Politiques de santé dans un modèle macroéconomique

V. Touzé, B. Ventelou
OFCE, Fondation Nationale des Sciences Politiques

Version au 15/09/00

Résumé :

Cet article propose une approche théorique de l'incidence des politiques de santé (contingentement, ticket modérateur, contrôle qualitatif) ; la voie envisagée est une analyse micro-macroéconomique. Le comportement de l'offre médicale est construit sur un choix d'allocation du temps de travail entre deux activités : l'une utile et l'autre inutile. Ce fondement microéconomique des arbitrages médicaux est ensuite intégré dans un modèle de croissance avec accumulation de capital ; les conclusions concernent alors les évolutions des dépenses de santé, du bien-être et de l'accumulation de patrimoine au regard de différents plans de politique de santé.

***Abstract* :**

This article offers a micro-macro-economic approach for the incidence of policies concerning health care. The medical supply behaviour is built using a working-time allocation choice between various qualities of care. This choice is assimilate to a decomposition of health care in two activities : one effective and another non-effective. This micro-economic foundation of health supply is then integrated in a model of growth with capital accumulation. The conclusions concern the evolution of health expenditures, the population's welfare and wealth accumulation, with respect to various policies.

1. INTRODUCTION

Les réformes de santé incluent souvent (en France et en Allemagne notamment) des mesures de contingentement du chiffre d'affaire des professionnels de santé, visant à diminuer la partie *inutile* de la consommation de santé. A défaut de pouvoir la contrôler directement, on cherche à réduire la "demande induite" liée à des comportements stratégiques présumés de la part des offreurs du secteur, qui sont suspectés, en information asymétrique[1], de provoquer une demande de soins supérieure à celle qui existerait si l'information était parfaite. Cependant, le législateur ne s'interroge que rarement sur ***l'élasticité*** de la "demande induite" à sa mesure de contingentement (ce qui nous fait supposer que, dans son esprit, l'élasticité est -/au moins- égale à 1 : tout contingentement réduit la dépense de santé inutile -/au moins- dans la même proportion, …sans réduire pour autant la partie utile de consommation de santé). Une mesure précise de *la "réponse" des offreurs de*

[1] Voir ROCHAIX et JACOBZONE 1997, pour une définition générale de la demande induite. La distorsion est due à une asymétrie d'information sur la qualité réelle du produit vendu : le médecin dispose d'une information meilleure que le patient sur l'efficacité des soins. On peut croiser cette justification informationnelle avec le fait que, par ailleurs, une part importante de la consommation est financée collectivement, ce qui engendre une très faible internalisation individuelle du coût réel des soins par les patients.

soins aux contingentements est pourtant cruciale pour qui ne se limite pas à des considérations purement financières et veut évaluer l'impact *sur le bien-être des consommateurs* de cette catégorie de réformes.

Il faut dire que la question apparaît complexe lorsque les offreurs de soins sont supposés bénéficier d'une marge de manœuvre importante dans l'allocation de leurs efforts entre les deux types d'activité (utile et inutile ; avec, on peut en faire l'hypothèse, des arbitrages de temps de travail alloués aux deux activités), et lorsque, par ailleurs, l'évolution des technologies modifie, périodes après périodes, les équilibres établis. Pour reprendre la distinction de WEISBROD (1991), le progrès technique est, en matière médicale, à la fois un progrès technique de productivité et un progrès technique de gamme. Parfois l'extension de la gamme permet d'envisager de traiter, enfin, les maladies précédemment incurables, au bénéfice certain des malades sinon de la société toute entière ; parfois la gamme s'étend simplement vers plus de confort de soins, voire vers une complexification cosmétique des soins. Ce qui accroît encore la difficulté d'un pronostic monochrome ; dans quelle mesure, en effet, le contingentement des quantités vient-il, ou ne vient-il pas, perturber les logiques technico-économiques en œuvre, tendant à proposer des soins meilleurs et plus nombreux ? Quelles sont alors les conséquences du contingentement médical lorsqu'on cherche à étudier l'ajustement d'une économie – analysée sur le long terme et sur ses différents marchés – présentant des gains de productivité variables dans ses secteurs de production ?

On veut donc pouvoir disposer d'un modèle d'équilibre général afin de traiter de l'incidence "complète" de l'évolution des dépenses maladie, notamment via ces conséquences sur le taux d'épargne macroéconomique.

Nous proposons dans ce papier une analyse micro-macroéconomique de cette question. En premier lieu, une formalisation microéconomique de l'offre médicale se basera sur un modèle de choix d'allocation du temps de travail entre deux *activités : l'une utile* et *l'autre inutile* (on suppose qu'en raison d'imperfection d'information propre au secteur médical, une production de bien inutile est possible). Nous intégrerons ensuite cette microéconomie médicale dans un modèle de croissance avec accumulation de biens capitaux, ceci afin de pouvoir traiter de l'équilibre général dynamique de l'économie sous l'effet des réformes.

Ce modèle nous permettra :

- sur le plan microéconomique appliqué, de traiter de la fonction de réponse optimale du secteur de santé à une mesure de contingentement et de discuter des propriétés incitatives respectives des différentes modalités du contingentement (blocage purement quantitatif, système de pénalités financières… etc.) ;
- sur le plan macroéconomique, en intégrant les raisonnements dans un modèle de croissance avec épargne endogène, on cherchera à étudier sur le long terme le comportement dynamique de l'économie et ses réactions face à des décisions de politique économique ; le modèle nous permettra notamment d'estimer l'évolution de la part *soins utiles* / *soins inutiles* sous l'effet d'une mesure durable de contingentement (l'analyse prendra en compte : des mesures de l'évolution du progrès technique médical et des mesures de l'élasticité revenu de la consommation de santé des ménages) ; puis, le modèle estimera les effets des mesures sur la propension à épargner de l'économie.

- on tentera enfin de donner une évaluation de l'évolution du bien-être des consommateurs en fonction des paramètres décrits ci-dessus. On proposera aussi des estimations de l'effet de quelques principes d'intervention alternatifs : augmentation des co-paiements ; intervention qualitative de l'assurance maladie.

1. L'OFFRE DE SANTÉ

Le secteur de santé de cette économie est composé de I(t) biens et services élémentaires. La valeur I(t) est supposée varier avec le temps, afin de rendre compte d'un progrès technique spécifique au secteur de santé (un progrès technique de gamme). En outre, pour chaque bien i, nous supposons que des éléments de l'offre de soins peuvent être vendus sans néanmoins procurer une utilité au consommateur (l'offreur de soin est capable "*d'induire*" une consommation de soin qui ne donne pas d'utilité réelle [2]) ; une partie du temps de travail médical est consacré à ce type de soins de santé. En posant h_1 et h_2, respectivement activité de soins utile et inutile, $l_{i,1}$ et $l_{i,2}$ temps de travail, on aura:

$$h_i(l_{i,1}, l_{i,2}) = h_{i,1}(l_{i,1}) + h_{i,2}(l_{i,2}) \quad \forall\, i \in 1 \ldots I(t)$$

Pour des raisons de simplicité, nous choisissons d'emblée de réduire l'hétérogénéité des soins de santé : on suppose que les différentes technologies des soins de la gamme sont identiques et qu'à l'équilibre les prix respectifs des différents soins de santé sont identiques (voir plus loin, l'utilité sociale des soins est identique). On conserve néanmoins le paramètre I_t, progrès technique de gamme, en considérant que l'extension porte alors simplement sur la "variété" des maladies traitables (plutôt que sur les méthodes et technologies de traitement). On aura pour $\bar{h}$ bien médical moyen et h un agrégat de consommation médicale :

$$h(l_1, l_2) = I_t \cdot \bar{h}\left(\frac{l_1}{I(t)}, \frac{l_2}{I(t)}\right) = I_t \cdot \left[\bar{h}_1\left(\frac{l_1}{I(t)}\right) + \bar{h}_2\,\frac{l_2}{I(t)}\right]$$

Avec, par définition,

$$\sum_{i=1}^{I(t)} l_{i,1} \,(= I_t \bar{l}_1) = l_1 \quad \text{et} \quad \sum_{i=1}^{I(t)} l_{i,2} \,(= I_t \bar{l}_2) = l_2$$

Nous supposons que les offreurs de soins (vivant deux périodes, activité et retraite) ont un comportement d'optimisation intertemporelle de leur consommation :

$$\text{Max. } \ln c_m + \beta_m \ln z^+_m - \delta_1 l_1 - \delta_2 l_2$$
$$\text{s.c. : } c_m + \frac{z^+_m}{R^+} = p\, I(t)\left[\bar{h}_1\left(\frac{l_1}{I(t)}\right) + \bar{h}_2\left(\frac{l_2}{I(t)}\right)\right]$$

Les valeurs c_m et z^+ sont les consommations du bien agrégé. β est le facteur d'escompte du temps. Les paramètres δ_1 et δ_2 sont les coûts – subjectifs[3] – associés à l'effort de travail médical, respectivement, pour les activités de soin h_1 et h_2.

---

[2] Voir annexe 1 et 2 sur la demande induite (l'annexe 1, non rédigée ici, cherchera à fonder l'incontrôlabilité par les contrats optimaux de l'imperfection d'information).

[3] C'est par cette voie qu'on peut envisager de traiter de la déontologie médicale : un paramètre δ_2 très élevé correspondra, toute chose égale par ailleurs, à une forte "déontologie" – les soins inutiles sont psychologiquement coûteux.

On obtient les comportements optimisés ($S_m = \frac{z^+_m}{R^+}$) :

$$S_m = s_m\, p\, h(l_1, l_2) \qquad \text{avec } s_m = \frac{\beta_m}{1+\beta_m}$$

$$\frac{\bar{h}_1'}{\bar{h}} \frac{1}{I(t)} = \delta_1 . \frac{1}{1+\beta_m}$$

$$\bar{h}_1' \left(\frac{l_1}{I(t)}\right) = \frac{\delta_1}{\delta_2}\ \bar{h}_2' \left(\frac{l_2}{I(t)}\right)$$

On peut définir α la proportion de biens médicaux utiles sur l'ensemble des biens médicaux. On aura :

$$\alpha = \frac{\bar{h}_1(l_1^*)}{\bar{h}(l_1^*, l_2^*)} .$$

Sous notre jeu d'hypothèse, cette proportion est totalement déterminée par les offreurs du secteur. On peut définir un plan optimal de production : $(h^*, \alpha^*) = (h , \alpha)_{(\delta_1, \delta_2, I, \ldots)}$, unique.

2. LE RISQUE MALADIE ET LES CHOIX DES MÉNAGES

Concernant les ménages, nous partons d'un modèle à générations imbriquées d'agents vivant deux périodes, recevant un salaire w en première période et épargnant un montant S entre leur première période et leur seconde période de vie (conformément à la spécification de DIAMOND, 1965). L'originalité du modèle consiste à introduire une dépense maladie survenant aux deux périodes de vie des agents. A la deuxième période de vie, nous supposons que la morbidité et la dépense de santé sont certaines (ce qui, en négligeant les problèmes posés par l'aléa et l'assurance, est une manière de "normer" – à l'unité – la probabilité de morbidité des "vieux"). A la première période de vie, nous autorisons un aléa maladie : les jeunes peuvent être malades ou non malades ; dans le premier cas, ils pratiquent des dépenses maladies réparatrices ; dans le second cas, leur consommation des santé est nulle[4]. On aura :

Pour des agents malades en première période :
Un état de santé $d_y = f(h_y)$
Pour des agents non malades :
Un état de santé $d'_y = d^*$

Où d^* est un état de santé initial, exogène pour les non malades[5].

---

[4] Outre l'hétérogénéité donnée par l'age, on aura donc deux types d'agents dans l'économie, définis selon leur historique concernant l'aléa maladie.

[5] On négligera la contrainte $d_y < d'_y$, ou encore $f(h_y) < d^*$, en supposant :

i) soit qu'on peut, en se soignant, atteindre un niveau de santé même supérieur à l'état initial (*la contrainte n'existe pas)* ;

ii) soit que l'état de santé initial, ou la dégradation de l'état de santé liée à la maladie, sont suffisamment forts pour que la contrainte ne soit jamais effective *(la contrainte existe mais elle n'est jamais saturée).*

La contrainte d'équilibre de l'assurance maladie :

$$\pi \geq [\psi (p . \lambda . I . \bar{h}_y + J)]$$

Où π est la prime d'assurance, λ est un taux de remboursement des soins ($1-\lambda$ est le "ticket modérateur") et J un (éventuel) paiement forfaitaire. On supposera la saturation de cette contrainte. Cette équation, qui détermine *ex post* la valeur π des contributions, n'est pas intégrée dans les choix *ex ante* des ménages[6].

2.1 Ménages malades en première période

Les ménages malades – proportion ψ – vont être décrits par :

$$U(c, z^+, h_y, h_o^+) = [\ln c + \gamma \sum_{i=1}^{I(t)} \ln(h_{y;i} - h_{y;i;2})] + \beta [\ln z^+ + \gamma^{\#} \sum_{i=1}^{I(t)} \ln(h_{o;i} - h_{o;i;2})]$$

$$\text{s.c. : } c = w - S - \pi - (1-\lambda) \sum_{i=1}^{I(t)} p_i (h_{y;i}) + J$$

$$\text{s.c. : } z^+ = R^+ S - \sum_{i=1}^{I(t)} p_i^+ (h_{o;i})$$

$h_{y,i}$ et $h_{o,i}$ représentent les différentes dépenses médicales, respectivement quand l'individu est jeune, puis vieux. On note que seules les dépenses médicales de type h_1 fournissent une utilité : les consommateurs – pour des raisons d'asymétrie informationnelle – achètent (payent) une quantité supérieure à la quantité réellement utile à leur bien-être ; h_2 est considérée comme une "consommation fatale" au même titre qu'on définit le concept de "productions fatales" (voir par exemple : BENARD, 1984, voir annexe 1 pour plus de détail). On procède ensuite à la simplification déjà proposée sur la gamme[7] :

$$U(c, z^+, h_y, h_o^+) = [\ln c + \gamma I \ln (\bar{h}_y - \bar{h}_{y,2})] + \beta [\ln z^+_m + \gamma^{\#} I^+ \ln (\bar{h}^+_o - \bar{h}^+_{o,2})]$$

$$\text{s.c. : } c = w - S - \pi - (1-\lambda) p I \bar{h}_y + J$$

$$\text{s.c. : } z^+ = R^+ S - p^+ I^+ \bar{h}^+_o$$

L'introduction de la contrainte nous conduirait à introduire une fonction min.

[6] La contribution π individuelle (et agrégée, pour une population normée à 1) est perçue comme un prélèvement forfaitaire : les agents n'internaliseront pas la part **sociale** de leurs consommations de soins au moment de leur choix, mais uniquement la part **individuelle** donnée par le ticket modérateur (myopie du patient).

[7] Il faut remarquer que la simplification préserve un 'effet variété' pour la gamme des soins : le facteur I d'accroissement de la gamme intervient en effet ***avant*** la concavité (donné par la fonction ln) de la fonction d'utilité. Autrement dit, alors qu'il existe une utilité marginale décroissante sur *chaque type de soins* ($\bar{h}$), il existe, simultanément, une non-décroissance de l'utilité marginale de *l'ensemble* des soins (l'augmentation de I – l'extension de la gamme – accroît linéairement l'utilité). Cet 'effet variété' a été exploité en théorie de la croissance du coté fonctions de production (voir ROMER, 1991).

Les conditions du premier ordre du programme sont :

$$z^{+} = \beta\, R^{+}\, c$$

$$\bar{h}_{y} = \frac{\gamma}{\alpha p(1-\lambda)}\, c$$

$$\bar{h}^{+}_{o} = \frac{\gamma^{\#}}{\alpha^{+} p^{+}}\, z^{+}$$

Le comportement d'épargne optimisé sera :

$$S = s\,(w - \pi + J) \quad \text{pour } s = \frac{\beta}{\frac{1+(\gamma/\alpha)I}{1+(\gamma^{\#}/\alpha^{+})I^{+}} + \beta}$$

Et les comportements optimisés de consommation:

$$c = (1-s)\,\frac{1}{1+\frac{\gamma I}{\alpha}}\,[w - \pi + J]$$

$$z^{+} = s\,\frac{1}{1+\frac{\gamma^{\#} I^{+}}{\alpha^{+}}}\, R^{+}\,[w - \pi + J]$$

$$\bar{h}_{y} = \frac{\gamma}{\alpha(1-\lambda)p}\,(1-s)\,\frac{1}{1+\frac{\gamma I}{\alpha}}\,[w - \pi + J]$$

$$\bar{h}^{+}_{o} = \frac{\gamma^{\#}}{\alpha^{+} p^{+}}\, s\,\frac{1}{1+\frac{\gamma^{\#} I^{+}}{\alpha^{+}}}\, R^{+}\,[w - \pi + J]$$

2.2 Ménages non malades en première période

Les ménages non malades – proportion $(1-\psi)$ – vont être décrits par:

$$U(c', z'^{+}, h'^{+}_{o}) = [\ln c' + d^{*}] + \beta\,[\ln z'^{+}_{m} + \gamma^{\#} I^{+} \ln(\bar{h}'^{+}_{o} - \bar{h}'^{+}_{o,2})]$$

$$\text{s.c. : } c' = w - S' - \pi$$

$$\text{s.c. : } z'^{+} = R^{+} S' - p^{+} I^{+}\, \bar{h}'^{+}_{o}$$

Les dépenses de santé de première période ne sont pas enclenchées (et le consommateur dispose d'une utilité liée à son état de santé initial d*). Les comportements optimisés sont :

$$S' = s' \,.\, (w - \pi) \quad \text{pour } s' = \frac{\beta}{\frac{1}{1+(\gamma^{\#}/\alpha^{+})I^{+}} + \beta}$$

Et :

$$c' = (1-s')\,[w - \pi]$$

$$z'^{+} = s'\,\frac{1}{1+\frac{\gamma^{\#} I^{+}}{\alpha^{+}}}\, R^{+}\,[w - \pi]$$

$$\bar{h}'^{+}_{o} = \frac{\gamma^{\#}}{p^{+}\alpha^{+}}\, s'\,\frac{1}{1+\frac{\gamma^{\#} I^{+}}{\alpha^{+}}}\, R^{+}\,[w - \pi]$$

3 EQUILIBRE EN LAISSEZ FAIRE

L'équilibre sur le marché du capital permet d'établir la dynamique du modèle. On aura :

$$k^{+} = [\ \psi\ s\ (w-\pi+J) + (1-\psi)\ s'\ (w-\pi)\ ]\ \frac{1}{G^{+}} + s_m\ q\ .\ p^*\ .\ h^*$$

p* et h* sont les prix et quantités d'équilibre du marché de la santé. Sur ce dernier marché, on aura :

$$[\ \psi^{-}\ h_o + (1-\psi^{-})\ h'_o\ ]\ G + \psi\ h_y = q\ .\ h^*$$

et donc (sachant que l'offre est fixée à h* –indépendante du prix–, et en exprimant les fonctions de demande h_o, h'_o, et h_y du côté des ménages) :

$$p^* = I\ \frac{[\psi^{-}\ z^{*} + (1-\psi^{-})\ z'^{*}]\ \gamma^{\#}\ G + \frac{\gamma}{1-\lambda}\ \psi\ .\ c^{*}}{I\ .\ \alpha^{*}\ .\ q\ .\ [\ \bar{h}_1\ (\frac{l_1}{I}^{*}) + \bar{h}_2\ (\frac{l_2}{I}^{*})]}$$

Les paramètres représentent respectivement : G le taux de croissance de la population (donc : la part respective des jeunes et des vieux dans la population[8]) ; et q la part des offreurs de soins dans la population. Commentaires :

- Les effets de λ (taux de remboursement), ψ (paramètre de morbidité), γ (paramètre de goût pour la santé) sont conformes à l'intuition première : ils déforment à la hausse la consommation de soins de la société et contribuent positivement au prix de santé. On relève particulièrement l'effet de λ qui, lorsqu'il s'approche de l'unité fait tendre la consommation de santé vers l'infini (myopie sur le coût des soins). Il y a néanmoins une force de rappel liée au fait que, dans ce cas, la prime d'assurance π tendrait vers l'infini, et les valeurs c*, z* et z'* vers zéro.
- Le paramètre α (1–α, paramètre de "demande induite") donne à première vue des effets ambigus (on peut vérifier qu'il joue au dénominateur mais aussi au numérateur *via* les variables c, z et z'). Plus α est grand, moins les consommations de santé ont *a priori* à être élevées, une unité donnée d'offre de santé fournissant déjà un service de qualité ; cette première réflexion joue en faveur d'un effet négatif de α sur p (visible dans l'équation de prix). Mais d'un autre coté, plus α est grand, plus le bien santé peut être considéré comme attractif pour un consommateur ; on aurait alors un effet positif de α sur p. En l'état (i.e. : compte tenu du choix de la fonction d'utilité), c'est le premier effet qui domine.
- L'effet de I (PT de gamme) est lui aussi ambigu : l'augmentation de I conduit à une augmentation de la demande, et aussi à une augmentation de l'offre (on a fait apparaître deux I qui s'éliminent, au numérateur et au dénominateur). Du côté de l'offre, avec l'extension de la gamme, il faut aussi répartir les ressources

[8] On constate, de ce point de vue, que la démographie **ne** modifie **pas** la proportion 'épargne - dépense de santé' des agents (hors effet de la population médicale). Il s'agit d'un handicap du modèle, lié au fait que les agents n'ont que deux périodes de vie, jeunes / vieux (et seuls les jeunes épargnent). L'effet suivant : 'vieillissement – hausse de la morbidité – hausse des dép. de santé – baisse de l'épargne' n'est pas rendu dans le modèle faute d'une capacité d'épargne en seconde période.

du secteur médical sur plus de biens produits, la tension augmente ; mais, en même temps, les incitations à produire augmentent (la hausse de I augmente le chiffre d'affaire potentiel) et conduisent le secteur à travailler plus. En l'état (i.e. : compte tenu des choix des fonctions objectif du secteur), c'est le premier effet qui l'emporte (l'incitation à travailler n'est pas suffisante pour compenser l'effet de dilution du travail sur les I sous-secteurs de santé[9]).

4. IMPACT SUR LE BIEN-ÊTRE DES MENAGES DE QUELQUES PRINCIPES D'INTERVENTION

Le recherche de l'optimum parétien d'une telle économie est difficile[10]. Il y a, dans ce modèle, plusieurs sources d'inefficacité pour les ménages. On peut notamment citer : les choix d'épargne peuvent présenter une inefficience dynamique ; l'assurance contre le risque n'est pas complète ; le mode de remboursement des soins n'est pas parfaitement incitatif[11] (mais il correspond néanmoins à une certaine réalité, en France tout au moins). Ici donc, nous limitons l'étude aux problèmes posés par la demande induite h_2 et, plus précisément, à la sous-efficacité qu'elle entraîne sur le montant des soins h, mais aussi, sur l'évolution de l'épargne k^+ et sur le bien-être des agents malades et non malades. On choisit de s'intéresser au bien-être des consommateurs uniquement[12]. On veut savoir si une mesure de contingentement, appliquée dans une économie en équilibre général tel qu'elle a été décrite plus haut, conduit à une amélioration de leur satisfaction. On peut distinguer plusieurs types de contingentement :

Contingentement quantitatif pur : le législateur contrôle la quantité de soins produits et échangés, fixée à h^c, mais laisse libre le prix p. On a la contrainte :

$$h^* \leq h^c$$

La moindre disponibilité de soins sur le marché de la santé augmente le prix (le long de la demande de soin) : il en résulte que l'évolution, à la hausse ou à la baisse, de la dépense de santé est déterminée par l'élasticité de la demande au prix. Pour des élasticités particulières, on peut même craindre une hausse en valeur des dépenses de santé. Cette technique n'est pas, sauf exceptions (contingentement par *numerus*

---

[9] On raisonne ici pour un accroissement fini de I. Lorsque I tend vers l'infini, les propriétés du modèles sont délicates à interpréter. Un travail est en cours sur cette question.

[10] Voir Touzé, Ventelou (2000) pour une réflexion sur l'optimum de Pareto d'une telle économie.

[11] Avec le mode de remboursement décrit ci-dessus les agent n'internalisent que partiellement les coûts social de leur consommation de soin.

[12] Pour être complet et obtenir un vrai optimum de Pareto, il faudrait disposer d'un critère utilitariste comprenant l'utilité des médecins. Dans notre modèle, en l'absence d'un critère utilitariste discriminant, on ne peut pas vraiment dire que, par exemple, les surconsommations médicales liées à la prise en charge collective des dépenses sont a priori socialement inefficaces ; elles conduisent simplement à un partage de la richesse nationale en faveur des médecins plus avantageux que celui qui s'observerait si le remboursement était incitatif (un transfert forfaitaire paramétrique par type de maladie).

clausus du secteur à honoraire libre), celle utilisée en France : les contingentements sont généralement associés à un contrôle de prix.

Contingentement quantitatif et contrôle du prix (cas administré, cas français appliqué à l'hôpital public, partiellement appliqué au secteur ambulatoire) : le législateur contrôle les quantités produites et impose un prix, par exemple le long de l'offre de santé (prix minimum nécessaire pour obtenir la quantité contingentée). On obtient une moindre consommation médicale (en volume : h^c), ceci associée, cette fois, à une dépense de soin inférieure (en valeur : $p^c\ h^c$, car le prix associé est, lui aussi, inférieur ou égal à sa valeur de marché). Les sommes dégagées peuvent être alors réallouées à la consommation présente et à l'épargne : *effets de bien-être positif*. En supposant que le rationnement s'adresse aux ménages jeunes $\psi.I.\overline{h}_y = h_y^c$, on aura, concernant l'accumulation de capital :

$$k^+ = \{ s' [w_{(k)} - \pi - \psi (1 - \lambda) p^c h_y^c + \psi J] \} \frac{1}{G^+} + s_m [p^c (h_y^c + h_o)]$$

En exprimant π par la contrainte budgétaire de l'ass. maladie :

$$k^+ = \{ s' [w_{(k)} - p^c h_y^c] \} \frac{1}{G^+} + s_m [p^c (h_y^c + h_o)]$$

Primo, le contingentement aurait l'avantage d'élever la propension à épargner des ménages malades (qui vient s'établir à la valeur de la propension à épargner des ménages non malades[13]) ; *secundo*, l'accentuation du contingentement (réduction de h_y^c) libère du pouvoir d'achat, qui se retrouve affecté à l'épargne dans la proportion s' ; *tertio*, on fait aussi apparaître un effet prix du contingentement : si le prix p^c est inférieur au prix notionnel, on obtient là aussi une libération de pouvoir d'achat (qui touche d'ailleurs aussi les "vieux" et leur consommation de seconde période). Toute chose égale par ailleurs, donc, l'accumulation de capital de la société s'accroît, …à condition toutefois que le prélèvement opéré sur le secteur médical ne réduise pas, dans les mêmes proportions, sa propre contribution à l'épargne macroéconomique (c'est la valeur de « $s' - s_m$ » qui détermine le sens global de l'effet). Il s'agit là d'une première réserve sur l'efficacité possible d'une politique de contingentement.

En outre, on sait aussi que le contingentement aura des *effets de bien être négatifs* via une réduction vraisemblable des consommations de santé utiles (h_1), dans la mesure où il n'y a aucune raison pour que la demande induite (h_2) soit la seule atteinte par le contingentement. C'est cette seconde réserve sur l'efficacité du contingentement qui nous paraît déterminante. On peut voir en annexe 2 une évaluation de l'effet du contingentement sur la partition h_1 / h_2.

Enveloppe globale avec reversements (cas français souhaité pour le futur de la médecine libérale[14]) : le législateur fixe une dépense agrégée E_y ; si la dépense

[13] La consommation de santé n'étant plus un choix (mais une allocation administrée), l'épargne des ménages malades n'est plus affectée par leur préférence en matière de santé (en première période).

[14] On peut voir BATIFOULIER ET TOUZÉ, 2000, MOUGEOT 1999, ou VENTELOU, 1999, pour une description des mesures de politique économique récentes envisagées par les payeurs institutionnels (CNAM et ministère). Les objectifs quantifiés nationaux (OQN) ont été instaurés officiellement en 1992. Ils sont

médicale dépasse ce seuil, il modifie les lettres clés du praticien en fixant un paramètre τ défini de la manière suivante : $(1-\tau).\psi.p.h_y = E_y$. La première différence avec le contingentement décrit ci-dessus est que cette technique contient des effets de redistribution des malades vers les non malades : les reversements ($\tau.\psi.p.h_y$) constituent de nouvelles ressources pour l'assurance maladie et le montant π de la contribution payée par tous peut alors être réduit à l'équilibre (on peut, de fait, parler d'augmentation implicite du ticket modérateur). On aura en effet l'égalité comptable 'emploi – ressource' de la branche maladie de la Sécurité sociale :

$$\pi^E + (\tau.\psi.p.h_y) = \psi\ (p\ .\ \lambda\ .\ h_y + J)$$

ou encore :

$$\pi^E = \psi\ (p\ .\ (\lambda - \tau)\ .\ h_y + J)$$

La seconde différence est que cette technique laisse place à un équilibre de marché : les quantités de soins sont effectivement choisies (on est dans un système de médecine libérale) et la décision administrative ne fait que re-transférer un pouvoir d'achat *ex post* des médecins vers les ménages. On s'aperçoit alors que la contrainte d'enveloppe admettrait, dans ce modèle, une propriété séduisante : elle permet à l'équilibre du marché de la santé de se fixer aussi haut que les agents le voudraient[15]; on n'assistera donc pas à l'effet de contraction de h_1 montré précédemment. Cette dernière propriété est néanmoins contestable dans la mesure où elle est très dépendante de la propriété d'inélasticité prix de l'offre médicale (propre à la forme des fonctions d'utilité). *A contrario*, lorsque l'offre est élastique au prix et qu'un véritable calcul a lieu sur les quantités totales produites en santé (par le secteur médical), il en ressort une contraction de la courbe d'offre et, donc, une baisse des quantités d'équilibre (la décision d'offre de santé se basant sur la rémunération réelle $(p-\tau)$ et non plus sur la dépense du malade p). On peut prévoir dans ce cas les mêmes ambiguïtés que dans l'intervention précédente : il y a un gain de bien-être pour les ménages lié à un transfert en provenance du secteur médical ; mais il existe aussi une perte de bien-être résultant de l'impact de la contrainte d'enveloppe sur h_1.

Concernant l'accumulation du capital, on obtient la relation suivante[16] :

$$k^+ = [\ \psi\ s\ (w_{(k)} - \pi^E + J) + (1-\psi)\ s'\ (w_{(k)} - \pi^E)\ ]\ \frac{1}{G^+} + s_m\ [p.q.h^* - \tau.\psi.p.h_y]$$

$$\text{avec } \pi^E = \pi - (\tau.\psi.p.h_y)$$

Les reversements se retrouvent en positif dans le profil d'épargne des ménages, et en négatif dans le profil d'épargne du secteur de santé. Comparé à la politique précédente (contingentement quantitatif), la politique d'enveloppe n'aura toutefois pas la propriété d'augmenter la propension à épargner des malades au niveau de celle des non malades ; cette hétérogénéité des taux d'épargne rend plus complexe

---

négociés entre les payeurs et les professionnels ; ils constituent une enveloppe fermée pour les cliniques privées, les infirmiers et les laboratoires d'analyses : tout dépassement entraîne une baisse des prix en remboursement du trop-perçu. Ils restent indicatifs pour les autres activités.

[15] c'est-à-dire ici au niveau de plein emploi des ressources médicales.

[16] Ici, on considère que la modification du tarif ne se reporte pas sur le prix des consommation de santé des "vieux" (l'enveloppe ne concerne que les assurés). Dans le cas contraire, les "vieux" contribueraient à l'assurance maladie via les reversements, ce qui rend les choses beaucoup plus complexes à analyser (transferts *ex post*, avec effets intertemporels).

l'analyse de l'effet agrégé (Cette fois, ce sont les valeurs de ψ.s, $(1-\psi)$.s' et s_m qui déterminent le résultat global de la réforme sur l'accumulation de capital).

Enfin, dernier point, il existe un "effet richesse" de la politique d'enveloppe globale qui, *via* une augmentation induite du prix des soins, tend à atténuer son efficacité. De fait, les reversements, en constituant une nouvelle ressource du bloc 'ménages + assurance' (qui s'enrichit), conduisent à une hausse des différentes consommations des agents, *y compris, donc, la consommation médicale*. Ce bouclage, typiquement macroéconomique, neutralise *ex post* une partie des effets attendus d'une politique de reversement.

5. CONCLUSION

Dans cet article, nous avons proposé de traiter, dans un contexte de bouclage macroéconomique, de deux questions d'économie de la santé : les politiques de contingentement réduisent-elles la demande induite ? Quelles sont leurs incidences sur le niveau de vie et le bien-être des ménages ?

Plusieurs raisons justifient ce cadre de réflexion. La première et la plus évidente est celle relative à *la désirabilité* des évolutions observées sur le ratio 'dépenses de santé/ PIB'. Dans la mesure où la richesse nationale est limitée, la santé a réellement un coût d'opportunité, dont on peut vouloir, précisément, calculer la valeur, et qu'il faut, ensuite, mettre en perspective avec le prix effectivement payé par la société pour sa fonction santé. La seconde raison est la nécessité de situer la régulation du marché des soins de santé dans un contexte général, qui tient compte à la fois des besoins sociaux de consommation médicale, des modes de financement et de remboursement, des incidences intertemporelles de l'allocation du revenu, ainsi que, bien sûr, des éventuels bouclages macroéconomiques existants entre ces différentes dimensions du choix social.

Les conclusions obtenues à l'issue de ce premier travail sont multiples : le gonflement des dépenses médicales est un facteur qui réduit l'accumulation de patrimoine des ménages et, donc, leur niveau de vie à plus long terme, mais, pas forcément leur bien-être ; le mode de remboursement des soins influençant la perception du coût réel des traitements puis le niveau de la demande, le prix de marché est d'autant plus élevé que la "myopie" du malade est importante ; la part des dépenses inutiles étant peu influençable, tout rationnement de l'offre conduit aussi à un rationnement de l'offre utile (l'asymétrie d'information dont bénéficie le secteur médical ne peut se corriger qu'à l'aide de politique de contrôle *ad hoc*) ; les logiques de contingentement quantitatif ne sont pas très efficaces et/mais, parallèlement, d'autres mesures apparaissent clairement de nature à réduire les inefficacités (par exemple, le remboursement forfaitaire corrige complètement la myopie du malade et il est possible d'offrir une couverture financière complète du risque maladie) ; les politiques d'enveloppe budgétaire allouée à la santé bénéficient à l'ensemble des ménages puisque la prime d'assurance maladie pratiquée *ex ante* est réduite des reversements.

Toutes ces conclusions sont bien évidemment liées aux spécifications adoptées. Ce travail n'est donc qu'une première étape dans projet plus général de modélisation systématique des comportements de santé par un modèle structurel macro-économique. De fait, de nombreuses améliorations seraient nécessaires : les notions de 'capital-santé', de consommations ostentatoires, de sensibilité aux vieillissement de la santé ne sont pas abordées. Enfin, les extensions immédiates du papier

renvoient à des questions d'économie publique : il serait souhaitable dans une logique normative d'identifier le premier rang et le second rang du modèle ; dans le même cadre d'analyse, il serait dès lors possible de déterminer un jeu de politiques économiques qui soit plus efficace : assurance complète avec paiement forfaitaire, transferts optimaux entre catégories d'agents pour réparer les distorsions, etc..

BIBLIOGRAPHIE

BATIFOULIER, P., 1999, "Ethique professionnelle et activité médicale : une analyse en termes de conventions", *Finance, Contrôle, Stratégie*, vol 2. juin.

BATIFOULIER, P., TOUZÉ, V., 2000, *La protection sociale*, Dunod.

BENARD, J., 1984, *Economie Publique*. Economica.

DIAMOND, P., 1965, "National Debt in a Neoclassical Growth Model", *American Economic Review*.

MOUGEOT, M., 1999, *Régulation du Système de Santé*, Rapport du Conseil d'Analyse Economique, La Documentation Française. Paris.

ROMER, P., 1991, "Progrès technique endogène", *Annales d'Eco et Stat*, n°22.

ROCHAIX, L., JACOBZONE, S., 1997, "L'hypothèse de demande induite : un bilan économique", *Economie et Prévision*, n°129-130.

TOUZÉ, V., 1999, "Insurance, Prevention and Macroeconomics", working-paper THEMA, 99-25.

TOUZÉ, V., VENTELOU, B., 2000, "The Macroeconomics of Health Care", OFCE working paper.

VENTELOU, B., 1999, "Les dépenses de santé des français : une maladie d'amour ?", *Revue de l'OFCE*, n° 71, Octobre.

WEISBROD, B.,1991, "The Health Care Quadrilemma : an Essay on Technological Change, Insurance, Quality of Care, and Cost Containment", *Journal of Economic Literature*, vol. XXIX, pp. 523-52.

Index des notations utilisées dans le modèle

X^+, X^-	Valeur de X en t+1 (resp. en t-1)
X^*	X a fait l'objet d'un arbitrage de l'agent
X'	X concerne les agents non malades en première période
X_m	X concerne les offreurs de soins
X_y	X concerne les agents en activité
X_o	X concerne les agents retraités
c	Consommation de l'agent lors de sa période d'activité
z	Consommation de l'agent lors de sa période d'inactivité
s	Epargne
k	Stock de capital agrégé
$h_{i,j}$ avec i =1, …, I et j=1,2	Pour un soin de type i, pour j = 1, activité de soins utiles (resp. inutile j = 2)
$l_{i,j}$ avec i =1, …, I et j=1,2	Pour un soin de type i, pour j = 1, temps de travail consacré aux soins utiles (resp. inutiles j = 2)
I(t)	Nombre de biens et services composant le secteur santé
I_{t+1}/I_t	Progrès technique de gamme
d_y, d'_y	Etat de santé des malades (resp. non malades) en première période
p	Niveau général des prix
w	Salaire
β	Facteur d'escompte du temps
δ_j, avec j=1,2	Coût subjectif associé à l'effort de travail médical utile (resp. inutile)
α	Proportion des biens médicaux utiles sur l'ensemble des biens médicaux
π	Prime d'assurance-maladie
λ	Taux de remboursement des soins ((1-λ) ticket modérateur)
J	Paiement forfaitaire de l'assurance-maladie
ψ	Proportion des ménages malades en première période
γ, $\gamma^{\#}$	Goût pour la santé des agents actifs (resp. des agents inactifs)
q	Proportion des offreurs de soins dans la population totale (1+q).
h^c	Offre de soins contingentée
E_y	Enveloppe de dépense médicale
R	Valeur des reversements par l'offreur de soins en cas de dépassement de l'enveloppe globale
τ	Taux de reversement

Annexe 1 : La contrôlabilité de soins hétérogènes. Qualité-Quantité.

D'une manière maintenant classique, la "demande induite" est définie comme la capacité des offreurs de soins à vendre *un bien de peu d'utilité* (si l'acheteur avait connaissance de la qualité du bien, il ne l'achèterait pas). Il s'agit fondamentalement d'une asymétrie d'information.

Au niveau macroéconomique, pour les besoins du modèle, l'absence d'information sur la *qualité* de l'offre médicale est ramenée à une incontrôlabilité, par l'acheteur, de la *composition* des soins achetés, chaque transaction de santé h comportant, en effet, une partie de *soins utiles* h_1 et *une partie inutile* h_2 ($h = h_1 + h_2$). Cette composition est déterminée par les producteurs du secteur sur un choix d'allocation du temps de travail. L'aspect inutile du soin se traduit par le fait que les consommateurs n'intègre pas la partie h_2 dans leur fonction d'utilité. On a $U(\text{santé}) = U(h-h_2) = U(h_1)$. Tout se passe comme si la distorsion entraînait une "production fatale" de soins, avec un gaspillage social lié à cette production : il y a d'abord un perte sèche de facteur travail (orientée vers une activité inutile) ; il y a ensuite un transfert de revenu des ménages vers les producteurs de soin qui n'est pas "optimal".

La distorsion est de nature microéconomique. On pourrait donc s'attendre à ce qu'elle engage une réflexion, de nature *microéconomique* elle aussi, à l'aide, notamment, de la *théorie des contrats* (visant à contrôler h_2, par exemple par des techniques de bonus-malus et/ou de contrats révélants). Nous n'avons pas exploité cette piste. Plusieurs arguments ont sous-tendu ce choix :

1. Ce n'est pas l'objet du présent papier[17], qui s'emploie à analyser la pertinence de mesures *macroéconomiques*, prédominantes dans certains pays. En France par exemple, on semble considérer que la surconsommation médicale n'est pas vraiment traitable par les outils microéconomiques (d'ailleurs, il est possible que cela soit, de fait, partiellement le cas, voir suite), et on recoure de manière systématique à ce qu'il est convenu de nommer la 'maîtrise macroéconomique' des dépenses.

2. Le second point tendant à limiter l'intérêt d'une démarche microéconomique incitative est *la difficulté à objectiver* au niveau individuel *une mesure de la qualité des soins de santé*. L'output du secteur de santé est extrêmement particulier : de fait, contrairement aux biens "*lemons*", l'acheteur peut très bien méconnaître durablement (voire toujours) la qualité réelle des soins achetés ; ce qui rend impossible l'élaboration d'un contrat de garantie. De même, si l'on veut appliquer la théorie des incitations, il est difficile de considérer que le soignant s'assimile à une entreprise publique en monopole (produisant un bien en quantité connue mais dont l'effort n'est pas observable) : l'entreprise n'est pas en monopole, elle ne produit pas un bien en quantité observable. Enfin, le fait que la relation soit tripartite (soigné, soignant, assureur) impliquerait de modifier la théorie classique des contrats en passant, par exemple, à des modèles *multi-principaux*.

3. Dernier point, la réforme des schémas de paiement en santé n'est pas, en l'état, réalisable. Les professions de santé sont sur une logique d'autocontrôle et de savoir d'experts. Cette modalité de régulation se substitue, de fait, à une régulation "microéconomique" par les contrats incitatifs. Il faudrait alors savoir si on peut croiser les deux modalités de régulation : incitations microéconomiques et règles déontologiques. En voulant gagner sur un terrain, le risque est de perdre de l'autre (voir Batifoulier, 1999).

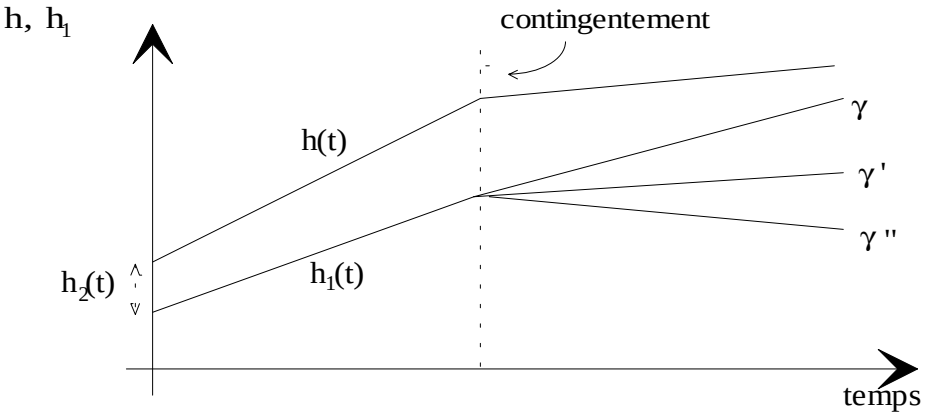
---

[17] Touzé, Ventelou, 2000 envisage plus précisément certaines techniques d'économie publique : first best et second best, contrôlabilité de la qualité.

Annexe 2 : La réaction de la "demande induite" à un contingentement

Les soins de santé de catégorie h_2 ne contribuent pas à l'état de santé de l'individu ; ils sont cependant, par hypothèse, nécessairement achetés par le consommateur de soins. La variable h_2 est une sorte de "produit fatal" de la consommation de santé, entrant dans le bien-être comme une externalité négative.

C'est, bien sûr, cette catégorie de dépenses h_2 – la "demande induite" – que le législateur cherche à réduire, notamment quand il utilise des politiques de contingentement. On peut définir une politique de contingentement de la manière suivante : on pose une valeur h^c et on considère que le législateur est capable d'imposer $p \, . \, h^c$ comme un plafond des dépenses totales[18]. On a : $p\,(h_1+h_2) \leq p\,h^c$. La contrainte de politique économique et l'évolution du panier de consommation médicale peuvent se représenter sur un graphique :



L'évolution γ est, bien évidemment, l'évolution souhaitée par le législateur : les dépenses de santé utiles h_1 sont sur un trend, trend qui n'est aucunement affecté par la mesure de contingentement ; les évolutions γ' et γ'' sont d'autres évolutions possibles : il existe des ré-allocations internes de l'offre médicale suite au contingentement, qui se traduisent par une "réponse" effective du trend de dépenses h_1 à la mesure de contingentement ; dans le cas γ'', la mesure de contingentement, loin de remplir son objectif (réduire la proportion de h_2 par rapport à h_1), conduit paradoxalement à une hausse de h_2/h_1. On peut établir le lemme 1.

Lemme 1. *L'élasticité des soins utiles à une politique de contingentement n'a aucune raison d'être égale à 0.*

Preuve : sous l'effet du contingentement, la maximisation du programme du secteur médical donne deux conditions :

$$\bar{h}_1'\left(\frac{l_1}{I}\right) = \frac{\delta_1}{\delta_2}\,\bar{h}_2'\left(\frac{l_2}{I}\right)$$

$$\left(I\,\bar{h}_1 + I\,\bar{h}_2\right) = h^c$$

La première équation est la condition d'arbitrage sur les choix d'allocation du travail médical ; la seconde équation reprend la contrainte de contingentement. Ces deux

[18] On définira plus tard diverses techniques de contingentement : contingentement pur, enveloppe globale, etc. (voir section 4).

équations nous permettent de calculer les niveaux de production h_1 et h_2 en fonction de h^c et des paramètres technologiques.

En prenant par exemple, $\bar{h}_1 = A_1 \log(\frac{l_1}{I})$; $\bar{h}_2 = A_2 \log(\frac{l_2}{I})$. On obtient :

$$l_2 = \frac{\delta_1}{\delta_2} \cdot \frac{A_2}{A_1} \cdot l_1$$

et

$$I \cdot A_1 \log(\frac{l_1}{I}) + I \cdot A_2 \log(\frac{\delta_1}{\delta_2} \cdot \frac{A_2}{A_1} \cdot l_1) = h^c$$

Où l'on voit clairement que la dérivée $\frac{\partial l_1}{\partial h^c}$ n'est pas nulle, mais positive.

Tout aggravation du contingentement (baisse de h^c) se traduit aussi par une réduction de la quantité de travail utile (baisse de l_1). CQFD.