Hétérogénéité des patients, tarification hospitalière et autres instruments de régulation. Un cadre d'évaluation économique

Dominique Bureau[1]

Résumé

Les approches macro-budgétaires en termes d'objectifs globaux de dépenses ne sont pas suffisantes pour orienter efficacement l'évolution du système de santé, dont la performance dépend des décisions de nombreux acteurs. Il faut des mécanismes microéconomiques de responsabilisation, incitant ceux-ci à minimiser le coût de fourniture des soins pour leur niveau de qualité visé, sans risquer de compromettre l'accès aux soins. Ceci nécessite des modes de paiements des offreurs de soins appropriés, abordant de front l'hétérogénéité des patients.

Avec en perspective les conclusions du « Ségur de la santé », qui pose le principe d'une nouvelle politique de financement des hôpitaux associée une rénovation des mécanismes d'encadrement des dépenses (« ONDAM »), on rappelle les références économiques pour cela. La tarification à l'activité (T2A) constituant la toile de fond des controverses, ses fondements constituent le point de départ de l'analyse, qui examine ensuite les inconvénients de la prolifération de la nomenclature et les justifications à de nouveaux modes de paiements. À cet égard, on souligne l'importance d'un *design* bien conçu, abordant de front l'hétérogénéité des patients ou des *case-mix*, en mobilisant les apports de la théorie de la régulation incitative.

Mots-clefs : puissance des incitations, concurrence par comparaison, hétérogénéité, sélection des patients

[1] Dominique Bureau est membre du HCAAM et professeur chargé de cours à l'école polytechnique. Les vues exprimées n'engagent que son auteur.

Introduction

Le 25 mars 2020, le Président de la République avait décidé, qu'à l'issue de la crise sanitaire, « un plan massif d'investissement et de revalorisation de l'ensemble des carrières serait construit pour notre hôpital ». Les conclusions du processus de concertation mis en place à cette fin par le ministère des solidarités et de la santé, dit « Ségur de la santé » (2020), mettent en exergue les rigidités de la gestion et le manque d'attractivité de l'hôpital public. Elles posent le principe d'une nouvelle politique de financement des hôpitaux, associée une rénovation des mécanismes de l'objectif national de dépenses de l'assurance-maladie (« ONDAM ») pour y intégrer « les enjeux de santé à long-terme ». Les conditions pour cela restent cependant à définir, étant noté que le diagnostic et les orientations du projet « Ma santé (2022) » sont confirmées.

Celui-ci établissait que : « parmi d'autres leviers, celui du financement est essentiel pour favoriser la transformation du système de santé et permettre de réorienter celui-ci vers les besoins des patients. Les modalités de financement sont en effet un puissant moteur pour permettre l'évolution des comportements et des organisations, via les ressources qu'elles permettent de consacrer à un patient donné et le signal qu'elles adressent aux différents professionnels ». Mais il décrivait par ailleurs une situation insatisfaisante, caractérisée par des inégalités qui demeurent, malgré les mécanismes d'assurance universelle mis en place. Étaient aussi pointés l'orientation insuffisante vers la prévention, des modes de prise en charge inadaptés à l'augmentation des pathologies chroniques, et des modes de tarification qui poussaient à accroître la quantité des soins produits plutôt que leur qualité.

Pour y remédier, l'idée générale était de diversifier les modalités de rémunération des offreurs de soins, en introduisant des paiements forfaitaires au suivi, des paiements à la qualité et à la pertinence, des paiements reposant sur des dotations populationnelles et des paiements groupés pour favoriser de meilleures coordinations entre médecine de ville et hôpital ; et, simultanément de réformer les mécanismes existants pour les épisodes uniques de soin.

Enrichir la panoplie d'instruments est utile pour alléger les conflits entre les différents objectifs que doit considérer la régulation économique des systèmes de santé doit prendre en compte trois grands objectifs (Newhouse, 1996, Mougeot et Naegelen, 2011): établir les bons arbitrages entre les bénéfices sociaux de la qualité des soins ou de l'état de santé des populations, et le coût imposé à la société pour les obtenir; minimiser le coût de fourniture des soins pour atteindre le niveau de qualité visé; et assurer l'équité d'accès aux soins.

Cependant, il n'y a pas de miracle. Certains problèmes rencontrés avec la tarification hospitalière à l'activité (T2A) réapparaitront peu ou prou si les questions sous-jacentes aux dysfonctionnements ne sont pas traitées. Les nouveaux modes de tarification appellent donc une évaluation précise des impacts incitatifs des mécanismes de rémunération des services fournis par un hôpital ; et ils nécessitent un cadre d'ensemble cohérent.

La fragmentation du cadre actuel, tiraillé entre des mécanismes macro-budgétaires, de planification des différents moyens, les contraintes financières au niveau des services et leur « médicalisation » souhaitée est manifeste. Paradoxalement, la T2A, qui relève théoriquement d'une approche faisant de la liberté de gérer un principe cardinal, concentre les critiques alors que beaucoup des reproches mettent plus directement en cause les mécanismes d'allocation des

moyens relevant plutôt de la planification[2]. Les critiques plus spécifiques portant sur la T2A sont par ailleurs contradictoires. Les uns stigmatisent les incitations excessives à réduire les coûts, d'autres les rentes excessives laissées à certains acteurs, d'autres enfin le maintien de mécanismes suivant *in fine* les coûts, et donc en réalité peu incitatifs.

Les controverses à son propos révèlent des incompréhensions profondes entre : des médecins qui ne comprennent pas pourquoi ils sont « en déficit » alors qu'ils considèrent suivre parfaitement les recommandations de l'*evidence-based-medicine* , ont des prétentions de rémunération limitées et s'interdisent de sélectionner les patients, conformément à la déontologie ; et des gestionnaires qui ont tendance à évacuer les questions de sélection des patients, considérant que, si les dépenses augmentent, c'est qu'il y a une inflation d'actes inutiles, la maitrise des dépenses demeurant la première priorité. Les mécanismes budgétaires d'enveloppes fermées annuelles sont alors privilégiés, sans mesurer les risques de reports de charges, de rétention ou de contournements, qu'ils induisent. Pointer les symptômes de dysfonctionnements ou la conflictualité entre les différents objectifs ne suffit donc pas. Leurs causes et les solutions pour y remédier nécessitent des évaluations rigoureuses : si le bon *design* des modes de paiement est reconnu comme un moyen pour améliorer la performance globale de notre système de soins, il importe donc de préciser comment…

Dans cette perspective, on rappelle d'abord les principes économiques sur lesquels s'appuyer pour l'élaboration de la tarification hospitalière. Compte-tenu des controverses associées à la T2A, les raisonnements qui avaient motivé l'introduction de la tarification forfaitisée par pathologie par groupe homogènes de malades (les DRG[3]), à la suite de l'expérience du programme *Medicare* aux Etats-Unis à partir de 1983, constitueront le point de départ.

Cependant, dans la mesure où l'articulation entre les différents outils de régulation et la question des rémunérations à consentir pour assurer la continuité de certains services sont aujourd'hui au cœur des débats[4], on considérera un modèle élargi, intégrant à la fois le rôle et les attentes des patients et l'offre des professions médicales, ainsi que l'existence de différents niveaux de coûts de traitement potentiels selon les facteurs de risque de chaque malade puis de l'impact des traitements (aléa thérapeutique). En effet, les mécanismes introduits jusqu'à présent au sein de la T2A pour résoudre ce problème n'apparaissent pas satisfaisants alors qu'il met en cause l'architecture de la tarification, au-delà du cas d'*outliers*[5]. On examinera finalement l'introduction envisagée de nouveaux modes de paiements, ce qui conduira à souligner que la définition de ceux-ci devra aussi anticiper des problèmes d'hétérogénéité.

---

[2] Contrôle de la démographie médicale, regroupements et mutualisations opérés de manière centralisée sous l'hypothèse d'économies d'échelle illimitées et dont l'exploitation serait mécaniquement corrélée à la qualité sanitaire, accent mis essentiellement sur la tenue à court-terme des enveloppes budgétaires sans mesurer les impacts de plus long-terme…

[3] Pour *diagnosis related group*, terme que l'on privilégiera car la traduction française GHM suggère, au-delà de la notion de groupe tarifaire construit pour la tarification en lien avec le diagnostic, une homogénéité absolue des groupes de malades qui est à l'origine de certaines incompréhensions et erreurs de *design* dans la conception des nomenclatures, trop détaillées par rapport à des caractéristiques non-contrôlables.

[4] Avec à la fois des flux de départs importants et la nécessité de payer des vacations à des niveaux « théoriquement illégaux » pour assurer la continuité des gardes ou l'opérationnalité des blocs (cf. Conclusions du Ségur p.13, mais ce problème ne peut être résolu sans en considérer les données économiques sous-jacentes d'offre et de demande).

[5] Les dispositifs complémentaires pour éviter les situations de blocage par rapport à la prise en charge de ces cas extrême (type assurance *stoploss* pour les services, à combiner avec des mécanismes d'affectation régulés des patients correspondants) ne sont pas examinés ici mais c'est aussi une question importante.

À l'encontre des présentations habituelles, qui suivent la logique théorique de construction d'une tarification, on considérera des types de formules tarifaires posées a priori, adaptées au contexte considéré. En effet, notre objectif n'est pas d'apporter une contribution théorique originale, mais de mettre en exergue certains principes que l'on peut tirer de l'analyse économique, sans cacher que les problèmes « d'agence » à gérer dans ce domaine sont particulièrement complexes (Choné et Ma, 2011), compte-tenu de la multiplicité des paramètres inobservables concernant les patients, la qualité des soins, les objectifs des offreurs de soins…

I-Tarification hospitalière et incitations à l'efficacité dans la production des soins

I-1- *Medicare* 1983, retour sur la tarification par « *DRG* », …

À l'origine des mécanismes de tarification forfaitisée par pathologie, il y avait le constat que, souvent, les pratiques médicales s'écartent des recommandations. En conséquence, une meilleure responsabilisation des médecins aux coûts induits par leurs décisions thérapeutiques et prescriptions serait de nature à réduire les coûts des traitements sans remettre en cause l'accès aux soins. Les principes sous-jacents aux paiements hospitaliers qui, à partir de l'expérience de Medicare, se sont développées avec cet objectif peuvent être rappelés à partir du cadre stylisé suivant, dans lequel on examine la question du paiement à effectuer à un offreur de soins (un service hospitalier, par exemple) pour le traitement d'un patient classé dans un DRG particulier.

On suppose que celui-ci devrait impérativement être soigné (par des moyens adaptés eu égard aux connaissances et évaluations faisant autorité pour le traitement de la pathologie considérée). Par ailleurs, on fait l'hypothèse que les coûts à engager pour cela se décomposent en deux :

-des coûts potentiellement observables par le régulateur (C), en termes d'actes, examens, type d'opération ou traitement prescrits attribuables à chaque cas, dont on note (w) l'indice de prix,

-et des coûts non observables, dépendant de l'effort[6] réalisé pour réduire les volumes d'actes. Pour une réduction (e) de celui-ci, l'effort au-delà de ce qui est strictement nécessaire, lié par exemple à la profondeur ou au temps consacré à l'examen clinique, à des besoins de formation ou de recherche, ou encore à l'intensité du suivi à assurer, est noté $\psi(e)$. Celui-ci est fortement croissant avec le niveau d'effort ($\psi(0) = 0, \psi' > 0, \psi" > 0$). S'agissant principalement de temps médical, on suppose que son indice de prix est aussi (w).

Par ce biais, on cherche seulement à rendre compte des écarts souvent constatés dans les pratiques médicales, en intégrant la possibilité d'arbitrages entre ces deux types de coûts : le choix d'une stratégie réclamant plus « d'effort » de la part de l'offreur de soins permettrait potentiellement de réduire le volume « d'actes » ; *a contrario*, des modes de paiement ne

[6] L'ensemble de l'exposé s'appuyant sur le chapitre I-1 de l'ouvrage de Laffont et Tirole (1993), on en reprend les notations, les hypothèses (ex. neutralité au risque des différents agents) et la terminologie, notamment ce terme « d'effort ». Mais il ne faut voir aucune connotation normative à celle-ci dans le contexte où nous réinterprétons leur modèle, seulement l'expression de la possibilité de substituer différents coûts, plus ou moins observables. Évidemment, cet effort à réduire les actes inutiles a un coût, par exemple en termes d'investissement ou stress des soignants, qui est donc intégré dans le modèle, au niveau des choix individuels et de l'objectif collectif. Ce coût ne pouvant être ignoré dans le contexte de crise qui motive les projets de réforme, ceci rend en fait leur modèle particulièrement attractif pour étudier les différents modes de paiements des offreurs de soins.

responsabilisant pas aux coûts conduisent à des répétitions d'examens inutiles ou des durées de séjour excessives.

Outre de ce niveau d'effort, le nombre d'actes dépend de la pathologie, représentant un certain volume d'actes en l'absence d'effort (noté β). Ainsi, après le diagnostic ayant conduit à classer le patient dans ce DRG, un niveau d'effort est choisi, qui déterminera un coût de traitement observable, en termes de quantités d'actes :

$$C = w(\beta - e)$$

Par ailleurs, on note R le paiement total versé à l'hôpital pour un épisode de soins relevant de ce DRG, t ce paiement, net des coûts des actes et prescriptions observables qu'il supporte ($t = R - C$) ; et $U = t - w\psi(e)$, son « bénéfice », net du coût de « l'effort ». Ce bénéfice doit être positif pour assurer la pérennité de l'offre de soins par rapport à ce DRG et donc éviter le risque d'exclusion.

Le coût total de traitement valant $w[(\beta - e) + \psi(e)]$, le niveau idéal d'effort (e^*) serait tel que :

(1) $\psi'(e^*) = 1$

L'effort correspondant est efficace, par rapport à l'arbitrage entre son coût et les gains permis sur les volumes d'actes, avec égalité, à l'optimum, à la marge, entre le coût de l'effort supplémentaire nécessaire pour diminuer le nombre d'actes et le coût de ceux-ci. On note $C^* = w\ (\beta - e^*)$.

Ce niveau d'effort ne sera pas réalisé si la formule de paiement est de type *cost-plus* ($ie. R = C + constante$), car l'offreur de soins a alors intérêt à choisir l'effort minimal ($\psi'(e) = 0, e = 0, C = w\beta$). Dans ce cas, l'absence de responsabilisation des offreurs de soins aux coûts des actes qu'ils génèrent ne les incite pas à choisir les stratégies les plus efficaces.

C'est ce type de formule (par le biais de prix de journée, par exemple) qui prévalait avant la mise en place de la T2A et prévaut encore en ambulatoire. Elle pousse à la multiplication des actes ou la duplication de certains de ceux-ci, tout accroissement des durées de séjour, par exemple, étant compensé dans un tel cadre quelle que soit leur utilité. Le strict *cost-plus* ($R = C; t = 0$) a cependant l'avantage de ne pas abandonner de rente à l'offreur de soins.

La tarification forfaitisée à l'activité visait à remédier à ce défaut d'incitations à l'efficacité, grâce à l'utilisation de formules de paiements incitant à réduire les coûts. À cet égard, un paiement forfaitaire $R = A$ (soit, $t = A - C$), type T2A, laisse à l'offreur de soins le bénéfice des réductions de coûts qu'il réalise, et celui-ci est pénalisé s'il ne s'attache pas à le diminuer. Dès lors, il a intérêt à choisir l'effort optimal $e = e^*$, qui maximise $U = A - C - w\psi(e) = A - w\ (\beta - e + w\psi(e))$.

À l'encontre des formules de type *cost-plus*, les formules à prix fixes conduisent donc spontanément les offreurs à choisir la stratégie la plus efficace en termes de combinaison « actes versus effort », ceux-ci bénéficiant intégralement des économies à la marge qu'ils permettent à la société de réaliser sur le premier terme.

C'est ce type de raisonnement qui a présidé à la mise en place de la tarification à prix fixe par DRG, à un moment où, dans l'ensemble de la gestion publique, le besoin de renforcer l'efficacité était mis en exergue.

En théorie, la mise en place du paiement forfaitisé au niveau $R^* = A^* = w\psi(e^*) + C^*$ permettrait même de réaliser le niveau d'effort optimal, sans laisser de rente inutile ($e = e^*, U = 0$). Dans ce cas, les coûts seraient juste remboursés, ceci assurant l'accès aux soins (pas de « *dumping* ») puisque traiter les patients est profitable. Mais ce résultat est obtenu par l'ajustement du montant du forfait et non par un mécanisme de remboursement automatique des coûts. La formule de paiement à prix fixe correspondante s'écrit :

$$t^* = w\psi(e^*) + (C^* - C) \qquad (2)$$

Elle compense forfaitairement le coût de l'effort optimal et laisse à l'offreur de soins l'écart entre le coût réalisé et le coût optimal escompté, bénéfice d'un coût réduit ou charge du dépassement. L'offreur de soins bénéficie ainsi d'un paiement à prix fixe, compensant son effort et lui remboursant ses actes utiles mais il est incité à choisir la combinaison d'actes et d'effort la plus efficace (schéma 1).

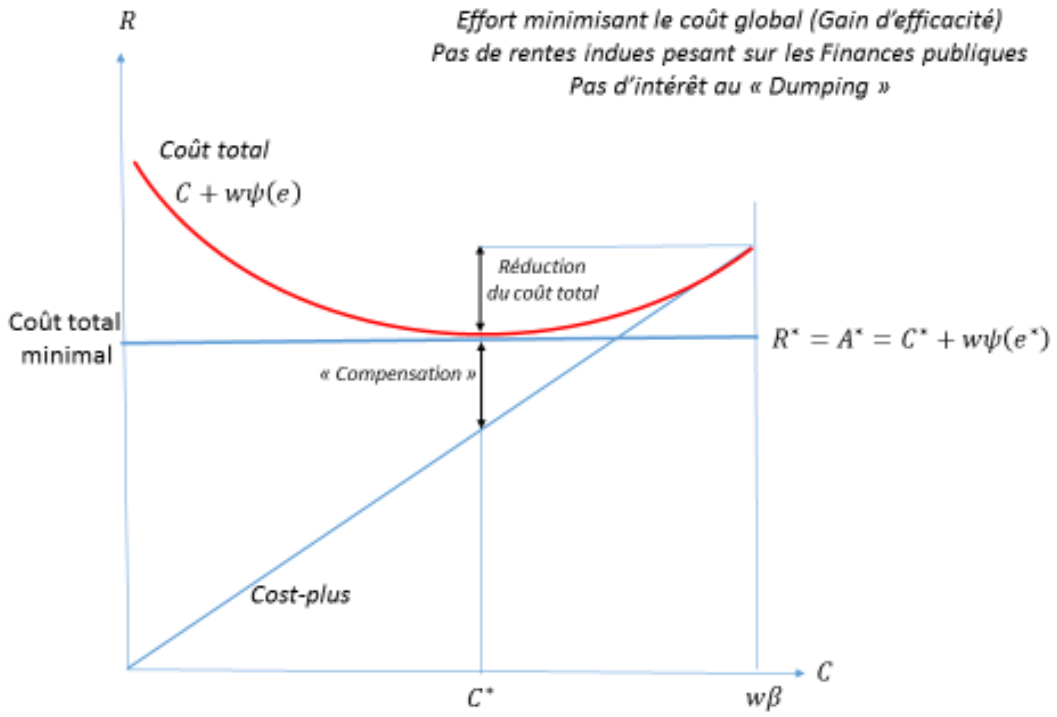


Schéma 1. Bénéfices de la forfaitisation optimale

Pour ne pas laisser de rente, il faut cependant que le régulateur connaisse β ou, de manière équivalente, soit capable d'estimer C^*. A priori, ce n'est pas le cas car, alors, il pourrait aussi contrôler indirectement l'effort, celui-ci pouvant se déduire des coûts observables. Dans ces conditions, les régulateurs qui ont été établis dans les différents secteurs nécessitant une régulation tarifaire mettent généralement en place des tarifications à prix fixes fondées sur l'estimation qu'ils peuvent faire de ce paramètre. Les opérateurs sont alors incités à choisir le niveau d'effort optimal, mais il faut leur laisser une rente pour cela.

C'est ici qu'intervient l'idée de la « mise en concurrence par comparaison ». En effet, si cette pathologie est traitée dans des conditions similaires dans différents services, il devient possible d'obtenir l'information manquante en considérant les performances relatives des offreurs de soins et en disant à chacun de ceux-ci que leur sera appliqué la formule précédente, dans laquelle C^* sera remplacé par le coût moyen, estimé « par comparaison », i.e. observé pour le traitement de cette pathologie dans les autres services.

De cette manière chaque service est confronté à une formule de paiement à prix fixe, ce qui le conduit à choisir le niveau d'effort optimal. Et les coûts moyens observés constituent des estimateurs sans biais de l'espérance du coût efficace de traitement. Par ce moyen, l'offreur de soins est bien payé pour ses efforts et remboursé de ses coûts, mais : forfaitairement par rapport aux coûts efficaces, ceux-ci étant appréciés par « parangonnage », en observant les coûts des autres offreurs de soins pour ce même DRG ; et non par rapport aux coûts observés pour le cas considéré, ce qui ne responsabiliserait pas à leur maitrise.

En résumé, la tarification forfaitisée par groupe homogène de malades procède de deux idées :

- le recours à une tarification à prix fixe pour inciter à supprimer les actes inutiles, réduire les durées de séjour excessives, bien sélectionner les examens demandés ;

- et celle de concurrence par comparaison (*yardstick*) pour ajuster son niveau à ce qui est nécessaire pour éviter tout déficit net, mais sans laisser de rente inutile.

I-2-Complémentarités et affectation des instruments de régulation

Afin d'avoir une vision plus globale de la régulation des dépenses hospitalières et pouvoir en comparer les différentes approches, deux autres éléments doivent être incorporés au modèle :

-le comportement du patient dans sa manière à recourir au système de soins. À cet égard, on suppose, qu'au-delà du fait d'avoir été « soigné », les conditions dans lesquelles le patient a accès aux soins, possibilité d'avoir plusieurs avis par exemple, compte. On le formalise par un surcoût de traitement x, lié par exemple à la duplication des examens ou des contraintes accrues du côté de l'offre de soins, d'autant plus élevé que le patient a la possibilité d'accéder sans aucune restriction au système de soins. Et on suppose qu'à celui-ci, qui est supporté par l'assurance-maladie, est associé une valeur pour le patient $S(x)$ avec $S' > 0,\ S" < 0$;

-l'offre de ressources médicales, pour en endogénéiser le prix w. En ce domaine, on se contentera de la formulation la plus simple, d'une courbe d'offre reflétant un coût d'opportunité des ressources médicales croissant avec sa quantité totale employée (L), soit $\widetilde{w}(L), \widetilde{w}' > 0$. On notera (η) l'élasticité-prix correspondante de l'offre médicale. Évidemment, une analyse plus approfondie devrait considérer les spécificités de ces emplois en termes de formation et de fonctionnement des marchés du travail correspondants, externes et internes, ainsi que la diversité des ressources mobilisées, plus ou moins substituables.

Finalement, on considère que la nécessité de recourir aux prélèvements obligatoires pour financer le système de santé est associé à un (sur)coût social des fonds public λ, reflétant les inconvénients macroéconomiques des prélèvements obligatoires et rendant en particulier socialement coûteuses les rentes laissées aux offreurs de soins. Si l'on considère que l'accès

aux soins doit être absolument garanti et qu'on note N le volume total de cas à traiter, la quantité de ressources utilisée est donc :

(3) $$L = N[\beta + x - e + \psi(e)]$$

Le coût social net associé au traitement de tous les patients, dans les conditions associées à des valeurs quelconques de (e, x), vaut alors :

(4) $$CS = \int_0^L \widetilde{w}(l).dl + \lambda N(U + wL) - NS(x)$$

En information parfaite, c'est-à-dire si le régulateur connait β et est à même de contrôler les choix des offreurs de soins (e) et des patients (x), il imposerait la stratégie thérapeutique et le niveau de recours aux soins minimisant ce coût social. Pour cela, il éviterait de laisser des rentes inutiles, et minimiserait (L), parce que ceci est ici bénéfique en soi et que cela permet en plus, dans ce modèle bouclé, de réduire le prix des ressources médicales car, pour disposer d'une offre accrue, il faut accroître leur rémunération.

La politique optimale en information parfaite vérifie donc :

(5) $$\begin{cases} w = \widetilde{w}(L) \\ e = e^* \\ U = 0 \end{cases}$$

Si le marché du travail médical fonctionne efficacement, ces conditions seraient réalisées en appliquant la tarification forfaitaire par pathologie ajustée par comparaison définie précédemment. Par ailleurs, le choix des patients doit vérifier :

(6) $$S'(x) = w.(1 + \lambda'); \quad \lambda' = \lambda(1 + 1/\eta)]$$

Ainsi, le principe d'accès garanti aux soins n'empêche pas que les patients doivent être responsabilisés par rapport aux coûts induits par leurs comportements, ceux-ci générant des coûts pour la collectivité qui doivent être justifiés. C'est ce qu'exprime cette condition, qui intègre le financement public de ces coûts et le fait que l'accroissement des moyens mobilisés requiert une hausse de leur prix pour que l'offre suive.

Concrètement, la bonne orientation des patients dans le système de soins est un enjeu critique, comme le montrent les problèmes d'utilisation des urgences, par exemple. Plus généralement, la non-responsabilisation des patients qui résulte de mécanismes d'assurance posant en principe l'absence de ticket modérateur « d'ordre public », renforcée par le fait que l'affectation des risques est diluée entre l'assurance maladie obligatoire et les assurances complémentaire, les deux types d'assurances couvrant le même panier de soins, demeure l'obstacle majeur à une meilleure régulation du système de santé français (Dormont et al., 2014).

L'éducation des assurés et la gestion des risques sont donc essentielles. Les enjeux correspondants ont été cernées depuis une trentaine d'années maintenant (cf. encadré 1).

Encadré 1. Assurance publique et responsabilisation des patients.

L'assurance - maladie, en solvabilisant la demande de soins, introduit une différence entre le coût payé par l'assuré et le coût pour la société. Il en résulte un arbitrage à réaliser, entre la

valeur de cette mutualisation du risque, et le souci de responsabiliser les assurés aux coûts qu'ils induisent. L'arbitrage ne fait évidemment aucun doute pour le « gros risque », pour lequel la solidarité prime. En revanche, il peut être justifié de laisser une part du « petit risque » à la charge de l'assuré.

Cet arbitrage peut être évalué en termes d'analyse coût-avantage, la notion économique de prime de risque reflétant en effet ce que les agents sont prêt à payer pour sa mutualisation. Suivant cette approche, le « Health Insurance Experiment » de la Rand avait étudié dès 1987, empiriquement par un essai randomisé, l'impact des modalités d'assurance des patients sur la demande soins. Il montrait que, globalement, la gratuité des soins entraîne un niveau de consommation supérieur de 46% à celui que l'on observe lorsque les frais sont intégralement supportés par les ménages. La moitié de cet écart est obtenue en passant de la gratuité totale à un ticket modérateur (d'ordre public) de 25%. L'essentiel de la différence entre ces montants de dépenses résultait de la fréquence des consultations, puis, dans une moindre mesure, de celle des hospitalisations.

Face à ces résultats la question fût, bien évidemment, celle de l'impact éventuel en termes de santé : les patients couverts à 100 % reçoivent-ils trop de soins, ou ceux soumis à un fort taux de responsabilisation pas assez ? L'étude complémentaire réalisée pour répondre à cette question ne mettait pas en évidence d'amélioration significative des indicateurs de santé du fait d'un taux de couverture de 100 %, pour les personnes de caractéristiques moyennes. Le principal impact notable concernait la correction de la vision, l'écart sur l'hypertension artérielle étant à la limite de la significativité. Par contre, ces effets apparaissaient amplifiés pour les personnes cumulant un faible revenu et un handicap initial important en ces domaines. Ces résultats ne constituent donc pas un argument pour écarter, en général, un certain taux de participation des assurés à la couverture des petits risques des dépenses maladie, complété par des programmes spécifiques pour les enjeux et populations spécifiques.

I-3- Retour d'expérience et conditions d'efficacité

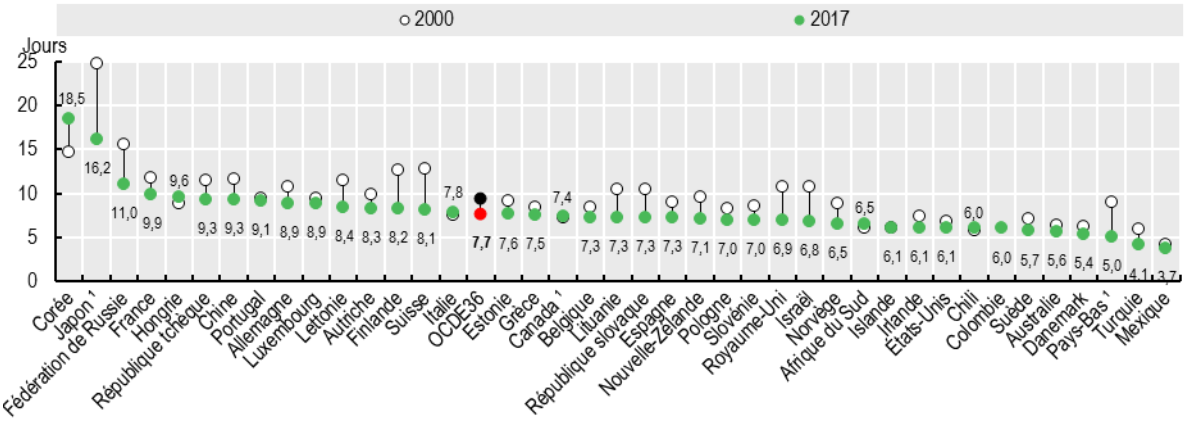
L'évolution des durées de séjours suite à la mise en place de paiements à prix fixe montre que la forfaitisation joue bien le rôle responsabilisant que l'on peut en attendre (cf. encadré 2).

Encadré 2. Durées moyenne de séjour (source Panorama de la santé, OCDE, 2019)

La durée moyenne de séjour à l'hôpital est souvent considérée comme un indicateur d'efficience de la prestation des services de santé. Toutes choses égales par ailleurs, une hospitalisation de plus courte durée diminuera le coût par sortie et transfèrera la prise en charge des patients à des structures moins onéreuses. Les séjours de longue durée peuvent être le signe d'une mauvaise coordination des soins, ce qui a pour effet de laisser certains patients attendre inutilement à l'hôpital que des soins de rééducation ou de longue durée soient organisés. Dans le même temps, il arrive que certains patients sortent trop tôt, alors qu'un séjour plus long aurait pu améliorer leur état de santé ou réduire le risque de ré-hospitalisation (…)

Depuis 2000, la durée d'hospitalisation moyenne a diminué dans la plupart des pays ; les reculs les plus marqués ont été constatés au Japon, en Finlande, en Suisse, au Royaume-Uni, en Israël et aux Pays-Bas.(…)

Outre les disparités dans la durée d'hospitalisation moyenne dues aux différents types de pathologies traités, d'autres facteurs, dont les structures de paiement, peuvent expliquer les écarts entre pays. Le raccourcissement du séjour moyen a été attribué à la mise en place de systèmes de paiement prospectif qui encouragent les prestataires à réduire le coût des épisodes de soins, comme les groupes homogènes de malades (GHM). La France, l'Autriche et la Suède comptent parmi les pays qui ont adopté des mécanismes de cette nature et qui ont enregistré, dans le cadre de ce processus, une diminution de la durée moyenne des hospitalisations (…)



Pour autant, la tarification hospitalière demeure souvent imparfaite. En effet, pour que les incitations recherchées jouent à plein, il faut que la forfaitisation du paiement soit crédible :

-si le service, *too big to fail*, anticipe que ses déficits seront en fait comblés, on en revient à un cadre de type *cost-plus*, avec des incitations à l'efficacité faibles ;

-il en va de même si le service peut coder ses patients *ex-post* au vu des coûts réalisés, en choisissant alors dans la nomenclature les niveaux de gravité lui permettant de couvrir ses coûts ;

-par ailleurs, un service qui anticipe que les moyens dont il disposera seront réduits dès qu'il révélera ses potentiels de gains d'efficacité aura intérêt à choisir un faible niveau d'effort, sachant que l'ajustement précipité de ses moyens revient à lui confisquer le bonus potentiel qui lui était promis par une formule de paiement sensée être à « prix fixe ».

Ces problèmes de crédibilité sont exacerbés dans le secteur de la santé car on ne ferme pas un hôpital comme une entreprise. Mais ils n'y sont pas propres. Ils expliquent que la plupart des régulateurs sectoriels se soient orientés vers la mise en place de plafonds de prix pluriannuels (*price-cap* de type « indice de prix-X%) pour les tarifs d'accès aux grands réseaux : la forfaitisation visant à inciter les opérateurs en charge de ceux-ci à l'efficacité ; et la pluri-annualité pour que les incitations mises en place pour cela soient crédibles.

Plus généralement, les réformes correspondantes au niveau de la tarification s'inscrivaient dans une vision globale, dans un contexte où l'offre des secteurs régulés était généralement jugée trop rigide, soumise à trop d'injonctions contradictoires quant aux objectifs à privilégier et tendanciellement décalée par rapport aux attentes de la société.

Reconnaissant que la bonne gestion des secteurs correspondants nécessite de constituer des cadres d'action (« *level-playing-field* ») solides, fondés sur des règles plutôt que sur l'illusion de la vertu des décisions discrétionnaires héroïques, la remise à plat des structures tarifaires concernées a été confiée à des régulateurs indépendants, pour orienter les choix de long-terme des différents acteurs. L'analyse sous-jacente considérait que les gains d'efficacité à en attendre était bien supérieurs au coût social des rentes devant être abandonnées en contrepartie.

Jusqu'à présent, ces principes n'ont pas prévalu pour la régulation hospitalière en France. Mais le système s'avère peu performant, les difficultés rencontrées pour instaurer la T2A ayant conduit à complexifier à l'excès les nomenclatures, à multiplier les cadres spécifiques, par référence au statut des offreurs de soins plutôt que par rapport à la nature des services produits, et à substituer à la concurrence par comparaison des mécanismes budgétaires annuels de points flottants (Mougeot et Naegelen, 2104).

I-4-ONDAM ou T2A ?

En fait, le cadre régulatoire français essaye de combiner une forfaitisation de la tarification hospitalière (T2A) avec, plutôt que l'ajustement des forfaits par comparaison, une contrainte d'enveloppe budgétaire globale (l'ONDAM), les deux étant supposées réconciliées par un mécanisme de points flottants. Schématiquement, on se donne donc une enveloppe budgétaire (B) pour financer l'ensemble des dépenses. Il faut donc que :

$$N(R + wx) \leq B$$

Pour cela, le niveau de paiements de la T2A $(R = A)$ est donc ajusté automatiquement, tel que :

$$A = \left(\frac{B}{N}\right) - wx$$

Si cette tarification est crédible et que les offreurs de soins se comportent effectivement en « *price-taker* », on peut en attendre un niveau d'effort optimal $(e = e^*)$, d'où une demande totale de facteurs de production $L = N(\beta - e^* + \psi(e^*) + x)$, et un prix de ceux-ci au moins égal à $\widetilde{w}(L)$. Enfin, il faut que les offreurs de soins ne soient pas en déficit, donc :

$$A \geq w\,[\beta - e^* + \psi(e^*)]$$

Il en résulte :

Proposition 1 : ONDAM et T2A ne sont compatibles que si :

$$\boldsymbol{B \geq\ N(\beta - e^* + \psi(e^*) + x).\widetilde{w}(N(\beta - e^* + \psi(e^*) + x))}$$

En information imparfaite, l'égalité stricte ne peut-être que fortuite. De plus, si le souci au moment du vote de l'ONDAM est de contenir les évolutions budgétaires en évitant d'abandonner des rentes, l'incompatibilité est plus probable que l'abondance budgétaire. Toutefois, différents instruments de politique de santé sont susceptibles de réduire le fossé éventuel, en agissant sur les quatre variables déterminantes restantes de cette condition :

-une meilleure orientation du comportement des patients (x), dont on a vu ci-dessus l'enjeu intrinsèque, mais qui revêt une importance particulière si, comme cela a été spécifié, la sous-enveloppe hospitalière de l'ONDAM tend à être un solde, qui doit compenser les éventuels dérapages sur l'ambulatoire, par exemple,

-des politiques réduisant intrinsèquement les coûts de traitement (β), par l'innovation, une meilleure gouvernance ou des structures hospitalières plus efficaces. Mais il faut, pour cela, que les gains organisationnels soient effectifs, donc, par exemple, que l'informatisation décharge les médecins de tâches administrative ou de saisie, et que les regroupements hospitaliers soient synergiques,

-la fonction d'offre $\widetilde{w}$. Mais là encore, il ne faut pas se tromper. Des restrictions malthusiennes accroissent les rentes.

-ou réduire le nombre de cas traités (N), mais on a supposé que cette variable reflétait justement le nombre de cas devant être soignés.

Ceci ne saurait surprendre : en information imparfaite, on ne peut concilier incitation à l'effort efficace et absence de déficit des offreurs de soins, donc traitement de tous patients, sans laisser de rentes à ceux-ci. Tentative pour concilier efficacité des coûts sans laisser de rentes, la combinaison T2A/ONDAM, telle qu'elle a été organisée, ne peut intrinsèquement atteindre les objectifs escomptés, sauf à remettre en cause l'objectif sanitaire qui sous-tend notre système de soins. En revanche, la gestion de notre système de santé gagnerait à :

- mobiliser un mécanisme de concurrence par comparaison bien conçu, révélateur de l'information manquante pour fixer les forfaits de la T2A et sans lequel on ne peut escompter lever le dilemme entre rentes et incitations ;

-ou, à défaut, confier la régulation tarifaire à un régulateur indépendant, à l'instar de ce qui a été fait d'en d'autres secteurs, avec pour mission de créer les conditions pour tirer tous les bénéfices sociaux d'une tarification incitative (avec les rentes résiduelles nécessaires eu égard à l'information dont il dispose).

II-Tarification hospitalière incitative et hétérogénéité des patients

II-1-Inconvénients de la T2A « détaillée »

Faut-il conclure de ce qui précède que c'est l'ONDAM qu'il faut seulement incriminer ? Sans doute pas, car il faut aussi considérer les critiques qui concernent plus directement l'orientation des pratiques médicales résultant de la T2A et la manière dont celle-ci affecte, par exemple, la prise en charge des cas plus sévères ou la répartition des patients entre l'hôpital public et les établissements privés.

En effet, l'efficacité de la forfaitisation ajustée par comparaison suppose des DRG parfaitement homogènes par rapport aux malades considérés (ceux-ci relevant d'un même β).

Si[7] le niveau de forfait reflète plutôt les coûts en moyenne de populations en fait hétérogènes, il y a un risque évident de sélection des patients par les offreurs de soins : les services traitant une population moins lourde feront alors indument des bénéfices et ceux ayant des patientèles plus difficiles se trouveront en déficit. Dès lors, les établissements privés (par exemple) vont chercher à ne jamais avoir à traiter les cas les plus lourds; et les établissements publics devront, pour équilibrer leurs comptes, essayer d'attirer à tout prix des patients « rentables ». On voit poindre ainsi la situation des urgences, ou l'âpreté des débats sur la mise en place des Groupements hospitaliers territoriaux, dans un contexte où les patients ont la liberté de choix…

Pour répondre à ce problème, l'idée qui vient est de « découper » les DRG, en introduisant différents niveaux de forfait selon la sévérité, comme cela a été fait (particulièrement en France, (cf. encadré 3) avec la mise en place des nomenclatures extrêmement détaillées. Mais ce n'est une solution satisfaisante que si le diagnostic de sévérité est parfaitement observable par le régulateur. Sinon, ce qui est l'hypothèse réaliste en général, les offreurs de soins ont intérêt à faire passer les malades pour plus graves qu'ils ne le sont, pour bénéficier d'un meilleur prix (« surcodage »).

Encadré 3. Eléments sur la « T2A »

La tarification hospitalière à l'activité pour le court-séjour (médecine, chirurgie, obstétrique ou MCO) a été introduite progressivement à partir de 2005, dans un environnement marqué par ailleurs par un renforcement de la contrainte financière associée à l'objectif global de dépenses (ONDAM), avec pour objectif de fonder le financement des établissements publics et privés en fonction de leur activité, telle que décrite par groupe homogène de séjours. Le passage a été intégral en 2005 pour le secteur privé et graduel pour les hôpitaux publics. Pour permettre un ajustement des pratiques de gestion et d'organisation, ces derniers ont reçu une tarification mixte, comportant encore un pourcentage de « budget global », jusqu'en 2008.

Pour la mettre en œuvre, une base de coûts, « l'étude nationale des coûts », a été constituée dans le cadre du programme de médicalisation des systèmes d'informations (« PMSI »). Initialement, la classification retenue comportait environ 800 groupes. Chaque année, des modifications étaient réalisées à la marge pour prendre en compte des changements dans la prise en charge ou les évolutions thérapeutiques. En 2009, le nombre de groupes atteignait ainsi 2200. Ceci a conduit alors à une systématisation, la nomenclature étant désormais construite sur une base arborescente, aboutissant à 600 racines caractéristiques des pathologies, mais aussi des pratiques ou traitements, auxquelles s'appliquent en plus quatre niveaux de sévérité.

En dépit de ce niveau de détail extrême, les études de Milcent (2017, 2019) observent que « les effets néfastes de sélection des patients ou de diminution du niveau de qualité ne sont pas évités par les forfaits actuels ». Surtout, elles constatent que le raffinement des nomenclatures a conduit sans ambiguïté à du surcodage et des transferts importants entre les différents types d'hôpitaux sans bénéfice en termes de santé publique.

En effet, cette manière de traiter l'hétérogénéité non observable au sein des DRG aura des effets différents selon le mode de régulation retenue, mais toujours néfastes.

Dans le cadre d'une régulation tarifaire où primerait la stabilité des forfaits, seul le forfait destiné aux patients les plus graves tendra à être utilisé finalement. Certes, le risque d'exclusion

[7] « idiosyncracies often prevail over common features »…

sera écarté et les efforts à réduire les coûts resteront optimaux, mais les offreurs de soins tireront des rentes indues quand ils traitent les patients moins graves puisqu'ils bénéficient alors du niveau de prix calibré pour les patients plus lourds. À la limite, le surcoût financier induit peut remettre en cause la préférence à donner à la régulation incitative par rapport aux formules de remboursement des coûts.

Cependant, si, pour réduire ces rentes indues, on réduit le niveau de paiement pour les cas sévères, soigner les malades les plus graves est source de déficit et on doit s'attendre à ce que les offreurs de soins cherchent à s'en débarrasser, compromettant l'objectif d'accès aux soins, de plus à propos des cas les plus graves…ou qu'ils parient que ce déficit leur sera comblé pour éviter cet inconvénient. Mais ceci signifie alors qu'ils anticipent un paiement *cost-plus* et l'effort qu'ils choisiront sera insuffisant. De même, le contrôle des dépenses par le biais d'un mécanisme de points flottants ne serait pas un « rabotage » neutre, mais précipitera *in fine* la sélection des patients et le surcodage. Au bout du bout, seuls les malades moins coûteux seront soignés, mais en utilisant, de plus, toute l'enveloppe qui était prévue pour l'ensemble des cas potentiels : les inconvénients des deux mécanismes, ONDAM et T2A « détaillée », tendent alors à se renforcent…

Si les forfaits sont ajustés par comparaison, l'évolution tendancielle va aussi être cette exclusion des patients les plus lourds, les patients plus légers demeurant toutefois traités efficacement et sans laisser de rente indue aux offreurs de soins.

Ainsi, dès lors que l'information dont dispose le régulateur n'est pas assez fine pour empêcher du « surcodage », des arbitrages vont devoir être réalisés entre les différents objectifs. Cependant, il est possible de les alléger. Pour le montrer, nous considérerons la contrainte de ne pas remettre en cause l'accès aux soins nécessaires comme intangible, et examinerons les possibilités d'optimiser l'arbitrage entre rentes et efficacité. Les travaux de Laffont et Tirole (1993) fournissent les concepts pour cela.

II-2- Comment prendre en compte l'hétérogénéité au sein d'un DRG ?

Considérons, pour simplifier, une pathologie comportant deux niveaux de gravité[8], que l'on schématisera donc par deux niveaux β, notés $\underline{\beta}, \overline{\beta}$ avec $(\overline{\beta} > \underline{\beta})$. On suppose par ailleurs que l'offreur de soins les distingue au moment du diagnostic, mais que le régulateur ne peut le contrôler. Pour gérer une telle situation, plutôt que de chercher à construire un barème unique complexe de remboursement, il faut combiner les deux idées suivantes :

- proposer des « menus » de formules de paiement, conçues pour que les offreurs de soins choisissent entre celles-ci en fonction du diagnostic de gravité, sans surcodage. Comme vu précédemment, pour ne pas détruire les incitations qu'il cherche à établir ainsi, le régulateur doit être crédible sur le fait qu'il effectuera ensuite le paiement correspondant à la formule choisie,

-une telle révélation indirecte du diagnostic sur le niveau de gravité est possible mais n'est pas sans coût, avec un arbitrage à réaliser entre les rentes à consentir à l'offreur de

[8] Pour simplifier l'exposé. Le chapitre I-1 de l'ouvrage de Laffont et Tirole (op.cit.) traite aussi du cas de distributions continues. Par ailleurs, il aborde des questions qui seront laissées de côté ici : opportunité de compléter les dispositifs par des plafonds de coûts ou de profits ; cas où les différences de coûts sont à relier à des choix d'investissements…

soins pour le traitement des patients les moins graves et les surcoûts en termes de volume d'actes s'agissant des malades plus graves. Cependant, la rente informationnelle, c'est-à-dire le profit que pourrait escompter un offreur de soins en choisissant une formule de paiement destinée à des patients plus lourds, croit avec le niveau « d'effort » auquel seront soumis les offreurs de soins pour les patients plus graves. Il est donc possible de réduire ces rentes en distordant celui-ci.

Ensuite, il faut préciser les formules de paiement envisageables au sein de ces menus. À cet égard, il faut intégrer le fait que les patients sont hétérogènes *ex ante*, au moment du diagnostic et choix de la stratégie thérapeutique, mais aussi *ex post*, du fait de l'aléa thérapeutique. Ceci réduit les schémas de paiements envisageables car le coût observable *ex post*, que l'on notera (c), est affecté par l'aléa thérapeutique, ce qui ne permet pas d'estimer indirectement les niveaux d'effort, même si β était connu ; et donc de fixer directement des coûts par niveau de sévérité.

En revanche, les contrats linéaires de type $t = a - bc$ (soit un paiement fixe a complété par un taux de remboursement partiel $(1 - b)$ des coûts observables) sont robustes dans ce contexte. Ceci fournit par ailleurs une panoplie plus large que celle limitée aux deux formules polaires considérées jusqu'à présent, du prix fixe et du *cost-plus*, qui en sont des cas particuliers. Cette classe de contrats permet en fait de contrôler le niveau d'effort qui sera mis en œuvre selon la gravité des patients. En effet, face à une formule de paiement de ce type, l'offreur de soins choisira un niveau d'effort tel que (cf.schéma 2) :

(7) $\psi'(e) = b.$

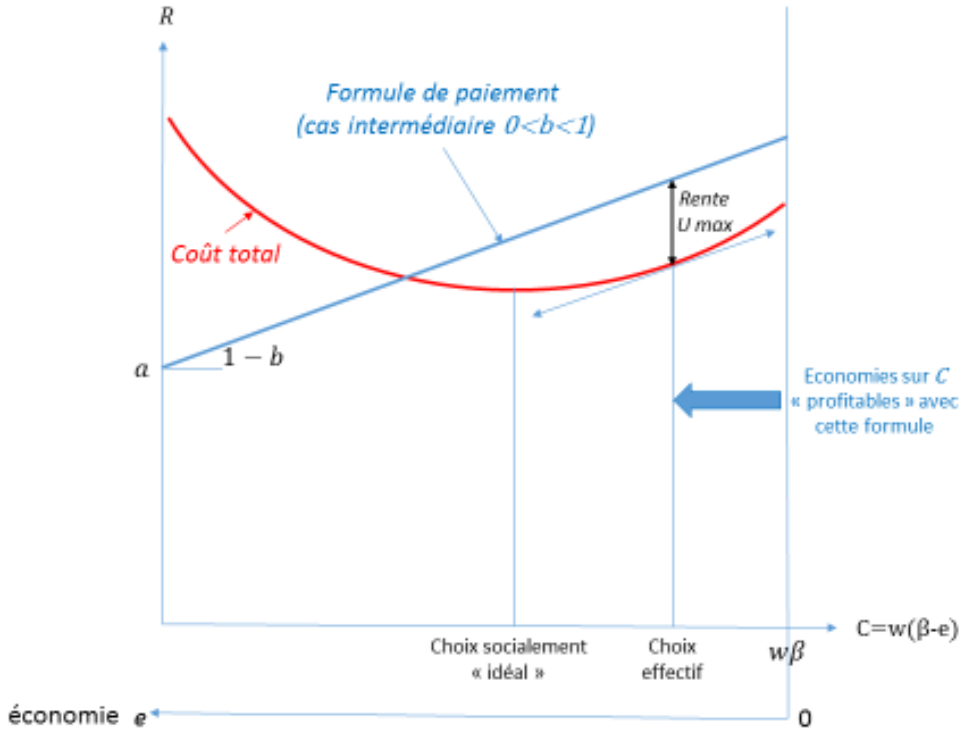


Schéma 2. Choix du niveau d'effort face à une formule de paiement linéaire

Ce schéma, qui illustre ce résultat général en économie de la régulation, conserve les notations de la partie précédente, mais les majuscules $(C = w(\beta - e), R \ldots)$ désignent désormais les

espérances des variables correspondantes, au moment du choix du niveau d'effort. Pour $b = 1$ (prix fixe), les incitations à réduire le nombre d'actes sont donc puissantes, faibles quand $b \to 0$ (*cost-plus*).

Appliquant ces principes, on cherche le menu de deux contrats linéaires à mettre en place par un régulateur qui ne peut observer le diagnostic de gravité mais a seulement une information sur la proportion ν de cas moins graves. Pour caractériser ces contrats, on notera respectivement $\underline{t}, \overline{t}$ les contrats destinés à être choisis, selon que le patient est de type $\underline{\beta}$, $resp. \overline{\beta}$. On notera $\underline{e}, \overline{e}$ les niveaux d'efforts associés, et $\underline{U}, \overline{U}$ les espérances de bénéfice associées (compte-tenu de l'aléa thérapeutique) si l'option choisie est bien celle destinée au type de malade considéré.

II-3-Menus d'options tarifaires optimaux

Comme dans la première partie, le bouclage avec l'offre de ressources médicales renforce l'exigence d'efficacité. Le menu d'options tarifaires proposé aux offreurs de soins doit donc minimiser le volume moyen anticipé de ressources pour traiter un cas, au sein du DRG hétérogène comprenant les deux niveaux de gravité, sous deux contraintes :

-assurer que le traitement des patients les plus graves ne génère pas de perte systématique, sinon leur accès aux soins serait compromis. Il faut pour cela que l'option conçue pour les patients les plus graves ne soit pas déficitaire, donc $\overline{U} = 0$. On aboutit donc à une formule de paiement, génériquement du type :

$$\overline{t} = w\psi(\overline{e}) + \psi'(\overline{e})[w(\overline{\beta} - \overline{e}) - c)] \quad (8)$$

Celle-ci combine la compensation de l'effort visé pour ce degré de sévérité et un « bonus-malus » par rapport au cout moyen des actes observables pour ce niveau d'effort et ces patients. Face à une telle formule de paiement, l'offreur de soins choisira effectivement le niveau d'effort visé ($\bar{e}$, restant à déterminer) et le second terme sera donc égal à zéro en moyenne pour les patients les plus graves.

-articuler les deux types de paiements de manière à ce que l'offreur de soins ne soit pas incité à « surcoder », donc choisisse effectivement l'option qui lui est destinée (« autosélection » du type de contrat correspondant à la gravité du malade) dans le cas de patients moins sévères. Sachant que, s'il surcodait, c'est-à-dire s'il choisissait la formule de paiement précédente alors qu'il traite un cas plus léger, il aurait encore intérêt à choisir le niveau d'effort $\overline{e}$, et réaliserait donc un bénéfice escompté égal à $\psi'(\overline{e})w(\overline{\beta} - \underline{\beta})$. Dès lors, l'autre formule de paiement, destinée aux cas moins graves, doit assurer un bénéfice $\underline{U}$ équivalent pour éviter le surcodage. En plus de la compensation de l'effort visé dans ce cas ($\underline{e}$, à déterminer aussi), il faut donc admettre de laisser la rente informationnelle correspondante, d'où une formule du type:

$$\underline{t} = w\psi(\underline{e}) + \psi'(\overline{e})w(\overline{\beta} - \underline{\beta}) + \psi'(\underline{e})[w(\underline{\beta} - \underline{e}) - c)] \quad (9)$$

Le coût de l'effort étant supposé fortement croissant ($\psi'' > 0$), cette rente informationnelle $\psi'(\overline{e})w\left(\overline{\beta} - \underline{\beta}\right)$ croit avec l'effort demandé pour le traitement des malades les plus lourds. Dans ces conditions, il est souhaitable, pour limiter les rentes abandonnées aux offreurs, de

soins, de distordre les choix de traitement pour les malades les plus lourds dans le sens d'un accroissement relatif des actes et d'une diminution de l'effort, par rapport à ce que serait optimal en information parfaite.

Ainsi la définition des deux options du menu se trouvent liées. De manière plus précise, le régulateur a intérêt à ce que les niveaux d'efforts qui seront induits par le menu de contrats qu'il propose minimise :

$$Min_{(\underline{e},\overline{e})}\ \nu\left[(1+\lambda)\left(\underline{\beta}-\underline{e}+\psi(\underline{e})\right)+\lambda\psi'(\overline{e})(\overline{\beta}-\underline{\beta})\right]+(1-\nu)(1+\lambda)\left(\overline{\beta}-\overline{e}+\psi(\overline{e})\right)$$

D'où :

Proposition 2 : le menu d'options optimal est associé aux niveaux d'effort :

$$\begin{cases}\underline{\boldsymbol{e}}^{**}=\boldsymbol{e}^{*}=\mathbf{1}\\ \boldsymbol{\psi}'(\overline{\boldsymbol{e}}^{**})=\mathbf{1}-\dfrac{\boldsymbol{\lambda\nu}}{(\mathbf{1}+\boldsymbol{\lambda})(\mathbf{1}-\boldsymbol{\nu})}(\overline{\boldsymbol{\beta}}-\underline{\boldsymbol{\beta}})\boldsymbol{\psi}"(\overline{\boldsymbol{e}}^{**})\end{cases}$$

Le menu d'options intégrant ces niveaux d'effort dans les formules de paiement (8) et (9), conduira les offreurs de soins à ne pas surcoder et ils choisiront alors « spontanément » les niveaux d'effort optimaux sous les deux contraintes à prendre en compte.

En particulier, il conduit à l'optimisation de premier rang du nombre d'actes pour le traitement des patients moins graves, le contrat correspondant restant de type prix-fixe. Toutefois, son niveau est ajusté pour écarter le surcodage, en ajoutant à la compensation de l'effort, la rente informationnelle ajustée :

$$(10)\qquad \underline{t}=w\psi(e^{*})+\psi'(\overline{e}^{**})w\left(\overline{\beta}-\underline{\beta}\right)+\left(w\left(\underline{\beta}-e^{*}\right)-c\right)$$

Pour limiter le relèvement du forfait nécessaire pour éviter les contournements, l'incitation à réduire les volumes d'actes est donc choisie moins puissante pour les malades plus graves, par rapport à ce que ce qui serait idéalement souhaitable. Toutefois, le niveau de responsabilisation à la réduction des coûts demeure relativement élevé si le coût social des fonds publics ou la part des malades moins graves est faible. En effet, le premier paramètre détermine le coût social des rentes et le second le volume de rentes à consentir. En revanche, plus ceux-ci sont élevés, plus l'option de paiement dans le cas des malades graves tend vers le *cost-plus* pur ($\overline{t}=0$).

II-4-Ajustement des barèmes

Le dispositif ainsi construit peut s'interpréter : soit comme la mise en place d'un menu d'options au sein d'un DRG conçu globalement, tous niveaux de sévérité inclus ; soit comme un éclatement du DRG, le codage du niveau de sévérité déterminant la formule de paiement qui sera appliquée. L'analyse qui précède souligne cependant que les deux formules de paiements doivent être conçues de manière cohérente pour éviter le surcodage. Il faut donc privilégier la première interprétation. Dans cette perspective, l'inflation des nomenclatures à laquelle a conduit la T2A reflète un processus de création de menus d'options « de fait », mais dont le *design* est très imparfait.
Par ailleurs, l'argument de complexité qui est souvent mis en avant pour écarter de tels « menus d'options » est à relativiser car nous sommes tous habitués à choisir entre différents niveaux de

couverture et les primes associées dans le domaine des assurances privées, par exemple. Surtout, la logique sous-jacente aux mécanismes proposés ici est strictement symétrique à celle conduisant les gestionnaires d'équipements tels que les musées, les équipements sportifs ou les réseaux de transport à proposer le choix entre des formules d'abonnement et des tickets à la visite, selon qu'ils visent les usagers à forte demande ou ceux plus occasionnels : dans ce cas, le prix à l'unité distord la demande de ces derniers pour éviter le contournement des abonnements dont on attend la contribution principale à l'équilibre financier des services. La démarche suivie pour établir une tarification hospitalière efficace ne fait donc que transposer ces méthodes d'une situation de vendeur à celle d'un acheteur.

Notant $b^{**} = \psi'\left(\overline{e}^{**}\right)$ le taux de forfaitisation des coûts s'agissant des malades les plus graves, et respectivement $\overline{C}^{**} = w\left(\overline{\beta} - \overline{e}^{**}\right)$, $\underline{C}^{*} = w(\underline{\beta} - e^{*})$ les coûts observables moyens de traitement qui seront ainsi induits pour les deux niveaux de sévérité, on peut réécrire ce menu d'options par rapport aux paiements ($r = t + c$).

Proposition 3 : les formules de paiement optimales s'écrivent respectivement :

$$\begin{cases} \overline{\boldsymbol{r}} = \boldsymbol{c} + \boldsymbol{b}^{**}\,(\overline{\boldsymbol{C}}^{**} - \boldsymbol{c}) + \boldsymbol{w}\,\boldsymbol{\psi}\left(\overline{\boldsymbol{e}}^{**}\right) \\ \underline{\boldsymbol{r}} = \boldsymbol{w}\boldsymbol{\psi}(\boldsymbol{e}^{*}) + \underline{\boldsymbol{C}}^{*} + \boldsymbol{b}\left(\overline{\boldsymbol{C}}^{**} - \underline{\boldsymbol{C}}^{*} - \boldsymbol{w}(\boldsymbol{e}^{*} - \overline{\boldsymbol{e}}^{**})\right) \end{cases}$$

Les contraintes informationnelles et les arbitrages associés sont illustrés ci- dessous : plus de forfaitisation (b) pour les malades sévères réduit le coût de traitement des cas graves mais implique de laisser une rente plus élevée pour les moins sévères ; la pondération entre ces deux effets dépend de la proportion des deux types de cas pour chaque DRG (cf. schéma 3).

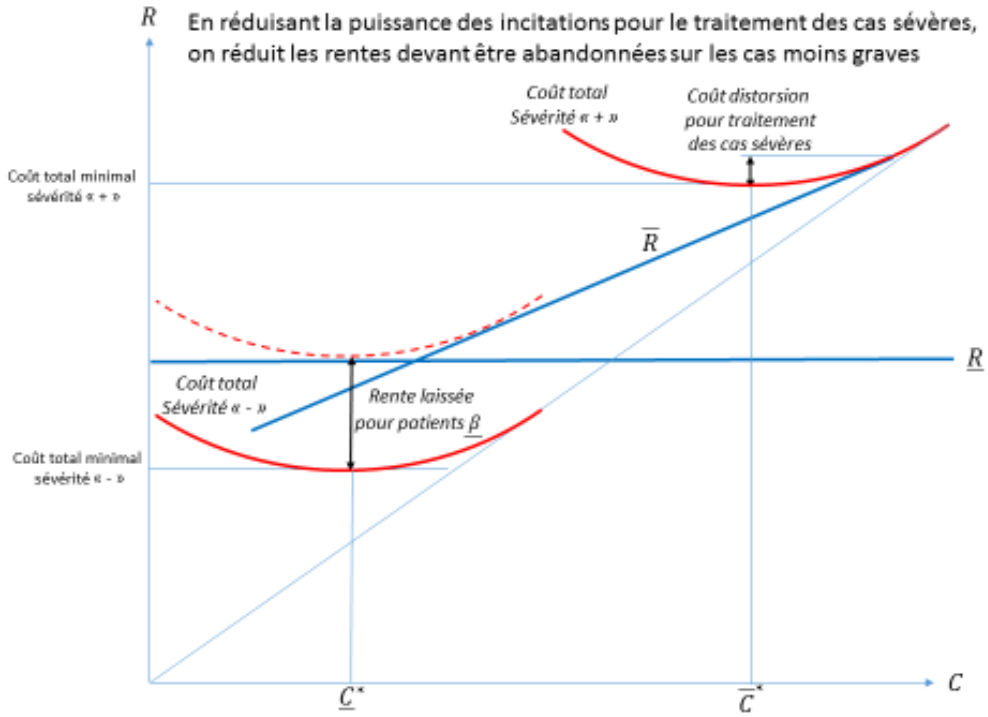


Schéma 3. Arbitrages associés à la construction du menu d'options

À nouveau, la mise en œuvre nécessite de connaitre les valeurs de référence pour les coûts $(\overline{\beta}, \underline{\beta})$. Mais on peut adapter les principes de la concurrence par comparaison pour cela[9].

L'adaptation du mécanisme de concurrence par comparaison consisterait à proposer pour l'ensemble du DRG : le menu défini par les propositions 2 et 3 dans lequel toutes les grandeurs seraient exprimées en fonction de $\overline{C}^{**}$ et $\underline{C}^{*}$; celles-ci étant estimées à partir des coûts observables réalisés en moyenne par les offreurs de soins ayant choisi respectivement la première ou la seconde option.

Cependant, annoncer qu'un tel mécanisme s'appliquera serait sans doute peu lisible, les opérateurs devant notamment anticiper le taux de forfaitisation b qui s'appliquera dans la formule de paiement destinée aux cas sévères.

Dans ces conditions, il faut sans doute envisager plutôt un mécanisme en deux étapes dans lequel le régulateur commence par fixer, à partir de l'estimation qu'il peut faire de $\overline{e}^{**}$ (cf. proposition 2), le niveau d'effort $\overline{e}$ qu'il vise dans ce cas, le taux de forfaitisation $b = \psi(\overline{e})$ s'en déduisant, et donc l'espérance du coût de traitement observable $\overline{C} = w(\overline{\beta} - \overline{e})$.

Ce paramètre étant fixé, le menu d'options optimal est :

$$\begin{cases} \overline{r} = c + b\,(\overline{C} - c) + w\,\psi(\overline{e}) \\ \underline{r} = w\psi(e^{*}) + \underline{C}^{*} + b\big(\overline{C} - \underline{C}^{*} - w(e^{*} - \overline{e})\big) \end{cases}$$

Il peut être mis en œuvre en remplaçant $\overline{C}$ et $\underline{C}^{*}$ par les coûts observés en moyenne selon la formule de paiement choisie. La première formule, destinée aux cas sévères, correspond à un *cost-plus* corrigé, intégrant un terme incitatif (bonus-malus selon que coût est inférieur ou supérieur aux coûts moyens observés pour cette formule de paiement), le supplément d'effort ainsi induit étant par ailleurs compensé. La seconde formule correspond à un paiement à prix fixe, qui comprend en plus de la compensation forfaitaire du coût optimal, un bonus pour inciter à ne pas surcoder, ajusté sur l'écart de coût observé entre les deux formules et réduit si l'effort pour les cas sévères est relâché.

III- Diversification des modes de paiements. Quelques points d'attention

II-1-Enjeux et actualité

Constatant que le financement actuel « favorise insuffisamment la qualité, la prévention et la coordination et peut inciter à la réalisation de soins non pertinents », le rapport Aubert (2019) qui avait servi de base au projet « Ma santé 2022 » envisageait une diversification et une restructuration profonde des modes de paiements (cf. schéma 4). L'approche générale en a été confirmée par le « Ségur de la santé », qui réaffirme l'objectif d'accroitre globalement le niveau de qualité des prises en charge et, en particulier, d'améliorer la pertinence des soins en réduisant les soins inadéquats ou inutiles. Le schéma 4 ci-dessous illustre ces orientations.

[9] Cf. Laffont-Tirole, I-1.7, op.cit.

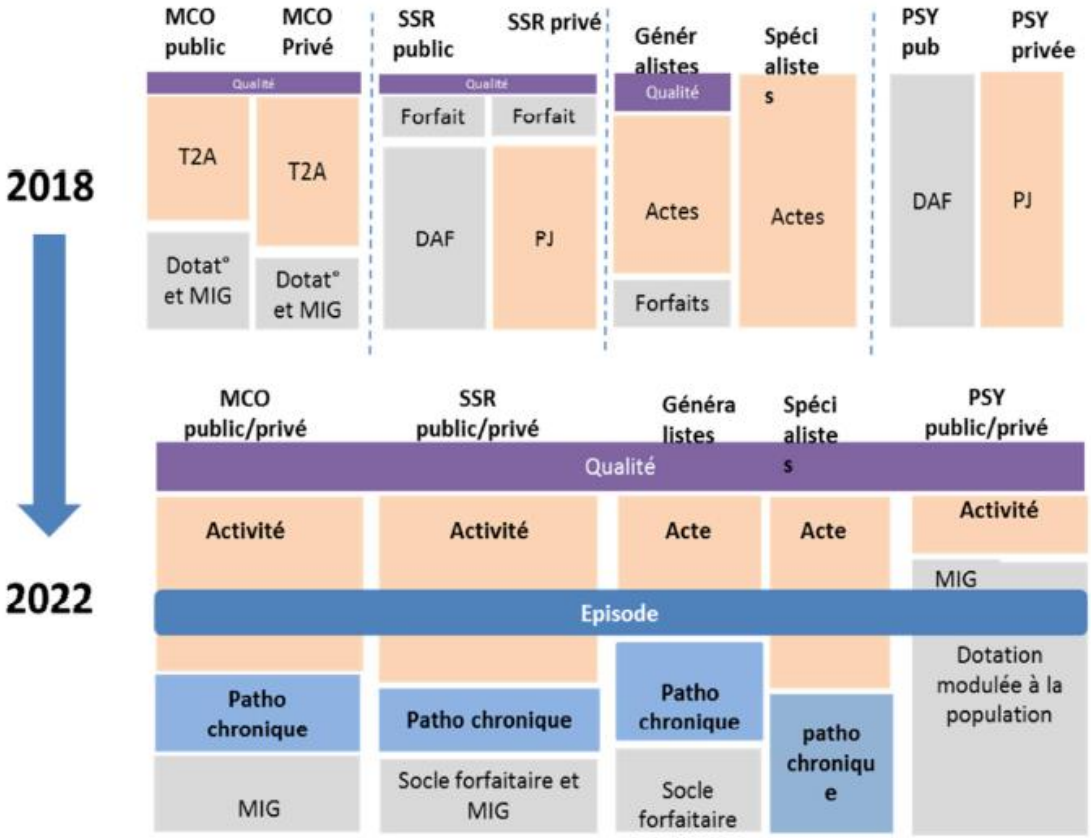


Schéma 4. Architecture du projet

L'objectif de favoriser le suivi au long cours des patients fonde les propositions de paiement au suivi ainsi que les dotations populationnelles, dans un contexte de progression continue des maladies chroniques, Par ailleurs, l'objectif de responsabiliser les acteurs à la qualité globale délivrée et celui de renforcer la prévention se concrétisaient dans la proposition par des paiements à la qualité et à la séquence de soins : par ce type de paiement dit « groupé », il s'agit de rémunérer conjointement des acteurs aujourd'hui financés séparément, pour qu'ils se coordonnent plus efficacement, ceux-ci s'accordant sur la répartition du financement commun.

Dans ces conditions la part des paiements à l'acte en ambulatoire diminuerait. S'agissant de la tarification hospitalière, il était noté que la T2A demeurerait une modalité importante de financement dès lors que les épisodes uniques de soins représentent environ 55% des séjours hospitaliers. Outre les questions de nomenclatures, il était pointé des besoins d'évolutions pour favoriser des pratiques ambulatoires, insuffisamment rémunératrices en l'état. Le rapport insiste par ailleurs sur la nécessité d'établir une régulation lisible, compréhensible et prospective, ainsi que sur le besoin de faciliter l'accès à l'innovation.

L'analyse qui précède conduit à souligner le besoin de construction articulée de ces différents modes de paiements, en prenant bien en compte les hétérogénéités, pour qu'il soient effectivement utilisés dans les domaines visés, sans contournements et exclusion. De plus, quand différentes solutions peuvent être envisagées par rapport à un même problème, il convient de les comparer.

Par exemple, l'argument déterminant en faveur des dotations populationnelles est celui du suivi des patients. Mais il faudrait préciser alors si l'on cherche : à résoudre ainsi que, dans certains

contextes, l'historique est intrinsèquement difficile à faire partager, donc que le besoin de suivi ne peut être résolu seulement par la régulation de l'accès au dossier médical ; ou plutôt à corriger le risque d'attention insuffisante à certaines conséquences potentielles à moyen terme dans le cadre d'une gestion par épisode de soins, si ce risque ne peut aisément être contenu par l'introduction de paiements à la qualité appropriés, ce qui constituerait un moyen alternatif pour responsabiliser les offreurs de soins à ces enjeux, ré-hospitalisations, par exemple. L'analyse de l'information disponible et des besoins de responsabilisation en résultant sont donc essentiels pour faire la balance des avantages et des inconvénients des différentes solutions envisageables dans les différents contextes.

II-2- Dotations populationnelles et hétérogénéité des *case-mix*

Comme pour la tarification hospitalière à l'activité, la définition des dotations populationnelles sera confrontée à des problèmes d'hétérogénéité, d'autant plus aigus que les variables socio-démographiques ne permettent en général de capturer qu'une part limitée de l'hétérogénéité des *case-mix*. Dès lors, un paiement unique à prix fixe, de type capitation et *fundholding* pour les examens et prescriptions, même ajusté sur ces variables, ne suffit pas : ceux qui traitent des populations plus lourdes vont se trouver en déficit, avec comme échappatoire de sélectionner *ex ante* leur patientèle ou le renvoi des épisodes de soins trop coûteux sur les urgences hospitalières, par exemple ;…et l'on retrouvera donc, comme pour la T2A, les tensions entre rentes, efficacité et risque de sélection des patients…

Il faudra donc examiner aussi cette forme de paiement sous l'angle de l'hétérogénéité des *case-mix*. Ceci peut d'ailleurs fournir des points de repères pour l'évolution de la tarification hospitalière, notamment en médecine, ou pour concevoir des mécanismes de responsabilisation plus efficaces en ville, incitant à réduire les actes, prescriptions et examens non justifiés. A cet égard, le cadre d'analyse développé ci-dessus conduit en effet à souligner que la mise en œuvre de la tarification à l'activité nécessite que le volume d'actes réalisés, notamment ceux pour traiter les malades plus graves, fasse l'objet de comptabilité détaillée, dans des conditions garantissant l'impossibilité d'y imputer les coûts de malades légers. Dès lors, des dotations populationnelles peuvent constituer une réponse à chaque fois que cela est trop coûteux à mettre en œuvre.

Le problème d'hétérogénéité est donc reporté au niveau des *case-mix*. S'appuyant sur les indicateurs demeurant observables au niveau de la patientèle de chaque offreur de soins, les formules de paiement peuvent être conçues pour les différencier par un mécanisme révélateur, similaire à celui décrit dans la partie précédente. Cependant, le régulateur ne peut alors contrôler l'effort qu'au niveau global. Sachant qu'à coût moyen donné des actes générés, l'offreur de soins a intérêt à allouer son effort entre les différents niveaux de gravité de malades là où il sera le plus profitable, ceci le conduit à égaliser les réductions marginales de coût des actes permises par un supplément d'effort. Avec les spécifications précédentes, l'offreur de soins choisira donc le même niveau d'effort pour les différents niveaux de gravité. L'analyse précédente reste donc qualitativement pertinente, non plus au niveau de l'épisode unique de soin d'un patient, mais des *case-mix* des différents offreurs de soins.

Elle peut alors être reprise pour définir les modes de paiements optimaux, en considérant non plus les β correspondants aux deux niveaux de gravité, mais les β moyens des populations auxquelles sont confrontées les différents offreurs de soins. Si l'on considère deux types de *case-mix* possibles, il conviendra donc de proposer un menu d'options, combinant un paiement à prix fixe pour les offreurs de soins exerçant sur des populations plus favorables et un contrat

linéaire comportant une part de remboursement du coût moyen observable, pour les populations plus lourdes, distordant donc le niveau d'effort dans ce cas de manière à limiter la rente abandonnée aux précédents.

En fait, dans ce contexte, il serait plus pertinent de considérer un modèle dans lequel il n'y a pas un ensemble discret de niveaux de sévérité mais plutôt un continuum de *case-mix* et donc de coûts de référence au sein des offreurs de soins traitant les populations concernées. Si chacun de ces offreurs de soins a une patientèle suffisamment large, telle que l'effet de l'aléa thérapeutique se compense au sein de celle-ci, le mécanisme de tarification optimale à considérer consiste en un barème complexe de type capitation par personne suivie, dont le montant serait fonction du coût observable moyen constaté pour chaque offreur de soins, soit : $\Re(C)$.

Cependant, si la distribution des *case-mix* est régulière, on sait que ce barème peut être remplacé par un menu d'options de formules de paiements linéaires proposé aux offreurs de soins. Procéder ainsi n'est pas restrictif (cf. schéma 5). La seule question est d'apprécier le nombre d'options intermédiaires à introduire en plus des deux options extrêmes considérées jusqu'à présent pour avoir ainsi une approximation satisfaisante du barème optimal, quand ceci apparaît nécessaire.

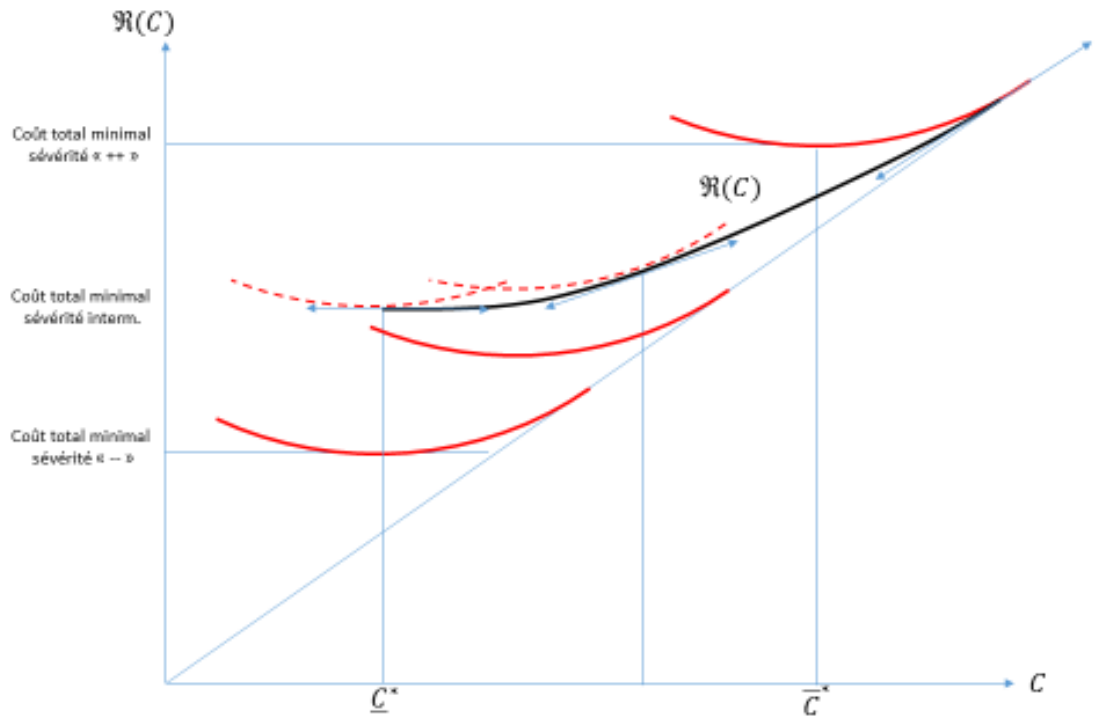


Schéma 5. Approximation du barème de capitation par un menu d'options linéaires

Cette simple adaptation du modèle de la tarification par DRG pour l'appliquer aux dotations populationnelles suppose toutefois que le processus de sélection des patients n'intervient que si l'offreur de soins est globalement en déficit, en d'autres termes que son « éthique » le conduit à accepter des pertes sur les malades les plus graves dès lors qu'elles se compensent par des gains suffisants sur les malades plus légers.

Si ce n'est pas le cas, le niveau de la part fixe dans le cas des malades plus graves doit être relevé, avec donc des rentes pour tous les offreurs de soins. Ceci n'est cependant pas à proprement parler un avantage déterminant d'un paiement plus « populationnel » : dans le cas de référence, à l'épisode de soins, la contrainte de profitabilité est aussi allégée si le service hospitalier se contente d'un équilibre global, avec alors des subventions croisées entre niveaux de sévérité.

Le point est donc d'ordre plus général : l'éthique médicale permet de limiter les rentes devant être abandonnées aux offreurs de soins. En conséquence, il y a lieu, quand on introduit des incitations plus puissantes à la maitrise des coûts, de prendre garde à ne pas bouleverser des éléments de comportement qui sont socialement précieux, d'où l'importance de construire une vision partagée des cadres tarifaires.

II-3-Paiements à la qualité

Pour apprécier les enjeux de tels paiements, il est nécessaire de revenir sur les relations entre coûts et qualité des soins. En effet, la référence à la « maitrise médicalisée des dépenses » nourrit les controverses entre ceux qui en déduisent que la santé ne saurait avoir de prix versus l'idée qu'il n'y aurait pas réellement d'arbitrage, la maitrise des coûts étant sensée toujours bénéficier à l'état de santé des populations. À l'encontre de ces débats, nous avons pris soin de préciser que l'objectif de minimisation des coûts que nous considérions s'entendait « à qualité donnée », le modèle étudié posant une obligation de traitement, dans des conditions définies par ailleurs, par l'état de l'art des connaissances et des techniques médicales (cf. encadré 4).

Encadré 4. La maitrise des coûts, de quoi parle-t-on ?

Aspects économiques de l'évaluation des traitements médicaux

Le souci de développer une médecine fondée sur des bases objectives (« evidence-based »), hiérarchisant rigoureusement les priorités, s'est progressivement imposé depuis la fin des années 80. La réalisation de cet objectif passe d'abord par la disponibilité d'observations suffisantes sur les impacts des traitements, et la caractérisation des niveaux de preuve, l'idéal épidémiologique étant l'essai thérapeutique randomisé. Le problème a cependant aussi une dimension économique. Le risque zéro n'existe pas. Les ressources que la collectivité peut consacrer à la santé sont rares, et évincent d'autres actions publiques, potentiellement légitimes et importantes. Enfin, au sein même des dépenses de santé, il faut arbitrer entre différents types de dépenses. Ceci oblige donc à préciser le seuil pour lequel on admettra que le coût par année de vie gagnée d'une mesure sera jugé acceptable ou non. Une telle valeur de référence commune est la condition pour établir efficacement des priorités sur l'ensemble du système de santé. Ainsi, la quantification en termes d'analyse coûts-bénéfices aide à déterminer si la variation nette des risques pour la santé associée à une intervention justifie le coût d'opportunité des ressources utilisées pour l'atteindre. L'élaboration des « recommandations » des sociétés savantes ou des autorités de santé en charge de définir les normes médicales doit donc s'y référer, pour éviter, par exemple que des équipements rares ne se trouvent saturés par des cas ne le méritant pas, évinçant ou allongeant les listes d'attente pour ceux qui en auraient besoin. Les paiements aux offreurs de soins doivent être compatibles avec ces normes dont l'élaboration se situe cependant en amont des choix de tarification.

Approches macroéconomiques vs microéconomiques

La France a privilégié une approche macroéconomique de la régulation des dépenses de santé combinant le contrôle de la démographie médicale et la mise en place d'enveloppes budgétaires. Cette approche s'inspirait notamment de « l'hypothèse de la demande induite » établissant un lien entre la densité médicale et le niveau des dépenses de santé, hypothèse fortement débattue dans les années 80, par rapport au sens des causalités impliquées. Le débat s'était cependant déplacé dès les années 90, le doute s'étant installé, à la fois sur les résultats de beaucoup de ces études, et sur les conséquences à en tirer pour les politiques sanitaires. Notamment, un doute définitif est venu d'études trouvant des effets de demande induite dans des contextes de rémunération ou de spécialité, où celui-ci est très difficile à expliquer. La plus spectaculaire à cet égard est celle celle de Dranove et Wehner (1994) « montrant » un effet de « demande induite » de la densité d'obstétriciens sur le volume des naissances ! Dans ces conditions, la question de « l'aléa moral » du côté des offreurs de soins devait être abordée différemment, d'un point de vue plus microéconomique, c'est à dire à partir des divergences éventuelles entre leurs prescriptions et ce qui serait socialement souhaitable, avec en perspective la mise en place de cadres « responsabilisateurs »…

Dans ce contexte, il reste cependant que la forfaitisation des paiements peut être incompatible avec le respect de certaines normes médicales définies sous des pressions incitant à l'évitement des questions de coûts. Leur non-respect peut aussi constituer une alternative à la sélection des patients en cas de tension. La définition de modes de paiement prenant mieux en compte l'hétérogénéité des patients est donc en soi une première réponse à ce problème, comme pour les risques de sélection des patients, à accompagner cependant d'un contrôle renforcé du risque de déviation de la norme, d'autant plus élevé que les incitations à réduire les coûts sont fortes.

Par ailleurs, il est des cas où certains aspects de qualité sont importants pour les patients ou le système de soins, mais spontanément relativisés par celui qi gère l'épisode de soins. Comment assurer alors que les dimensions correspondantes sont suffisamment prises en compte?

Pour répondre à cette question, on peut adapter notre cadre d'analyse en introduisant une dimension de qualité des soins au-delà de ce qui est fourni obligatoirement (mesurée par une[10] variable notée q) ayant une valeur pour le patient mais qui évidemment coûte pour l'offreur de soins. On supposera de plus que l'incitation à réduire les coûts est, toutes choses égales par ailleurs, antagoniste avec cette qualité, pouvant alors directement mesurée par son coût. En d'autres termes, on considère des fonctions :

$$S(x, \beta, q, e) \text{ avec } S_q > 0, S_{qq}, S_e < 0 \text{ et } C = \beta + q - e)$$

Si l'on reprend la démarche de la première partie et que l'on suppose dans un premier temps un DRG parfaitement homogène, la politique optimale vérifierait maintenant :

$$\begin{cases} (e): & \psi'(e) = 1 + S_e(\beta, q, e)/(1 + \lambda') \\ (q): & S_q(\beta, q, e) = (1 + \lambda') \end{cases}$$

La première équation exprime que l'effort optimal de réduction des coûts observables sera réduit, par rapport à ce que l'on avait ci-dessus, pour tenir compte de son impact défavorable

[10] En pratique, une difficulté importante est le caractère très multidimensionnel de la qualité des soins, le risque étant en effet que des politiques focalisées sur des aspects partiels conduisent essentiellement à l'abandon des efforts dans les dimensions non prises en compte.

sur cette qualité. La seconde fixe le niveau optimal de qualité à fournir compte-tenu de la balance coûts-avantages (sociaux) qui y est associée et du financement public de ceux-ci.

Un contrat linéaire de type $t = a - bC$ avec l'offreur de soins n'est ici plus suffisant pour atteindre cet optimum, car il ne fournit pas d'incitation à produire la qualité q : dès lors, la production de cette qualité au-delà de la norme est seulement une source de coûts pour l'offreur de soins, il n'y a pas de raison de s'y engager pour celui-ci. Ceci affecte en retour le niveau d'effort à réduire les coûts que l'on peut viser. En effet, la première équation reste valide, mais avec comme valeur pour la qualité $q = 0$.

L'hypothèse la plus raisonnable est que l'effort à réduire les coûts que l'on devrait viser alors est réduit, d'autant plus que l'enjeu de la qualité est fort. En effet, la variable d'effort doit contrôler deux objectifs antagonistes : réduire les coûts vs préserver une qualité suffisante. En conséquence, si cette variable n'est pas directement contrôlable par le régulateur ou si celui-ci se limite aux formules de paiement précédentes, sans paiement à la qualité, il devra s'écarter de la tarification à prix fixe et appliquer une formule avec des incitations moins puissantes.

Dans certaines conditions (Mougeot et Naegelen, 2018), ce conflit d'objectifs complique la résolution des problèmes de sélection des patients, lorsque l'on ré-introduit leur hétérogénéité. Ce sera le cas si le problème de qualité est essentiel pour les niveaux de sévérité plus faibles (les protocoles pour les malades plus graves intégrant déjà des normes strictes contrôlées) et si la proportion de cas sévères domine.

En effet, il devient alors préférable de proposer un seul type de contrats plutôt qu'un menu d'options construit comme précédemment, qui bute sur le fait que les coûts pris en compte dans la formule de paiement peuvent devenir plus élevés pour les malades moins sévères. Dès lors, il faut aussi prendre en compte la contrainte supplémentaire[11], que le contrat conçu pour stimuler la qualité par rapport aux malades moins sévères ne soit pas *in fine* plus intéressant pour les offreurs de soins confrontés aux malades plus lourds. Évidemment, le contrat unique correspondant est un compromis très contraint par le conflit d'objectifs entre réduction des coûts, qualité des soins, rentes et accès aux soins[12] : les DRG sont plus larges et les incitations appliquées uniformément au sein de ceux-ci sont à un niveau inférieur.

Sous réserve que la qualité soit bien observable, l'introduction de paiements à la qualité (τ), soient des formules de paiement de type $t = a - bC + \tau q$, permet d'alléger ces conflits : le paiement à la qualité orientant le choix de qualité au niveau approprié ; et le coefficient de partage des coûts déterminant essentiellement la puissance des incitations, qui peut donc être relevée.

En présence d'hétérogénéité, les paiements à la qualité doivent cependant être différenciés si la valeur de celle-ci diffère selon les niveaux de sévérité. De plus, l'ajustement des termes forfaitaires réalisé au moment où ce nouveau type de paiement est introduit doit prendre en compte que les revenus des paiements à la qualité seront hétérogènes selon les *case-mix* : si l'hypothèse précédente concernant les enjeux de qualité visés prévaut, un ajustement seulement en moyenne suite à l'introduction des paiements à la qualité risquant de générer des déficits sur les patients les plus lourds…

[11] Qui était automatiquement vérifiée dans ce qui précède.

[12] Dans ce cas, cela signifie qu'il faut renoncer à l'idée de raffinement des DRG, assurer la cohérence entre ceux-ci devenant trop coûteux ou impossible. Choné et Ma (op.cit.) aboutissent à des résultats similaires lors que les dimensions d'hétérogénéité sont multiples.

II-4-Paiements pour une meilleure coordination des soins

L'introduction de paiements à la qualité permet de faire « internaliser » par l'offreur de soins des dimensions non prises en compte spontanément par celui-ci, s'apparentant, du point de vue économique, à des externalités. La qualité pour le patient en est une dimension mais d'autres externalités sont entre offreurs de soins : il faut inciter, par exemple, à prendre des mesures qui éviteront les ré-hospitalisations, ou pour que les soignants prennent le temps de l'éducation des patients pour la prévention[13].

Pour les dimensions observables, la référence pour construire les éléments de tarification correspondants est donc l'approche « pigouvienne » : selon les cas, on peut procéder par bonus ou par malus, mais il importe que les assiettes soient assises le plus directement possible sur la qualité ou la non-qualité ; et les barèmes établis par rapport à la valeur de celle-ci ou au coût de celle-là.

Cependant, la mise en place de tels paiements, s'apparentant à un « prix » de la qualité, n'est pas le seul instrument possible pour régler ce type de problème. S'agissant, d'externalités répétées impliquant de fortes de dimensions de qualité et de risque, le développement de certaines formes de relations plus intégrées -contractuelles ou par intégration effective- entre certains offreurs de soins peut constituer une alternative, préférable si cela permet de favoriser un investissement commun qui ne pourrait l'être autrement.

Dans cette perspective, la diversification des modes de paiements qui est envisagée dans le projet « ma santé 2022 » peut servir le besoin de mieux orienter les choix des offreurs de soins par rapport aux multiples externalités qui sont associées à leur pratique.

Cependant, les solutions les plus intégrées peuvent aussi avoir des inconvénients sur le fonctionnement d'ensemble, d'où des arbitrages entre : le risque de perdre une réactivité de l'offre précieuse sur des segments où la concurrence par la qualité serait précieuse ; et la duplication excessive de certains coûts fixes, là où une concurrence entre filières n'est pas vraiment possible. En tout état de cause, il convient : de bien articuler les paiements pour des modes d'exercice coordonnés avec les mécanismes de paiements conçus dans un contexte d'intervenants indépendants, pour ne pas permettre ainsi la multiplication d'actes inutiles mais bien rémunérés ; et d'évaluer la cohérence d'ensemble au regard de l'hétérogénéité des patients concernés.

Conclusion

Comme l'avait soulignée la présentation du projet « Ma santé 2022 », les modalités de financement sont un puissant moteur pour faire évoluer les comportements et les organisations, via « les ressources qu'elles permettent de consacrer à un patient donné et le signal qu'elles adressent aux différents professionnels ». Différents types de paiements nouveaux sont envisagés dans cette perspective, pour accroitre les incitations à la qualité ou inciter à une meilleure coordination entre acteurs ou niveaux de recours, ou pour faire évoluer les modes de paiements de base.

[13] Dans un contexte marqué par les pénuries de certaines compétences, dont certaines sont appelées à durer, il pourrait être intéressant d'envisager aussi des paiements pour les pratiques qui libèrent des capacités rares et des pénalités pour celles qui accroissent en fait les goulots d'étranglement.

Dans tous les cas, l'hétérogénéité inobservable des malades est un élément essentiel à prendre en compte, pour concevoir des dispositifs ne risquant pas de générer des incitations non désirables à l'exclusion de certains patients. La théorie de la régulation incitative fournit les références pour définir les tarifications correspondantes, allégeant les antagonismes entre maitrise des coûts, qualité et accès aux soins. En particulier, la définition de menus d'options au sein des DRG apparaît préférable à la construction d'un barème unique complexe. Mais la multiplication des DRG, telle qu'elle a été mise en œuvre jusqu'à présent au sein de la T2A, a ignoré la question du surcodage. La combinaison d'un forfait et d'une option linéaire, à l'instar de ce qui existe dans beaucoup de tarifications publiques, serait préférable et la leçon est à retenir pour concevoir de nouveaux modes de paiements.

Bibliographie

Aubert J.M. (2019), « Modes de financement et régulation », rapport final, « *Stratégie de transformation du système de santé* », DREES

Choné P. et C.A. Ma (2011), « Optimal health care contract under physician agency », Annals of economics and statistics, 101-102

Dormont B., Geoffard P-Y. et J.Tirole (2014), « Refonder l'assurance-maladie », note du CAE, n°12, Conseil d'analyse économique

Laffont J-J. et J.Tirole (1993), « A theory of Incentives in Procurement and Regulation », ch. I.1 *Cost-reimbursement rules*, The MIT Press

Ministère des solidarités et de la santé (2019), « Vers un modèle de paiement combiné », rapport « Ma santé 2022, un engagement collectif »

Ministère des solidarités et de la santé (2020), « Ségur de la santé. Les conclusions », dossier de presse, juillet 2020

Milcent C. (2017), « Premier bilan de la T2A sur la variabilité des coûts hospitaliers », Paris School of Economics, wp, 2017-54

Milcent C. (2019), « From downcoding to upcoding : DRG based payments in hospitals », Paris School of Economics, wp, 2019-53

Mougeot M. et F.Naegelen (2011), « Régulation et tarification des hôpitaux ». Economica

Mougeot M. et F.Naegelen (2014), « La tarification à l'activité : une réforme dénaturée du financement des hôpitaux». Revue française d'économie, 3, 111-141

Mougeot M. et F.Naegelen (2018), « Non-reponsiveness, severity auditing and upcoding deterrence », Journal of institutional and theoretical economics

Newhouse J-P. (1996), « Reimbursing Health Plans and Health Providers : Selection versus Efficiency in Production », Journal of economic literature, 34(3), 1236-1263