Impact du passage à la T2A: une modélisation pour l'hopital public

David Crainich
CNRS et LEM (UMR 8179),
Université Catholique de Lille.
email: dcrainich@cresge.fr

Hervé Leleu
CNRS et LEM (UMR 8179),
Université Catholique de Lille.
email: hleleu@cresge.fr

Ana Mauleon
FNRS, CEREC (Facultés Universitaires Saint-Louis)
et CORE.
email: mauleon@fusl.ac.be

November 22, 2007

Abstract

Nous analysons les conséquences du passage d'un financement par dotation globale à un financement prospectif lié à l'activité et tentons de déterminer en quoi la réforme pourrait être préjudiciable ou bénéfique à la santé financière des hôpitaux publics. Nous développons un modèle théorique qui lie la demande des patients aux décisions conjointes des médecins et des managers d'hôpitaux pour analyser la façon dont le nouveau mode de financement affecte les décisions prises par l'ensemble des agents et donc l'équilibre budgétaire des établissements.

Classification JEL: I18, D21

Mots-clés: Tarification à l'activité, Hôpital public, Equilibre budgétaire

1 Introduction

Avant 2004, le système de la dotation globale concernait les hôpitaux publics et les hôpitaux privés participant au service public (PSPH). D'après la Mission Tarification à l'activité du Ministère de la Santé, ce système de financement générait des effets pervers en ne liant que très faiblement le budget à l'activité réalisée et en conduisant soit à la constitution de rentes de situation, soit à un manque de financement pour les structures les plus actives. Depuis lors, le système de tarification à l'activité (T2A) a été introduit en France en 2004 avec une mise en place progressive du financement prospectif qui devrait aboutir en 2012.

Aux Etats-Unis, les systèmes de paiements prospectifs sont considérés dans le secteur hospitalier comme supérieurs aux autres systèmes (Ma 1994, Newhouse 1996). En France, les bénéfices attendus sont de plusieurs ordres: une plus grande médicalisation du financement, une responsabilisation des acteurs et une incitation à s'adapter, une équité de traitement entre les secteurs et le développement des outils de pilotage médico-économique (contrôle de gestion) au sein des hôpitaux publics et privés. Au-delà de ces objectifs qui visent à améliorer l'efficacité et la gestion des hôpitaux publics, rien ne garantit dans la réforme T2A la pérennité financière des établissements.

Dans ce travail, nous analysons les conséquences du passage d'un financement par dotation globale à un financement prospectif lié à l'activité et tentons de déterminer en quoi la réforme pourrait être préjudiciable ou bénéfique à la santé financière des hôpitaux publics. Nous développons un modèle théorique qui lie la demande des patients aux décisions conjointes des médecins et des managers d'hôpitaux pour analyser la façon dont le nouveau mode de financement affecte les décisions prises par l'ensemble des agents et donc l'équilibre budgétaire des établissements.

Nous proposons, à la suite des travaux de Custer et al. (1990), Dor et Watson (1995) et de Boadway et al. (2004), une modélisation adaptée au cas Français. L'analyse est centrée sur un couple composé d'un hôpital et d'un médecin représentatif coordonnant leurs actions afin de produire conjointement des séjours hospitaliers homogènes. Ces séjours hospitaliers s'adressent à des individus sensibles à l'effort déployé par le médecin, à la durée d'hospitalisation et aux services fournis par l'hôpital pour améliorer la prise en charge et le confort des patients.

Le médecin, qui décide de son niveau d'effort et de la durée de séjour, cherche à maximiser sa propre fonction objectif qui dépend du niveau de demande et de sa désutilité à l'effort. L'utilité du médecin est donc liée positivement à la satisfaction des patients via la demande et négativement au niveau d'effort qu'il fournit pour soigner les patients. Nous considérons également que cette désutilité est une fonction du niveau d'input mis à disposition par l'hôpital. Pour son activité à l'hôpital public, le médecin est supposé être complètement salarié de sorte que son utilité n'est pas liée à sa rémunération.

L'hôpital public, au travers de son manager, décide du niveau d'input qu'il met à disposition des médecins et des patients. Sa fonction d'utilité est supposée dépendre de la demande exprimée par les patients. Son objectif est donc la maximisation de la satisfaction des patients sous la contrainte de maintenir son équilibre budgétaire. Les coûts supportés par l'hôpital sont liés au niveau d'effort du médecin, à la durée d'hospitalisatione et au niveau d'input qu'il met à disposition des patients. Les décisions de l'hôpital et du médecin sont considérées comme simultanées et étudiées dans le cadre d'un équilibre à la Cournot/Nash. La dépense totale de l'hôpital est soumise à une contrainte budgétaire qui dépend du mode de financement imposé par les pouvoirs publics. Le budget est considéré comme fixe et exogène au modèle dans le cas d'un financement par dotation globale et comme variable et endogène au modèle dans le cas d'une tarification à l'activité.

Aprés une présentation du modèle dans la première section, nous analysons l'impact du passage à la T2A sur les décisions du médecin et de l'hôpital. Nous étudions en particulier les conditions dans lesquelles l'hôpital est capable d'accroître son activité tout en maintenant son équilibre budgétaire. Pour analyser uniquement les effets liés au changement de financement, nous neutralisons dans cette section les effets dus au volume budgétaire (en considérant que le tarif T2A est égal au coût moyen du séjour sous le financement précédent). Nous relâchons cette hypothèse dans la dernière section dans laquelle nous étudions le passage à la T2A sans neutralité budgétaire avec un tarif favorable ou défavorable à l'hôpital public. Nous étudions également l'impact de chocs exogènes comme une réduction de la demande ou une augmentation des coûts sur l'équilibre budgétaire de l'hôpital après la réforme T2A.

2 Présentation du modèle

A l'image de Dor et Watson (1995), la modélisation est centrée sur un couple composé d'un hôpital et d'un médecin représentatif coordonnant leurs actions afin de produire conjointement des séjours hospitaliers homogènes. Ces séjours hospitaliers s'adressent à des individus sensibles à l'effort déployé par le médecin pour les soins prodigués (noté e et qui résume les actes médicaux et/ou chirurgicaux, les consultations pré ou post opératoires...), à la durée d'hospitalisation (h) et au montant des inputs mis à disposition par l'hôpital pour améliorer la prise en charge et le confort des patients lors de leur séjour hospitalier (noté q et qui inclut les équipements des salles et ceux associés aux lits ainsi que les personnels soignants, administratifs ou médico-techniques). L'activité de l'hôpital est soumise à une contrainte budgétaire.

Nous considérons une fonction de production de services de santé $S(e, h, q)$ qui dépend du niveau d'effort du médecin, de la durée de séjour et du niveau d'input mis à disposition par l'hôpital. Cette fonction est considérée comme croissante, continue et concave sur e et q. Ce qui signifie que le service de santé proposé au patient s'améliore à mesure que e et q augmentent, mais leur apport est d'autant moins important que leur niveau initial est déjà élevé. D'où:

$$S(e, h, q) \text{ avec } S_e > 0, S_q > 0, S_{ee} < 0, S_{qq} < 0 \tag{1}$$

Concernant la durée d'hospitalisation, nous considérons qu'il existe une durée de séjour au-delà de laquelle le niveau de production de santé diminue (par exemple par l'augmentation du risque d'infection nosocomiale, d'apparition d'escarres...). D'où:

$$\overline{h} = \left\{ h \in \mathbb{R}^+ \mid S_h = 0, S_{hh} < 0 \right\} \tag{2}$$

Cette durée de séjour correspond à la durée d'hospitalisation optimale du point de vue du patient.

La demande à laquelle font face le médecin et l'hôpital dépend directement du niveau des services de santé qu'ils proposent à chaque patient, de sorte que :

$$D = D(S(e, h, q)) \tag{3}$$

Puisque D est une fonction croissante et concave en S et que S est une fonction croissante et concave en e et q alors nous pouvons réécrire la demande des patients comme:

$$D(e, h, q) \; avec \; D_e > 0, D_q > 0, D_{ee} < 0, D_{qq} < 0 \tag{4}$$

Par rapport à la durée d'hospitalisation, la demande est d'abord croissante en h, atteint un maximum en $\overline{h}$ et diminue ensuite. Nous pourrions l'illustrer par exemple par un graphe apparenté à une loi normale.

Le médecin, qui décide de son niveau d'effort et de la durée d'hospitalisation, cherche à maximiser sa propre fonction objectif qui dépend du niveau de demande et de sa désutilité à l'effort. L'utilité du médecin est donc liée positivement à la satisfaction des patients via la demande et négativement au niveau d'effort qu'il fournit pour soigner les patients. Cette désutilité à l'effort $R(e(q))$ comprend le coût d'opportunité du temps passé à l'hôpital et qu'il ne peut consacrer à une activité libérale ou à ses loisirs. Ce coût dépend du niveau d'input de l'hôpital. La désutilité est supposée être convexe en e de telle sorte que le supplément de désutilité engendré par un effort supplémentaire est d'autant plus élevé que le niveau initial d'effort est déjà important ($R_e > 0$ et $R_{ee} > 0$). La désutilité est également liée au niveau d'input q mis à disposition par l'hôpital. Nous supposons en effet que les médecins valorisent éventuellement davantage leur travail dans un environnement favorable et que la durée des services médicaux peut être réduite si le médecin bénéficie d'un plateau technique plus développé ($R_{eq} \leq 0$). Pour son activité à l'hôpital public, le médecin est supposé être complétement salarié de sorte que son utilité n'est pas liée à sa rémunération. Sous ces hypothèses, sa fonction objectif peut être exprimée comme suit:

$$U^M(e, h, q) = D(e, h, q) - R\left(e(q)\right) \tag{5}$$

De son côté, l'hôpital public décide du niveau d'input qu'il met à disposition des médecins et des patients. Sa fonction d'utilité est supposée dépendre de la seule demande exprimée par les patients. Son objectif est donc directment lié à la satisfaction des patients:

$$U^H(e,h,q) = D(e,h,q) \tag{6}$$

Les coûts supportés par l'hôpital et liés à la durée d'hospitalisation d'un patient, $CH(h)$, (frais d'hébergement, frais de blanchisserie, frais de restauration...) sont considérés comme linéaire en h ($CH(h) = c_h.h$). Les mêmes hypothèses sont faites pour les coûts liés aux soins prodigués par les médecins à un patient: $CM(e)$ est une fonction linéaire en e ($CM(e) = c_e.e$). L'hôpital supporte également des coûts fixes liés au niveau d'input qu'il met à disposition des patients ($CF(q)$, avec $CF(q) = c_q.q$). La dépense totale de l'hôpital est soumise à une contrainte budgétaire imposée par les pouvoirs publics. Elle est définie comme suit:

$$D(e,h,q)\,[CH(h) + CM(e)] + CF(q) \leq B \tag{7}$$

Le budget, B, dépend crucialement du mode de financement de l'hôpital. Il peut être considéré comme fixe et exogène au modèle dans le cas d'un financement par dotation globale ou variable et endogène au modèle ($B = D(e,h,q) * \overline{T}$) dans le cas d'une tarification à l'activité avec $\overline{T}$ comme tarif associé à un séjour. Nous évaluons le passage d'un financement par budget global vers un financement à l'activité sur les décisions du médecin et de l'hôpital. Ces décisions sont considérées dans un cadre non coopératif, dans lequel à la fois le médecin et l'hôpital adoptent un comportement passif, c'est-à-dire ne tirent pas parti de leur capacité à influer, à travers leur propre choix, sur ceux de leur partenaire. Ceci conduit à un équilibre de Cournot/Nash.

3 Impact du passage à la T2A

Nous commençons par l'analyse de l'équilibre de Cournot/Nash dans le cadre d'un financement par budget global. Nous étudions ensuite quel sera l'impact du passage à la T2A pour le médecin et l'hôpital.

3.1 Equilibre de Cournot/Nash sous un financement par budget global

Le programme d'optimisation du médecin peut être défini comme suit:

$$\underset{e,h}{Max}\, U^M(e,h,q) = D(e,h,q) - R\left(e(q)\right) \tag{8}$$

Les conditions de premier ordre correspondant à ce programme d'optimisation sont alors les suivantes:

$$D_e = R_e \tag{9}$$

$$D_h = 0 \tag{10}$$

Les deux conditions de second ordre nous indiquent bien que nous sommes en un maximum:

$$D_{ee} - R_{ee} < 0 \tag{11}$$

$$D_{hh} < 0 \tag{12}$$

Le médecin ne faisant face à aucune contrainte maximise son utilité en arbitrant entre la demande marginale liée à son niveau d'effort et la désutilité marginale produite par cet effort. Il maximise également son utilité s'il choisit la durée d'hospitalisation qui maximise la demande des patients ($\widehat{h}_{bg} = \overline{h}$). Notons qu'il considère le niveau d'input de l'hôpital comme donné.

Le programme d'optimisation de l'hôpital peut, quant à lui, être défini de la façon suivante:

$$\begin{gathered} \underset{q}{Max}\, U^H(e,h,q) = D(e,h,q) \\ D(e,h,q)\,[CH(h) + CM(e)] + CF(q) \leq B \quad (\lambda) \end{gathered} \tag{13}$$

Les conditions de premier ordre correspondant à ce programme d'optimisation sont alors les suivantes:

$$D_q - \lambda\,[D_q\,[CH(h) + CM(e)] + CF_q] = 0 \tag{14}$$

$$D(e,h,q)\,[CH(h) + CM(e)] + CF(q) \leq B \tag{15}$$

De la première CPO, on tire:

$$\lambda = \frac{D_q}{D_q\,[CH(h) + CM(e)] + CF_q} > 0 \tag{16}$$

La dernière inégalité provient des hypothèses faites sur la fonction de demande et sur les fonctions de coût. La positivité du multiplicateur indique clairement que la contrainte budgétaire sera toujours saturée:

$$D(e,h,q)\,[CH(h) + CM(e)] + CF(q) = B \tag{17}$$

L'interprétation de cette égalité est claire: l'hôpital a toujours intérêt à augmenter son niveau d'input pour maximiser sa fonction objectif et donc il utilisera la totalité de son budget.

Notons que les conditions de premier ordre du médecin et de l'hôpital dépendent en partie des variables de décision de l'autre acteur. Pour le médecin, un exercice de statique comparative peut nous fournir le sens de la relation qui lie le niveau d'input sous contrôle de l'hôpital à ses deux variables de décision. Concernant l'hôpital, les choix du médecin et le budget défini par les pouvoirs publics vont aussi influencer son niveau d'input. Nous calculons en premier lieu les dérivées partielles des CPO du médecin par rapport à chacune des variables:

$$\frac{\partial CPO}{\partial e} = D_{ee} - R_{ee} < 0 \tag{18}$$

$$\frac{\partial CPO}{\partial h} = D_{hh} < 0 \tag{19}$$

$$\frac{\partial CPO}{\partial q} = -R_{eq} > 0 \tag{20}$$

Le signe des dérivées partielles découle directement des hypothèses faites sur l'allure des fonctions de demande et de désutilité.

L'impact d'une variation de q sur e est définie par:

$$de = -dq \left(\frac{\frac{\partial CPO)}{\partial e}}{\frac{\partial CPO}{\partial q}} \right) \geq 0 \tag{21}$$

Une augmentation du niveau d'input de l'hôpital induit toute chose égale par ailleurs une augmentation du niveau d'effort du médecin.

L'impact d'une variation de q sur h est donc définie par:

$$dh = -dq \left(\frac{\frac{\partial CPO)}{\partial h}}{\frac{\partial CPO}{\partial q}} \right) = 0 \tag{22}$$

Le médecin fixe la durée de séjour indépendemment du niveau d'input de l'hôpital en ne cherchant qu'à maximiser la production de santé.

Pour l'hôpital, nous calculons les dérivées partielles de la contrainte budgétaire par rapport à chacune des variables en posant $C_{bg}(e,h,q) = D(e,h,q)\,[CH(h) + CM(e)] + CF(q) - B$:

$$\frac{\partial C_{bg}(e,h,q)}{\partial e} = CM_e D(e,h,q) + D_e(CH(h) + CM(e)) > 0 \tag{23}$$

$$\frac{\partial C_{bg}(e,h,q)}{\partial h} = CH_h D(e,h,q) + D_h(CH(h) + CM(e)) > 0 \tag{24}$$

$$\frac{\partial C_{bg}(e,h,q)}{\partial q} = CF_q + D_q(CH(h) + CM(e)) > 0 \tag{25}$$

La positivité des dérivées partielles découle directement des hypothèses faites sur l'allure des fonctions de demande et de coût.

Sachant que la contrainte doit toujours être respectée, nous en déduisons que la différentielle totale est nulle:

$$de\frac{\partial C_{bg}(e,h,q)}{\partial e} + dh\frac{\partial C_{bg}(e,h,q)}{\partial h} + dq\frac{\partial C_{bg}(e,h,q)}{\partial q} = 0 \tag{26}$$

L'impact d'une variation positive du niveau d'effort du médecin sur le niveau des inputs en considérant une durée d'hospitalisation fixe est mesuré par:

$$dq = -de \left(\frac{\frac{\partial C_{bg}(e,h,q)}{\partial e}}{\frac{\partial C_{bg}(e,h,q)}{\partial q}} \right) < 0 \tag{27}$$

Une augmentation du niveau d'effort de la part du médecin induit toute chose égale par ailleurs une diminution du niveau d'input engagé par l'hôpital.

L'impact d'une variation positive de la durée d'hospitalisation sur le niveau des inputs en considérant un niveau d'effort du médecin fixe est mesuré par:

$$dq = -dh \left(\frac{\frac{\partial C_{bg}(e,h,q)}{\partial h}}{\frac{\partial C_{bg}(e,h,q)}{\partial q}} \right) < 0 \tag{28}$$

Une augmentation de la durée d'hospitalisation induit toute chose égale par ailleurs une diminution du niveau d'input engagé par l'hôpital.

L'impact conjoint d'une variation positive de la durée d'hospitalisation et du niveau d'effort du médecin sur le niveau des inputs est mesuré par:

$$dq = -\left(\frac{de\frac{\partial C_{bg}(e,h,q)}{\partial e} + dh\frac{\partial C_{bg}(e,h,q)}{\partial h}}{\frac{\partial C_{bg}(e,h,q)}{\partial q}} \right) < 0 \tag{29}$$

Une augmentation de la durée d'hospitalisation et du niveau d'effort du médecin induit toute chose égale par ailleurs une diminution du niveau d'input engagé par l'hôpital. L'effet d'une variation contraire de la durée d'hospitalisation et du niveau d'effort du médecin reste cependant ambigu.

En considérant simultanément les deux programmes d'optimisation du médecin et de l'hôpital, nous pouvons obtenir l'équilibre de Cournot sous budget global. La solution est fournie par le triplet $(\widehat{e}_{bg}, \widehat{h}_{bg}, \widehat{q}_{bg})$ qui satisfait à la fois aux conditions d'optimisation du médecin et de l'hôpital. Les propriétés de courbure des fonctions de demande, de désutilité et de coût permettent d'assurer l'unicité de cet équilibre de Cournot. Ce modèle fournira le benchmark pour apprécier l'impact du passage à une tarification à l'activité. En fait, nous allons étudier en quoi la T2A va modifier la solution du modèle de Cournot en budget global pour mesurer ses effets sur les variables d'intérêts (e, h, q) et sur les fonctions d'utilité et de demande.

3.2 Equilibre de Cournot/Nash sous T2A

Nous abordons dans cette section la question de l'impact du passage d'un financement par budget global à une tarification à la pathologie sur le comportement du couple hôpital-médecin dans le cadre d'un équilibre de Cournot. Le passage à la T2A considéré ici est totalement neutre du point de vue de la rémunération par séjour hospitalier. En effet, nous nous intéressons principalement aux modifications de comportement des acteurs face au changement de mode de tarification et non face à un éventuel changement du niveau global de dotation. Nous retenons donc comme tarif à l'activité le budget initial de l'hôpital divisé par la demande optimale dans le modèle sous budget global:

$$\overline{T} = \frac{B}{D(\widehat{e}_{bg}, \widehat{h}_{bg}, \widehat{q}_{bg})} \tag{30}$$

Le passage à la T2A concerne en premier lieu l'hôpital qui est soumis à la contrainte budgétaire. Le programme d'optimisation de l'hôpital devient :

$$\begin{gathered} \underset{q}{Max}\, U^H(e,h,q) = D(e,h,q) \\ D(e,h,q)\left[CH(h) + CM(e)\right] + CF(q) \leq D(e,h,q)\overline{T} \quad (\lambda^{t2a}) \end{gathered} \tag{31}$$

La première CPO est donnée par:

$$D_q - \lambda^{t2a}\left[D_q\left[CH(h) + CM(e) - \overline{T}\right] + CF_q\right] = 0 \tag{32}$$

Tout comme pour le cas d'un financement par budget global, l'hôpital a toujours intérêt à saturer sa contrainte budgétaire et sa condition d'optimisation est donc:

$$D(e,h,q)\left[CH(h) + CM(e)\right] + CF(q) = D(e,h,q)\overline{T} \tag{33}$$

Le changement induit par la T2A sur le comportement de l'hôpital porte sur cette relation où le budget exogène est remplacé par un financement endogène dépendant de la demande. La proposition suivante résume l'impact d'un passage du budget global à une tarification à l'activité.

Proposition 1 *L'impact de la T2A sur le comportement de l'hôpital public dépend de l'élasticité de la demande par rapport au niveau d'input. Si cette élasticité est supérieure à l'unité alors l'hôpital augmente son niveau d'input et tire profit du passage à une tarification à l'activité. Cela a pour conséquence d'augmenter l'effort du médecin et la satisfaction des patients. L'impact pour les pouvoirs publics est une augmentation des dépenses de santé. A l'inverse, si l'élasticité de la demande par rapport au niveau d'input est inférieure à l'unité alors le statu quo prévaut et la T2A n'a pas d'impact sur l'hôpital public.*

Preuve:

La variable de contrôle de l'hôpital étant le niveau d'input q, nous allons étudier la variation de la contrainte budgétaire par rapport à une variation de q. En posant:

$$SF_{t2a}(e,h,q) = D(e,h,q)\left[\overline{T} - CH(h) - CM(e)\right] - CF(q)$$

$SF_{t2a}(e,h,q)$ s'interprète comme le solde financier de l'hôpital (revenu - coût). La contrainte budgétaire impose que ce solde soit toujours positif ou nul. La variation de ce solde par rapport à une augmentation du niveau d'input est donné par :

$$\frac{\delta SF_{t2a}(e,h,q)}{\delta q} = D_q\left[\overline{T} - CH(h) - CM(e)\right] - CF_q \tag{34}$$

Or, d'après(33), nous avons à l'équilibre budgétaire:

$$\left[\overline{T} - CH(h) - CM(e)\right] = \frac{CF(q)}{D(e,h,q)} \tag{35}$$

et d'après la définition des coûts fixes, on a:

$$CF(q) = c_q.q \tag{36}$$

$$CF_q = \frac{\delta CF(q)}{\delta q} = c_q \tag{37}$$

En insérant ces expressions dans (34), on obtient:

$$\begin{aligned}
\frac{\delta SF_{t2a}(e,h,q)}{\delta q} &= D_q \left(\frac{c_q.q}{D(e,h,q)} \right) - c_q \qquad (38)\\
&= c_q \left(D_q \frac{q}{D(e,h,q)} - 1 \right) \qquad (39)\\
&= c_q \left(\epsilon_{D/q} - 1 \right) \qquad (40)
\end{aligned}$$

où $\epsilon_{D/q}$ est l'élasticité de la demande par rapport au niveau d'input de l'hôpital. Finalement, comme le coût fixe unitaire est strictement positf, $c_q > 0$, on obtient le résultat suivant:

$$\begin{aligned}
\frac{\delta SF_{t2a}(e,h,q)}{\delta q} &> 0 \Leftrightarrow \epsilon_{D/q} > 1 \qquad (41)\\
\frac{\delta SF_{t2a}(e,h,q)}{\delta q} &= 0 \Leftrightarrow \epsilon_{D/q} = 1 \qquad (42)\\
\frac{\delta SF_{t2a}(e,h,q)}{\delta q} &< 0 \Leftrightarrow \epsilon_{D/q} < 1 \quad \blacksquare \qquad (43)
\end{aligned}$$

Par conséquent, par rapport à l'équilibre en budget global, l'hôpital peut augmenter son utilité (en augmentant son niveau d'input car la demande est sensible à ce dernier) si l'élasticité de la demande par rapport au niveau d'input est supérieure à l'unité. L'intuition économique de ce résultat est relativement simple: en augmentant son niveau d'input, l'hôpital agit positivement sur la demande et donc obtient un supplément de revenu mais il augmente également son coût variable et son coût fixe. Dans le cas où le revenu supplémentaire généré par l'augmentation de la demande est supérieur au coût supplémentaire, l'hôpital à intérêt à augmenter son niveau d'input puisqu'il satisfera sa contrainte budgétaire tout en augmentant son utilité. La question qui suit est naturellement jusqu'à quel point l'hôpital peut augmenter son niveau d'input.

Corollaire 1 . Après le passage à la T2A, le niveau d'équilibre de l'input de l'hôpital public est tel que l'élasticité de la demande par rapport à ce niveau d'input est inférieure à l'unité.

Preuve:

Si l'élasticité de la demande par rapport au niveau d'input est inférieure à l'unité avant le passage à la T2A, nous avons vu que le statu quo prévalait. Si l'élasticité est supérieure à l'unité alors il faut s'intéresser à la variation de l'élasticité par rapport à la variation du niveau d'input:

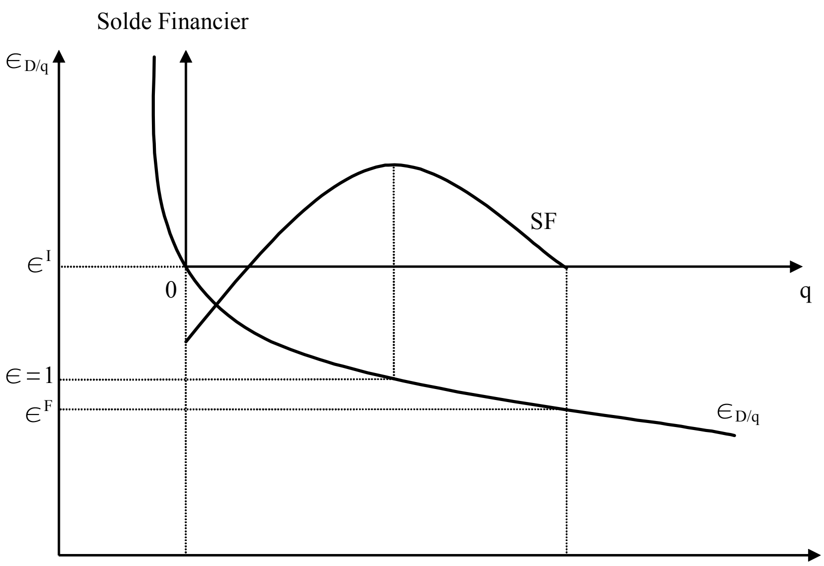
$$\frac{\delta \epsilon_{D/q}}{\delta q} = \frac{\delta \left(D_q \frac{q}{D(e,h,q)} \right)}{\delta q} = D_{qq} \frac{q}{D(e,h,q)} + D_q \frac{D(e,h,q) - D_q q}{D(e,h,q)^2} = D_{qq} \frac{q}{D(e,h,q)} + \frac{D_q}{D(e,h,q)} (1 - \epsilon_{D/q}) \tag{44}$$

Comme au départ l'élasticité est supérieure à l'unité, cette variation est négative car le premier terme de la somme est aussi négatif. Par conséquent, l'élasticité diminue avec l'augmentation du niveau d'input et deviendra inférieure à l'unité. Dans cette zone, le signe de l'expression (44) devient indéterminé et l'élasticité pourrait augmenter avec q. Cependant, elle ne repasse jamais au-dessus de 1. En effet, dans la zone croissante de (44), il est facile de montrer que:

$$\epsilon_{D/q} < 1 + D_{qq}\frac{q}{D_q} \text{ avec } D_{qq}\frac{q}{D_q} < 0 \text{ d'où } \epsilon_{D/q} < 1 \tag{45}$$

Le niveau d'input optimal sera celui qui rééquilibrera la contrainte budgétaire de l'hôpital. En effet, tant que l'élasticité est supérieure à l'unité, l'augmentation du niveau d'input améliore le solde de l'hôpital (revenu marginal supérieur au coût marginal). Lorsque l'élasticité devient inférieure à l'unité, le solde marginal est négatif. L'hôpital poursuit donc l'augmentation du niveau d'input jusqu'à équilibrer son solde. Au niveau d'input optimal, l'élasticité de la demande est donc inférieure à l'unité. ■

Ce résultat est illustré par la figure suivante (cas 1). Cette figure nous servira d'outil d'analyse pour étudier les chocs exogènes considérés dans la section suivante. Nous représentons à la fois l'évolution de l'élasticité et du solde financier avec le niveau d'input q.



Effet du passage à la T2A - cas 1

La figure comporte donc un double système d'axe, le premier lié aux variations de $\epsilon_{D/q}$ (axe extérieur), le second associé au solde financier de l'hôpital (axe intérieur). L'interprétation des courbes doit se faire par rapport à l'élasticité unitaire ($\epsilon = 1$). La situtation initiale est ici caractérisée par une élasticité supérieure à 1 ($\epsilon^I > \epsilon = 1$) associée à un équilibre budgétaire (revenu net global nul) au niveau initial q^I. Lorsque q augmente, l'élasticité diminue, mais tant qu'elle reste supérieure à 1, le revenu net global de l'hôpital augmente et atteint son maximum pour

($\epsilon = 1$). Ensuite, dans la zone où l'élasticité est inférieure à l'unité, le revenu net global diminue pour retrouver l'équilibre budgétaire en q^F. Nous vérifions donc qu'en ce point l'élasticité finale est inférieure à l'unité ($\epsilon^F < 1$).

Après avoir traité de l'impact du passage à la T2A pour l'hôpital, il faut s'interroger sur son impact concernant le médecin. A priori, la contrainte budgétaire n'affecte pas le médecin et il n'a donc pas de réaction directe au passage à la T2A. Cependant, l'hôpital a modifié son comportement en augmentant son niveau d'input et donc la demande. Le médecin pourrait donc éventuellement réagir à cette augmentation de la demande en ajustant son effort et la durée d'hospitalisation. Au vu des conditions (9) et (10), la demande n'intervient pas directement mais elle pourrait avoir un impact au travers des demandes marginales par rapport à e et h. Il faut donc analyser les variations de e et de h par rapport à la variation de q. D'après (21) et (22), on a le résultat suivant:

$$\frac{de}{dq} \geq 0 \text{ et } \frac{dh}{dq} = 0$$

Par conséquent, le médecin ne modifie pas la durée d'hospitalisation des patients mais peut augmenter son niveau d'effort. Deux situations sont envisageables. D'une part, la désutilité du médecin peut ne pas être sensible au niveau d'input de l'hôpital et dans ce cas, nous aurons $\frac{de}{dq} = 0$ et le comportement du médecin sera insensible au passage à la T2A et à la réaction de l'hôpital. La condition sur l'élasticité de la demande déterminera donc entièrement l'issue du passage à la T2A. D'autre part, le médecin peut être sensible au niveau d'input de l'hôpital ce qui l'incite à faire plus d'effort $\frac{de}{dq} > 0$. Dans ce cas, la contrainte budgétaire de l'hôpital s'en trouve modifiée puisque l'augmentation de l'effort du médecin va augmenter le coût associé pour l'hôpital. Il faut donc analyser si un nouvel équilibre se produit en étudiant les fonctions de réaction des deux agents. Nous présentons cette analyse en annexe. Il en ressort qu'un nouvel équilibre s'établit dès lors que la réaction du médecin n'est pas excessivement sensible à la modification du niveau d'input de l'hôpital.

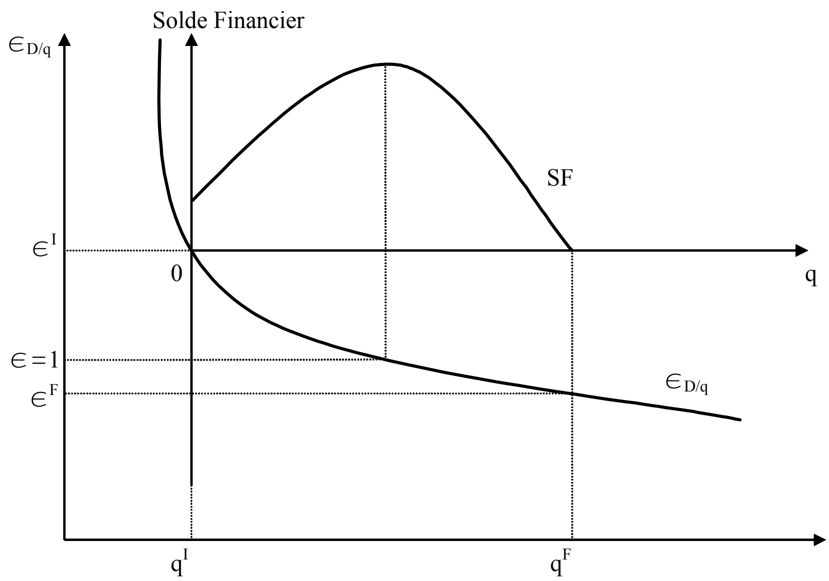
4 T2A et chocs exogènes

Dans cette section, nous analysons le comportement de l'hôpital public face à des chocs exogènes. Nous étudions plus particulièrement le passage à la T2A avec un tarif qui ne respecterait pas la neutralité budgétaire, la capacité de l'hôpital à faire face à une variation exogène de sa demande (qui peut provenir par exemple d'une perte de parts de marché face au secteur privé) ou à une variation exogène des coûts de production (résultant par exemple d'une augmentation du coût d'un traitement après une innovation technologique ou thérapeutique sans ajustement du tarif).

4.1 Impact du passage à la T2A sans neutralité budgétaire

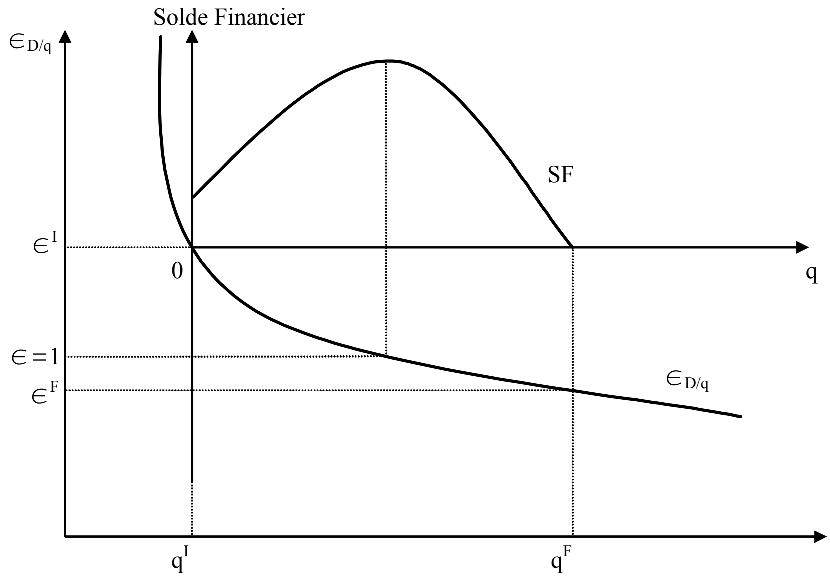
4.1.1 Le passage à la T2A est budgétairement favorable

L'élasticité de la demande par rapport à l'input de l'hôpital est supérieure à 1 Dans le cas où le tarif est supérieur au tarif en neutralité budgétaire ($\overline{\overline{T}} > \overline{T} = \frac{B}{D(e,h,q)}$), la situation est favorable à l'hôpital qui bénéficie d'un revenu net global positif lors du passage à la T2A. Ce cas est relativement similaire au cas 1 mise à part la situtation initiale. Comme décrit à la figure représentant le cas 2, l'hôpital réagit en augmentant son input jusqu'au niveau q^F. Dans ce cas, la demande augmente et l'hôpital est à son équilibre budgétaire. Par rapport au cas de référence (cas 1 avec neutralité budgétaire), le niveau d'input proposé par l'hôpital est plus important.



Effet du passage à la T2A - cas 2

L'élasticité de la demande par rapport à l'input de l'hôpital est inférieure à 1 Ici encore, le passage à la T2A est budgétairement favorable à l'hôpital public ($\overline{\overline{T}} > \overline{T} = \frac{B}{D(e,h,q)}$). Puisqu'il se trouve dans une zone où son élasticité est inférieure à l'unité, l'hôpital se contente d'épuiser son budget. Il en résulte une augmentation du niveau d'input jusqu'au niveau q^F (voir figure représentant le cas 3). La demande augmente et l'hôpital est à son équilibre budgétaire. Par rapport aux deux cas précédents (cas 1 avec neutralité budgétaire et cas 2 avec une élasticité de départ supérieure à l'unité), le niveau d'input proposé par l'hôpital est moins important.



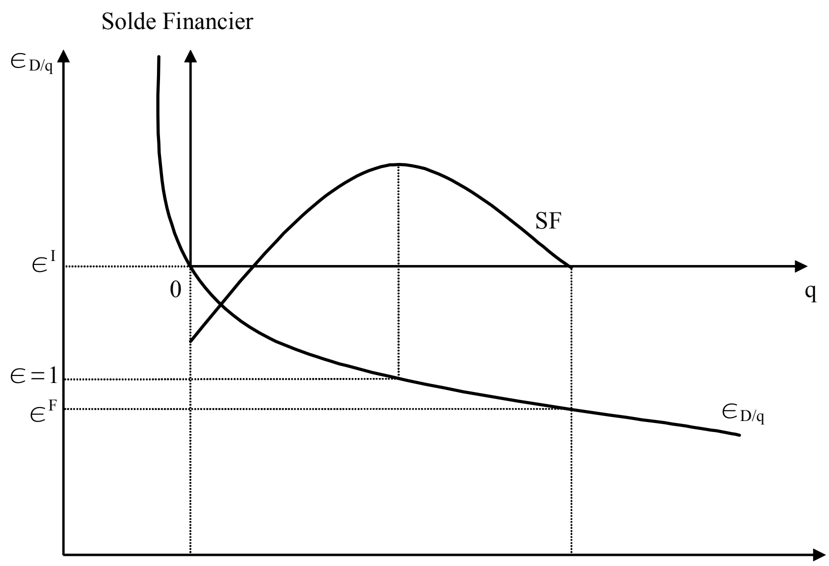
Effet du passage à la T2A - cas 3

4.1.2 Le passage à la T2A est budgétairement défavorable

Le tarif est inférieur au coût variable Lorsque le passage à la T2A est budgétairement défavorable ($\overline{\overline{T}} < \overline{T} = \frac{B}{D(e,h,q)}$), c'est-à-dire lorsque l'hôpital ne couvre pas son coût total compte tenu des valeurs d'équilibre de départ et que de surcroît le tarif est tel que l'hôpital ne couvre même pas ses coûts variables ($\overline{\overline{T}} < CH(h) + CM(e)$), l'hôpital se trouve dans l'impossibilité de retrouver son équilibre budgétaire et est donc en déficit.

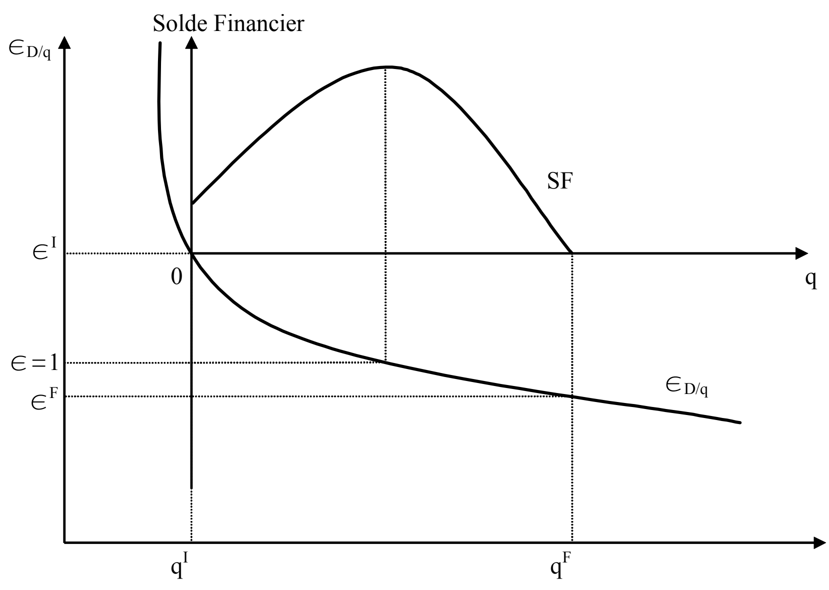
Le tarif est supérieur au coût variable Nous considérons maintenant la situation où le passage à la T2A est budgétairement défavorable à l'hôpital ($\overline{\overline{T}} < \overline{T} = \frac{B}{D(e,h,q)}$) qui peut néanmoins couvrir ses coûts variables grâce au tarif proposé ($\overline{\overline{T}} > CH(h) + CM(e)$). Quatre cas sont à envisager.

1. L'élasticité de la demande par rapport à l'input est supérieure à 1 et l'hôpital récupère son équilibre budgétaire Au passage à la T2A, l'hôpital se retrouve en déficit budgétaire s'il maintient son input initial (q^I). Mais puisqu'il démarre avec une élasticité de la demande par rapport à l'input qui est favorable ($\epsilon^I > \epsilon = 1$), il peut augmenter son niveau d'input en vue de dégager des marges budgétaires. S'il parvient à retrouver l'équilibre budgétaire dans une zone où l'élasticité de la demande par rapport à l'input est supérieure à l'unité (voir figure représentant le cas 4), l'hôpital peut continuer à augmenter son input jusqu'en q^F où son budget redevient équilibré. Par rapport aux deux premiers cas traités (budget favorable ou neutre), le niveau d'input proposé par l'hôpital est moins important.



Effet du passage à la T2A - cas 4

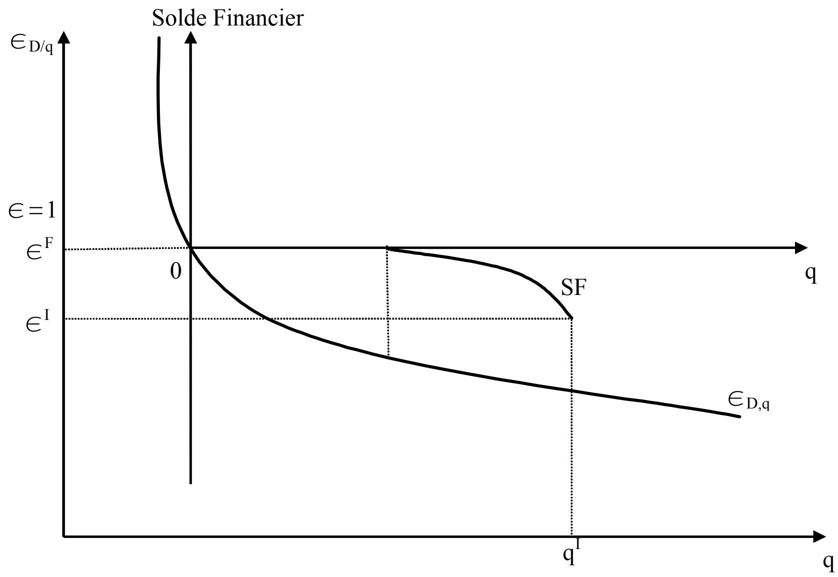
2. L'élasticité de la demande par rapport à l'input est supérieure à 1 et l'hôpital ne récupère pas son équilibre budgétaire Comme dans la situation précédente, l'hôpital se retrouve en déficit budgétaire au passage à la T2A s'il maintient son input initial (q^I).



Effet du passage à la T2A - cas 5

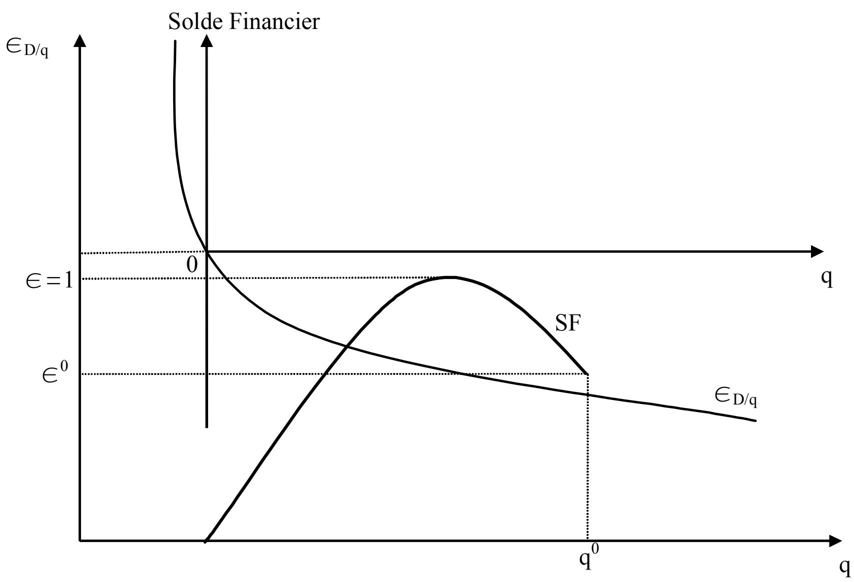
Mais puisqu'il démarre avec une élasticité de la demande par rapport à l'input qui est favorable ($\epsilon^I > \epsilon = 1$), il peut augmenter son niveau d'input en vue de dégager des marges budgétaires. Il se peut néanmoins qu'il atteigne une élasticité égale à l'unité avant d'atteindre l'équilibre budgétaire (voir figure représentant le cas 5). Augmenter son niveau d'input dans cette zone le conduit à creuser son déficit. Par conséquent, l'hôpital ne parvient pas à retrouver l'équilibre budgétaire et se retrouve en déficit.

3. L'élasticité de la demande par rapport à l'input est inférieure à 1 et l'hôpital récupère son équilibre budgétaire L'hôpital se retrouve en déficit budgétaire au passage à la T2A s'il maintient son input initial (q^I). Mais puisqu'il démarre avec une élasticité de la demande par rapport à l'input qui est défavorable ($\epsilon^I < \epsilon = 1$), il est contraint de diminuer son niveau d'input en vue de dégager des marges budgétaires ce qui a pour effet d'augmenter l'élasticité de la demande à laquelle il fait face. La situation décrite à la figure représantant le cas 6 est celle où il parvient à rétablir l'équilibre budgétaire avant d'arriver à élasticité de la demande par rapport à l'input unitaire.



Effet du passage à la T2A - cas 6

4. L'élasticité de la demande par rapport à l'input est inférieure à 1 et l'hôpital ne récupère pas son équilibre budgétaire Ici encore, le maintien de l'input d'avant le passage à la T2A (q^I) conduit l'hôpital à un déficit budgétaire qu'il peut réduire s'il diminue son niveau d'input puisque qu'il se situe initiallement dans la zone où l'élasticité de la demande par rapport à l'input est inférieure à l'unité ($\epsilon^I < \epsilon = 1$). Cela augmente l'élasticité de la demande à laquelle il fait face. Mais lorsqu'il atteint une élasticité unitaire avant d'avoir résorbé son budget (voir voir figure représentant le cas 7), l'hôpital ne peut pas rétablir la situation initiale de déficit.



Effet du passage à la T2A - cas 7

4.2 Impact de chocs exogènes sur la demande ou sur les coûts

Nous analysons l'impact d'une diminution de la demande et d'une augmentation des coûts sur l'équilibre et mettons en évidence le danger que ces chocs exogènes représentent pour l'équilibre budgétaire de l'hôpital. Il est important de garder à l'esprit que, quelle que soit l'élasticité de départ, le passage à la T2A conduit l'hôpital à un niveau d'input associé à une élasticité de la demande par rapport à l'input inférieure ou égale à l'unité.

4.2.1 Une diminution de la demande

Pour l'hôpital, une diminution de la demande peut conduire au déficit. En effet, l'hôpital se trouve dans une zone où l'élasticité de la demande par rapport à l'input est inférieure à l'unité. L'hôpital doit donc réduire son input pour éventuellement récupérer son équilibre budgétaire. Ce qui n'est pas nécessairement le cas si, en diminuant q, il atteint une élasticité de la demande égale à l'unité avant d'avoir rééquilibré son budget. Ces deux cas correspondent aux cas 3 et 4 analysés à la section 4.1.2. Notons par ailleurs que cette diminution de l'input de l'hôpital renforce cette tendance à la baisse de la demande qui lui est adressée ($D_q > 0$). Effet qui sera d'autant plus marqué si le médecin est sensible à l'input de l'hôpital ($R_{eq} < 0$), ce qui l'amène à réduire son effort et éloigne d'autant plus les patients de l'hôpital public.

4.2.2 Une augmentation des coûts

Une augmentation des coûts fixes ou une augmentation des coûts variables telle que $\overline{T} > CH(h) + CM(e)$ (le tarif reste supérieur aux coûts variables) peuvent aussi conduire l'hôpital soit à rétablir son équilibre budgétaire en réduisant son input soit au déficit. L'effet sur l'équilibre de l'hôpital est le même qu'une réduction de la demande. Nous renvoyons donc le lecteur au cas précédent.

Lorsque l'augmentation des coûts variables est telle que $\overline{T} < CH(h) + CM(e)$ (le tarif proposé est inférieur au coûts variables), l'hôpital ne peut réagir et est donc conduit au déficit budgétaire. Cette situation renvoie à la section 4.1.2.

5 Conclusion

La comparaison d'un système de financement de type budget global versus un financement à l'activité a été l'objet d'une littérature abondante ces dix dernière années. Cependant, comme la plupart des pays de l'OCDE ont opté pour un mode de financement à l'activité, la question aujourd'hui n'est plus tant de montrer que ce système est supérieur aux autres mais de réfléchir à l'impact de ce mode de financement sur le comportements des acteurs du secteur hospitalier. Il nous semble intéressant d'analyser notamment en quoi le passage à la T2A en France va modifier le comportement économique des médecins et des gestionnaires hospitaliers à l'hôpital dans la mesure où ces changements vont avoir un impact sur la demande, le niveau d'effort de chaque agent ou le montant des dépenses de santé.

Notre travail offre une contribution à la compréhension de cette nouvelle problématique en proposant une modélisation de la coordination entre médecins et gestionnaires hospitaliers pour la production de services de santé. Comme dans tout travail de modélisation, des hypothèses simplificatrices servent à mettre en lumière et à comprendre les interactions entre les principales variables d'intérêt mais des limites existent pour tenter de faire des prédictions ou de généraliser les conclusions du modèle.

En considérant une neutralité budgétaire dans laquelle le tarif de la T2A est identique au coût moyen sous budget global, l'hôpital est directement affecté par la réforme au travers de sa contrainte budgétaire. Cependant, l'hypothèse de neutralité budgétaire ne l'oblige pas à modifier ses décisions au passage de la réforme puisque son équilibre budgétaire est maintenu. La question étant : a-t-il intérêt à modifier son niveau d'input pour attirer davantage de patients tout en respectant sa nouvelle contrainte budgétaire ? Nous montrons dans ce travail que la réponse dépend de l'élasticité de la demande des patients par rapport à l'input de l'hôpital. Si cette élasticité est supérieure à l'unité, l'hôpital tire profit du passage à une tarification à l'activité puisqu'en augmentant son niveau d'input, il induit une augmentation de la demande tout en continuant à satisfaire sa contrainte budgétaire. L'impact pour le médecin est positif puisque son niveau d'utilité dépend positivement de la demande et du niveau d'input de l'hôpital. L'impact pour les patients est également positif puisque la demande traduit la satisfaction des patients. A contrario, si le niveau de départ de cette élasticité est inférieur à l'unité, l'hôpital ne peut pas tirer profit d'une augmentation de son niveau d'input et le statu quo prévaut.

Nous montrons également que la situation est davantage contrastée dans le cas où le tarif de la T2A ne respecte pas la neutralité budgétaire. L'hôpital profite évidemment d'un tarif favorable mais un tarif défavorable peut dans certains cas l'amener à un déficit. La situation est identique si, après le passage à la T2A, des chocs exogènes viennent diminuer la demande de l'hôpital ou augmenter ses coûts de production. Dans ce cas, l'hôpital peut au mieux s'adapter en dimininuant ses ressources pour retrouver son équilibre budgétaire ou au pire subir un déficit qu'il ne pourra résorber seul.

References

[1] Boadway, R., Marchand, M., Motohiro, S., 2004. An optimal contract approach to hospital financing. Journal of Health Economics 23, 85-110.

[2] Custer W.S., J.W. Moser, R.A. Musacchio, R.J. Willke, 1990, The production of health care services and changing hospital reimbursement – The rôle of hospital-medical staff relationships, Journal of Health economics, 9 : 167-192.

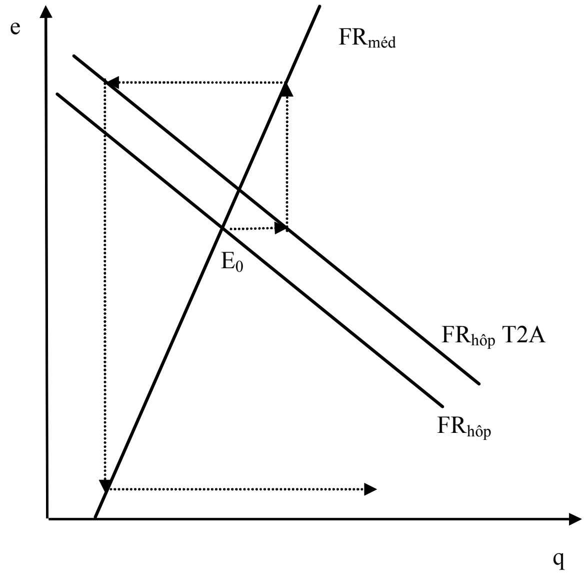
[3] Dor A., H. Watson, 1995, The hospital-physician interaction in U.S. hospitals : evolving payment schemes and their incentives, European Economic Review, 39-3/4 : 795-802.

[4] Ma C.A., 1994, Health care payment systems: cost and quality incentives. Journal of Economics and Management Strategy 8-2, 93-112.

[5] Newhouse J.P., 1996, Reimbursing health plans and health providers : selection versus efficiency in production. Journal of Economic Literature 34, 1236-1263.

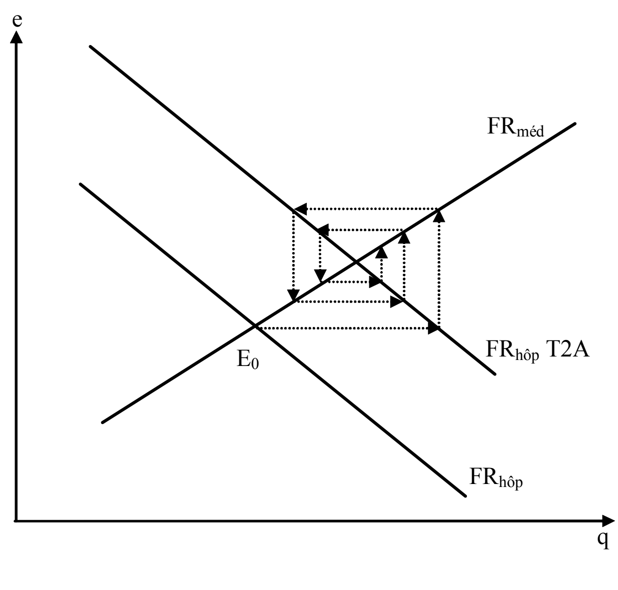
A Appendix 1

Lorsque le médecin est sensible à l'input de l'hôpital ($R_{eq} < 0$), l'obtention d'un équilibre suite au passage à la T2A est soumis à une condition supplémentaire. Le passage à la T2A va en effet, si l'élasticité de la demande par rapport à l'input est initialement supérieure à l'unité, conduire l'hôpital a augmenter son input. Ce qui incite le médecin à réagir en augmentant son effort et donc les coûts de l'hôpital. Ce dernier pourrait donc être amené à réduire son input pour respecter sa contrainte budgétaire si le médecin est extrêmement sensible à l'input de l'hôpital. Mais cette diminution de l'input va elle même amener le médecin à réduire son effort et nous nous trouvons là dans une situation où il n'y a pas de convergence vers l'équilibre. La situation est représenté ci-dessous.



Absence d'équilibre

Nous supposons - sans que cela ne change le résultat - que la fonction de réaction de l'hôpital est décroissante ($\frac{dq}{de} < 0$) et nous partons de l'équilibre initial E_0. Si l'élasticité de la demande par rapport à l'input de l'hôpital est supérieure à l'unité, le passage à la T2A amène l'hôpital à augmenter son input q quel que soit le niveau de e. L'hôpital passe donc de la fonction de réaction $FR_{hôp}$ à la fonction de réaction $FR_{hôp}T2A$. Compte tenu du fait que le médecin est assez sensible à l'input de l'hôpital (la fonction de réaction du médecin $FR_{méd}$ est plutôt verticale), le passage à la T2A ne permet pas de converger vers un nouvel équilibre. Ce nouvel équilibre peut être atteint si le médecin n'est pas trop sensible à l'input de l'hôpital comme représenté à la figure 9.



Equilibre convergeant

Formellement, une condition suffisante pour qu'il y ait convergence vers un nouveau point d'équilibre et que, pout toutes valeurs de e et de q:

$$\frac{de}{dq}_{méd} < \left|\frac{dq}{de}\right|_{hôp}$$

Ce qui se traduit par:

$$\frac{R_{eq}}{D_{ee} - R_{eq}} < \left|\frac{-D_e\left[CH(h) + CM(e)\right] - D(e,h,q)CM_e}{D_q\left[CH(h) + CM(e)\right] + c_q}\right|$$