frac{c}{1-t}$, and $q_L = q_L(c, t)$ is defined implicitly by :

$$V_H(q_L, \pi_L q_L + \frac{c}{1-t}) = u\left(W - \pi_H D - \frac{c}{t}\right). \tag{13}$$

By totally differentiating (13) with respect to t, we get :

$$\left(\frac{\partial V_H}{\partial q} + \pi_L \frac{\partial V_H}{\partial p}\right) \frac{\partial q_L}{\partial t} = \left(\frac{c}{t^2} u'\left(W - \pi_H D - \frac{c}{t}\right) - \frac{c}{(1-t)^2} \frac{\partial V_H}{\partial p}\right) > 0.$$

Since V_H increases with q on Δ_L, this implies that $\frac{\partial q_L}{\partial t} > 0$. Similarly, V_L is also increasing with q on Δ_L so that $V_L(q_L, p_L)$ is an increasing function of t. Moreover, by

(12), $V_L(q_E, p_E)$ is a maximum of decreasing functions of t, thus it is also decreasing in t. Therefore ψ is an increasing function of t, so for all c there is at most one value $t(c)$ such that

$$\psi(c, t(c)) = 0.$$

We now examine when such a value exists. From the proof of lemma 4, we already now that it does not exist when $c < c_1(t)$ (since this implies $q_A < q_K$). When $c = c_1(t)$, L coincides with K (and A), so that

$$V_L(L) \geq V_L(E) \Leftrightarrow V_L(K) \geq V_L(E) \Leftrightarrow c \geq c_2(t).$$

Therefore, by contraposition,

$$(c = c_1(t), c < c_2(t)) \Rightarrow \psi(c, t) < 0. \tag{14}$$

consider now the symmetric case where $c = c_2(t)$ (therefore $E = K$) and $c < c_1(t)$ (therefore $q_A > q_K$). It is easy to see that, in this case, the indifference curve of low risks that goes through $K = E$ does not interest Δ_L in the region $q < D$. Since L belongs to this region, we have established $V_L(L) > V_L(E)$, and therefore

$$(c = c_2(t), c < c_1(t)) \Rightarrow \psi(c, t) > 0.$$

To finish the proof it remains to use the fact that the two curves $c = c_1(t)$ and $c = c_2(t)$ have a unique intersection (t^*, c^*) (see proposition 5). Therefore :

$$\forall t < t^*, c_1(t) < c_2(t) \text{ and } \psi(c_1(t), t) < 0.$$

Also,

$$\forall t > t^*, c_1(t) > c_2(t) \text{ and } \psi(c_2(t), t) < 0.$$

Thus, for all c in $[0, c^*]$, there is a (unique) $t(c)$ such that : $\psi(c, t(c)) = 0$ (see figure 7). For $t < t(c)$ there is no equilibrium. For $t > t(c)$ and $c \leq c_1(t)$ there exists a separating equilibrium.

■

The conditions for the existence of a separating equilibrium can be summarized as follows :

Proposition 4 : *A separating equilibrium exists if and only if* $c \leq (c_1(t)$ *and*

$$t > t(c).$$

To sum up, we have obtained three functions $c_1(t)$, $c_2(t)$ and $t(c)$ such that:

$$q_A \leq q_K \quad \Leftrightarrow \quad c \geq c_1(t)$$

$$q_E \leq q_K \quad \Leftrightarrow \quad c \geq c_2(t)$$

A pooling equilibrium (proposition 3) exists if and only if q_A and q_E are less than q_K, i.e., $c \geq \max(c_1(t), c_2(t))$. A separating equilibrium (which can only be (H, L)) (proposition 4) exists if $c \leq \max(c_1(t), c_2(t))$, and $t > t(c)$. This is represented in the following figure:

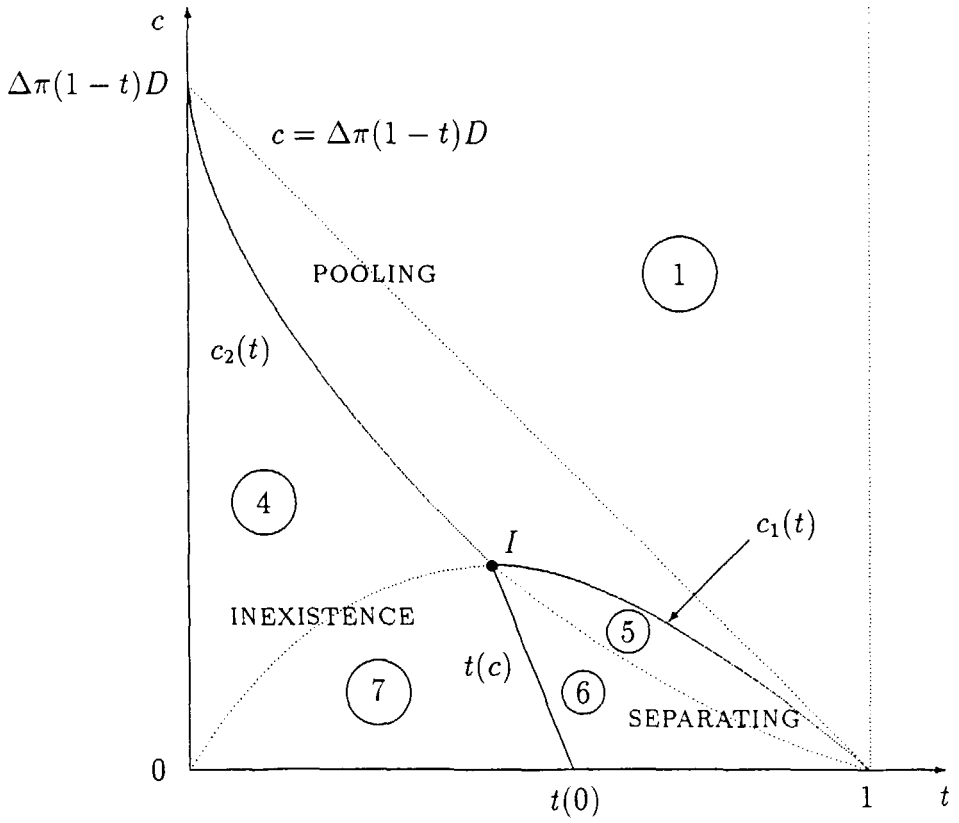


Figure 8: Equilibrium zones: ① pooling, ④ and ⑦ inexistence, ⑤ and ⑥ separating.

The different regions in figure 8 are numbered like the respective lemmas. $t(0)$ represents the value of t below which there is no equilibrium in the Rothschild-Stiglitz model. As suggested by figure 7, for any $c > 0$, the inexistence is less likely (that is $t(c)$ is decreasing) and the possibility of pooling exists. As suggested also by figure 7, the three curves intersect in I, which corresponds to the values of t and c such that all the points

A, E, K, L in the (q,p) plane coincide. More formally, the last property can be stated as follows :

Proposition 5 : *The curves $c = c_1(t)$ and $c = c_2(t)$ have a unique intersection $I = (t^*, c^*)$. Moreover this intersection belongs to the curve $t = t(c)$.*

Proof : If $c^* = c_1(t^*) = c_2(t^*)$ then by definition $q_A = q_E = q_K$, and therefore $A = E = K = L$ (see fig 1) which implies $V_L(E) = V_L(L)$ and thus $t^* = t(c^*)$. We now prove that such a point is unique. Let $q^* = \frac{c^*}{\Delta\pi(1-t^*)}$ and $\pi_m = \pi_L + t^*\Delta\pi$. We know that $E = K$, therefore

$$\pi_m = \left(-\frac{\partial V_L}{\partial q} \Big/ \frac{\partial V_L}{\partial p}\right)(q^*, \pi_H q^*) \stackrel{def}{=} A(q^*). \tag{15}$$

Moreover $A = K$, which implies :

$$V_H(q^*, \pi_H q^*) = u\left(W - \pi_H D - \frac{c^*}{t^*}\right). \tag{16}$$

But

$$\begin{aligned} & c^* = \Delta\pi(1-t^*)q^*, \text{ and} \\ & t^* = \frac{\pi_m - \pi_L}{\Delta\pi}, \quad 1 - t^* = \frac{\pi_H - \pi_m}{\Delta\pi} \\ \Rightarrow \quad & \frac{c^*}{t^*} = \Delta\pi q^* \frac{\pi_H - \pi_m}{\pi_m - \pi_L}. \end{aligned}$$

Finally, (16) becomes

$$V_H(q^*, \pi_H q^*) = u\left(W - \pi_H D - \Delta\pi q^* \frac{\pi_H - \pi_m}{\pi_m - \pi_L}\right). \tag{17}$$

We can now incorporate (15) :

$$V_H(q^*, \pi_H q^*) = u\left(W - \pi_H D - \Delta\pi q^* \frac{\pi_H - A(q^*)}{A(q^*) - \pi_L}\right). \tag{18}$$

The left hand side of (18) increases from $V_H(0,0)$ to $u(W - \pi_H D)$ when q^* increases from 0 to D. The right hand side of (18) is not always monotonic but its extreme values are easy to compute : when $q^* = 0$, its value is $u(W - \pi_H D)$ and when q^* approaches D it tends to $-\infty$ (since $A(D) = \pi_L$). By continuity (18) has at least one solution q^*. Moreover in such a point, one must have $A(q^*) < \pi_H$ (since $V_H(q^*, \pi_H q^*)$ is always less than $u(W - \pi_H D)$. But A is a decreasing function of q, therefore $q \to q\frac{\pi_H - A(q)}{A(q) - \pi_L}$ has a positive derivative at q^*. Thus in any solution of (18) the left hand side has a positive derivative, while the right hand side has a negative derivative. This establishes the uniqueness of q^*.

■

6 Conclusion

Following the Rothschild and Stiglitz influential contribution, the theoretical literature on insurance markets has studied in a lot of detail the consequences of adverse selection. However in practice, economies of scale in the distribution technology seem to be an equally important issue (if not more). This paper is a first step in the exploration of the interaction between these economies of scale (which favor pooling contracts) and adverse selection (which tends to promote separating contracts). We have introduced an additional ingredient with respect to Rothschild and Stiglitz (1976): a fixed distribution cost c per contract offered, and we have characterized the set of equilibria:

- like in Rothschild and Stiglitz, there is a Pareto dominating separating pair of contracts (H, L), that gives full insurance to the high risks. It is an equilibrium if and only if the proportion of high risks is higher than a threshold that depends on c.
- However, contrary to Rothschild and Stiglitz, (H, L) can be Pareto optimal without being an equilibrium.
- Moreover, pooling equilibria always exist when c is large enough.

In two companion papers, we have started exploring further the interactions between adverse selection and increasing returns to scale:

- insurance markets with a continuous distribution of types (Allard et al. (1997a)),
- proliferation of products in a differentiated industry (Allard et al. (1997b)).

REFERENCES

Allard, M., J.P. Cresta and J.C. Rochet (1997a), "Interaction between Adverse Selection and Increasing Returns in an Insurance Market", (in preparation).

Allard, M., J.P. Cresta and J.C. Rochet (1997a), "Product Proliferation in a Differentiated Industry: A Reconsideration", (in preparation).

Diamond, Peter, (1992) "Organizing the Health Insurance Market" Econometrica, 60 (6), 1233-1254.

Neipp, J. and Zeckhauser R.J. (1985) "Persistence in the Choice of Health Plans", in Richard M. Scheffler and Louis F. Rossiter ed., pp. 47-72.

Newhouse, Joseph P., (1996), "Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection", *Journal of Economic Litterature*, Vol. XXXIV (September 1996), pp. 1236-1263.

Rothschild, Michael and Stiglitz, Joseph E., (1976), "Equilibrium in Competitive Insurance Markets: and Essay on the Economics of Imperfect Information", Quarterly Journal of Economics, 629-649.

IV .2 -Consultation médicale : l 'influence du revenu et de l 'assurance complémentaire

La documentation Française : "Coût et efficacité des dépenses de santé : rapport / Université des sciences sociales, GREMAQ ; Commissariat général du Plan ; Jean-Charles Rochet, Helmuth Cremer, Jean-Paul Cresta, \(et al.\)."

CONSULTATION MEDICALE: L'INFLUENCE DU REVENU ET DE L'ASSURANCE COMPLEMENTAIRE*

Gwenaël PIASER† Denis RAYNAUD‡

Novembre 1998

*Nous remercions Yves Aragon, Farid Gasmi, Agnès Gramain, Pierre-Yves Geoffard, Jean-Charles Rochet, et Diego Rodriguez pour leur aide précieuse ainsi que les participants du séminaire économie de la santé à Toulouse et les participants du colloque d'économie publique appliquée à Quimper. Nous remercions aussi Pascale Genier qui nous a permis d'utiliser ses données et le CREDES qui nous a fourni les indicateurs de morbidité. Toutes les erreurs et imprécisions ne doivent être reprochées qu'a nous.

†GREMAQ

‡GREMAQ, Université des Sciences Sociales, Manufacture des Tabacs, MF 005, 21, Allées de Brienne, 31 000 Toulouse. Tel. 05 61 12 87 65.

Résumé

Nous étudions l'influence du revenu et de la couverture complémentaire des ménages sur leur fréquence de consultation médicale à partir de données tirées de l'enquête santé 1991-92 de l'INSSE.

Dans une première partie nous estimons la probabilité de consultation d'un médecin en autorisant des effets non monotones pour le revenu. Les résultats obtenus montrent que le revenu a un effet positif sur la probabilité de consultation pour les individus disposant d'une couverture complémentaire (assurance ou mutuelle), mais cet effet devient négatif à partir d'un certain seuil pour les individus ne disposant pas d'une couverture complémentaire.

Dans une seconde partie nous estimons l'influence de différentes variables sur l'état de santé tel qu'il est perçu par les individus eux-mêmes. Nous montrons que les individus les plus "pauvres" considèrent généralement qu'ils sont en plus mauvaise santé alors que toutes choses égales par ailleurs, les individus les plus "riches" se considèrent en meilleure santé.

1 Introduction

1.1 Motivations et Résultats

Nous montrons dans la première partie que l'effet du revenu sur la probabilité de consultation médicale selon le type de couverture complémentaire n'est pas toujours positif. On ne peut faire raisonnablement l'hypothèse que la santé est un bien inférieur, et donc en conséquence la probabilité de consulter qui peut s'apparenter à une fonction de demande, toutes choses égales par ailleurs devrait augmenter avec le revenu. Ce n'est pas empiriquement toujours le cas. Ce résultat paradoxal peut être expliqué de façon théorique par le modèle proposé par Raynaud et Rochet (1998). Ils introduisent un coût psychologique à consulter, variant de façon négative avec le revenu, et ils montrent que la probabilité de consultation pour les personnes n'ayant pas d'assurance peut-être non monotone dans le revenu.

L'autre type de résultat empirique auquel nous arrivons est plus difficile à expliquer mais non moins surprenant. Dans la littérature économique il est souvent fait l'hypothèse que les individus les plus "riches" sont aussi ceux qui font le plus attention à leur santé. Sen (1998) prétend qu'il est difficile de se fier aux indicateurs de morbidité pour faire des comparaisons internationales puisque souvent ces indicateurs sont issus de déclarations que font les individus et que les individus les plus pauvres ne font pas attention à leur santé de la même façon que les personnes les plus riches. A l'appui de son argumentation il donne des chiffres montrant que les Américains se pensent en plus mauvaise santé que les habitants de l'Inde rurale. Strauss et Thomas (1996) font état du même type de problème.

Nous disposions de données relatives à l'état de santé objectif des individus mais aussi pour certains d'entre eux de données concernant leur état de santé tel qu'ils le déclarent, c'est à dire leur état de santé subjectif. Si l'intuition de Sen est vraie nous devrions trouver que "toutes choses égales par ailleurs" les riches se déclarent en moins bonne santé que les pauvres. Or nous trouvons exactement l'inverse.

1.2 Les Données

Les données dont nous disposons sont celles de l'Enquête Santé de l'INSEE de 1991, auxquelles s'ajoutent des données sur l'état de santé objectif calculées par les médecins du CREDES.

Le principe de cette enquête est de recueillir le maximum d'information sur les consommations médicales de 11 500 ménages (21 500 personnes) sur une période de trois mois, avec pour objectif de comprendre comment les individus se soignent, c'est-à-dire connaître les types de soins consommés selon les catégories sociales, mettre en évidence les maladies pour lesquelles les individus décident de consulter, et déterminer les motifs de recours à l'hôpital ou à une clinique. Chaque ménage enquêté à été suivi pendant douze semaines, au cours desquelles l'enquêteur a effectué cinq visites espacées de trois semaine.

L'Enquête Santé contient des données précises sur le revenu des ménages. La notion de revenu retenu est celle du revenu mensuel par unité de consommation qui est une pondération du revenu total par la structure du ménage afin de tenir compte des économies d'échelle au sein des ménages.

L'enquête santé 1991-1992 de l'INSEE a déjà donné lieu à plusieur études. C'est ainsi que Crety et Wencker (1995) tentent de calculer à partir de ces données la prime actuariellement équitable pour des contrats d'assurance maladie. Ils proposent ensuite une série de mesures dans l'esprit du "plan Juppé" pour lutter contre le risque moral.

Eeckhoudt, Cales et Dervaux (1998) estiment les probabilités de souscription d'assurance, de recours aux soins et la demande de prévention. Ils trouvent une légère preuve de sélection adverse en comparant les assurés qui ont librement choisi de souscrire une assurance complémentaire et ceux à qui elle est imposée par leur contrat de travail.

Genier (1998) cherche à distinguer et à mesurer l'effet de l'anti-sélection et du risque moral. Les résultats qu'elle obtient ne peuvent lui permettre de rejeter l'hypothèse d'anti-sélection alors que Caussat et Glaude (1993) avaient en utilisant les mêmes méthodes rejeté cette hypothèse à partir des données de l'enquête santé 1980. En régressant les indicateurs de consommation pour différentes populations, Genier montre que l'influence de la couverture sur la consommation est positive et conclut à la prédominance du risque moral sur la sélection adverse[1].

[1] Ces conclusions ddevrait à notre avis être tempérées: les données de l'enquête santé sont des données de coupe et on ne peut tester des relations de causalité avec ces données. Donc sauf à faire l'hypothèse que les variables sur l'état de santé sont bien les variables d'anti-sélection, on ne peut distinguer risque moral et sélection adverse. Pourtant cette distinction est fondamentale pour pouvoir déterminer une politique de la santé. (Pour une estimation empirique du risque moral voir Manning *et al* (1987) et Chiappori *et al* (1998)).

2 Statistiques descriptives

La population a été divisée en sept tranches de revenu, la première tranche étant celle des individus disposant d'un revenu mensuel par unité de consommation inférieur à 3 000 F, la septième tranche correspond aux individus disposant d'un revenu mensuel par unité de consommation supérieur à 10 000 F mensuel.

La population est aussi divisée en trois parties selon le type d'assurance dont dispose l'individu : sécurité sociale uniquement, sécurité sociale et une couverture complémentaire, ou prise en charge à cent pour cent par la sécurité sociale.

2.1 Probabilité de Consultation

Les probabilités de consultation sont calculées en contrôlant par âge, sexe et assurance.

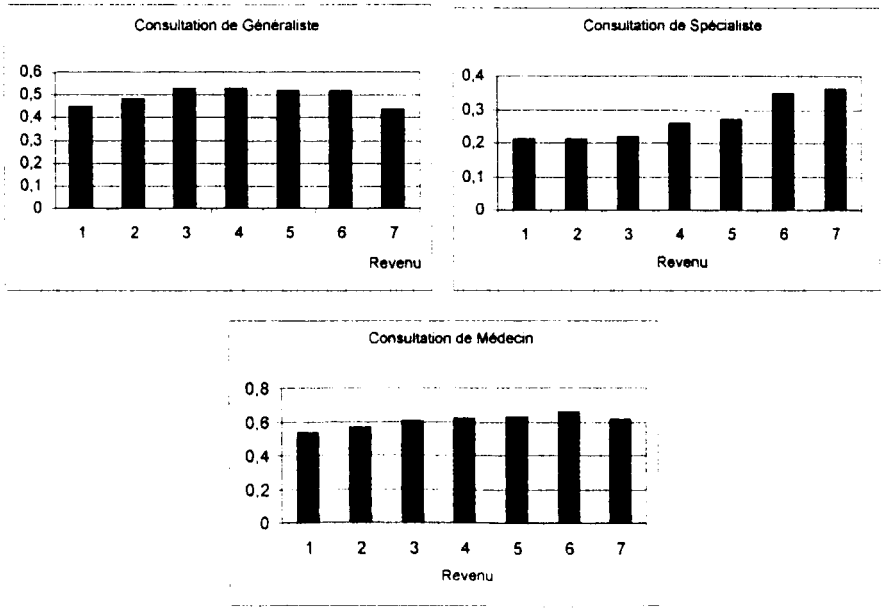


Figure 1: Probabilité de consultation

On constate que l'effet du revenu est non monotone pour les consultations de généraliste, alors qu'il est toujours positif pour les consultations de

spécialiste. On peut soupçonner l'existence d'un effet de substitution entre consultation de généraliste et consultation de spécialiste pour les individus d'un revenu plus élevé. Mais la probabilité de consultation d'un médecin en général diminue à partir d'une certaine tranche de revenu.

Dans tous les cas les très faibles revenus ont une probabilité plus faible de consultation que les gros revenus. Trois interprétation sont possibles: soit les bas revenus ont un meilleur état de santé général, soit il y a un problème d'accès aux soins pour les plus pauvres, soit les personnes avec un revenu élevé surconsomment.

	Sécurité Sociale	Complémentaire	100% S.S.
Généraliste	33%	50%	64%
Spécialiste	15%	26%	44%
Médecin	41%	60%	79%

Tableau 1: Probabilité de consultation selon le mode d'assurance

Les assurés complémentaires ont une plus forte probabilité de consulter que les personnes qui payent le ticket modérateur. Deux explications sont possibles :

- la présence d'aléa moral qui peut prendre deux formes, les individus assurés ont moins d'incitations à faire de la prévention (aléa moral ex-ante), et d'autre part les individus mieux couverts ont tendance, à maladie comparable, à plus consulter que des individus ayant une mauvaise couverture (aléa moral ex-post). Bien qu'il soit difficile de distinguer les deux types empiriquement, l'aléa moral ex-post semble être plus important que le risque moral ex-ante.(voir par exemple Genier et Jacobzone (1998))

- le phénomène d'anti-sélection dans le choix de la complémentaire. Les personnes qui ont une plus grande probabilité d'être malade et donc de consulter ont tendance à plus s'assurer que les autres.

Comme on ne peut pas distinguer facilement entre les deux effets (au contraire de la RAND qui avait éliminé l'anti-sélection en affectant aléatoirement differentes assurances aux individus) on séparera les échantillons selon la couverture dans nos estimations. De plus à cause de la difficulté que pose

la modélisation de leurs comportements, les individus disposant d'une couverture de la sécurité sociale à cent pour cent ne seront pas inclus dans les populations à partir desquelles on fera les estimations.

2.2 Etat de santé subjectif et état de santé objectif

Pour chaque individu "kish" (Dans chaque ménage a été choisi aléatoirement un adulte, l'individu "kish", auquel il a été proposé un questionnaire plus complet) nous disposons d'une évaluation subjective que cet individu a de son état de santé étant donné son âge. Cette variable peut prendre cinq valeurs différentes : Très bon, bon, moyen, médiocre, très mauvais.

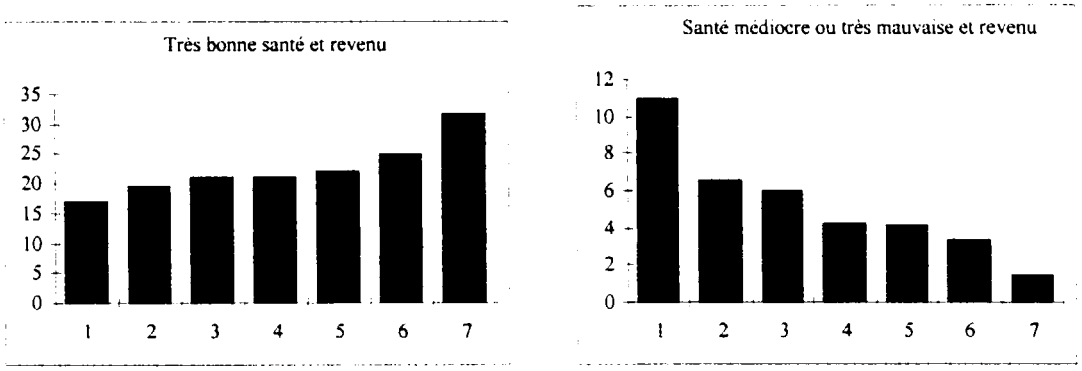


Figure 2: Perception de l'état de santé selon le revenu

La proportion d'individus se déclarant en bonne santé augmente avec le revenu et inversement la proportion d'individus se déclarant en mauvaise santé diminue avec le revenu. Deux interprétations sont possibles : les individus ayant un meilleur revenu sont en meilleure santé, ou bien la tendance à sous estimer son état de santé diminue avec le revenu.

Pour chaque individu nous disposons aussi d'un indice de santé objectif calculé à partir du risque vital par les médecins du CREDES. L'indice 1 correspond à l'état de santé moyen pour l'ensemble de la population. Plus l'indice est élevé, plus l'état de santé est mauvais. Si on compare l'indice de santé moyen par tranche de revenu il semble que l'indice soit décroissant avec le revenu.

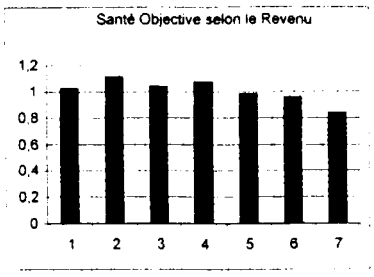


Figure 3: Indice de santé moyen selon le revenu

3 Effet du Revenu et Décision de Consulter

3.1 Méthodologie

Notre premier ojectif est d'étudier empiriquement les comportments de consultation d'un médecin (généraliste ou spécialiste) en fonction du revenu, en contrôlant par le type de couverture sociale, l'état de santé et diverses variables socio-démographiques.

Afin de contourner le problème de l'endogénéité de la couverture complémentaire, on estime les probabilités de consultation d'un médecin pour trois ensembles d'individus : l'ensemble des individus qui ne disposent que de la couverture de la sécurité sociale, l'ensemble de ceux qui disposent d'une couverture complémentaire, et enfin l'ensemble des individus qui disposent d'une "bonne" couverture complémentaire, c'est-à-dire l'ensemble des individus qui disposent d'une assurance complémentaire ou d'une mutuelle qui prend en charge tous les dépassements d'honoraires.

La décision de consulter au moins une fois le médecin (y=1) ou ne pas consulter (y=0) est estimée à partir d'un modèle PROBIT :

$$y = \begin{cases} 1 & si\ y^* \succeq 0 \\ 0 & si\ y^* < 0 \end{cases}$$

où

$$y^* = \beta x + \alpha_1 \log(revenu) + \alpha_2 \log^2(revenu) + u$$

avec

$$u \sim N(0;1)$$

où β est un vecteur de parametres à estimer et x un vecteur de variables explicatives parmi lesquelles on retrouve les variables socio-démographiques

usuelles, le nombre de maladies prévalentes (le nombres de maladies déclarées au départ de l'enquête santé), l'état de santé objectif. Le revenu est mis sous la forme $\log(revenu)$ et $\log^2(revenu)$ pour capturer d'éventuels effets non monotones.

L'effet du revenu sur la probabilité de consulter est donnée par l'expression :

$$\alpha_1 \log(revenu) + \alpha_2 \log^2(revenu)$$

Pour mettre en évidence les effets non monotones du revenu nous avons procédé de la façon suivante:

- Si $\widehat{\alpha}_2$ (estimation du coefficient α_2 par le modèle PROBIT) n'est pas significativement différent de zéro, alors les estimations sont refaites en imposant la contrainte $\alpha_2 = 0$
- Si $\widehat{\alpha}_1$ et $\widehat{\alpha}_2$ sont tous les deux significativement différents de zéro, alors on calcule le niveau de revenu $\overline{\mathrm{R}}$ (s'il existe) tel que l'effet revenu s'annulle :

$$\widehat{\alpha}_1 + 2\widehat{\alpha}_2 \log \overline{\mathrm{R}} = 0$$

$$\Rightarrow \overline{\mathrm{R}} = \exp\left(-\frac{\widehat{\alpha}_1}{2\widehat{\alpha}_2}\right).$$

La population est alors séparée en deux sous-populations : une constituée par les individus dont le revenu est inférieur à $\overline{\mathrm{R}}$ et l'autre par les individus dont le revenu est supérieur à $\overline{\mathrm{R}}$, et une estimation est refaite sur chacune de ces populations en ne gardant que le terme $\alpha_1 \log(revenu)$ pour capturer l'effet revenu.

3.2 Résultats

Tout d'abord, les personnes consultent plus si elles ont un nombre important de maladies prévalentes, ce qui semble logique. De plus les personnes consultent plus si leur indicateur de santé (santé à long terme) est mauvais et ceci quel que soit le type de médecin considéré (généraliste, spécialiste,médecin en général) et quel que soit la population considérée (tout type de couverture sociale).

Pour les autres variables explicatives (sauf revenu) les effets peuvent être significatifs ou pas selon les régressions considérées. Cependant, quelques grandes tendances se dégagent :

- **Effet du niveau de diplôme du chef de famille** : si le chef de famille n'a pas de diplôme, alors la probabilité de consulter un spécialiste sera faible. Inversement, si le chef de famille a un diplôme au moins égal au baccalauréat, alors les individus consulteront plus fréquemment un médecin spécialiste. L'effet du diplôme est non significatif pour les consultations du généraliste.

- **Taille de l'agglomération** : les individus habitant une petite ville (unité urbaine de moins de 20 000 habitants) consultent plus fréquemment le médecin généraliste, alors que les habitants de Paris consultent plus fréquemment les médecins spécialistes.

- **Catégorie socioprofessionnelle** : les enfants, les inactifs n'ayant jamais travaillé et les agriculteurs ont une plus grande propension à consulter un médecin. Le phénomène est plus marqué pour les généralistes que pour les spécialistes.

- **Etranger** : les étrangers soit consomment moins que les Français, soit n'ont pas une consommation significativement différente de celle des Français (d'origine ou d'adoption).

- **Age** : Les enfants et les personnes âgées ont tendance à plus consulter.

3.2.1 Population sans couverture complémentaire

La population qui est sans couverture complémentaire paye le ticket modérateur, donc on pourrait s'attendre à ce que l'effet du revenu sur la probabilité de consulter au moins une fois soit positif, la santé ne pouvant être considérée comme un bien inférieur.

	$\overline{R}$	$R < \overline{R}$		$R > \overline{R}$
Généraliste	4800	0.125 (1.53)		-0.458^{***} (7.67)
Spécialiste			0.205^{***} (9.45)	
Médecin	5500	0.251^{***} (8.49)		-0.185 (1.08)

Tableau 2: Effet du revenu sur la probabilité de consultation
*** signifie que le coefficient[2] est significatif à à 99%.

[2] Le test effectué est un test du chi-2.

Cet effet positif est clairement obtenu pour les consultations de spécialiste, mais pour les consultations d'un médecin généraliste, on a au contraire un effet négatif à partir d'un revenu (mensuel et par unité de consommation) de 4800 F.

Une première explication de ce phénomène est la possible substitution des visites de généraliste par les visite de spécialistes pour les individus ayant les plus forts revenus. Mais si on considère les consultations de médecin en général (généraliste ou spécialistes) le revenu n'a plus d'effet significatif à partir 5500 F (mensuel et par unité de consommation)

Le comportement contre-intuitif des individus ayant un revenu supérieur à 4800 F en ce qui concerne les consultations de généraliste peut s'expliquer par l'endogénéité du choix de la couverture maladie.Les gens de revenu confortable qui ont choisi de ne pas avoir de complémentaire savent qu'ils consultent rarement le médecin.

3.3 Population avec couverture complémentaire

On considère l'ensemble des individus ayant une couverture complémentaire, sans que l'on distingue entre les différents types de complémentaire.

	$\overline{R}$	$R < \overline{R}$		$R > \overline{R}$
Généraliste	6500	0.139^{***} (10.86)		-0.234^{***} (15.67)
Spécialiste			0.238^{***} (81.00)	
Médecin	15000	0.159^{***} (36.30)		-0.68 (0.391)

Tableau 3: Effet du revenu sur la probabilité de consultation
*** le coefficient est significatif à 99%

La probabilité de consulter un généraliste est croissante avec le revenu jusqu'à 6500F par mois, puis décroissante, alors que cette probabilité est toujours croissante pour les consultations de spécialiste. Là encore l'effet de substitution généraliste/spécialiste peut être une explication à la décroissance de la probabilité de consultation du généraliste: certains assurés complémentaires ne sont au total remboursés que d'une partie du coût des soins et donc ils hésiteront à aller chez le généraliste s'ils pensent que celui-ci les orientera vers un spécialiste.

La probabilité de consulter un médecin en général est croissante avec le revenu (jusqu'à 15000 F, mais seuls 2% des individus ayant une complémentaire ont un revenu mensuel supérieur à 15000F) pour les assurés complémentaires. On remarque que cet effet revenu est positif même au delà d'un revenu de 5500F pour les assurés complémentaires alors que l'effet revenu était non significatif à partir de ce seuil pour les individus sans complémentaire. Ce résultat étonnant s'explique par le comportement de consommation atypique des individus n'ayant pas de complémentaire et ayant un revenu moyen ou élevé.

3.4 Population avec une bonne couverture complémentaire

On s'attend ici à ce que le revenu n'influence pas la probabilité de consultation puisque tous les frais médicaux sont pris en charge pour ces individus.

	$\overline{R}$	$R < \overline{R}$		$R > \overline{R}$
Généraliste	7500	0.241^{***} (14.35)		-0.024 (0.03)
Spécialiste			0.333^{***} (49.13)	
Médecin			0.253^{***} (49.13)	

Tableau 4: Effet du revenu sur la probabilité de consultation
*** le coefficient est significatif à 99%

Contrairement à ce que l'on pourrait penser, l'effet revenu est positif pour les individus qui ont une assurance complémentaire qui rembourse les possibles dépassements d'honoraires. Pour les généralistes à partir de 7500F mensuel cet effet est non significativement différent de zéro. Par contre l'effet est toujours positif pour les consultations de spécialiste et de médecin en général.

Si on considère simplement un échantillon sur lequel on a retiré les enfants, alors il n'y a pas d'effet revenu pour les consultations de généraliste, mais celui-ci apparaît pour les consultation de spécialiste et donc pour les consultations de médecin en général.

4 Perception de l'Etat de Santé

4.1 Perception de l'état de santé

L'enquête santé comporte beaucoup de données sur les habitudes de vie, les conditions de travail, les problèmes psychologiques, pour une sous population: la population "kish" constituée d'adultes tirés aléatoirement par ménage. Pour cet échantillon kish on dispose notamment de l'état de santé subjectif. Nous avons voulu déterminer les facteurs qui infuencent la réponse d'un individu.

Pour cela nous avons regroupé les modalités de la variable "état de santé subjectif" en deux grandes modalités qui correspondent à "bonne santé" et "mauvaise santé". A partir de cette nouvelle variable nous avons procédé à une estimation d'un modèle PROBIT en prenant comme variables explicatives les variables socio démographiques habituelles, le revenu et l'état de santé objectif.

On obtient alors que toutes choses égales par ailleurs, un individu a tendance à se déclarer en meilleure santé:

- Son état de santé objectif bon;
- S'il est un homme;
- S'il est artisan, commerçant, chef d'entreprise ou s'il a une profession libérale;
- S'il a un diplôme équivalent ou supérieur au bac;
- S'il a un revenu élevé;

Si on pense que quelqu'un qui déclare être en mauvaise santé ira plus volontiers consulter le médecin, alors ce dernier résultat est important pour analyser la relation entre revenu et consultation. Pour bien comprendre les motifs de la décision de consulter il faut tenir compte de "l'effet psychologique" lié à la perception de l'état de santé. Ainsi, dans le cas des individus sans complémentaire, la plus faible consultation du généraliste pour les individus ayant un revenu élevé pourrait s'expliquer par le fait que ces derniers ont une perception de leur état de santé particulièrement biaisée par rapport à la perception de l'ensemble de la population.

On peut contester la validité de la réponse donnée par les individus au questionnaire. Tout d'abord on peut se demander si les individus ont une bonne appréciation sur le moment de leur état de santé, Loftus *et al* (1991) remarquent que les personnes interrogées sur leurs visites médicales ont de sérieux problèmes de mémoire et ne s'en souviennent qu'avec difficulté et de multiples erreurs. De plus Pedhazur et Pedhazur Schmelkin (1991) montrent trois types de problèmes que peut engendrer ce type de questionnaire où la réponse est un chiffre porté sur une échelle. Tout d'abord "l'effet halo": Les individus interrogés hésitent entre deux classes de réponse, autrement dit, il est difficile d'interpréter la signification de la différence entre "bon" et "moyen". Ensuite on observe un biais "vers le centre": les individus évitent de se situer dans les classes extrêmes. Enfin lorsque la définition de la question est douteuse (par exemple donner une note de 1 à 5 sur la capacité démocratique d'un homme politique) la réponse de l'individu le sera aussi. En ce qui concerne le troisième type de problème, la question posée par les enquêteurs de l'INSEE est parfaitement claire "Actuellement, compte tenu de votre âge, comment estimez-vous votre état de santé ?" et les réponses proposées le sont aussi "Très bien, Bon, Moyen, Médiocre, Franchement mauvais". Les deux autres types de biais possibles sont fortement réduits, puisque pour la régression nous avons regroupé les modalités en deux catégories. Enfin en ce qui concerne les problèmes de mémoire ou d'échelle propre à chaque individu, nous définissons ci-dessous une nouvelle variable (que nous appelons "Psycho"). Le fait que dans les régressions suivantes la variable "Psycho" soit significative nous laisse penser que les résultats que nous avons obtenus plus haut sont robustes.

4.2 La variable "psycho"

A partir de l'estimation précédente de l'état de santé subjectif, nous avons défini une variable "psycho" qui prendra la valeur 1 si l'individu sous-estime son état de santé (Dans ce cas il est alors considéré comme "pessimiste") et 0 sinon.

La régression précédente nous donne une valeur estimée pour la santé d'un individu étant données ses caractéristiques. On note $\widehat{s}$ cette variable qui peut prendre deux valeurs :

$$\widehat{s} = \begin{cases} 1 \text{ si l'état de santé estimé par la regression est "bon"} \\ 0 \text{ si l'état de santé estimé par la regression est "mauvais"} \end{cases}$$

Nous avons aussi son état de santé tel qu'il l'a déclaré, on note s cette variable qui peut prendre aussi deux valeurs :

$$s = \begin{cases} 1 \text{ si l'individu déclare être en bonne santé} \\ 0 \text{ si l'individu déclare être en mauvaise santé} \end{cases}$$

On considère la variable ε représentant la différence entre l'état de santé subjectif estimé par la régression et l'état de santé subjectif déclaré par l'"individu :

$$\varepsilon = \widehat{s} - s = \begin{cases} -1 \text{ si } \widehat{s} < s \\ 0 \text{ si } \widehat{s} = s \\ 1 \text{ si } \widehat{s} > s \end{cases}$$

ε prend trois modalités :

- $\varepsilon = 1$ l'individu se déclare en mauvaise santé alors que d'après la régression il devrait déclarer être en bonne santé: comportement "pessimiste".
- $\varepsilon = 0$ comportement "normal".
- $\varepsilon = -1$ comportement "optimiste".

Les estimations PROBIT de la décision de consulter doivent être refaites en introduisant la variable ε.

De plus, des résultats intermédiaires révèlent qu'il est pertinent de regrouper les modalités 0 et -1 de la variable ε. On définit donc la variable psycho ainsi :

$$\text{psycho} = \begin{cases} 1 \text{ si } \varepsilon = 1 \\ 0 \text{ si } \varepsilon \in \{-1; 0\} \end{cases}$$

4.3 Décision de consulter

L'état de santé subjectif n'étant connu que pour les individus "kish", pour pouvoir comparer les résultats des différentes estimations il faut refaire les estimations sur la population "kish".

Les estimations ont été refaites pour les consultations de généralistes.

	$\overline{R}$	$R<\overline{R}$	$R>\overline{R}$
Sans complémentaire	3400	0.241 (0.75)	-0.310^{*} (2.72)
Avec complémentaire	5900	0.218^{***} (7.89)	-0.077 (0.86)
Bonne complémentaire	6150	0.308^{**} (5.46)	0.069 (0.18)

Tableau 5: Estimations sans la variable psycho.
* significatif à 90%.
** significatif à 95%.
*** significatif à 99%.

Par rapport aux estimations faites sur la population totale, on perd un peu de significativité. l'échantillon est plus réduit, l'effet revenu négatif concerne plus d'individus parmi ceux qui n'ont pas de complémentaire (R>3400) et il n'y pas pas d'effet revenu négatif significatif pour les individus disposant d'une couverture complémentaire.

Variable		Revenu	Revenu	Psycho	Psycho
	$\overline{R}$	$R<\overline{R}$	$R>\overline{R}$	$R<\overline{R}$	$R>\overline{R}$
Sans complémentaire	3500	0.332 (1.44)	-0.268 (1.87)	0.36^{**} (3.85)	0.33^{*} (3.00)
Avec complémentaire	6700	0.142^{**} (4.58)	-0.155 (2.61)	0.27^{**} (9.9)	0.31^{***} (14.1)
Bonne complémentaire	6150	0.306^{**} (5.32)	0.086 (0.27)	0.36^{**} (9.46)	0.67^{***} (21.5)

Tableau 6: Estimations avec "Psycho"
* significatif à 90%
** significatif à 95%
*** significatif à 99%

Quelque soit la régression considérée, la variable psycho a un coefficient significativement différent de zéro. A chaque fois le signe du coefficient va dans le sens d'une plus grande probabilité consultation lorsque l'individu est "pessimiste" (psycho = 1). Cette variable est donc pertinente pour l'analyse du comportement de consultation.

On constate que pour les échantillons de population avec une couverture complémentaire l'introduction de la variable psycho modifie peu les résultats.

Sans la variable psycho l'effet revenu était significatif jusqu'à 5900F mensuels contre 6700F avec la variable psycho. Néanmoins ce changement va dans le sens attendu dans la mesure où les hypocondriaques étant peu représentés dans les revenus élevés, cela atténuait un effet revenu positif (ou accentuait un effet revenu négatif) lorsque l'on ignorait ce contrôle.

Le changement le plus spectaculaire concerne les individus sans complémentaire. Alors que l'effet revenu était négatif à partir de 3400F sans la variable psycho, avec elle on ne peut plus conclure qu'il existe un effet revenu négatif significatif pour les revenus supérieur à 3500F.

Ce résultat semble montrer qu'une partie du fort effet revenu négatif observé dans la première partie pour les régressions sur la population totale pour les individus n'ayant pas de couverture complémentaire peut-être expliquée par le fait que les individus ayant un fort revenu et qui décident de ne pas s'assurer ont une appréciation de leur état de santé optimiste par rapport à l'ensemble de la population. C'est donc une nouvelle preuve de l'endogénéité de la couverture.

5 Conclusion

Notre principale motivation était de montrer deux résultats non intuitifs. Ce type de résultat surprenant au vu de la théorie économique classique se retrouve dans le domaine de l'économie de la santé. Phelps (1992) trouve empiriquement que les choix des agents en matière d'assurance complémentaire ne sont pas ceux prédits par la théorie classique. Il reprend l'argument de Arrow (1963): On suppose que l'assureur fait face à des coûts d'immobilisation du capital tel par exemple des coûts administratifs et on suppose que l'assureur fait supporter ces coûts aux assurés en appliquant un pourcentage fixe au-dessus du tarif actuariellement équitable, alors le contrat que préfèrent des assurés averses aux risques est tel qu'au-dessus d'un certain seuil de pertes ils sont assurés complètement et qu'au-dessous de ce seuil ils ne sont pas assurés complètement. Or à partir de données américaines de 1977 Phelps montre que les assurés ne choisissent pas leur assurance complémentaire de cette façon, ils cherchent plutôt à réduire la franchise sur les petits risques qu'à s'assurer complètement sur les gros risques. De manière parallèle, Genier (1998) remarque à partir des données de l'INSEE que parmi les assurés ceux qui consomment le moins sont aussi ceux qui sont le mieux couverts.

Comme dans toute étude économétrique, nos résultats doivent être à la fois confirmés par d'autres études pour être considérés comme valables et doivent aussi être interprété à la lumière de la théorie.

Par contre le lien entre revenu et santé subjective reste plus mystérieux. Nos résultats ne sont peut être pas contradictoire avec l'intuition générale sur ce sujet. En général l'argument consiste à faire remarquer que les individus les plus riches sont aussi ceux qui sont le plus éduqués, et il est logique que les personnes les plus éduqués soient aussi les plus attentive à leur santé. Or dans notre régression nous contrôlons par rapport au niveau d'éducation (scolaire) des individus. L'effet du revenu que nous trouvons est un effet à niveau d'étude identique. Il serait donc souhaitable de refaire le même type d'étude sur des données plus complètes afin de pouvoir déterminer le lien entre état de santé (objectif et subjectif) revenu et niveau d'éducation. Etablir ce lien serait particulièrement important dans l'optique de faire de la redistribution à travers l'assurance maladie.

Un troisième type de résultat serait important à obtenir à partir des données dont nous disposons, connaître le lien entre revenu et état de santé objectif. Cela est parfaitement possible à obtenir en tenant compte du fait que les deux variables sont endogènes.

Enfin un domaine reste à étudier, celui de la signification et de l'utilisation de la variable "Psycho" que l'on a utilisée ici comme un indicateur psychologique des individus en ce qui concerne leur risque de maladie. On pourrait étudier son influence dans le choix des individus à s'assurer ou non.

Enfin du point de vu économétrique les modèles PROBIT pourraient être remplacés par des modèles de comptage ce qui permettrait d'affiner encore nos résultats.

References

[1] **Arrow K.J, (1963)** "Uncertainty and the Welfare Economics of Medical Care" American Economic Review, 53, 941-973.

[2] **Caussat L, M. Glaude (1993)** "Dépenses Médicales et Couverture Sociale" Economie et Statistique, 265.

[3] **Chiappori P.A, F. Durand, P. Y. Geoffard (1998)** "Moral Hazard and Demand for Physician Services : First Lessons from a French Natural Experiment" European Economic Review, 42, 499-511.

[4] **Crety L, Wencker A. (1995)** "Frais de Santé: de la Tarification à la maîtrise des Dépenses" Mémoire pour l'Obtention du Diplôme de l'Institut des Acturaires Français, Institut des Actuaires Français.

[5] **Eeckhoudt L, E. Cales, B. Dervaux (1998)** "Identification des Variables Explicatives du Recours aux Soins en Lien avec les Comportements de Prévention et en Présence de risques Multiples" Recherche effectuée pour la Mission Recherche Expérimentation (MIRE), Université Catholique de Lille.

[6] **Genier P. (1998)** "Assurance et Recours aux Soins" Revue Economique, à paraître.

[7] **Genier P, S. Jacobzone (1998)** "Comportement de Prevention, Consommation d'Alcool et Tabagie: Peut-on Parler d'une Gestion Globale du Capital Santé?" Document de travail INSSE n°G9605.

[8] **Holly A, L. Gardol, G. Domenlighetti, B. Bisig (1998)** "An Econometric Model of Health Insurance in Switzerland" European Economic Review, 42, 513-522.

[9] **Hurd M.D, K. McGary (1997)** "Medical Insurance and the Use of Health Care Services by the Elderly" Journal of Health Economics, 16, 129-154.

[10] **Loftus E.F, K.D. Smith, M.R. Klinger, J. Fiedler (1991)** "Memory and Mismemory for Health Events" in "Questions About Questions" Judith M. Tanur Editor, Russel Sage Foundatin, New York.

for Medical Care : Evidence from a Randomized Experiment" American Economic Review, 77(3), 251-277.

[12] **Newhouse J. (1996)** "Free for All ? : Lessons from the Rand Health Insurance Experiment" Harvard University Press.

[13] **Pedhazur E.J, L. Pedhazur Schmelkin (1991)** "Measurement, Design, and Analysis" Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

[14] **Phelps C, (1992)** "Health Economics" Harper-Collins Publishers Inc, New York.

[15] **Raynaud D. (1998)** "Santé et Accès aux Soins" Université de Toulouse.

[16] **Raynaud D, J.C. Rochet (1998)** "Consultation Médicale et Coût Indirect d'Accès aux Soins" Université de Toulouse.

[17] **Rotschild M, J. Stiglitz (1976)** "Equilibrium in Competitive Insurance markets", Quarterly Journal of Economics, 90, 629-649

[18] **Sen A. (1998)** "Mortality as an Indicator of Economic Success and Failure" Economic Journal, 108, 1-25.

[19] **Strauss J, D. Thomas (1996)** "Measurement and Mismeasurement of Social Indicators" American Economic Review, 86, 30-34.

6 Annexes

6.1 Statistiques Descriptives : Population Totale

Variable	*N*	*Mean*	σ	*Min*	*Max*
généraliste	21586	0.4941	0.4863	0	1
spécialiste	21586	0.2607	0.4271	0	1
médecin	21586	0.597	0.4771	0	1
mutuelle	19350	0.8540	0.3426	0	1
bonne_mut	21586	0.2550	0.4240	0	1
$\ln(revenu)$	19076	8.5914	0.5645	6.4171	10.5966
$\ln^2(revenu)$	19076	74.1342	9.7176	41.180	112.2886
âge	21586	37.1876	22.0193	0	100
$(âge)^2$	21586	1895.36	1845.3	0	10000
$âge \times femme$	21586	19.821	247838	0	100
Nbr_maldie	21586	2.7275	2.6835	0	19

Sante	*N*	%
*rv*0	8458.52	41.5
*rv*1	3691.38	18.1
*rv*2	4585.164	22.5
*rv*3	3149.15	15.4
*rv*45	519.048	2.5

Diplome	*N*	%
$Bac, Bac+2$	3197.716	15.7
$< Bac$	15170.4	74.3
Supérieures	2054.90	10.1

TailleVille	*N*	%
$+20000h$	8249.33	40.4
$-20000h$	8756.15	42.9
Paris + *Banl*	3417.537	16.7

Nationalité	*N*	%
Franais	18825.29	92.5
Naturalisé	315.52	1.6
Etranger	1211.05	6.0

Ménage	*N*	%
couple	4180.931	20.5
$couple + 1enfant$	3524.571	17.3
$couple + 2enfants$	4636.574	22.7
$couple + 3enfants$	3451.972	16.9
isolé	2190.979	10.7
autre	2437	11.9

6.2 Statistiques Descriptives : Individus Kish

Variable	*N*	*Mean*	σ	*Min*	*Max*
généraliste	7666	0.5094	0.4999	0	1
spécialiste	7666	0.2703	0.4441	0	1
médecin	7666	0.6109	0.4875	0	1
mutuelle	6607	0.8568	0.3521	0	1
bonne_mut	7666	0.2414	0.4282	0	1
$\ln(revenu)$	7155	8.6214	0.5615	6.4171	10.5966
$\ln^2(revenu)$	7155	74.6466	9.7121	41.1796	112.2886
âge	7666	45.6150	18.2487	17	95
$(\hat{a}ge)^2$	7666	2413.7	1827.75	289	9025
$\hat{a}ge \times femme$	7666	23.9652	26.5522	0	95
Nbr_maladies	7666	3.3729	2.8806	0	18
Psycho	7666	0.1492	0.3563	0	1

Sante	*N*	%
*rv*0	2033.471	26.7
*rv*1	1636.543	21.5
*rv*2	2206.636	29.0
*rv*3	1485.5	19.5
*rv*45	241.63	3.2

Sante3	*N*	%
*rv*0	2033.471	26.7
*rv*1	1636.543	21.5
*rv*2	2206.636	29.0
*rv*345	1727.16	227

TailleVille	*N*	%
$+20000h$	3109.754	40.6
$-20000h$	3266.011	42.6
$Paris + Banl$	1290.239	16.8

Diplome	*N*	%
$Bac, Bac + 2$	1213.143	15.8
$< Bac$	5703.75	74.4
Supérieures	749.113	9.8

Ménage	*N*	%
couple	2036.476	26.6
$couple + 1enfant$	1376.128	18.0
$couple + 2enfants$	1430.381	18.7
$couple + 3enfants$	879.5569	11.5
isolé	1070.968	14.0
autre	872.4949	11.9

6.3 Probabilité de Consultation

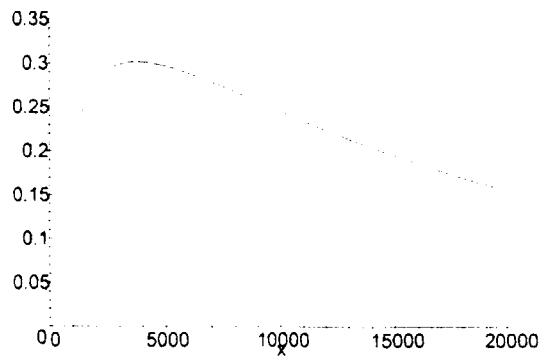
Ces probabilités sont calculées à partir du modèle probit sur la population d'assuré considérée, avec la possibilité d'un effet non monotone.

6.3.1 Sécurité sociale uniquement

Les probabilités de consultation sont calculées à partir de la régression faite sur la population des personnes ayant une couverture complémentaire(bonne ou mauvaise)avec comme variable de décision la consultation d'un médecin (spécialiste ou généraliste). L'individu type considéré appartient à une famille constituée d'un couple et de deux enfants, dont le "chef de ménage" n'a pas de diplôme équivalent au bac, il est français d'origine, habite une ville de moins de 20 000 habitant, est de sexe masculin et a environ 40 ans, il a 2.727 maladies prévalantes et a un excellent état de santé général.

Décision de consulter un généraliste L'impact du revenu est non monotone pour les individus ne bénéficiant que des remboursements de la sécurité sociale.

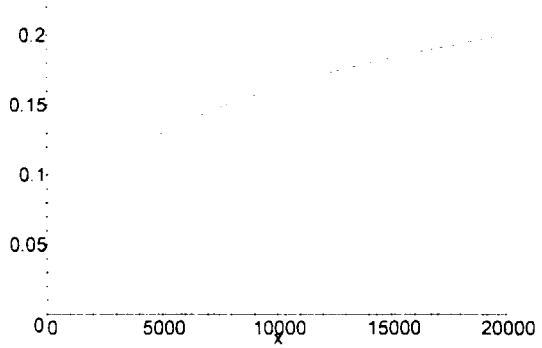
La probabilité de consulter un médecin peut être représentée par le graphique suivant:



Pour les faibles revenus (inférieur à 4000 F par mois et par unité de consommation), qui représentent 40% des assurés sans couverture complémentaire alors qu'ils ne représentent que 27% de la population, une augmentation du revenu aurait un effet positif sur la probabilité de consulter. Passé un seuil d'environ 4800 F par mois et par unité de consommation l'effet s'inverse, une hausse du revenu ayant un effet négatif sur la probabilité de

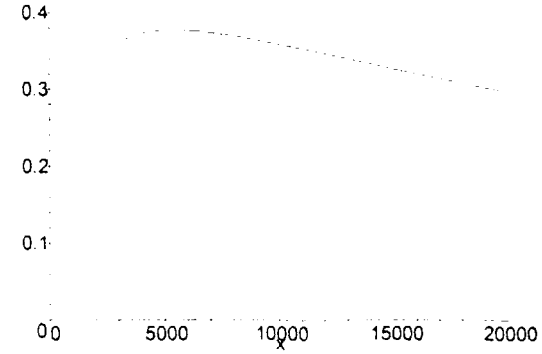
consultation. L'explication est vraisemblablement un effet de substitution avec les consultations de spécialiste. Les personnes ayant un bon revenu et pas de couverture complémentaire iront plus facilement chez le spécialiste sans passer au préalable chez le généraliste.

Décision de consulter un spécialiste



La probabilité de consulter un spécialiste est toujours croissante avec le revenu ce qui semble confirmer l'hypothèse de substitution entre généraliste et spécialiste.

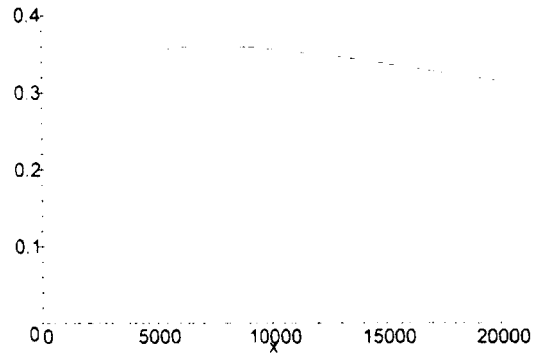
Décision de consulter un médecin (généraliste ou spécialiste)



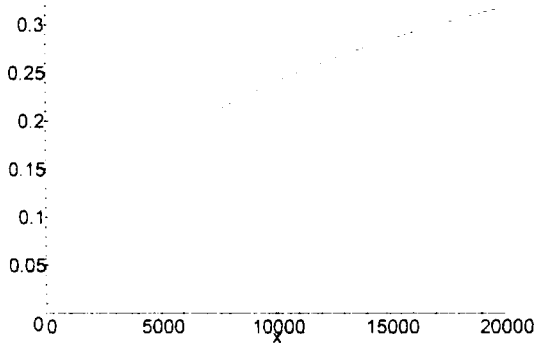
La probabilité de consulter un médecin est décroissante à partir d'un certain seuil, mais ce phénomène soit statistiquement significatif.

6.3.2 Sécurité sociale et bonne complémentaire

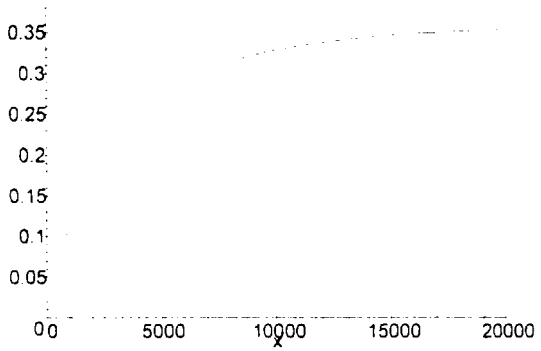
Décision de consulter un généraliste



Décision de consulter un spécialiste

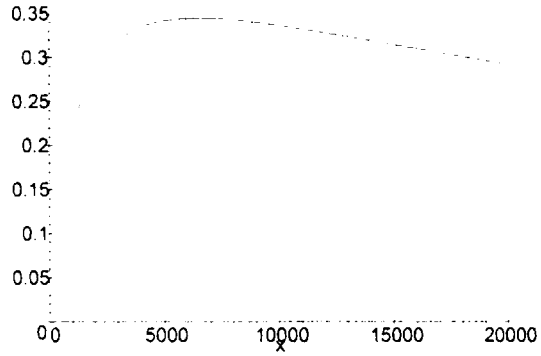


Décision de consulter un médecin (généraliste ou spécialiste)

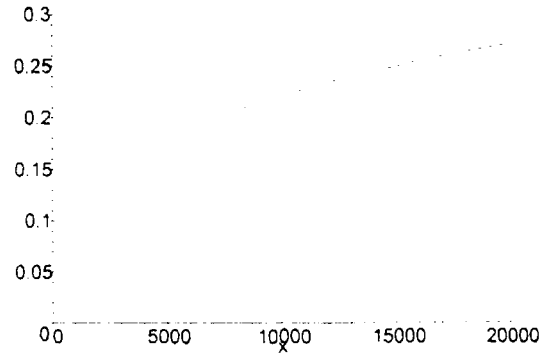


6.3.3 Sécurité sociale et une complémentaire

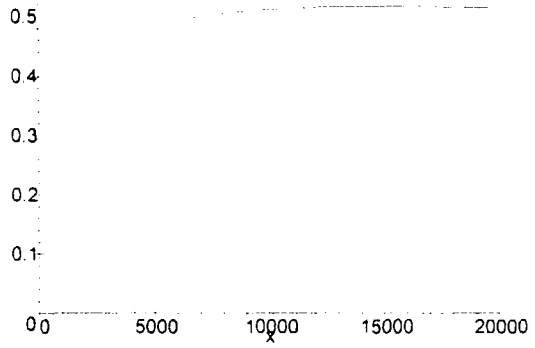
Décision de consulter un généraliste



Décision de consulter un spécialiste



Décision de consulter un médecin (généraliste ou spécialiste)



V - Couverture maladie universelle ou accès gratuit aux soins

La documentation Française : "Coût et efficacité des dépenses de santé : rapport / Université des sciences sociales, GREMAQ ; Commissariat général du Plan ; Jean-Charles Rochet, Helmuth Cremer, Jean-Paul Cresta, \(et al.\)."

COUVERTURE MALADIE UNIVERSELLE OU ACCES GRATUIT AUX SOINS[1]

Agnès COUFFINHAL[2]
et
Jean-Charles ROCHET[3]

1ère version : 5 Octobre 1998
Cette version : Novembre 1998

RESUME

Par un modèle aussi simple que possible, nous essayons de capturer les raisons pour lesquelles plus de 40 millions d'Américains (Hellander et al., 1995) renoncent à s'assurer contre la maladie. Nous utilisons ensuite notre modèle pour évaluer les effets qu'aurait eue l'introduction, prévue dans le plan Clinton, d'une couverture maladie universelle aux USA.

[1]Nous remercions Dominique Henriet de ses commentaires précieux. Toutes les erreurs nous sont entièrement imputables.

[2]CREDES, Paris, Email : agnes.couffinhal@hol.fr.
[3]GREMAQ-IDEI, Université Toulouse 1, Email :rochet@cict.fr.

1 Introduction

Les grandes réformes entreprises dans les systèmes de santé européens depuis les années quatre-vingt se sont pour la plupart traduites par l'introduction d'une dose de concurrence dans des systèmes caractérisés par une couverture ou un accès universel aux soins. A l'inverse, aux Etats-Unis, les interrogations sur les conséquences néfastes de la concurrence en termes d'accès à l'assurance ont été progressivement replacées au centre du débat. **En effet, dans l'ensemble des systèmes où les individus sont responsables de l'achat d'une assurance maladie, on constate que des proportions importantes de la population restent sans couverture.** Les Etats-Unis constituent l'exemple le plus frappant et le mieux documenté de ce phénomène : à l'heure actuelle, plus de 15% de la population ne bénéficie d'aucune assurance. Cette proportion augmente au fil du temps (elle était de 13.5% en 1989), alors même que la proportion de la population couverte par MEDICAID[1] a aussi augmenté sur la période (Hellander et al., 1995). Nous nous proposons ici d'étudier les causes de ce phénomène et d'analyser l'impact qu'aurait l'introduction d'une assurance maladie universelle, telle qu'envisagée par l'ex-réforme Clinton. Cette réflexion fait en outre écho à un débat d'actualité en France, car le problème de l'accès à une couverture universelle pour les prestations de base et complémentaires est au coeur de la politique d'insertion engagée par la loi votée le 9 juillet 1998.

Rappelons tout d'abord que le modèle de base de la théorie de la décision (Von Neumann-Morgenstern) ne permet pas d'expliquer ce phénomène car ce modèle prédit que, confronté à des primes d'assurance concurrentielles, tout individu riscophobe choisira de se couvrir complètement. La théorie économique suggère alors, en s'écartant du modèle de base, trois principales explications au phénomène de non-assurance.

En premier lieu, dès que les primes sont légèrement supérieures au niveau actuariel, une personne qui n'éprouve pas d'aversion pour le risque n'a pas intérêt à s'assurer. Il est toutefois peu vraisemblable que 40 millions d'Américains choisissent délibérément de ne pas se couvrir pour cette raison : dès que l'on réintroduit de l'aversion pour le risque, c'est l'assurance partielle qui redevient optimale, même avec des primes non concurrentielles.

Le second type d'explication fait peser plus largement sur les assureurs la responsabilité de cet état de fait : ils évitent ou refusent de couvrir certains risques ou certaines personnes qui présentent des niveaux de risque élevé. Cette hypothèse de sélection est partiellement validée par l'observation du comportement des assureurs : Dowd et Feldman (1992) citent une enquête réalisée en 1990 auprès de 48 assureurs : 29 d'entre eux refusent systématiquement d'assurer certains types d'entreprises (salons de coiffure et d'esthétique, compagnies de taxi par exemple). La première raison théorique à ce comportement est l'impossibilité (par exemple pour des raisons légales) pour les assureurs de moduler les primes en fonction du risque, même si ces assureurs peuvent évaluer précisèment le risque de leurs clients (voir Newhouse 1996). La seconde raison fournie par la théorie à ce comportement des assureurs est la présence d'aléa moral, dont la pertinence pour l'assurance maladie a été clairement démontrée par l'expérience de la Rand (voir Newhouse (1996) et, pour le cas français Caussat et Glaude (1993), Genier (1998) et Piaser et Raynaud

[1] Couverture publique destinée à certaines catégories de personnes à faible revenu.

(1998)). Pourtant, dans le contexte américain, avec l'influence croissante des HMOs, la consommation médicale est relativement bien contrôlée par les assureurs : on voit mal pourquoi le marché d'assurance maladie pourrait exister pour certains types d'assurés mais pas pour d'autres.

En réalité, ce que l'on entend souvent par sélection des risques ou refus d'assurance recouvre en fait une explication différente, que nous allons privilégier ici : la ponction que représente la prime d'assurance est trop élevée par rapport au revenu disponible de beaucoup de ménages. Aux Etats-Unis toujours, on constate que les personnes non couvertes appartiennent majoritairement à des ménages dont les revenus sont faibles : 66% des ménages non assurés ont un revenu annuel inférieur à 25 000 $. Les personnes appartenant à des ménages dont le revenu est intermédiaire (<50 000 $) semblent de plus en plus touchées et elles représentent une proportion de plus en plus élevée des personnes non couvertes (21% en 1989 et 24% en 1993, Hellander et al., 1995). En France, le constat est identique pour l'assurance complémentaire, quoiqu'avec des conséquences moins dramatiques, puisque l'essentiel du risque est couvert par la Sécurité Sociale. Les personnes appartenant au décile des revenus des plus faibles représentent 30% des personnes non assurées mais seulement 6% des personnes assurées[2]. Ce problème est d'autant plus crucial que les inégalités de santé sont corrélées avec les inégalités sociales (voir Judge et Whitehead 1995, pour une revue de littérature récente) : les personnes les moins favorisées sont en moyenne amenées à payer à cause de leur risque des primes relativement plus élevées que les personnes aisées. **Le revenu semble donc bien être une variable centrale expliquant l'accès à l'assurance.**

Le fait de n'être pas assuré a bien sûr des conséquences significatives en termes d'accès aux soins. Une étude empirique américaine (Berk et Schur, 1998) portant sur des données de 1994 met clairement en lumière les enjeux de l'accès à l'assurance. Les auteurs observent trois sous-populations âgées de moins de 65 ans[3] : les personnes qui ne bénéficient d'aucune assurance, celles qui sont couvertes par MEDICAID et celles qui sont couvertes par une assurance privée. Trois indicateurs d'accès aux soins sont étudiés : le fait de déclarer avoir une "source de soins habituelle", le fait d'avoir dû renoncer à des soins dans l'année précédente, et le nombre moyen de consultations chez un médecin de ville dans l'année précédente. A état de santé identique, on constate que les personnes qui ont une couverture (publique ou privée) ont la même probabilité d'avoir une source de soins habituelle et le même nombre moyen de consultations chez un médecin. Les personnes non assurées ont par contre une probabilité nettement plus faible d'avoir un lieu de recours habituel que les personnes couvertes et ont quasiment deux fois moins de consultations lorsqu'elles ont un mauvais état de santé. Les résultats sur le renoncement à des soins vont globalement dans le même sens, même s'ils semblent révéler que l'amélioration de l'accès que permet MEDICAID n'amène pas ses bénéficiaires au niveau des personnes couvertes de façon privée. Les personnes couvertes par MEDICAID ont une probabilité nettement plus élevée que des personnes couvertes par une assurance privée d'avoir dû renoncer à des soins au cours de l'année précédente et deux fois moins élevée que celles qui ne bénéficient d'aucune assurance. Une autre étude (Weinick, Zuckevas and Brilea,

[2]Revenu équivalent prenant en compte la taille du ménage (chiffres calculés à partir des enquêtes SPS 1994 et 1995 du CREDES).

[3]Les personnes de plus de 65 ans sont couvertes par MEDICARE.

1997) montre que les personnes qui n'ont pas d'assurance ont une probabilité trois fois plus élevée que des personnes assurées d'avoir des difficultés d'accès aux soins et de devoir attendre pour pouvoir bénéficier de soins dont elles perçoivent le besoin. Dans le cas de la France, les travaux issus de l'enquête santé INSEE-CREDES concluent dans le même sens (voir par exemple Genier (1998), Raynaud (1998)).

Pour autant, les personnes qui ne sont pas couvertes par une assurance finissent par avoir accès à des soins. Les hôpitaux publics et les médecins prennent en charge les personnes démunies, les Etats mettent parfois en place des financements spécifiques pour les besoins de santé des populations non couvertes. Le système sanitaire, y compris aux Etats-Unis, finit généralement par prendre en charge les patients qui ne bénéficient pas d'assurance. Ce système est critiqué en France car, "la mise en place de lieux de traitement des plus défavorisés, palliatifs utiles, maintient l'exclusion du droit commun" (rapport Boulard, 1998).

Ce mode de prise en charge des soins pose d'autres problèmes : les personnes non assurées n'étant pas suivies de façon régulière, leurs besoins sont exprimés plus tardivement. Lorsque ces besoins sont finalement pris en charge par la collectivité, ils peuvent se traduire par une prise en charge plus coûteuse. En France, les personnes qui ne bénéficient pas d'assurance complémentaire semblent ainsi avoir des niveaux de dépense et des modes de recours aux soins différents de ceux des personnes couvertes. A titre d'exemple[4], la dépense moyenne en 1995, des personnes non couvertes par une assurance est de 17 % plus élevée que celle des personnes couvertes. De plus, alors que le taux de personnes hospitalisées pendant l'année est comparable entre les populations couvertes et non couvertes, la dépense par personne hospitalisée couverte est quasiment deux fois plus élevée que cette d'une personne non couverte. Ce constat ne s'explique pas seulement par l'absence d'assurance, car l'avance de frais, exception française, constitue une barrière largement aussi importante pour les personnes dont le revenu est faible. Ceci dit, le renoncement aux soins pour l'optique, les prothèses dentaires ou auditives s'explique largement par l'absence d'assurance complémentaire en raison des taux de couverture dérisoires de la Sécurité Sociale. Comme le souligne le rapport Boulard[5] "les restriction d'accès à ces soins apparaissent d'autant plus préoccupantes qu'elles touchent fortement à l'insertion sociale, professionnelle et tout simplement humaine".

Outre une augmentation du coût de la prise en charge, un tel système de prise en charge des soins ex post introduit des distorsions de prix. Aux Etats-Unis, la pratique est connue sous le terme de "cost shifting" : les tarifs payés par les assureurs privés aux prestataires de soins sont surévalués car ces derniers reportent sur leur clientèle couverte et/ou aisée le coût de la prise en charge des personnes non solvables[6]. L'ensemble de ces préoccupations motive le débat législatif en cours. Citons encore le rapport Boulard : "Avec la couverture maladie universelle il ne s'agit pas de consolider l'existence d'une filière sanitaire des pauvres s'articulant autour de l'hôpital public et des centres de santé

[4]Enquête appariée SPS-EPAS 1995, échantillon de 4505 personnes (calcul de l'auteur).

[5]Boulard, J.-C., 1998.

[6]Remarquons toutefois que le contrôle croissant des coûts par les HMOs rend cette stratégie de plus en plus difficile. En conséquence, les médecins et les hôpitaux sont découragés de soigner les personnes non solvables.

mais d'ouvrir aux personnes en difficulté le système de soins de tous". Le présent travail vise à souligner les enjeux de l'introduction d'une telle couverture.

2 Le modèle

Dans une première étape, on fait l'hypothèse que les revenus sont exogènes et qu'il existe une prise en charge ex post par le système sanitaire des dommages de personnes non couvertes par une assurance privée. De plus, cette prise en charge n'est pas plus coûteuse que la prise en charge du dommage pour une personne qui bénéficie d'une assurance ex-ante. On montre alors que l'existence d'un tel "filet de sécurité" conduit certaines personnes à ne pas s'assurer alors même qu'elles en ont les moyens.

Hypothèses : Les individus ont deux caractéristiques exogènes : un revenu R (interprété comme le revenu disponible au-delà d'une consommation minimale incompressible), et une probabilité p de subir un dommage D, dont le montant est supposé très élevé par rapport à la moyenne des revenus de la population. Pour fixer les idées, disons que ce montant correspond par exemple au coût d'une hospitalisation de longue durée.

Les individus maximisent l'espérance d'une fonction d'utilité de type Von Neumann-Morgenstern, notée U et supposée croissante ($U' > 0$) et concave ($U'' < 0$).

Le marché de l'assurance privée est parfait et chaque individu peut acquérir la quantité d'assurance q qu'il souhaite à un prix (unitaire) actuariel p.

Si un individu n'est pas assuré de façon privée, il est pris en charge par l'Etat dans le cas où il subit un dommage et sa consommation est alors normée à zéro (autrement dit, l'Etat ne prend en charge que la part du dommage que l'individu ne peut pas couvrir par son revenu disponible).

Considérons, dans un premier temps, le cas où il n'y a pas de contrainte financière à l'accès à l'assurance.

Chaque individu maximise son espérance d'utilité V, par rapport au niveau de couverture q :

$$V(p, R, q) = (1 - p)U(R - pq) + pU(R - pq + q - D). \tag{1}$$

La solution est évidemment l'assurance complète. La consommation finale de l'individu est $R - pD$ et l'utilité individuelle $U(R - pD)$.

Considérons maintenant le cas où le revenu de certains individus ne leur permet pas de souscrire une assurance complète.

Si $R < pD$, les individus pris en charge par l'Etat en cas de dommage n'achètent pas d'assurance (même partielle), puisqu'ils sont pris en charge ex post. Ils ont alors l'espérance d'utilité suivante :

$$V(p, R) = (1 - p)U(R) + pU(0).$$

Nous noterons $R_0(p)$ le revenu minimum permettant à un individu de risque p de se couvrir : $R_0(p) = pD$.

3 Le renoncement à l'assurance

Montrons que cette possibilité de prise en charge ex post des soins par l'Etat incite certains individus à ne pas s'assurer sur le marché privé alors même qu'ils en ont les moyens. En effet, cette prise en charge implique que la fonction d'utilité des agents est en réalité[7] :

$$V(p, R) = \max\{U(R - pD), (1 - p)U(R) + pU(0)\}.$$

Sans perte de généralité, nous poserons $U(0) = 0$. Le choix des individus est alors dicté par le signe de la fonction :

$$\phi(R, p) = (1 - p)U(R) - U(R - pD).$$

Plus précisément, un individu de caractéristiques R (revenu) et p (risque) s'assure si et seulement si $\phi(R, p) \leq 0$.

Proposition 1 *Dans un système d'assurance maladie privée, la prise en charge par l'Etat des soins des ménages les plus démunis désincite certains ménages à s'assurer. Plus précisément il existe pour tout p un revenu critique $R^*(p)$ (supérieur au coût de l'assurance $R_0(p) = pD$, et éventuellement infini) en deçà duquel les ménages renoncent à s'assurer. Si $D < \lim_{R\to+\infty} \frac{U(R)}{U'(R)}$, $R^*(p)$ est fini pour tout p.*

On distingue alors trois catégories de ménages :

- *les ménages aux ressources insuffisantes pour s'assurer. Ils sont caractérisés par un revenu $R \leq R_0(p) = pD$,*
- *les ménages renonçant volontairement à l'assurance. Ils sont caractérisés par un revenu R tel que : $R_0(p) < R < R^*(p)$,*
- *les ménages qui s'assurent, dont le revenu est tel que $R \geq R^*(p)$.*

Démonstration : Notons $\phi(R, p)$ la différence entre l'utilité d'un ménage non assuré et celle d'un ménage assuré :

$$\phi(R, p) = (1 - p)U(R) - U(R - pD).$$

Remarquons tout d'abord que si $R = pD$, alors $\phi(R, p) = (1 - p)U(R) > 0$. (Un ménage ayant juste les moyens de payer son assurance préférera recourir à l'aide de l'Etat). D'autre part :

$$\frac{\partial \phi}{\partial R} = (1 - p)U'(R) - U'(R - pD) < U'(R) - U'(R - pD) < 0,$$

[7] Comme les primes d'assurance sont actuarielles, la quantité optimale d'assurance est toujours 0 ou D.

la dernière inégalité venant du fait que U' est décroissante. Cette inégalité signifie que, pour un niveau de risque donné, la tentation de ne pas s'assurer diminue quand le revenu augmente. Enfin :

$$\frac{\phi(R,p)}{U(R)} = \frac{U(R) - U(R-pD)}{U(R)} - p = pD\left[\frac{U'(R_1)}{U(R)} - \frac{1}{D}\right], \tag{2}$$

où R_1 est dans l'intervalle $[R - pD, R]$. La deuxième égalité découle du théorème des accroissements finis appliqué à la fonction U entre $R - pD$ et R.

U étant concave croissante, la fonction $\frac{U'(R)}{U(R)}$ est décroissante positive. Elle a donc une limite ≥ 0 quand R tend vers l'infini. Cette limite est aussi celle de $U'(R_1)/U(R)$ quand R tend vers l'infini. Si cette limite est supérieure à $\frac{1}{D}$ alors la formule (2) montre que ϕ est positive pour tout R. Si elle est $< \frac{1}{D}$ alors il existe un revenu R^* critique (dépendant de p) tel que $\phi(R,p) > 0 \Leftrightarrow R < R^*(p)$.

■

Remarquons que si $D > \lim_{R\to+\infty} \frac{U(R)}{U'(R)}$, le comportement de renoncement à l'assurance est total : aucun ménage ne s'assure. Toutefois, ce cas extrême ne peut se produire que si $\frac{U'(R)}{U(R)}$ a une limite non nulle en $+\infty$ et si D est assez élevé.

Etudions maintenant les propriétés de la fonction $R^*(p)$ (on suppose donc désormais que $D < \lim_{R\to+\infty} \frac{U(R)}{U'(R)}$).

Proposition 2 *La fonction $R^*(\cdot)$ est croissante sur $[0,1]$. Ses valeurs extrêmes sont caractérisées comme suit :*
$R^*(0)$ est la solution de $\frac{U}{U'}(R) = D$.
$R^*(1) = D$.

Démonstration : $R^*(p)$ est défini implicitement par $\phi(R^*(p),p) = 0$, avec $\phi(R,p) = (1-p)U(R) - U(R-pD)$. On a donc :

$$p\frac{\partial\phi}{\partial p}[R^*(p),p] = [DU'(R^*(p) - pD) - U(R^*(p))]\,p,$$

avec par définition :

$$pU(R^*(p)) = U(R^*(p)) - U(R^*(p) - pD).$$

Donc

$$p\frac{\partial\phi}{\partial p}[R^*(p),p] = pDU'(R^* - pD) - U(R^*) + U(R^* - pD).$$

La concavité de U implique que

$$U(R^*) - U(R^* - pD) > pDU'(R^* - pD),$$

donc $p\frac{\partial\phi}{\partial p}(R^*(p),p) < 0$, ce qui implique la croissance de R^* par rapport à p.

Pour le comportement en $p = 0$ et $p = 1$, il suffit de faire des développements limités :

$$\text{En } p = 0,\ \phi(R,p) \sim p[DU'(R) - U(R)] = 0 \Leftrightarrow \frac{U}{U'}(R^*(0)) = D$$

$$\text{En } p = 1,\ \phi(R,p) \sim (1-p)U(R) - U(R-D) + (1-p)DU'(R-D) \to 0$$

$$\Leftrightarrow \quad R^*(1) = D.$$

■

La figure ci-après illustre la Proposition 2 :

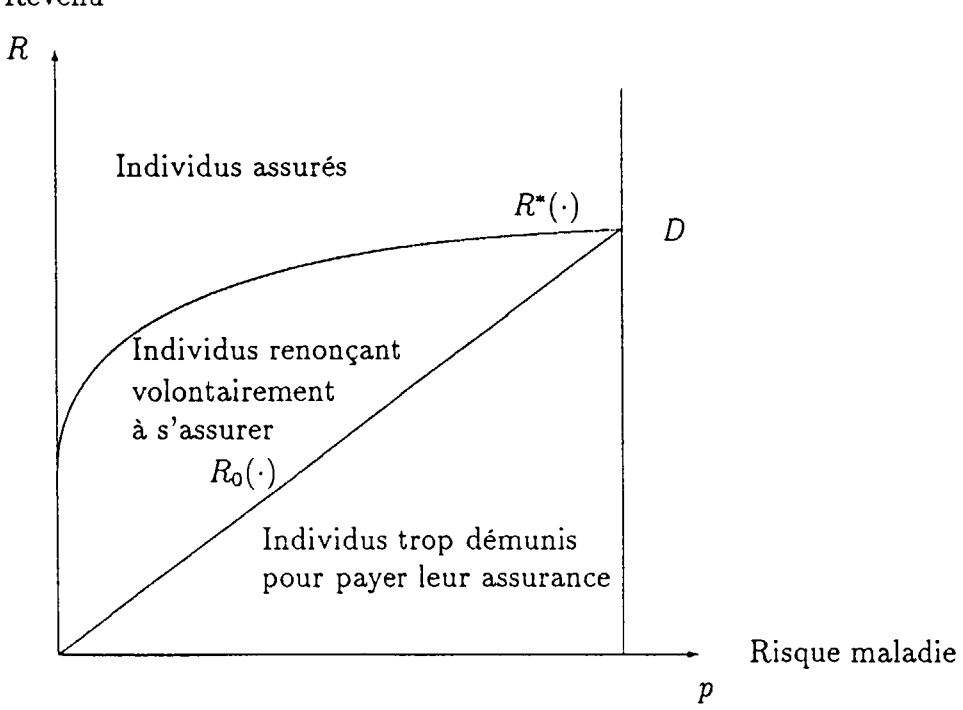


Figure 1 : Décision d'assurance en fonction du niveau de risque et de revenu.

Un remède classique au comportement de renoncement est l'obligation d'assurance, appliquée par exemple à la responsabilité civile en assurance automobile. Bien que des infractions à cette loi soient constatées très régulièrement dans la plupart des pays, il n'y a pas d'obstacle financier a priori à ce que la loi soit appliquée, dans la mesure où les personnes ayant fait l'acquisition d'une voiture ont en principe les moyens de payer aussi leur assurance responsabilité civile. Il en va tout autrement pour l'assurance maladie : la plupart des sociétés développées défendent le principe d'un accès universel aux soins de base alors même qu'une fraction non négligeable de la population n'a pas les moyens de couvrir le coût de son assurance maladie. Autrement dit même si les individus renonçant volontairement à s'assurer pourraient théoriquement être forcés à le faire, cette obligation d'assurance ne résoudrait pas le problème des ménages les plus démunis.

De fait, aussi bien les USA que les pays européens fournissent une aide médicale gratuite aux personnes non assurées. La différence c'est qu'en Europe ces personnes représentent une infime minorité, puisque la très grande majorité est couverte par le système public. Nous allons maintenant comparer la performance de ces deux systèmes (aide médicale gratuite et assurance publique universelle) en introduisant un autre élément fondamental du débat, à savoir la présence de désincitations à participer au marché du travail ("poverty trap").

4 La participation au marché du travail

La mise en place d'un système d'assurance maladie universelle peut également avoir des effets sur la participation au marché du travail. Plusieurs économistes américains ont donné des arguments convainquants dans ce sens. Par exemple, pour Cutler (1994, p. 20) "empirical estimates suggest that up to one-quarter of the approximately 4 million welfare recipients would enter the labor force if health insurance were available continuously". De même Newhouse (1994, p. 9) considère que "the loss of Medicaid benefits [for welfare mothers who start working] is a disincentive to work".

Nous allons maintenant endogénéiser la participation au marché du travail : chaque individu peut renoncer à travailler, auquel cas son revenu devient nul. R s'interprète désormais comme un revenu potentiel. On doit donc comparer 3 niveaux d'utilité :

- $U_A = U(R - c_1 - pD) - \gamma$ pour un actif assuré,
  où c_1 est une cotisation (supposée indépendante du revenu) destinée à financer l'aide médicale gratuite, et γ représente la désutilité du travail.
- $U_N = (1-p)U(R - c_1) - \gamma$ pour un actif non assuré.
- $U_I = 0$ pour un inactif.

Ces 3 niveaux d'utilité correspondent à 3 zones du plan (p, R), que nous allons maintenant caractériser. Nous supposerons que la désutilité du travail n'est pas trop forte : $\gamma < U(R^*(0))$, où $R^*(0)$ est donné implicitement par $\frac{U}{U'}(R^*(0)) = D$.

Proposition 3 *Dans un système d'assurance privé avec aide médicale gratuite des plus démunis, la population se répartit en 3 catégories :*

- *les actifs assurés (zone A), caractérisés par un revenu net* $R - c_1$ *supérieur à* $\max(R^*(p), pD + U^{-1}(\gamma))$.
- *Les actifs non assurés (zone N), dont le revenu net est compris entre* $U^{-1}\left(\frac{\gamma}{1-p}\right)$ *et* $R^*(p)$.
- *Les inactifs (zone I), correspondant au reste de la population.*

Démonstration : La zone A correspond aux inégalités
$U_A \geq U_N$ (caractérisée par $R - c_1 \geq R^*(p)$, voir proposition 1) et
$U_A \geq U_I$, qui équivaut à $R - c_1 \geq pD + U^{-1}(\gamma)$.

Le reste de la démonstration est évident.

■

Les trois zones sont représentées par la figure ci-après :

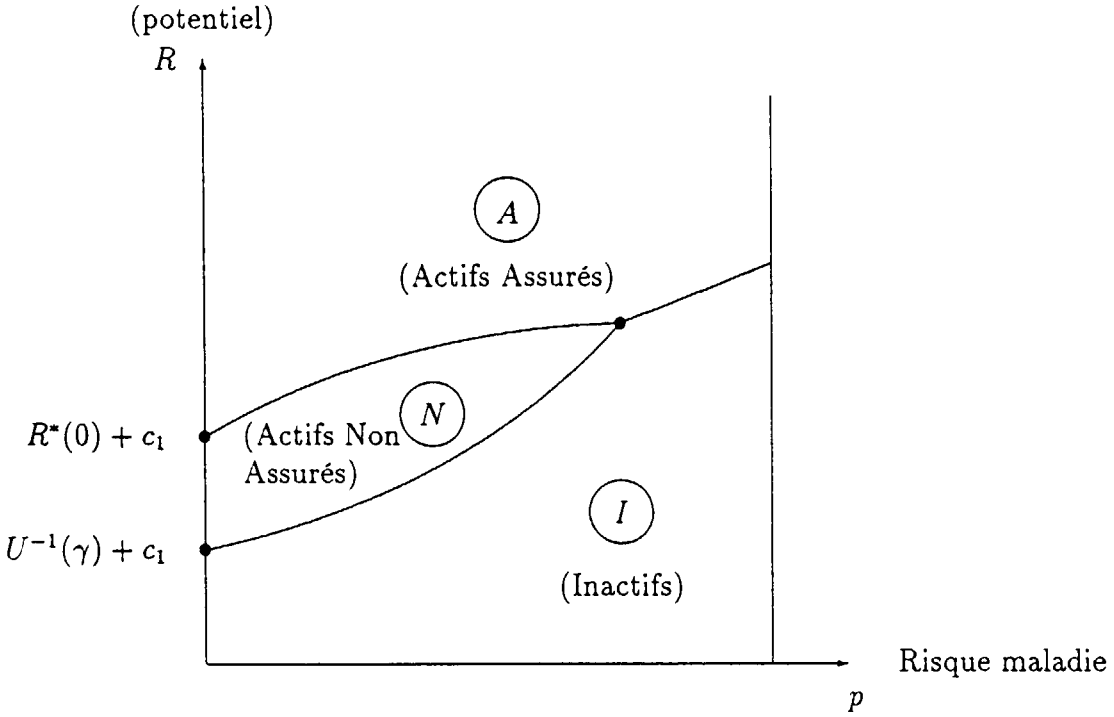


Figure 2 : Assurance et activité en fonction du niveau de risque et de revenu. (système privé)

Nous allons maintenant analyser l'impact qu'aurait une réforme introduisant une couverture maladie universelle, c'est-à-dire une assurance de tous les ménages, financée par une cotisation c_2 (supposée là encore indépendante du revenu)[8] prélevée sur l'ensemble des actifs. L'espérance d'utilité devient alors indépendante de la probabilité de dommage :

$$U = \max(U(R - c_2) - \gamma, 0).$$

Le niveau de c_2 est déterminé par l'équilibre budgétaire du système, qui s'exprime en disant que le coût unitaire moyen de la couverture maladie universelle doit être couvert

[8]Nous ne discutons pas de la dimension redistributive des cotisations d'assurance maladie, qui dépendent en général du revenu. Cette question est examinée dans Henriet et Rochet (1998).

par le produit du montant c_2 de la cotisation a la proportion d'assujettis :

$$p_m D = c_2 \left(1 - G(c_2 + U^{-1}(\gamma))\right)$$

(où p_m est la moyenne de p dans la population et G la fonction de répartition de R). c_2 est donc vraisemblablement plus élevé que c_1, qui ne correspond qu'à la couverture du coût beaucoup plus faible de l'aide médicale gratuite. L'impact de la réforme est alors représenté par la figure 3 ci-après :

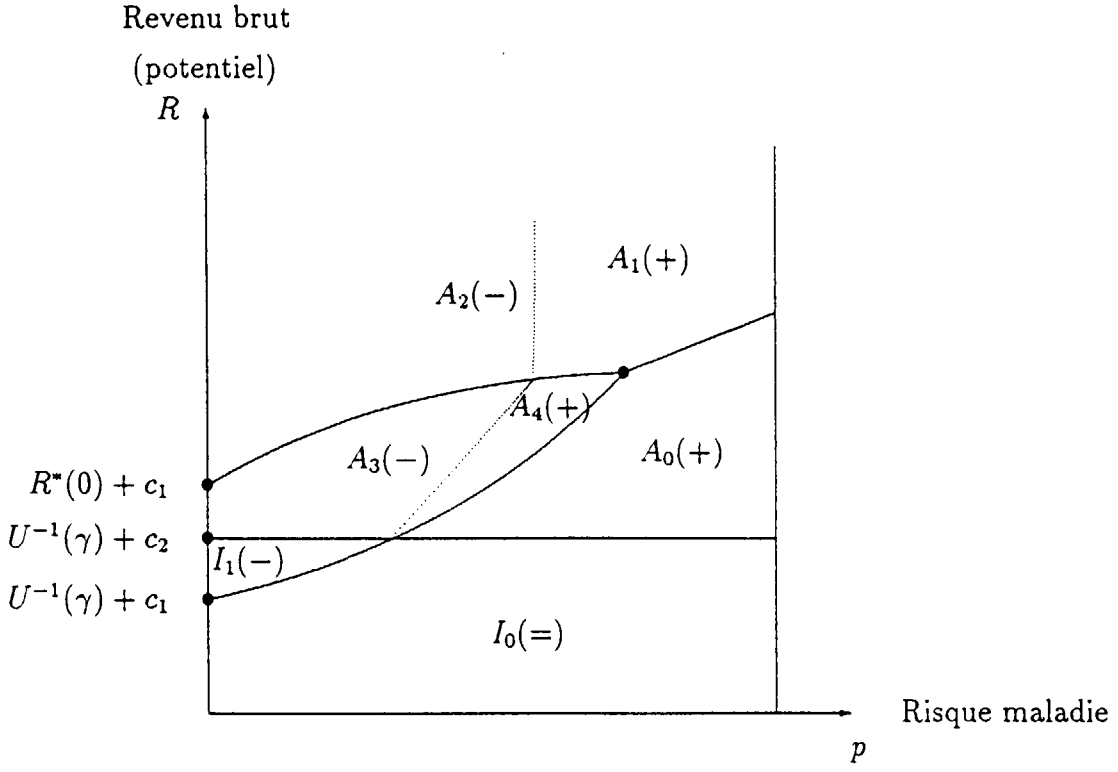


Figure 3 : Les conséquences de l'introduction d'une assurance maladie universelle (+ désigne les ménages qui y gagnent, —ceux qui y perdent).[9]

Comme nous l'avons remarqué, la séparation entre actifs (A) et inactifs (I) devient indépendante de p : la zone A est caractérisée désormais par : $R > c_2 + U^{-1}(\gamma)$.

A l'intérieur de la zone I, nous distinguons les individus que la réforme a désincité à travailler (I_1) et les autres (I_0), dont la situation est inchangée.

A l'intérieur de la zone A, les choses sont plus compliquées :

- la région A_0 correspond aux personnes qui (au contraire de I_1) ne travaillaient pas par peur de perdre le droit à l'aide médicale gratuite, mais que la réforme incite à travailler ;

[9]Nous avons représenté le cas où $U^{-1}(\gamma) + c_2 < R^*(0) + c_1$. Le cas où $U^{-1}(\gamma) + c_2 \geq R^*(0) + c_1$ est qualitativement similaire.

- la région A_1 correspond aux personnes actives et assurées dont la cotisation globale a diminué ($c_2 < c_1 + pD$) ;
- la région A_2 correspond aux personnes actives et assurées dont la cotisation globale a augmenté ($c_2 > c_1 + pD$) ;
- enfin les régions A_3 et A_4 correspondent aux personnes qui n'étaient pas assurées (mais qui le deviennent automatiquement) : certaines y gagnent (A_4) d'autres y perdent (A_3), la frontière entre ces deux régions (courbe en pointillé) étant déterminée par l'égalité :

$$U(R - c_2) = (1 - p)U(R - c_1).$$

5 Conclusion

A l'aide d'un modèle extrêmement simple, nous avons analysé l'impact qu'aurait, dans un pays comme les USA, le passage à une couverture maladie universelle financée par une cotisation forfaitaire payée par tous les actifs. Comme l'on pouvait s'y attendre, les ménages qui auraient à perdre d'une telle réforme sont ceux dont le revenu est élevé ou le risque maladie est faible. En fait notre analyse permet de distinguer 3 types de profils "perdants" :

- les ménages à revenu élevé et risque faible (zone A_2), dont la cotisation d'assurance augmente,
- les ménages à revenu moyen et risque faible (zone A_3), qui sont forcés de s'assurer contre leur gré,
- les ménages à revenu et risque faibles (zone I_1), qui sont incités à ne plus travailler.

Par contre, les autres ménages bénéficient de la réforme :

- les ménages à risque élevé et revenu élevé (zone A_1) dont la cotisation d'assurance baisse,
- les ménages à risque élevé et revenu moyen (zone A_0) qui sont désormais incités à entrer sur le marché du travail,
- les ménages à risque et revenu moyen (zone A_4) qui bénéficient désormais d'une couverture maladie.

L'échec du plan Clinton peut être probablement être en partie expliqué par l'arbitrage réalisé au Congrès US entre les intérêts des différents types de populations que nous venons d'évoquer et dont on se doute que la représentation n'est pas nécessairement assurée de façon proportionnelle.

En France, où la nécessité d'offrir à tous un accès minimum aux soins est plus consensuelle, on envisage une réforme de la protection sociale en trois volets : la création d'une

couverture maladie (vraiment) universelle, la garantie d'une protection complémentaire pour les plus démunis, et l'instauration de l'avance de frais pour les même catégories. Cette réforme doit cependant éviter un certain nombre d'écueils, en particulier en termes d'effets de seuil et de possibilités d'arbitrages que cette réforme pourrait instaurer, du côté des bénéficiaires comme du côté des employeurs.

Notre travail pourrait être étendu dans deux directions :

- une calibration du modèle, sur données US, permettant d'évaluer les importances relatives des différentes catégories de ménages évoquées ci-dessus,
- une adaptation du modèle au cas français.

BIBLIOGRAPHIE

Benzeval M., Judge K., et M. Whitehead Ed.(1995), Tackling Inequalities in Health: An Agenda for Action.

Berk, M.L. et C. L. Schur (1998), "Access to Care: How Much Difference does Medicaid Make?", *Health Affairs*, 17(3), 169-180.

Boulard, J.C. (1998), "Pour une Couverture Maladie Universelle Base et Complémentaire", Rapport Parlementaire, Août.

Caussat, L. et M. Glaude (1993), "Dépenses Médicales et Couverture Sociale", *Economie et Statistique*, 265.

Cutler, D. (1994), "A Guide to Health Care Reform", *Journal of Economic Perspectives*, 8(3), 13-29.

Dowd, B. et R. Feldman (1992), "Insurer Competition and Protection from Risk Redefinition in the Individual and Small Group Health Insurance Market", Inquiry 29(2), Summer, 148-57.

Genier, P. (1998), "Assurance et Recours aux Soins", *Revue Economique*, à paraître.

Hellander I., Moloo J., Himmelstein D. U., Woolhandler S. et S. Wolfe (1995), "The Growing Epidemic of Uninsurance: New Data on the Health Insurance Coverage of Americans", *International Journal of Health Services*, 25(3), 377-392.

Henriet, D. et J.C. Rochet (1998), "Is Public Health Insurance and Appropriate Instrument for Redistribution?", GREQAM, Université d'Aix-Marseille et GREMAQ, Université de Toulouse.

Newhouse, J.P. (1994), "Symposium on Health Care Reform", *Journal of Economic Perspectives*, 8(3), 3-11.

Piaser, G. et D. Raynaud (1998), "Consultation Médicale : l'Influence du Revenu et de l'Assurance Complémentaire", GREMAQ, Université de Toulouse.

Raynaud, D. (1998), "Santé et Accès aux Soins", GREMAQ, Université de Toulouse.

Weinick, R. M., Zuvekas S.H. et S.K. Drilea (1997), "Access to Health Care - Sources and Barriers", Agency for Health Care Policy and Research, Research Findings AHCPR pub. n^0 98-001, Oct.