Article

« Assurance-maladie : comment adapter les taux de remboursement aux dépenses individuelles de santé? »

Barbara Lipszyc et Maurice Marchand

L'Actualité économique, vol. 75, 1999, p. 447-473.

L'Actualité économique, Revue d'analyse économique, vol. 75, nos 1-2-3, mars-juin-septembre 1999

ASSURANCE-MALADIE : COMMENT ADAPTER LES TAUX DE REMBOURSEMENT AUX DÉPENSES INDIVIDUELLES DE SANTÉ?*

Barbara LIPSZYC
CREPP,
Université de Liège
Maurice MARCHAND
IAG
et *CORE,*
Université Catholique de Louvain

RÉSUMÉ – Nous considérons un modèle d'assurance-maladie dans lequel les agents ne se différencient que par la gravité de la maladie qui les atteint. L'État cherche à maximiser l'espérance d'utilité des assurés et décide en conséquence de rembourser une fraction des dépenses de santé. En l'absence d'aléa moral *ex post*, suivant lequel la décision individuelle de dépenses de santé est affectée par leur taux de remboursement, celui-ci pourrait être de 100 %. Cependant, avec aléa moral, la gratuité des soins n'est plus de mise. Après une brève présentation du cas d'un taux de remboursement uniforme, nous envisageons d'abord une structure de remboursement à deux taux, le premier s'appliquant en dessous d'un certain seuil de dépenses et le second au-delà du seuil. Nous voulons connaître la valeur relative de ces deux taux de remboursement, ainsi que le montant du seuil. Ensuite nous montrons les caractéristiques d'un remboursement non linéaire, qui nous rapproche un peu plus de la solution de premier rang. Des exemples numériques illustrent les développements analytiques et montrent comment le partage des risques entre bien-portants et malades et la perte d'efficacité due à l'aléa moral varient selon le schéma de remboursement.

ABSTRACT – *Health Insurance: Defining the Reimbursement Rates According to Individual Health Care Expenses.* We consider a health insurance model with heterogeneous agents who only differ in illness severity. The public insurer intends to maximize the expected utility of people insured taking into account the premium paid by them to balance the insurer's budget. Without any *ex post* moral hazard, the reimbursement rate would be set at

* Les auteurs tiennent à remercier P. Pestieau pour ses précieux conseils, ainsi que les participants au *Public Economics Workshop* du 14/03/97 à la KULeuven, et plus particulièrement Cl. d'Aspremont et Ph. Michel. Le premier auteur est aspirante auprès du FNRS (Fonds National de la Recherche Scientifique).

100 %. But with moral hazard, the individual decision, as regards medical expenses, varies with this rate. After a short presentation of the single rate case, we consider a two-rate reimbursement structure, with a threshold defining the scope of each rate, the one applying below the threshold and the other taking effect above this cut-off point. We want to determine the relative value of these two rates, as well as the amount of the threshold. Then we characterize a non-linear reimbursement, which is closer to the first-best solution. Some numerical simulations illustrate the analytical developments. They show how the reimbursement structure affects the risk sharing between the healthy and the sick and the efficiency loss caused by moral hazard.

INTRODUCTION

Dans nombre de systèmes d'assurance-maladie, les patients doivent prendre partiellement en charge le coût des soins curatifs qui leur sont prodigués. On parle de tickets modérateurs ou de coassurance. Cette participation financière des patients est expliquée par la volonté de les inciter soit à adopter un comportement préservant leur santé et à recourir aux soins préventifs (problème de l'aléa moral *ex ante*), soit à ne pas surconsommer les soins curatifs une fois la maladie survenue (aléa moral *ex post*). La préoccupation de fournir ainsi aux assurés des incitations appropriées – afin de contrôler les phénomènes d'aléa moral – doit être mise en balance avec la préoccupation de partage des risques qui justifie l'existence de l'assurance-maladie. En effet, une augmentation des tickets modérateurs a pour conséquence automatique d'accroître le risque financier supporté par les assurés. Il existe bien évidemment de nombreux travaux qui se sont penchés sur cette question[1]. Dans cet article, notre attention se concentre sur l'aléa moral *ex post*. D'un côté, une absence complète de participation des patients peut entraîner un excès d'utilisation des soins curatifs et donc un coût qui peut se révéler prohibitif pour l'assurance-maladie. Mais de l'autre, fixer la participation financière des assurés à un niveau élevé implique que l'on accepte de profondes inégalités entre leurs niveaux de vie, selon la gravité de la maladie à laquelle ils font face. Il s'agit dès lors de trouver un compromis entre ces deux considérations.

Nombre de travaux théoriques existant sur le choix des tickets modérateurs en présence d'aléa moral *ex post* font l'hypothèse d'un seul taux de ticket modérateur – ou d'un seul taux de remboursement, qui est son complément à l'unité – s'appliquant aux dépenses quel que soit leur montant. Ceci même si l'on reconnaît que les individus peuvent avoir des états de santé différents. Or, dans la réalité, nous observons des systèmes de remboursement qui sont davantage différenciés soit en fonction des catégories de soins, soit en fonction de l'importance des dépenses individuelles. La liaison du ticket modérateur aux dépenses individuelles est ainsi présente dans les systèmes de remboursement avec franchise

1. On peut mentionner, entre autres, l'article bien connu d'Arrow (1963), Spence et Zeckhauser (1971) et Winter (1992). Concernant plus particulièrement l'aléa moral *ex ante*, citons Henriet et Rochet (1991 : ch. 6), Pauly (1968) et (1974), Salanié (1994 : ch. 5), ou encore Shavell (1979). Quant à l'aléa moral *ex post*, le principal article à citer à ce sujet est celui de Zeckhauser (1970).

où le ticket modérateur marginal devient nul (gratuité des soins) dès que la participation financière de l'assuré (tickets modérateurs cumulés) dépasse un montant spécifié. En Belgique par exemple, un tel système a été récemment introduit dans l'assurance-maladie pour les soins ambulatoires (la franchise étant de surcroît liée aux revenus des assurés, à l'instar de ce qui fut réalisé dans l'expérimentation de la Rand Co.[2] aux États-Unis).

La différenciation des taux de remboursement en fonction des types de soins a fait l'objet de plusieurs travaux, parmi lesquels ceux de Zeckhauser (1970) et Besley (1988). Par contre, le recours à des systèmes d'assurance-maladie où le taux marginal de remboursement varie en fonction des dépenses occasionnées par un assuré n'a guère été étudié dans la littérature. C'est ainsi que Keeler, Newhouse et Phelps (1977) s'intéressent à l'effet sur le comportement des individus du système de remboursement suivant : le patient prend en charge la totalité de ses dépenses jusqu'à un certain seuil (franchise) au-delà duquel un taux de remboursement est appliqué. Mais seule la valeur du taux après franchise est analysée : l'opportunité d'avoir deux taux de remboursement positifs mais inférieurs à 100 % n'est pas envisagée. Dans le présent article, notre objectif est d'étudier, au moyen d'un modèle fort simple, le choix optimal de taux marginaux de remboursement différenciés selon l'importance de la consommation médicale.

Une double question se pose. Est-il optimal de différencier ces taux? Si oui, comment faut-il procéder? Faut-il amener les agents les plus malades à payer plus ou moins à la marge? L'intuition nous suggère deux réponses. D'une part, la préoccupation de partage des risques nous conduit à accorder un taux marginal de remboursement plus élevé aux agents dont les dépenses en soins de santé sont les plus importantes. D'autre part, la question de l'aléa moral montre qu'il peut être efficace de pénaliser les plus grands consommateurs de soins lorsque leur demande est plus sensible au taux de remboursement. Par conséquent, de ces deux effets nous devons déterminer quel est celui qui dominera. Après la présentation du modèle utilisé dans la section 1 et l'analyse du choix d'un taux unique dans la section 2, ces questions sont étudiées dans la section 3. Des exemples numériques illustrent le choix du taux de remboursement en fonction de l'aversion au risque des assurés et de la sensibilité de leur demande au taux de remboursement.

Un schéma de remboursement à deux taux correspond à une fonction de remboursement qui est linéaire par morceaux; il s'agit donc d'un cas particulier d'un schéma de remboursement non linéaire. La linéarité par morceaux impose bien entendu de fortes contraintes sur le choix de la fonction de remboursement et on peut chercher à en dépasser les limites. C'est ce qui est fait dans la section 4 où aucune forme n'est imposée *a priori* à la fonction de remboursement. Nous y expliquons comment construire un contrat d'assurance optimal dans le cas où le nombre d'états de maladie est fini. Enfin, les résultats obtenus avec les trois

2. Voir Manning *et al.* (1987).

schémas de remboursement (linéaire à taux unique, linéaire à deux taux et non linéaire) sont comparés dans la section 5, au moyen d'exemples numériques; ceux-ci montrent que le partage des risques s'améliore et que la perte d'efficacité due à l'aléa moral diminue à mesure que la fonction de remboursement se complexifie.

1. LE MODÈLE UTILISÉ

Comme nous l'avons déjà indiqué, dans le modèle d'assurance optimale utilisé dans cet article, l'autorité publique (dénommée dans la suite « l'assureur ») recherche la fonction de remboursement des dépenses de santé qui maximise l'espérance d'utilité des individus. Cette fonction de remboursement lie l'indemnité payée à l'assuré à ses dépenses de santé.

Au départ, l'assuré ne connaît pas le degré de gravité de la maladie à laquelle il peut être confronté, gravité que nous dénotons par g. La fonction de distribution de cette variable aléatoire est dénotée par $F(g)$ avec $g \in [0, \overline{g}]$ et elle est supposée identique pour tous les individus. Une fois qu'un assuré connaît la réalisation de cette variable aléatoire (c'est-à-dire la gravité de sa maladie éventuelle), il décide du montant de sa dépense en soins curatifs. Nous faisons ici l'hypothèse d'un « patient-décideur », en ce qui concerne les soins qui lui sont administrés, ou de manière équivalente l'hypothèse d'un médecin qui se comporterait en agent parfait du patient. En effet, l'assuré tire son utilité de la consommation d'un bien composite x (utilisé comme numéraire), mais aussi de son état de santé h, amélioré le cas échéant grâce à sa dépense en soins curatifs. Sa fonction d'utilité est caractérisée comme suit : $u = u(x + h)$, avec $u' > 0$ et $u'' < 0$. Suivant cette fonction, l'état de santé est exprimé dans les mêmes unités que le numéraire x, et l'aversion au risque de l'assuré porte à la fois sur sa consommation du bien composite et sur son état de santé.

L'état de santé dépend à la fois de la gravité de la maladie et des dépenses en soins curatifs y. On a donc $h(y; g)$, dont les propriétés sont par hypothèse les suivantes :

$$h_y > 0,\ h_{yy} < 0, \tag{1}$$

$$g_1 < g_2 \Rightarrow \begin{matrix} h(y; g_1) > h(y; g_2) \\ h_y(y; g_1) < h_y(y; g_2) \end{matrix} \tag{2}$$

et $h_y(y_1; g_1) = h_y(y_2; g_2) \Rightarrow h(y_1; g_1) > h(y_2; g_2)$,

$$h(y; 0) = h_0,\ \forall y \tag{3}$$

et $$h(\infty; g) < h_0,\ \forall\ g > 0. \tag{4}$$

Suivant (1), la productivité marginale des dépenses en soins curatifs est supposée positive et décroissante avec leur montant. Par contre, suivant (2), lorsque la gravité de la maladie augmente, cette productivité marginale est plus élevée pour un niveau de dépenses donné alors que l'état de santé après traitement est

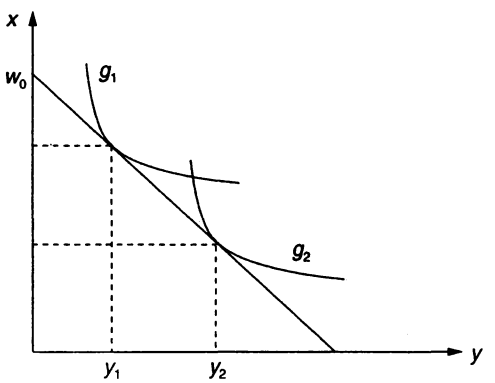
moins bon. De plus, en présence d'un taux de remboursement constant des soins curatifs, l'état de santé – après traitement au niveau optimal – se détériore avec la gravité de la maladie. La propriété (3) implique que le seul choix de dépenses dans l'état de bonne santé est un niveau nul, car des soins ne sont dans ce cas d'aucune utilité. Enfin, nous faisons l'hypothèse en (4) que les soins ne permettent jamais à un malade de retrouver le niveau de santé atteint en l'absence de maladie (h_0).

Il est intéressant d'examiner la représentation graphique du choix individuel dans l'espace (y, x) pour un degré de gravité donné. Une courbe d'indifférence y est le lieu des paires (y, x) telles que l'argument de la fonction d'utilité ($x + h$) reste inchangé, ce qui implique un taux marginal de substitution donné par : $TMS = -\left.\frac{dx}{dy}\right|_u = h_y$ pour un g donné. En conséquence, ces courbes sont verticalement parallèles et satisfont à la condition de monotonicité suivante : en un point quelconque (y, x) l'inclinaison des courbes d'indifférence est en valeur absolue croissante avec g ($h_{yg} > 0$, *cf.* propriété (2)). En d'autres termes, le taux marginal de substitution est croissant avec g en tout point de l'espace (y, x).

En l'absence de tout système d'assurance (solution de laisser-faire), la situation est facilement représentée par la figure 1. La droite de budget dans l'espace (y, x) prend une forme très simple, soit $x = w_0 - y$, où w_0 est le revenu initial, supposé identique d'un individu à l'autre. À l'optimum individuel, le taux marginal de substitution est égal à $h_y = 1$, ce qui conduit à $y_2 > y_1$ lorsque $g_2 > g_1$.

FIGURE 1

ARBITRAGE INDIVIDUEL ENTRE CONSOMMATION ET SOINS DE SANTÉ EN L'ABSENCE D'ASSURANCE : SOLUTION DE LAISSER-FAIRE



Introduisons à présent un système d'assurance. Dans la situation la plus réaliste, l'autorité publique n'a aucune possibilité d'observer la gravité de la maladie, g. L'assurance consiste alors en un couple [π, r (y)], où π est la prime d'assurance et

$r(y)$ la fonction de remboursement. L'individu maximise donc la fonction d'utilité suivante, conditionnellement à g – il s'agit bien d'une maximisation *ex post* – :

$$\underset{y}{\text{Max}}\, u(w_0 - \pi - y + r(y) + h(y; g)) \quad \forall g \in [0, \bar{g}],$$

ce qui nous permet d'obtenir une fonction de demande en soins de santé $y(g)$ qui dépend bien entendu du contrat d'assurance $[\pi, r(y)]$.

Il faut noter que si l'assureur pouvait observer le g de chacun des assurés (information parfaite), la solution de premier rang serait atteinte en faisant payer à ceux-ci une prime liée à g, $\pi(g)$, sans subventionner les dépenses de santé, et en les laissant choisir leur dépense y. Celle-ci – que nous dénoterons par $y^*(g)$ – satisferait la condition : $TMS = h_y = 1$ (et serait donc identique dans notre modèle à celle qui prévaudrait dans la solution de laisser-faire). Les primes seraient ajustées de manière à faire partager entièrement leurs risques par les individus (c'est-à-dire éliminer tout risque individuel). Ceci revient à égaliser le terme $(x + h(y^*(g)))$ quel que soit le degré de gravité de la maladie des individus (y compris l'état de bonne santé). En fait, puisque le risque porte à la fois sur x et h, un individu malade se voit compenser forfaitairement à la fois pour ses dépenses de santé et pour la baisse de son niveau de santé (du fait que $h(\infty; g) < h_0$, $\forall g > 0$)[3]. Le niveau relatif des primes doit donc satisfaire la condition suivante :

$$\pi(g_2) - \pi(g_1) = [h(y^*(g_2); g_2) - y^*(g_2)] - [h(y^*(g_1); g_1) - y^*(g_1)],$$

$$\forall g_1, g_2 \in [0, \bar{g}],$$

leur niveau absolu étant ajusté de manière à couvrir les indemnités totales par les primes.

En revanche, dans le cas où l'assureur est dans l'incapacité d'observer g mais doit se contenter d'observer le niveau des dépenses y, le remboursement doit être lié aux dépenses en soins curatifs, qui jouent en quelque sorte le rôle d'indicateur de la gravité de la maladie de l'individu. Ce schéma fait inévitablement apparaître le problème de l'aléa (ou « risque ») moral; l'assureur doit donc réaliser un arbitrage entre le partage des risques et l'ampleur de cet aléa moral.

2. REMBOURSEMENT LINÉAIRE À TAUX UNIFORME

Le plus simple en pratique est probablement de rembourser une proportion constante des dépenses de santé ou, en d'autres termes, d'imposer un taux de remboursement uniforme quelle que soit la gravité de la maladie. Dans ce cas, la fonction de remboursement prend la forme suivante : $r(y) = \tau y$, laissant à charge du patient le montant $(1 - \tau) y$, soit ce que nous appelons communément le « ticket modérateur ». L'individu observe g et choisit y. Le problème d'optimisation individuel, conditionnellement à chaque valeur de g, est donc :

3. Voir à ce sujet Arrow (1963).

$$\underset{y}{\text{Max}}\, u\,(w_0 - \pi - (1-\tau)\,y + h\,(y;\, g)), \tag{5}$$

qui conduit à la condition de premier ordre : $h_y = 1 - \tau$, ce qui implique que y est une fonction de τ, pour un g donné : $y\,(\tau;\, g)$[4]. La statique comparative permet d'obtenir : $\dfrac{\partial y}{\partial \tau} = -h_{yy}^{-1} > 0$; ce terme est notre indicateur d'aléa moral, puisqu'il représente la sensibilité des dépenses à une modification du taux de remboursement. Nous avons également : $\dfrac{\partial y}{\partial g} = -h_{yg} h_{yy}^{-1} > 0$.

Quant à l'assureur, il choisit le paramètre τ en solutionnant le problème suivant. Il maximise l'espérance d'utilité de l'individu :

$$\underset{\tau}{\text{Max}} \int_0^{\bar{g}} u\,(w_0 - \pi - (1-\tau)\,y + h\,(y;\, g))\, dF\,(g), \tag{6}$$

avec $\pi = \tau \int_0^{\bar{g}} y\,(\tau;\, g)\, dF\,(g)$, en tenant compte du fait que le choix de τ influence les décisions individuelles de soins curatifs : $y\,(\tau;\, g)$, comme nous venons de le voir.

Après substitution de l'expression de π dans la fonction objectif et utilisation du théorème de l'enveloppe, nous obtenons la condition du premier ordre suivante :

$$\left[\tau E \frac{\partial y}{\partial \tau} + Ey\right] Eu' = E(yu') \tag{7}$$

où E est l'opérateur « espérance mathématique » par rapport à la variable aléatoire g. Cette condition peut aisément être interprétée. Lorsque le taux de remboursement augmente, la prime augmente en proportion de l'expression entre crochets (à gauche) faisant intervenir l'indicateur d'aléa moral, et par franc de prime en supplément, l'utilité espérée diminue de Eu'; cependant, en contrepartie le montant déboursé par l'assuré conditionnellement à chaque gravité de maladie diminue proportionnellement à $y\,(\tau;\, g)$, ce qui provoque un effet-revenu bénéfique valorisé à l'utilité marginale du revenu u' contingente à ce degré de gravité (à droite). Suivant (7), le taux de remboursement est augmenté jusqu'au niveau où cette perte et ce gain d'utilité espérée s'équilibrent à la marge.

Cette condition nous donne l'expression suivante du taux de remboursement optimal[5] :

$$\tau^* = \frac{E\left(y \dfrac{u'}{Eu'}\right) - Ey}{E \dfrac{\partial y}{\partial \tau}} = \frac{\text{cov}\left(y, \dfrac{u'}{Eu'}\right)}{E \dfrac{\partial y}{\partial \tau}}. \tag{8}$$

4. S'il n'y avait pas de ticket modérateur, on aurait $\tau = 1$, et donc la condition $h_y = 0$.

5. Notons que l'apparition de Eu' dans le terme de covariance permet que celle-ci ne dépende pas de la cardinalisation de u.

Dans ce cas très simple, le taux optimal de remboursement dépend de l'arbitrage entre deux préoccupations. D'une part, le phénomène d'aléa moral est pris en compte au dénominateur, où $E\frac{\partial y}{\partial \tau}$ reflète la sensibilité de y au taux de remboursement, et donc la préoccupation d'efficacité. D'autre part, le numérateur est la covariance entre u' et y et reflète l'objectif de partage des risques[6]. Comme y et u' augmentent avec $g\left(cf.\ \frac{\partial y}{\partial g} > 0\right)$, la covariance $\text{cov}\left(y, \frac{u'}{Eu'}\right)$ est positive et elle augmente avec l'aversion au risque de l'individu. On retrouve donc le résultat bien connu selon lequel le taux optimal de remboursement augmente avec l'aversion au risque mais diminue avec la sensibilité des dépenses de santé au taux en question[7]. Par conséquent, l'objectif de partage des risques conduit à un taux de remboursement positif, ce qui implique une surconsommation de soins curatifs par rapport à la solution de premier rang. En résumé, il y a donc, lors du choix du taux de remboursement, arbitrage entre un meilleur partage des risques et une surconsommation à la marge par rapport à la solution de premier rang (*cf.* Besley, 1988).

Ces résultats sont illustrés dans le tableau 1, qui est basé sur la spécification suivante (également utilisée dans les autres exemples numériques de cet article) :

$$h(y; g) = 0{,}45 \,.\, 125 \qquad \text{pour } g = 0,$$

$$= 0{,}45 \,.\, (125 - g) + 1{,}2 \,.\, (y\varepsilon(g) - g)^{0{,}25} \qquad \text{pour } g > 0,$$

$$u(x + h) = (1 - \rho)^{-1} (x + h)^{1-\rho}$$

et $w_0 = 75.$

TABLEAU 1

DÉTERMINATION DU TAUX OPTIMAL UNIQUE τ^*
VARIATIONS CROISSANTES DE L'AVERSION AU RISQUE (ρ) ET DE L'ALÉA MORAL (k)

τ^*	$\rho = 0{,}60$	$\rho = \mathbf{0{,}75}$	$\rho = 0{,}90$
$k = 0{,}75$	0,8215	0,8340	0,8436
$\boldsymbol{k = 1{,}00}$	0,7776	**0,7929**	0,8047
$k = 1{,}25$	0,7399	0,7575	0,7712

6. Concernant la liaison du taux optimal à la covariance, voir Sheshinski (1972).

7. Bien entendu, si l'élasticité de la demande était nulle, on pourrait appliquer un taux de remboursement de 100 %.

La sensibilité des dépenses de soins curatifs au taux de remboursement (et donc l'aléa moral) augmente avec le paramètre ε (g) qui apparaît dans la spécification de h (y; g) pour $g > 0$[8]. D'autre part, l'aversion relative au risque des individus est constante et égale à ρ. En plus de l'état de bonne santé ($g_0 = 0$), nous distinguons deux états de maladie caractérisés par $g_1 = 2$ et $g_2 = 30$, avec les probabilités $p_0 = 1/6$, $p_1 = 3/6$ et $p_2 = 2/6$, associées respectivement aux états g_0, g_1 et g_2. Par ailleurs, nous spécifions $\varepsilon(g_1) = \varepsilon_1 = 0{,}25\,k$ et $\varepsilon(g_2) = \varepsilon_2 = 0{,}35\,k$ où k est un paramètre ajustable (voir tableau 1).

Le tableau 1 met bien en évidence que le taux optimal de remboursement s'accroît avec l'aversion au risque des assurés (ρ croissant) alors qu'il décroît avec la sensibilité de leur demande au prix qui reste à leur charge (k croissant).

3. REMBOURSEMENT LINÉAIRE PAR MORCEAUX : LE CAS DE DEUX TAUX MARGINAUX DE REMBOURSEMENT

Dans cette section notre objectif est de montrer qu'il est possible d'améliorer le résultat obtenu ci-dessus en distinguant deux taux marginaux de remboursement : le premier taux marginal (τ_1) s'applique aux dépenses en deçà d'un seuil (dénommé s), et le second (τ_2) aux dépenses au-delà de ce seuil. Autrement dit, la fonction de remboursement devient :

$$r(y) = \tau_1 y \qquad \text{si} \quad y \leq s,$$

$$= \tau_1 s + \tau_2 (y - s) \qquad \text{si} \quad y > s.$$

La question est de déterminer le seuil de dépenses s, le sens de la différenciation des taux marginaux ($\tau_1 \gtrless \tau_2$) ainsi que leurs valeurs optimales. Dès lors, deux situations sont à distinguer puisque l'optimum peut être tel que $\tau_2 \geq \tau_1$ (cas a) ou $\tau_2 \leq \tau_1$ (cas b)[9].

Au premier abord, l'intuition nous amène à pencher en faveur de taux croissants (cas a). En effet, quand la gravité de la maladie augmente, le montant des dépenses est plus élevé, ce qui pour l'individu le plus malade signifie un moindre revenu disponible à allouer à d'autres consommations. En termes de partage des risques, nous sommes donc conduits à recommander une croissance du taux de remboursement, afin d'atténuer l'effet-revenu que nous venons de décrire. Cependant, la préoccupation d'efficacité nous commande de prendre en compte le

8. Dans notre exemple numérique, c'est bien l'augmentation de ε avec g qui permet de poser l'hypothèse d'une sensibilité des dépenses croissante avec g. En effet, nous avons $y_{\tau g} = 0$ mais $y_{\tau\varepsilon} > 0$.

9. Dans le contexte de l'imposition sur les revenus du travail, l'article de Slemrod *et al.* (1994) montre qu'accroître la complexité du schéma fiscal – d'une manière similaire à celle que nous utilisons – permet d'augmenter le bien-être des deux classes d'individus (deux catégories de productivité). De plus, dépassant l'erreur commise par Sheshinski (1989), il suggère un taux marginal d'imposition du revenu inférieur dans la deuxième partie du schéma, soit pour la catégorie dont les revenus sont plus élevés.

phénomène d'aléa moral. Si la sensibilité des dépenses au taux de remboursement (telle que définie au dénominateur de l'expression 8) augmente avec la gravité de la maladie, cette préoccupation peut même conduire à inverser notre conclusion et donc à choisir le schéma de remboursement proposé par le cas *b*. Dans les développements qui suivent, nous verrons que même en supposant que la sensibilité des dépenses augmente avec la gravité de la maladie, il existe des situations où la préoccupation d'assurance domine le problème d'aléa moral, impliquant le choix du cas *a*, alors que le cas *b* s'impose lorsque l'aléa moral domine l'objectif de partage des risques. Nous pouvons escompter qu'*a fortiori*, si la sensibilité au prix diminue avec la gravité de la maladie, les taux seront choisis croissants.

Considérons d'abord la première situation, soit un taux de remboursement plus élevé pour les dépenses dépassant s : $\tau_2 \geq \tau_1$ (cas *a*). La situation est décrite par la figure 2. La contrainte budgétaire est convexe et composée de deux segments de droite, successivement d'inclinaison $(1 - \tau_1)$ et $(1 - \tau_2)$ en valeur absolue; elle présente un coude à la verticale de $y = s$. Nous avons représenté sur cette figure le choix optimal de dépenses de santé pour trois individus caractérisés par des gravités différentes de leur maladie. L'individu dont la courbe d'indifférence est dénotée par u_1 choisit un niveau de dépenses y_1 inférieur au seuil s (régime 1 de dépenses); par contre, celui caractérisé par la courbe d'indifférence u_2 – ayant une maladie plus grave $(g_2 > g_1)$ – choisit un niveau de dépenses y_2 supérieur au seuil s (régime 2). La troisième courbe d'indifférence, dénotée par $\hat{u}$, a deux points de tangence avec la contrainte budgétaire. Soit $\hat{g}$ la gravité de la maladie de l'individu caractérisé par cette courbe d'indifférence. Cet individu est indifférent entre les deux niveaux de dépenses $\hat{y}_1$ et $\hat{y}_2$ respectivement dans les régimes 1 et 2. Au-delà du degré critique $\hat{g}$, il est donc optimal pour le patient d'augmenter ses dépenses de manière à dépasser s et à bénéficier ainsi du régime 2. En résumé, les individus dont le g est en deçà de $\hat{g}$ ont leur point de tangence à gauche de s (régime 1), tandis que ceux dont le g dépasse $\hat{g}$ se situent à droite du seuil (régime 2)[10]. Comme nous le montrons ci-après, ce degré de gravité $\hat{g}$ dépend bien entendu de τ_1, τ_2 et s : $\hat{g} = \hat{g}(\tau_1, \tau_2, s)$.

En termes formels, les dépenses en soins curatifs satisfont les relations suivantes :

$$h_y(y; g) = 1 - \tau_1 \quad \text{si } g \leq \hat{g},$$
$$= 1 - \tau_2 \quad \text{si } g \geq \hat{g},$$

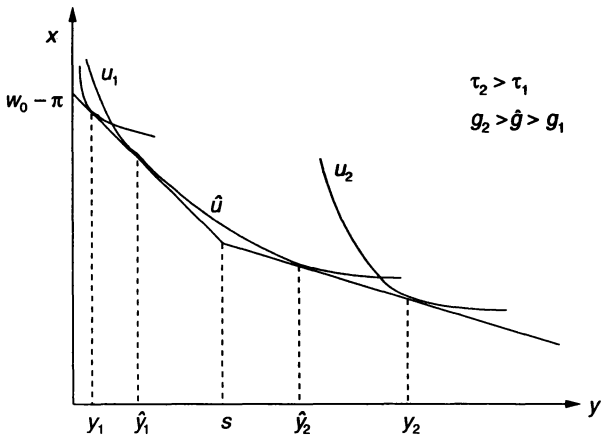
ce qui fournit les fonctions de demande

$$y = y(\tau_1; g) \quad \text{si } g \leq \hat{g},$$
$$= y(\tau_2; g) \quad \text{si } g \geq \hat{g}.$$

10. Remarquez que Keeler *et al.* (1977) décrivent une situation du même type, avec un seuil-franchise en deçà duquel le taux de remboursement est nul. Dans ce cas, ils montrent qu'aucun individu ne se positionnera à proximité du montant-seuil.

FIGURE 2

REMBOURSEMENT LINÉAIRE À DEUX TAUX : CAS *a*



On peut dès lors caractériser la condition d'indifférence qui détermine le degré de gravité $\hat{g}$:

$$h\,(y\,(\tau_1;\,\hat{g});\,\hat{g}) - (1-\tau_1)\,y\,(\tau_1;\,\hat{g}) = h\,(y\,(\tau_2;\,\hat{g});\,\hat{g}) - (1-\tau_1)s - (1-\tau_2)\,(y\,(\tau_2;\,\hat{g}) - s), \tag{9}$$

ce qui définit $\hat{g}$ comme une fonction des paramètres de remboursement : $\hat{g} = \hat{g}\,(\underset{-}{\tau_1}, \underset{-}{\tau_2}, \underset{+}{s})$, où le signe figurant sous chaque argument indique le sens de la variation de $\hat{g}$ lorsque cet argument augmente.

Ces signes sont issus de la différenciation totale de l'expression (9), compte tenu des conditions de premier ordre dérivées ci-dessus[11]. Il semble évident qu'une hausse de τ_2 pousse les individus à choisir plus aisément le deuxième régime, ce qui se traduit par une baisse de $\hat{g}$. Tout aussi clairement, une augmentation du seuil s produit l'effet inverse. Il est peut-être moins intuitif qu'une hausse de τ_1 induise une baisse de $\hat{g}$, ce qui signifie une plus grande proportion

11. En effectuant une différenciation totale de l'expression (9) et en utilisant le théorème de l'enveloppe, on obtient aisément :

$$\left.\begin{aligned} &\frac{d\hat{g}}{d\tau_1} = c(s-\hat{y}_1) < 0, \\ &\frac{d\hat{g}}{d\tau_2} = c(\hat{y}_2 - s) < 0 \\ \text{et}\quad &\frac{d\hat{g}}{ds} = c(\tau_1 - \tau_2) > 0 \end{aligned}\right\} \text{avec } c = \left(\frac{\partial h(\hat{y}_1;\hat{g})}{\partial \hat{g}} - \frac{\partial h(\hat{y}_2;\hat{g})}{\partial \hat{g}}\right)^{-1} < 0.$$

d'agents dans le régime 2. Cependant, il suffit de voir sur la figure 2 qu'une telle variation réduit la pente de la contrainte dans son premier segment; celui-ci pivote autour de son point d'intersection avec l'axe vertical (d'ordonnée $w_0 - \pi$), amenant un certain nombre d'individus auparavant dans le régime 1 à lui préférer le régime 2. En effet, une hausse du taux marginal de remboursement en régime 1 pousse vers le haut la consommation médicale, d'où un passage dans le régime 2 pour certains assurés.

Dès lors, l'assureur doit tenir compte de tous ces éléments lorsqu'il choisit les paramètres du système optimal d'assurance-santé, soit (τ_1, τ_2, s, π). Le problème d'optimisation est le suivant :

$$\begin{aligned}&\text{Max} \int_0^{\hat{g}(\cdot)} u\left(w_0 - \pi - (1-\tau_1)\, y\,(\tau_1; g) + h\left(y\,(\tau_1; g); g\right)\right) dF\,(g)\\ &+ \int_{\hat{g}(\cdot)}^{\bar{g}} u\left(w_0 - \pi - (1-\tau_1)s - (1-\tau_2)\left(y\,(\tau_2; g) - s\right)\right.\\ &\left. + h\left(y\,(\tau_2; g); g\right)\right) dF\,(g),\end{aligned} \tag{10}$$

$$\text{avec } \pi = \tau_1 \int_0^{\hat{g}(\cdot)} y\,(\tau_1; g)\, dF\,(g) + \int_{\hat{g}(\cdot)}^{\bar{g}} \left(\tau_1 s + \tau_2\left(y\,(\tau_2; g) - s\right)\right) dF\,(g) \tag{11}$$

et $\hat{g}\,(.) = \hat{g}\,(\tau_1, \tau_2, s)$.

En égalant à 0 la dérivée de (10) par rapport à τ_1, une condition similaire à (7) mais légèrement plus compliquée est obtenue :

$$\begin{aligned}&\left[F(\hat{g})\left(\tau_1 \underset{g\le\hat{g}}{E}\frac{\partial y}{\partial \tau_1} + \underset{g\le\hat{g}}{E} y\right) + (1-F(\hat{g}))s - f(\hat{g})\Delta R(\hat{g})\frac{d\hat{g}}{d\tau_1}\right] Eu'\\ &= F(\hat{g})\underset{g\le\hat{g}}{E}(yu') + (1-F(\hat{g}))\underset{g\ge\hat{g}}{E}(su')\end{aligned} \tag{12}$$

où

$$\Delta R\,(\hat{g}) = \tau_1\,(s - \hat{y}_1) + \tau_2\,(\hat{y}_2 - s) > 0 \tag{13}$$

représente l'augmentation de remboursement que l'assureur doit supporter lorsqu'un assuré ayant le degré de gravité de maladie $\hat{g}$ passe du régime 1 au régime 2. Cette modification accroît la prime que doivent payer l'ensemble des assurés, ce qui est pris en compte par le terme incluant $\Delta R(\hat{g})$ figurant dans la condition (12). Excepté ce nouveau terme, la condition (12) fait apparaître le même genre de termes que la condition (7). Il faut cependant garder à l'esprit que les effets-revenus d'une augmentation de τ_1 sont proportionnels à y et s respectivement pour les individus caractérisés par $g \le \hat{g}$ et $g \ge \hat{g}$. Il en est de même pour ses effets sur la prime.

La condition (12) conduit aisément à la formule suivante pour τ_1 :

$$\tau_1 = \frac{1}{\underset{g\le\hat{g}}{E}\dfrac{\partial y}{\partial \tau_1}}\left[\underset{g\le\hat{g}}{\operatorname{cov}}\left(y, \frac{u'}{Eu'}\right) + (1-F(\hat{g}))\frac{\underset{g\ge\hat{g}}{E}u' - \underset{g\le\hat{g}}{E}u'}{Eu'}\left(s - \underset{g\le\hat{g}}{E}y\right)\right.$$

$$\left. + F(\hat{g})^{-1} f(\hat{g}) \Delta R(\hat{g}) \frac{d\hat{g}}{d\tau_1}\right], \tag{14}$$

qu'il faut comparer à l'expression (8). À nouveau, les divers termes reflètent les préoccupations de partage des risques (au numérateur) et d'aléa moral (au dénominateur). Dans l'expression entre crochets figurant dans cette formule (numérateur), le premier terme est la composante intra-groupe de la covariance et est semblable à celui rencontré précédemment dans l'expression (8), sinon qu'il ne prend en compte que les individus dans le premier régime; tandis que le deuxième terme y représente la composante inter-groupes de la covariance (où les groupes sont ici définis suivant que g est inférieur ou supérieur à $\hat{g}$). Quant au troisième terme de l'expression entre crochets, il traduit l'effet sur la prime du déplacement de certains assurés du régime 1 vers le régime 2, comme nous venons de l'expliquer ci-dessus.

Une condition similaire pour τ_2 peut être interprétée de la même manière. En effet, nous pouvons obtenir :

$$\tau_2 = \frac{1}{\underset{g\ge\hat{g}}{E}\dfrac{\partial y}{\partial \tau_2}}\left[\underset{g\ge\hat{g}}{\operatorname{cov}}\left(y, \frac{u'}{Eu'}\right) + F(\hat{g})\frac{\underset{g\ge\hat{g}}{E}u' - \underset{g\le\hat{g}}{E}u'}{Eu'}\underset{g\ge\hat{g}}{E}(y-s)\right.$$

$$\left. + (1-F(\hat{g}))^{-1} f(\hat{g}) \Delta R(\hat{g}) \frac{d\hat{g}}{d\tau_2}\right]. \tag{15}$$

Lorsque $\tau_1 < \tau_2$, c'est donc qu'en passant d'un régime à l'autre, la variation du terme de covariance intra-groupe domine celle des autres effets, en particulier de l'aléa moral. En effet, nous avons

$$\underset{g\ge\hat{g}}{\operatorname{cov}}\left(y, \frac{u'}{Eu'}\right) > \underset{g\le\hat{g}}{\operatorname{cov}}\left(y, \frac{u'}{Eu'}\right),$$

puisque le montant des dépenses ainsi que l'utilité marginale sont toujours plus élevés dans le régime 2 que dans le régime 1; les deux autres termes entres crochets peuvent être considérés comme étant d'amplitude similaire dans chacune des expressions (14) et (15). Pour que la préoccupation de partage des risques impose ainsi un taux marginal de remboursement croissant avec y, il faut qu'au dénominateur la sensibilité des dépenses à ce taux (aléa moral) diminue – ou augmente de manière limitée – avec y et donc avec g.

Enfin, la condition sur le seuil s donne :

$$(1-F(\hat{g}))\,(\tau_1-\tau_2)\left(\underset{g\geq\hat{g}}{E}\,u'-Eu'\right)+Eu'f(\hat{g})\Delta R(\hat{g})\frac{d\hat{g}}{ds}=0, \qquad (16)$$

que l'on peut interpréter au moyen des considérations suivantes. Lorsqu'on augmente s, les assurés dans le régime 2 ($g \geq \hat{g}$) subissent une baisse de leur remboursement proportionnelle à ($\tau_2 - \tau_1$) – suite à l'application de τ_2 à un montant moindre d'unités de y –, ce qui leur est défavorable. En contrepartie, en bénéficient l'ensemble des assurés sous forme d'une réduction de leur prime. Par ailleurs, l'augmentation de s provoque un déplacement de certains assurés du régime 2 au régime 1, ce qui réduit à nouveau la prime d'assurance.

L'exemple numérique permet de présenter au tableau 2 les taux obtenus en fonction de différentes valeurs de l'aversion au risque et de l'aléa moral[12].

TABLEAU 2

DÉTERMINATION DES TAUX DIFFÉRENCIÉS
CAS a : $\tau_1 < \tau_2$
VARIATIONS CROISSANTES DE ρ ET DE k
$\varepsilon_1 = 0{,}25\,k$ ET $\varepsilon_2 = 0{,}35\,k$

	τ^*, τ_1, τ_2 $s = y_1$	$\rho = 0{,}6$	$\rho = \mathbf{0{,}75}$	$\rho = 0{,}9$
$k = 0{,}75$	τ^*	0,8215	0,8340	0,8436
	τ_1	0,4815	0,5077	0,5287
	τ_2	0,8644	0,8743	0,8819
	s	10,9426	10,9624	10,9801
$\mathbf{k = 1{,}00}$	τ^*	0,7776	**0,7929**	0,8047
	τ_1	0,3916	**0,4193**	0,4418
	τ_2	0,8305	**0,8427**	0,8522
	s	8,2454	**8,2611**	8,2753
$k = 1{,}25$	τ^*	0,7399	0,7575	0,7712
	τ_1	0,3246	0,3522	0,3749
	τ_2	0,8011	0,8153	0,8264
	s	6,6300	6,6431	6,6550

12. Pour rappel, nous avons intégré à ce tableau les valeurs correspondantes du taux uniforme optimal contenues dans le tableau 1. Par ailleurs, avec deux états de maladie, $s = y_1$ à l'optimum. En effet, la dérivée de l'espérance d'utilité par rapport au seuil s'écrit désormais :

$(1 - F(\hat{g}))\,(\tau_1 - \tau_2)\left(\underset{g\geq\hat{g}}{E}\,u'-Eu'\right)$, soit une expression négative. L'espérance d'utilité augmente lorsqu'on diminue le seuil.

Ce tableau nous permet d'observer que les résultats correspondent bien à notre intuition, comme c'était déjà le cas pour les variations du taux de remboursement unique (*cf.* tableau 1). En effet, lorsque l'aversion au risque (ρ) augmente, les deux taux augmentent, tandis qu'ils diminuent lorsque l'importance de l'aléa moral s'amenuise, induisant par là-même une baisse des dépenses médicales et, donc, une baisse du seuil s (ici égal à y_1). De plus, les deux taux se fixent désormais de part et d'autre de l'ancien taux uniforme, et ce pour chaque cas envisagé. Ceci correspond bien à ce que l'on pouvait escompter. Bien entendu, nous avons vérifié que l'espérance d'utilité est plus élevée lorsqu'on utilise le schéma de remboursement à deux taux et ce pour toutes les valeurs de ρ et de k envisagées. Dans les exemples numériques répertoriés dans le tableau 1, les valeurs de ε_1 et ε_2 sont telles que l'optimisation des taux conduit à $\tau_2 > \tau_1$ (cas *a*). Avec d'autres valeurs, on obtient des taux optimaux tels que $\tau_1 > \tau_2$ (cas *b*) comme nous le verrons dans le tableau 3.

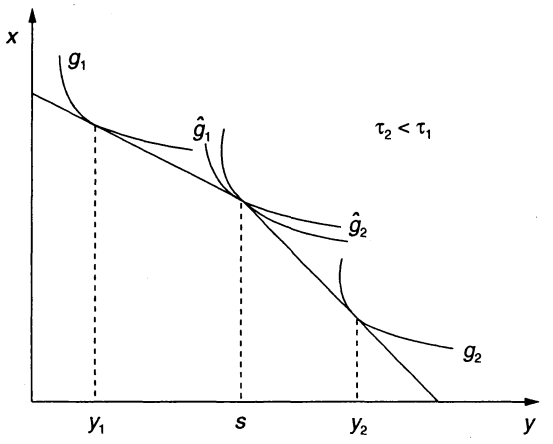
Considérons donc à présent le cas de taux décroissants : $\tau_2 < \tau_1$ (cas *b*), soit un taux de remboursement moins élevé pour les dépenses au-delà du seuil s fixé. En d'autres termes, la contrainte budgétaire devient concave et il n'y a plus un degré de gravité correspondant à un individu qui serait indifférent entre les deux régimes. Par contre, un troisième régime apparaît au niveau du coude, dans la mesure où certains individus caractérisés par des g de valeur intermédiaire trouvent plus intéressant de situer leur dépense en ce point s que dans les régimes 1 ou 2, comme on peut le voir sur la figure 3. Nous pouvons maintenant distinguer trois régimes :

$$\begin{aligned} &\text{régime } 1 : g < \hat{g}_1 && \text{avec } y < s, \\ &\qquad s : \hat{g}_1 \leq g \leq \hat{g}_2 && \text{avec } y = s, \\ &\qquad 2 : \hat{g}_2 < g, && \text{avec } y > s \end{aligned}$$

et les probabilités qui leur sont associées sont respectivement pour les régimes 1 et 2 : $F(\hat{g}_1)$ et $1 - F(\hat{g}_2)$, avec pour le régime s : $F(\hat{g}_2) - F(\hat{g}_1)$.

FIGURE 3

TAUX DÉCROISSANTS – CONTRAINTE CONCAVE



Les limites du régime s sont définies par les deux conditions de premier ordre suivantes :

$$h_y(s, \hat{g}_1) = 1 - \tau_1$$

et $h_y(s, \hat{g}_2) = 1 - \tau_2$,

qui expriment que la courbe d'indifférence de l'assuré ayant la gravité de maladie $\hat{g}_1$ (respectivement $\hat{g}_2$) est tangente en $y = s$ au segment de gauche (respectivement de droite) de la contrainte budgétaire. Une différenciation totale nous donne aisément : $\frac{d\hat{g}_1}{d\tau_1} < 0$, $\frac{d\hat{g}_1}{ds} > 0$, $\frac{d\hat{g}_2}{d\tau_2} < 0$ et $\frac{d\hat{g}_2}{ds} > 0$.

Le problème d'optimisation de l'assureur est similaire au problème décrit pour le cas précédent, sinon qu'il faut distinguer cette fois trois régimes. Par exemple, la prime se définit comme :

$$\pi = \tau_1 \int_0^{\hat{g}_1} y(\tau_1; \hat{g}_1)\, dF(g) + \tau_1 \int_{\hat{g}_1}^{\hat{g}_2} s\, dF(g) + \int_{\hat{g}_2}^{\bar{g}} [\tau_1 s + \tau_2 (y(\tau_2; \hat{g}_2) - s)]\, dF(g). \tag{17}$$

Les conditions du premier ordre donnent les expressions suivantes :

$$\tau_1 = \frac{1}{\underset{g<\hat{g}_1}{E} \frac{\partial y}{\partial \tau_1}} \left[\underset{g<\hat{g}_1}{\operatorname{cov}} \left(y, \frac{u'}{Eu'} \right) + (1 - F(\hat{g}_1)) \frac{\underset{g\geq\hat{g}_1}{E} u' - \underset{g<\hat{g}_1}{E} u'}{Eu'} \left(s - \underset{g<\hat{g}_1}{E} y \right) \right] \tag{18}$$

et

$$\tau_2 = \frac{1}{\underset{g>\hat{g}_2}{E}\dfrac{\partial y}{\partial \tau_2}}\left[\underset{g>\hat{g}_2}{\operatorname{cov}}\left(y, \frac{u'}{Eu'}\right) + F(\hat{g}_2)\frac{\underset{g>\hat{g}_2}{E}u' - \underset{g\le\hat{g}_2}{E}u'}{Eu'}\underset{g>\hat{g}_2}{E}(y-s)\right]. \tag{19}$$

Pour que ces expressions nous fournissent des taux optimaux tels que $\tau_1 > \tau_2$, il faut que le dénominateur de τ_1 en (18) soit à ce point supérieur à celui de τ_2 en (19) que l'effet de l'aléa moral domine la préoccupation d'assurance. Notez qu'il n'y a plus ici de termes équivalents à ceux incluant $\Delta R\,(\hat{g})$ qui apparaissaient dans (14) et (15). En effet, il n'y a pas ici de saut entre les deux régimes (*cf.* passage de $\hat{y}_1$ à $\hat{y}_2$ à la figure 2), alors que c'était le cas pour $\tau_1 < \tau_2$. Quant à la condition sur s, elle nous donne :

$$\begin{aligned}&(1-F(\hat{g}_2))\,(\tau_1-\tau_2)\left(\underset{g>\hat{g}_2}{E}u' - Eu'\right)\\ &+(F(\hat{g}_2)-F(\hat{g}_1))\left[\underset{\hat{g}_1\le g\le\hat{g}_2}{E}\left((h'-(1-\tau_1))u'\right)-\tau_1 Eu'\right]=0,\end{aligned} \tag{20}$$

et peut être interprétée de la manière suivante. Le premier terme a la même signification que celui qui apparaît dans la relation (16). Cependant, puisque $(\tau_1 - \tau_2)$ est ici > 0, une augmentation du seuil s est favorable aux assurés dans le régime 2 avec pour conséquence négative une croissance de la prime pour tous les assurés. Quant à l'expression entre crochets, elle reflète la perte d'utilité des assurés dans le régime s qui est causée par l'augmentation de s ainsi que la réduction de prime qui en résulte. Suivant la relation (20), le seuil s est augmenté jusqu'au niveau où les gains et pertes d'utilité espérée se compensent à la marge.

Le tableau 3 présente pour ce cas-ci le même type de résultats que le tableau 2. Le changement de cas est obtenu par l'accroissement de l'écart entre ε_1 et ε_2. Les commentaires faits sur les résultats du tableau 2 s'appliquent également ici[13].

13. Il faut remarquer qu'avec deux états de maladie, $s = y_2$ à l'optimum. Pour le voir, reprenons l'expression (20). Elle se réduit à présent à : $(1-F(\hat{g}_2))\,(\tau_2-\tau_1)\left(\underset{g>\hat{g}_2}{E}u' - Eu'\right)$, soit une expression positive; ce qui implique qu'une augmentation du seuil augmente l'espérance d'utilité. On imposera donc effectivement un seuil $s = y_2$.

TABLEAU 3

DÉTERMINATION DES TAUX DIFFÉRENCIÉS
CAS b : $\tau_1 > \tau_2$
VARIATIONS CROISSANTES DE ρ ET DE k
$\varepsilon_1 = 0{,}20\,k$ ET $\varepsilon_2 = 0{,}40\,k$

	τ^*, τ_1, τ_2 $s = y_2$	$\rho = 0{,}60$	$\rho = \mathbf{0{,}75}$	$\rho = 0{,}90$
$k = 0{,}75$	τ^*	0,8042	0,8178	0,8284
	τ_1	0,8380	0,8496	0,8585
	τ_2	0,1024	0,1222	0,1408
	s	100,155	100,160	100,165
$\mathbf{k = 1{,}00}$	τ^*	0,7578	**0,7744**	0,7873
	τ_1	0,7992	**0,8133**	0,8243
	τ_2	0,0844	**0,1015**	0,1178
	s	75,1664	**75,1707**	75,1749
$k = 1{,}25$	τ^*	0,7182	0,7371	0,7519
	τ_1	0,7657	0,7820	0,7947
	τ_2	0,0735	0,0891	0,1040
	s	60,1765	60,1805	60,1845

4. REMBOURSEMENT NON LINÉAIRE

Dans la section précédente, une fonction de remboursement non linéaire a été introduite : une forme particulière (linéaire par morceaux) lui a été imposée. Cette restriction peut être levée en supposant que la fonction de remboursement $r(y)$ peut prendre une forme quelconque. Notre objectif dans cette section est de mettre en évidence les propriétés que satisfait une fonction de remboursement non linéaire optimale. Bien entendu, l'utilité espérée de l'assuré s'en trouvera améliorée puisque les fonctions de remboursement considérées dans les deux sections précédentes sont des cas particuliers. Nous nous limiterons au cas d'un nombre fini d'états de maladie possibles où, à l'instar de ce qui est supposé dans les exemples numériques développés dans cet article, l'assuré peut se trouver dans trois états : l'état de bonne santé ($g_0 = 0$) et deux états de maladie de gravité différente (où $g_2 > g_1$).

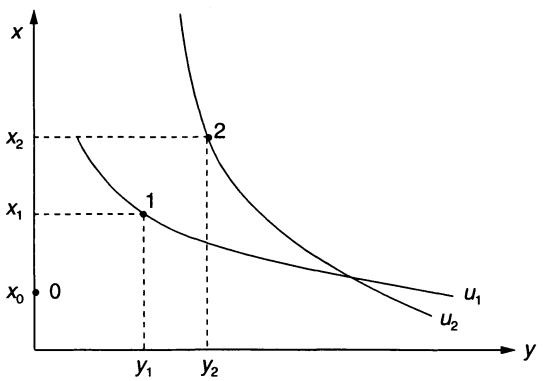
Dans la procédure à utiliser pour trouver une fonction de remboursement optimale, deux étapes peuvent être distinguées. *Dans la première étape*, trois paires $\{(y_i, x_i), i = 0, 1, 2\}$ – ou de manière équivalente trois points dans l'espace (y, x) – sont recherchées : chacune de ces paires définit le panier de soins curatifs et de bien composite qui sera consommé par l'assuré dans un état spécifique. Ces trois paires doivent satisfaire deux types de contraintes : il s'agit, d'une part, d'une

contrainte de ressources (ou de faisabilité) et, d'autre part, de contraintes d'incitation (ou d'autosélection). Ces dernières assurent que si le choix entre les trois paires proposées est laissé à l'assuré, il préférera celle qui a été prévue pour l'état de santé où il se trouve. Si ces contraintes d'incitation sont satisfaites il est possible, *dans la seconde étape de la procédure*, de trouver un contrat $(\pi, r(y))$ qui décentralise le choix des soins curatifs par les assurés eux-mêmes; pour qu'il en soit ainsi, ce contrat doit, en particulier, satisfaire les conditions suivantes : $x_i = w_0 + r(y_i) - y_i - \pi$ $(i = 0, 1, 2)$. C'est cette procédure en deux étapes que nous décrivons maintenant en détail.

Tout d'abord, il est important de remarquer que dans le cas présent, la solution de premier rang (esquissée précédemment) ne respecte certainement pas les contraintes d'incitation. En effet, puisque la solution de premier rang est caractérisée par $h'_i(y^*_i) = 1$ $(i = 1, 2)$, les hypothèses (2) à (4) impliquent que $y^*_2 > y^*_1 > y^*_0 = 0$ et $h_2(y^*_2) < h_1(y^*_1) < h_0$, avec la notation $h_i(y) \equiv h(y; g_i)$. Dès lors, avec $u(x + h)$ comme fonction d'utilité, le partage intégral des risques dans la solution de premier rang signifie que : $x^*_2 > x^*_1 > x^*_0$. C'est ce qui est représenté à la figure 4 : les transferts effectués pour compenser les pertes des individus de types 1 et 2 se traduisent par une incitation générale à choisir le contrat de type 2. Il faut donc définir des contraintes d'incitation telles que l'individu de type 0 choisisse le contrat 0 qui lui est destiné plutôt que les autres contrats et l'individu de type 1 le contrat 1 plutôt que le contrat 2.

FIGURE 4

SOLUTION DE PREMIER RANG

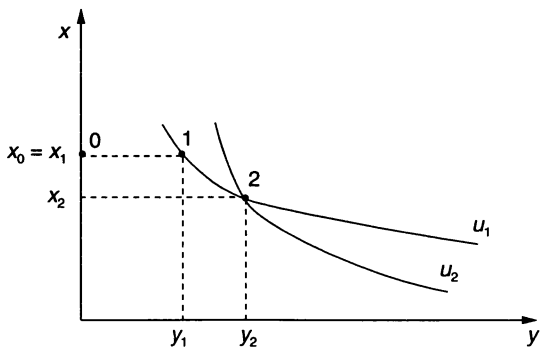


Pour que l'individu de type 0 (en bonne santé) préfère la paire (y_0, x_0) à la paire (y_1, x_1), il faut que $u(x_0 + h_0) \geq u(x_1 + h_0)$. Cette inégalité est équivalente à $x_0 \geq x_1$, ce qui définit la première contrainte d'incitation à prendre en compte. Par ailleurs, l'individu de type 1 doit préférer la paire (y_1, x_1) à la paire (y_2, x_2) : $x_1 + h_1(y_1) \geq x_2 + h_1(y_2)$, ce qui fournit la seconde contrainte d'incitation. En conséquence, les trois paires $\{(y_i, x_i), i = 0, 1, 2\}$ se présentent comme il est

indiqué à la figure 5 où les deux contraintes d'incitation ci-dessus sont satisfaites avec égalité (remarquez que les courbes d'indifférence de l'individu de type 0 y sont des droites horizontales). Il faut cependant noter que suivant cette figure, la paire (y_2, x_2) se trouve à droite de la paire (y_1, x_1) sur la courbe d'indifférence marquée u_1 de l'individu de type 1. C'est indispensable car une troisième contrainte d'incitation – à savoir, $x_0 \geq x_2$ – doit être satisfaite pour que l'individu de type 0 préfère la paire (y_0, x_0) à la paire (y_2, x_2). En comparant les positions relatives des trois paires à la figure 5 avec celles de la figure 4 (solution de premier rang), on peut mesurer l'effet de l'introduction des contraintes d'incitation.

FIGURE 5

REMBOURSEMENT NON LINÉAIRE



Pour trouver les valeurs optimales des paires $\{(y_i, x_i), i = 0, 1, 2\}$ nous devons maximiser l'espérance d'utilité de l'assuré sous les contraintes de ressources et d'incitation. Comme les deux premières contraintes d'incitation formulées ci-dessus devront être satisfaites avec égalité à l'optimum (en particulier, $x_0 = x_1$) et que $y_0 = 0$, le problème d'optimisation peut être formulé de la manière suivante :

$$\underset{x_1, y_1, x_2, y_2}{\text{Max}} \; p_0 u(x_1 + h_0) + p_1 u(x_1 + h_1) + p_2 u(x_2 + h_2), \tag{21}$$

sous les contraintes

$$w_0 - p_0 x_1 - p_1 (x_1 + y_1) - p_2 (x_2 + y_2) = 0 \tag{22}$$

et

$$x_1 + h_1 (y_1) - x_2 - h_1 (y_2) = 0, \tag{23}$$

avec p_0, p_1 et p_2 les probabilités de chaque état de santé. Dans cette formulation, nous avons omis d'inclure la contrainte d'incitation $x_1 (= x_0) \geq x_2$ car elle n'est pas active à l'optimum (comme la figure 5 le suggère). En associant aux contraintes (22) et (23) les multiplicateurs γ et μ, on obtient :

$$p_0 (u'_0 - \gamma) + p_1 (u'_1 - \gamma) + \mu = 0, \tag{24}$$

$$p_2 (u'_2 - \gamma) - \mu = 0, \tag{25}$$

$$p_1 (u'_1 h'_1 (y_1) - \gamma) + \mu h'_1 (y_1) = 0 \tag{26}$$

et

$$p_2 (u'_2 h'_2 (y_2) - \gamma) - \mu h'_1 (y_2) = 0 \tag{27}$$

où γ et μ sont positifs.

En additionnant les conditions (24) et (25), μ est éliminé :

$$p_0 (u'_0 - \gamma) + p_1 (u'_1 - \gamma) + p_2 (u'_2 - \gamma) = 0, \tag{28}$$

ce qui implique que $u'_0 - \gamma < 0$ car $u'_0 < u'_1 < u'_2$. Ces dernières inégalités s'expliquent par les contraintes d'incitation qui empêchent de réaliser le partage intégral des risques (qui était possible dans la solution de premier rang avec information parfaite).

Par ailleurs, si on élimine μ entre les conditions (24) et (26), le résultat est :

$$p_0 (u'_0 - \gamma) h'_1 (y_1) = p_1 \gamma (h'_1 (y_1) - 1). \tag{29}$$

Puisque $u'_0 - \gamma < 0$, cette condition implique :

$$h'_1 (y_1) < 1,$$

ce qui signifie que $y_1 > y_1^*$. Autrement dit, dans l'état de maladie g_1, il y a surconsommation de soins curatifs par rapport à la solution de premier rang (information parfaite)[14].

Si l'on s'intéresse maintenant au sort de l'individu de type 2, il suffit de soustraire la condition (25) de la condition (27) pour obtenir :

$$p_2 u'_2 (h'_2 (y_2) - 1) = \mu (h'_1 (y_2) - 1). \tag{30}$$

Nous venons de conclure que $h'_1 (y_1) < 1$, ce qui implique qu'*a fortiori* $h'_1 (y_2) < 1$ puisque $y_2 > y_1$. Le membre de droite de l'expression (30) est donc négatif, et nous en concluons également que $h'_2 (y_2) < 1$ et donc $y_2 > y_2^*$. Pour l'individu de type 1 comme pour celui de type 2, nous avons donc bien un taux marginal de remboursement ($r' (y)$) positif.

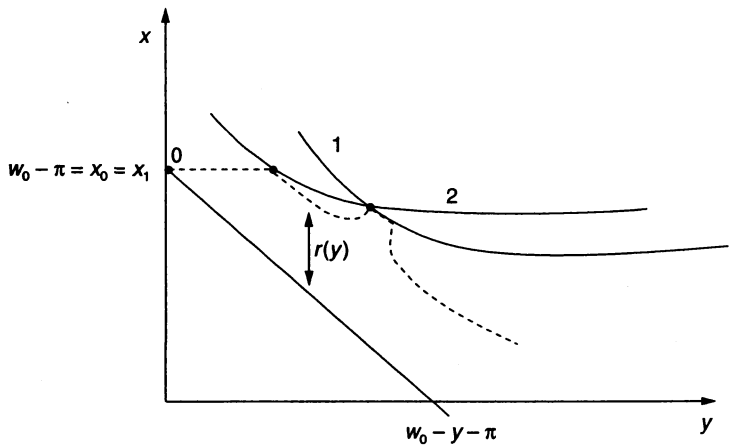
Nous avons ainsi terminé la première étape de la procédure décrite au début de cette section. La seconde étape consiste à trouver un contrat d'assurance $[\pi, r (y)]$ qui permette de mettre en oeuvre la solution optimale dérivée ci-dessus. On recherche donc dans l'espace (y, x) une courbe de budget à laquelle seront confrontés les individus de type 0, 1 et 2, telle que celle présentée en pointillés sur

14. En l'absence d'individus de type 0, le résultat $h'_1 (y_1) = 1$ aurait été obtenu (ce qui correspond dans le présent modèle au résultat classique « d'absence de distorsion au sommet »).

la figure 6 : elle doit être située en dessous des courbes d'indifférence atteintes par les différents types d'individus à l'optimum (excepté en certains points où elle se confond avec ces courbes). Cette courbe conduit bien chaque individu à choisir le point (y_i, x_i) qui lui est destiné. À partir de là, on définit $r(y)$ comme l'écart entre la courbe de budget et la droite $(w_0 - y - \pi)$. Remarquez que nous avons sur cette figure $r'(y_1) > 0$, et $r'(y_2) > 0$, soit des taux marginaux de remboursement positifs, comme nous l'avons dérivé ci-dessus.

FIGURE 6

REPRÉSENTATION GRAPHIQUE DE LA FONCTION $r(y)$



L'exemple numérique que nous avons utilisé dans les précédentes sections peut également servir ici à illustrer ce que nous venons d'examiner théoriquement. Des résultats numériques sont fournis dans la section suivante qui est consacrée à la comparaison des trois schémas de remboursement.

5. COMPARAISON DES TROIS SCHÉMAS DE REMBOURSEMENT

Dans les sections précédentes, nous avons voulu mettre en évidence comment le choix optimal des paramètres du contrat d'assurance résultait, dans chacun des trois schémas de remboursement envisagés, d'un arbitrage entre la préoccupation d'équité (partager les risques liés à la maladie) et le souci d'efficacité (endiguer l'aléa moral). Notre propos dans cette dernière section est de montrer que les préoccupations d'équité et d'efficacité sont d'autant mieux rencontrées que le schéma de remboursement est plus complexe. Dans notre modèle, il est aisé de mesurer le degré de réalisation de ces deux objectifs. En ce qui concerne l'équité, il suffit d'observer la dispersion des valeurs que prend l'argument de la fonction d'utilité des assurés $(x + h)$ suivant leur état de maladie (g). Quant à la perte

d'efficacité due à l'aléa moral, elle peut être mesurée par l'écart entre l'argument $(x + h)$ dans la solution de premier rang et sa moyenne dans celle de second rang (avec comme pondération les probabilités de survenance des états de maladie).

En même temps que les taux de remboursement marginaux à l'optimum, ces informations sont fournies pour deux spécifications des paramètres de l'exemple numérique dans les tableaux 4 et 5; on y distingue les résultats obtenus avec les trois schémas de remboursement (linéaire à taux unique, linéaire à deux taux et non linéaire) ainsi qu'à titre de référence, avec la solution de premier rang (information parfaite). Les deux spécifications des paramètres retenues se différencient uniquement par la manière dont la sensibilité de la demande en soins curatifs au taux de remboursement évolue avec la gravité de la maladie. En ce qui concerne le schéma linéaire à deux taux, ceci conduit dans le tableau 4 ($\varepsilon_1 = 0{,}25\ k$ et $\varepsilon_2 = 0{,}35\ k$) à un taux marginal de remboursement qui est croissant avec y, alors qu'il est décroissant dans le tableau 5 ($\varepsilon_1 = 0{,}20\ k$ et $\varepsilon_2 = 0{,}40\ k$).

TABLEAU 4

COMPARAISON DES DIFFÉRENTS SCHÉMAS

$\rho = 0{,}75,\ k = 1{,}00$

$\varepsilon_1 = 0{,}25\ k$ ET $\varepsilon_2 = 0{,}35\ k$

	Taux marginaux de remboursement	Argument de la fonction d'utilité	Perte d'efficacité (aléa moral)
Remboursement linéaire		$(= x_i + h_i)$	
à taux unique	$\tau^* = 0{,}7928$	g_0 : 104,71 g_1 : 102,80 g_2 : 74,17	0,4857 (0,52 %)
à deux taux	$\tau_1 = 0{,}4193$ $\tau_2 = 0{,}8427$ $(s = 8{,}2611)$	g_0 : 106,14 g_1 : 101,05 g_2 : 76,44	0,3646 (0,39 %)
Remboursement non linéaire	$r'(y_1) = 0{,}0139$ $r'(y_2) = 0{,}0749$	g_0 : 99,23 g_1 : 98,84 g_2 : 84,30	0,0002
Solution de premier rang (information parfaite)	$\tau_1 = \tau_2 = 0$	g_0, g_1 et g_2 : 94,06	

TABLEAU 5

COMPARAISON DES DIFFÉRENTS SCHÉMAS
$\rho = 0{,}75$, $k = 1{,}00$
$\varepsilon_1 = 0{,}20\ k$ ET $\varepsilon_2 = 0{,}40\ k$

	Taux marginaux de remboursement	Argument de la fonction d'utilité	Perte d'efficacité (aléa moral)
Remboursement linéaire		$(= x_i + h_i)$	
à taux unique	$\tau^* = 0{,}7744$	g_0 : 107,41 g_1 : 104,83 g_2 : 77,72	0,4009 (0,41 %)
à deux taux	$\tau_1 = 0{,}8133$ $\tau_2 = 0{,}1015$ $(s = 75{,}1707)$	g_0 : 106,36 g_1 : 104,21 g_2 : 79,44	0,3159 (0,33 %)
Remboursement non linéaire	$r'(y_1) = 0{,}0135$ $r'(y_2) = 0{,}0717$	g_0 : 101,74 g_1 : 101,31 g_2 : 87,04	0,0002
Solution de premier rang (information parfaite)	$\tau_1 = \tau_2 = 0$	g_0, g_1 et g_2 : 96,62	

L'un et l'autre de ces tableaux mettent en évidence que la dispersion des niveaux de bien-être décroît lorsque l'on passe du remboursement linéaire à taux unique au remboursement non linéaire. Pour s'en convaincre, il suffit de voir l'évolution de l'écart entre les valeurs de $(x_i + h_i)$ qui caractérisent les individus dans les états extrêmes g_0 et g_2. La perte d'efficacité due à l'aléa moral décroît également; la réduction est particulièrement forte lors de l'adoption d'un remboursement non linéaire, ce qui s'explique par la forte réduction des taux marginaux de remboursement.

CONCLUSION

Dans cet article, nous nous sommes limités à un aspect de la mise en oeuvre d'une assurance-maladie : le choix de son schéma de remboursement et ceci dans le cadre d'un modèle fort simple. Plusieurs extensions sont possibles pour rendre ce modèle plus proche de la réalité. Une synthèse des questions posées par l'assurance-maladie est fournie dans Marchand et Pestieau (1996) et Marchand (1997).

Tout d'abord, nous avons fait l'hypothèse que le malade décidait de sa propre utilisation des soins en étant informé des résultats qu'il pouvait en escompter ou,

alternativement et de manière plus réaliste, que le médecin était l'agent parfait de son patient (autrement dit, que le médecin prenait pour son patient les décisions que celui-ci aurait prises s'il avait les mêmes connaissances que son médecin). Dans la réalité, le médecin peut-être influencé par d'autres considérations que le seul intérêt de son patient. Ce point est important; dès lors que le patient a pris de son propre chef la décision de consulter son médecin, il s'en remet largement aux indications de celui-ci quant aux examens à réaliser et aux traitements à suivre. En conséquence, le phénomène d'aléa moral concerne autant le comportement du patient que celui de son médecin. Une extension possible est donc de développer un modèle qui chercherait à optimiser non seulement la fonction de remboursement de l'assuré, mais aussi les taux de rémunération des médecins en fonction de leur activité. Par ailleurs, il serait intéressant d'étudier un modèle dans lequel on distinguerait la décision de recours au médecin de la décision relative à l'intensité des soins. Ceci permettrait de mettre en place un paiement forfaitaire au moment où le patient initie la demande de soins, accompagné d'une coassurance portant sur le montant des dépenses encourues.

Une autre extension possible concerne une différenciation des taux de remboursement en fonction des catégories de soins. En effet, ces différentes catégories ne présentent pas la même élasticité de leur demande aux taux de remboursement, ce qui suggère une différenciation de ceux-ci. Comme déjà indiqué, cette question a été étudiée par Zeckhauser (1970) et Besley (1988). Son traitement correct nécessite que l'on prenne en compte les phénomènes de complémentarité et de substituabilité qui caractérisent les différentes catégories de soins diagnostiques et thérapeutiques, voire les différents secteurs de la santé (par exemple, les soins ambulatoires et hospitaliers). Ainsi en Belgique, le ticket modérateur forfaitaire dû par le patient par journée d'hospitalisation augmente après un certain nombre de jours. Ceci peut induire un déplacement des dépenses de soins de l'hôpital vers d'autres formes de prise en charge, telles que les maisons de repos et de soins ou les maisons de convalescence.

Par ailleurs, dès qu'on adopte une fonction de remboursement des dépenses qui est, comme dans cet article, plus complexe que celle basée sur un taux proportionnel (unique) de remboursement, il faut se fixer une période de temps (en général, une année) au terme de laquelle les dépenses totales de l'assuré sont déterminées pour calculer le remboursement auquel il a droit. Nous avons fait l'hypothèse qu'en début de période l'assuré connaissait son état de santé et donc les soins qu'il nécessiterait. En conséquence, par exemple dans le cas d'une fonction de remboursement à deux taux marginaux (section 3), il lui était possible en début de période de déterminer le régime 1 ou 2 (alternativement, régime 1, s ou 2) dans lequel se situerait le total de ses dépenses. Ce n'est pas complètement réaliste : l'assuré ne peut en général que faire des prévisions sur la façon dont son état de santé évoluera pendant la période de référence. Au départ, le régime de remboursement dans lequel il se trouvera en fin de période (et donc son taux marginal de remboursement) n'est connu qu'en probabilité et la précision de cette prévision s'améliorera au fil de la période (et au fur et à mesure que les épisodes

de maladie se présenteront). Comme nous l'avons déjà mentionné, le comportement de l'assuré face à cette incertitude a été analysé par Keeler, Newhouse et Phelps (1977), mais ses conséquences n'ont guère été étudiées sur le plan normatif de la fixation d'une fonction optimale de remboursement.

Enfin, dans les débats sur l'assurance-maladie l'incidence de celle-ci sur la distribution du bien-être figure en bonne place. Et la discussion va généralement bien au-delà du partage des risques entre bien-portants et mal-portants pour porter également sur la liaison du remboursement des assurés avec leur niveau de revenu[15]. D'aucuns mettent en évidence que des tickets modérateurs trop élevés pourraient conduire à une « médecine à deux vitesses ». Cette crainte est d'autant plus fondée qu'il existe en effet une corrélation positive qui est empiriquement observée entre le revenu des individus et leur état de santé (la causalité pouvant aller dans l'une ou l'autre direction). Le choix de fonctions de remboursement dépendant des revenus du travail des assurés (et donc indirectement de leur productivité) a été étudié conjointement avec celui du barème d'imposition par Blomquist et Horn (1984), Rochet (1991) et Cremer et Pestieau (1996). Mais ces travaux ne prennent pas en compte le phénomène d'aléa moral qui a été la préoccupation centrale dans le présent article (et aussi dans les débats sur l'assurance-maladie). Ceci suggère d'étendre le modèle développé ici en faisant varier le revenu individuel (w_0) d'un assuré à l'autre, ou mieux en le faisant dépendre de leur offre endogène de travail qui varierait en fonction de leur productivité.

BIBLIOGRAPHIE

ARROW, K. (1963), « Uncertainty and Welfare Economics of Medical Care », *American Economic Review*, 53 : 941-973.

BESLEY, T.J. (1988), « Optimal Reimbursement Health Insurance and the Theory of Ramsey Taxation », *Journal of Health Economics*, 7 : 321-336.

BLOMQUIST, A., et H. HORN (1984), « Public Health Insurance and Optimal Income Taxation », *Journal of Public Economics*, 24 : 353-371.

CREMER, H., et P. PESTIEAU (1996), « Redistributive Taxation and Social Insurance », *International Tax and Public Finance*, 3 : 281-295.

HENRIET, D., et J.-CH. ROCHET (1991), *Microéconomie de l'assurance*, Éd. Economica, Paris.

KEELER, E.B., J.P. NEWHOUSE, et C.E. PHELPS (1977), « Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule under Uncertainty », *Econometrica*, 45 (3) : 641-655.

15. Comme indiqué dans l'introduction, le montant de la franchise récemment introduite en Belgique dépend du revenu de l'assuré. L'expérimentation de la Rand Co. présentait la même caractéristique.

MANNING, W.G., J.P. NEWHOUSE, N. DUAN, E.B. KEELER, A. LEIBOWITZ, et M.S. MARQUIS (1987), « Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment », *American Economic Review*, 77 : 251-277.

MARCHAND, M., et P. PESTIEAU (1996), « L'État ou le marché dans l'assurance-maladie », *Revue française d'économie*, XI : 3-19.

MARCHAND, M. (1997), « Assurance-maladie publique et privée : coexistence bénéfique ou conflictuelle? », *Risques*, 30 : 45-55.

PAULY, M.V. (1968), « The Economics of Moral Hazard: Comment », *American Economic Review*, 58 : 531-537.

PAULY, M.V. (1974), « Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection », *Quarterly Journal of Economics*, 88 (1) : 44-62.

ROCHET, J.-CH. (1991), « Incentives, Redistribution and Social Insurance », *The Geneva Papers on Risk and Insurance Theory*, 16 (2) : 143-165.

SALANIÉ, B. (1994), *Théorie des contrats,* Éd. Economica, Paris.

SHAVELL, S. (1979), « On Moral Hazard and Insurance », *Quarterly Journal of Economics,* 11 : 541-562.

SHESHINSKI, E. (1972), « The Optimal Linear Income Tax », *Review of Economic Studies*, 39 : 297-302.

SHESHINSKI, E. (1989), « Note on the Shape of the Optimum Income Tax Schedule », *Journal of Public Economics*, 40 : 201-215.

SLEMROD, J., SH.YITZHAKI, J. MAYSHAR, et M. LUNDHOLM (1994), « The Optimal Two-Bracket Linear Income Tax », *Journal of Public Economics*, 53 : 269-290.

SPENCE, M., et R. ZECKHAUSER (1971), « Insurance, Information, and Individual Action », *American Economic Review*, Papers and Proceedings, 61 : 380-387.

WINTER, R.A. (1992), « Moral Hazard and Insurance Contracts », 61-96, in G. DIONNE (éd.), *Contributions to Insurance Economics*, Kluwer, Dordrecht.

ZECKHAUSER, R. (1970), « Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives », *Journal of Economic Theory,* 2 : 10-26.