RAND Journal of Economics
Vol. 34, No. 4, Winter 2003
pp. 647–669

Prior health expenditures and risk sharing with insurers competing on quality

Maurice Marchand*

Motohiro Sato**

and

Erik Schokkaert***

Insurers can exploit the heterogeneity within risk-adjustment classes to select the good risks because they have more information than the regulator on the expected expenditures of individual insurees. To counteract this cream skimming, mixed systems combining capitation and cost-based payments have been adopted that do not, however, generally use the past expenditures of insurees as a risk adjuster. In this article, two symmetric insurers compete for clients by differentiating the quality of service offered to them according to some private information about their risk. In our setting it is always welfare improving to use prior expenditures as a risk adjuster.

1. Introduction

■ To create incentives for cost containment, several countries relying on public funds to cover part or all of their health insurers' expenditures have introduced a prospective financing mechanism into their health insurance system. In this mechanism, competing health insurers do not get their costs reimbursed by the state, but receive capitation payments from the regulator. If insurers raise additional premiums, these have to be community rated. It is well known that the combination of prospective financing and community rating of additional premiums may create incentives for risk selection, i.e., "actions of economic agents on either side of the market to exploit unpriced risk heterogeneity and break pooling arrangements, with the result that some consumers may not obtain the insurance they desire" (Newhouse, 1996, p. 1236). To mitigate this problem, the capitation payments are adjusted on the basis of observable risk factors. However, existing risk-adjustment schemes are based mainly on demographic and socioeconomic variables, and they reflect only a small fraction of the heterogeneity in risks. Insofar as the insurers have better information than

* CORE, Université catholique de Louvain; mmarchand@core.ucl.ac.be.

** Hitotsubashi University; satom@econ.hit-u.ac.jp.

*** CES, Katholieke Universiteit Leuven.

We thank Ake Blomqvist, Nicolas Boccard, Miguel Gouviea, Hugh Gravelle, Jean Hindriks, and Tor Iversen for helpful comments. Maurice Marchand passed away while this article was in press. His death is a great loss for his friends and for the scientific economic community.

what is captured in the risk-adjustment system, there remains room for risk selection (van de Ven and Ellis, 2000).[1]

One way to reduce the incentives for risk selection is to introduce a mixed reimbursement scheme, in which payments from the regulator to the insurers are a combination of a prospective capitation and actual costs (Newhouse, 1986, 1996, 1998). Such a mixed reimbursement (or partial capitation) system can be interpreted as a form of risk sharing between the regulator and the insurers. Although its aim is to reduce the incentives for risk selection, it dilutes the incentives for efficiency. The optimal system will therefore depend on the weights given by the regulator to these two considerations (risk selection and efficiency) and on the behavioral reactions of the insurers. Risk sharing can take different forms, from a simple proportional rule in which the reimbursement is a linear combination of capitation and actual expenditures, to sophisticated systems of outlier or condition-specific risk sharing. The consequences of these different systems in terms of the tradeoff between efficiency and risk selection have been analyzed empirically by Beebe (1992) and van Barneveld et al. (2001). These articles do not contain an explicit description of insurer behavior and/or market equilibrium. In the present article, we want to model explicitly the tradeoff between efficiency and selection in the context of a theoretical model of insurer competition in which social welfare objectives are stated explicitly. To keep the model tractable, we focus on the simple proportional scheme, which has received much attention in the theoretical literature and is applied in countries such as the Netherlands and Belgium.

The tradeoff between selection and efficiency in the context of risk adjustment has been analyzed recently in a set of articles by Encinosa and Sappington (1997), Glazer and McGuire (2000), Frank, Glazer, and McGuire (2000), and Jack (2001). These articles model a situation of adverse selection, in which the patients have superior information about their own health status and insurers or HMOs differentiate benefits and/or quality to let the good risks self-select. We focus rather on the selection activities of the insurers, i.e., on the other side of the market. While quality differences are also central in our model, we will introduce them as an explicit risk-selection device in a model of competition between two profit-seeking insurers who can distinguish between different types of insurees. This is one of the main risk-selection issues in the European compulsory insurance schemes we have in mind, but it is also relevant in other systems when there is concern about the equal treatment of different risk groups.

Two assumptions are crucial to our approach. First, the risk-adjustment system is imperfect and the insurers are better informed than the regulator about the relative risks of different patients. To model this, we concentrate on a group of patients who are identical with respect to the observable variables appearing in the risk-adjustment formula but who differ with respect to other health characteristics about which the insurers have superior information. In this situation it will be profitable for insurers to attract good risks and discourage bad risks—where "good" and "bad" refer to the relative expected health expenditures of individuals within a given risk class as defined by the risk-adjustment scheme. Second, we assume that explicit cream skimming is forbidden and that the insurers use quality differentiation to discriminate between patients having different expected health expenditures.[2] Our concept of quality includes aspects, such as the timeliness of payments or friendliness of staff, that are related to the services provided by insurers and not to health care itself.

Our model has some similarities to the one proposed by Ellis (1998). He analyzes different forms of risk selection (including explicit dumping of patients) within a duopoly model. However, he does not allow explicitly for cost-reduction efforts by insurers and is therefore not explicit about the efficiency-selection tradeoff. We characterize the equilibrium in the model of quality competition and derive comparative statics with respect to the policy instruments. This is the first contribution of this article.

[1] Recently, diagnostic information has been introduced in the risk-adjustment schemes of Medicare (United States) and the Netherlands. Although this is a big improvement, even the use of diagnostic information does not remove all incentives for risk selection.

[2] Selection activities through quality differentiation are also modelled in Ma (1994), but he considers the case of one single firm.

More important, however, we analyze the consequences of introducing insurees' prior expenditures into the risk-sharing system. It is well known that prior expenditures are a good predictor of actual expenditures (Newhouse, 1986; van Vliet, 1992; van de Ven and Ellis, 2000; van Barneveld et al., 2001). Their use in the risk-adjustment scheme has not been very popular, however, because of their diluting effect on the incentives for efficiency.[3] It has been argued that, apart from a discount factor, including past expenditures is similar to a simple cost reimbursement. Although this may be true, the conclusion that prior expenditures should not be used no longer follows once some risk sharing is introduced. In that case we introduce actual expenditures anyway, and this is certainly (although perhaps marginally) worse from the point of view of efficiency. Even then, Newhouse (1986, 1994, 1998) has consistently advocated the use of actual rather than prior expenditures in the partial-capitation system. Apart from practical reasons, he argues that actual expenditures are more efficient to reduce risk-selection incentives because they yield a more sensitive measure of predictable variation in expected cost. The latter argument is not obvious, however. As a matter of fact, it is possible that the use of prior insurees' expenditures is more effective at reducing risk selection than the use of actual expenditures, because prior expenditures might be more strongly correlated with the (imperfect) signal used by insurers to distinguish different risk groups. We model this effect explicitly and derive the (broad set of) conditions under which the introduction of prior expenditures into the risk-sharing mechanism is welfare improving. To do so, we have to distinguish two periods in our model. This two-period specification will force us to solve a subgame-perfect Nash equilibrium by backward induction.

The details of the model are described in Section 2. Section 3 characterizes the symmetric equilibrium in the insurance market. The comparative statics with respect to the policy instruments are analyzed in Section 4. Section 5 formulates the conditions for the optimal government policy. We compare the second-best solution to the hypothetical first-best case in which the regulator has direct control on the qualities offered to the different types of insurees. Section 6 concludes.

2. Description of the model

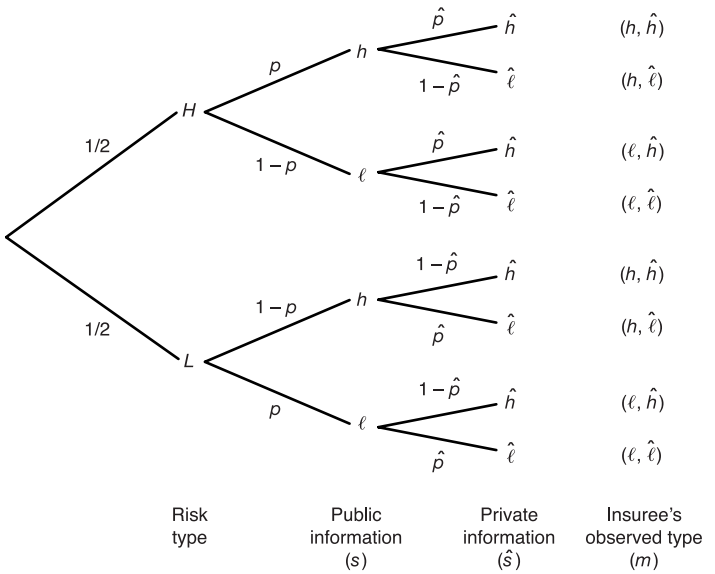
■ We consider two symmetric insurers, denoted $k = A, B$, who compete for clients in a regulated health insurance market. They have to cover the costs of all medical treatments supplied by health care providers to their insurees (or clients). To do so, insurers receive from the state some capitation and cost-based payments financed by tax revenue. There is no private premium or copayment paid by insurees.

The capitation payment is fixed according to a risk-adjustment scheme that classifies the insurers' clients into groups by means of a set of observable demographic and socio-economic risk-adjustment characteristics. In the following we shall concentrate on one risk-adjustment group of clients treated as homogeneous by the risk-adjustment scheme. Therefore, the same capitation payment is received by insurers for any client belonging to that group. But since the demographic and socioeconomic information is coarse, there remains some heterogeneity within each risk-adjustment group that can be exploited by insurers if they have superior information about the expected health care expenses of individual insurees.

To model this in the simplest manner, we assume that there are two unobservable risk types, labelled H (high) and L (low), within the demographic group that we consider. The size of these two risk types are equal and normalized to one. The size of the group is therefore equal to two. High-risk clients are more likely than low-risk ones to incur high health expenses. More precisely, individuals of risk type H have health expenditures $C_h(e)$ and $C_\ell(e)$ with probabilities p and $1-p$ respectively, where $C_h(e) > C_\ell(e)$ and $p > 1/2$, and e denotes some cost-reducing effort (see below). In a symmetric way, individuals of risk type L have expenditures $C_h(e)$ and $C_\ell(e)$ with probabilities now equal to $1-p$ and p respectively. This is depicted in Figure 1 (see the columns labelled "risk type" and "public information"). The conditional expected expenditures per insuree

[3] Glazer and McGuire (2000) note that past use may be used as a signal by the insurers (while also being known by the regulator) but explicitly reject the possibility that it could be used as a risk adjuster.

FIGURE 1
INFORMATION STRUCTURE



are therefore

$$E\big[C(e)/H\big] = pC_h(e) + (1-p)C_\ell(e), \tag{1a}$$
$$E\big[C(e)/L\big] = (1-p)C_h(e) + pC_\ell(e). \tag{1b}$$

For a given client, expenditures in different periods are assumed to be independently and identically distributed according to his risk type. Therefore our setting does not fit the case of chronic illnesses.[4]

Since insurers are financed in part by capitation, they are encouraged to put some effort into reducing treatment costs by providers. This can be done by, e.g., selective contracting with some providers or direct monitoring of the behavior of providers and patients, which requires time to implement and corresponds to long-run decisions. More precisely, denoting insurer k's cost-reducing effort by e_k, we posit that

$$C_h(e_k) = \theta_h - e_k, \tag{2b}$$
$$C_\ell(e_k) = \theta_\ell - e_k, \qquad k = A, B, \tag{2b}$$

where $\theta_h > \theta_\ell$.

Clients are assumed to live for two periods. In the first period there is no information available to both insurers and to the regulator about their risk type. Given our symmetry assumptions, it is natural to posit that in the first period, clients are distributed over the two insurers independently of their risks, i.e., the proportions of high- and low-risk types are identical across insurers. The expected medical expenditures per client to be reimbursed by insurer k in the first period can then be written as

$$E\big(C(e_k)\big) = \frac{1}{2}E\big[C(e_k)/H\big] + \frac{1}{2}E\big[C(e_k)/L\big] = \frac{1}{2}(\theta_h + \theta_\ell) - e_k, \qquad k = A, B. \tag{3}$$

[4] The assumption that expenditures in different periods are i.i.d. was made for technical reasons only. Introducing the possibility of chronic illnesses would increase the predictive power of past expenditures and therefore strengthen our case in favor of integrating them in the risk-adjustment scheme.

In the second period, the insurers and the regulator know which clients incurred either high or low health expenditures in the first period (i.e., either $C_h(e_k)$ or $C_\ell(e_k)$). This public information can be used as a (imperfect) signal of a client's risk type. Let s denote this information, with $s \in (h, \ell)$. In addition, insurers have some private information about the risk type of insurees. This is obtained through direct contact with providers or through detailed analysis of individual administrative files. This information is known to both insurers and to insurees but not to the regulator, which explains why we call it private. It is known to both insurers because they can costlessly collect this information at the moment a potential new member arrives (e.g., in an interview or in a questionnaire with personal information). We model it in a way analogous to the above public information: individuals of risk type H give signals $\hat{h}$ and $\hat{\ell}$ with probabilities $\hat{p}$ and $1 - \hat{p}$ respectively, while those of risk type L give signals $\hat{h}$ and $\hat{\ell}$ with probabilities $1 - \hat{p}$ and $\hat{p}$. Without loss of generality, we assume that $\hat{p} \geq 1/2$. Of course, this signal is informative only if $\hat{p} > 1/2$. Let $\hat{s}$ denote this information, with $\hat{s} \in (\hat{h}, \hat{\ell})$.

Given these signals, the insurers are now able to distinguish different observed types in the second period. Let m be the (privately) observed type (or message) of an insuree, with

$$m = (s, \hat{s}) \in M = \{(h, \hat{h}), (h, \hat{\ell}), (\ell, \hat{h}), (\ell, \hat{\ell})\}.$$

The proportions in the population of each observed client type can be inferred from the information structure given in Figure 1. We have

$$\begin{aligned}
\Pr(h, \hat{h}) &= \frac{1}{2} p\hat{p} + \frac{1}{2}(1 - p)(1 - \hat{p}),\\
\Pr(h, \hat{\ell}) &= \frac{1}{2} p(1 - \hat{p}) + \frac{1}{2}(1 - p)\hat{p},\\
\Pr(\ell, \hat{h}) &= \frac{1}{2}(1 - p)\hat{p} + \frac{1}{2} p(1 - \hat{p}),\\
\Pr(\ell, \hat{\ell}) &= \frac{1}{2}(1 - p)(1 - \hat{p}) + \frac{1}{2} p\hat{p},
\end{aligned} \tag{4}$$

with these proportions adding up to one. From the viewpoint of insurers, a client's expected expenditures in the second period will now depend upon his observed type m. This is because the conditional probabilities $\Pr(H/m)$ and $\Pr(L/m) = 1 - Pr(H/m)$ will vary across observed types according to

$$\begin{aligned}
\Pr(H/h, \hat{h}) &= \frac{p\hat{p}}{p\hat{p} + (1 - p)(1 - \hat{p})},\\
\Pr(H/h, \hat{\ell}) &= \frac{p(1 - \hat{p})}{p(1 - \hat{p}) + (1 - p)\hat{p}},\\
\Pr(H/\ell, \hat{h}) &= \frac{(1 - p)\hat{p}}{(1 - p)\hat{p} + p(1 - \hat{p})},\\
\Pr(H/\ell, \hat{\ell}) &= \frac{(1 - p)(1 - \hat{p})}{(1 - p)(1 - \hat{p}) + p\hat{p}}.
\end{aligned} \tag{5}$$

The expected expenditures of an insuree with observed type m can be written as

$$E\big[C(e_k)/m\big] = E[C(e_k)/H] \Pr(H/m) + E[C(e_k)/L] \Pr(L/m), \qquad m \in M, \tag{6}$$

which clearly depends upon m. Insurees with a bad signal (h and/or $\hat{h}$) are more likely to be of risk type H.

Since individuals differ by their observed type, in the second period insurers have an incentive to attract clients of the observed types that they expect to be the most profitable. They can do

so by differentiating the quality of the services they offer to different observed types, which amounts to resorting to risk selection. This can be implemented in different ways: as indicated in the Introduction, the possibilities range from differences in the speed with which clients' claims are processed to differences in the quality of personal contacts with the clients. These quality differences cannot be regulated. Let q_k^m denote the quality of services offered to clients of observed type m ($m \in M$) by insurer k ($k = A, B$). The benefit that clients derive from this quality of service is $B(q_k^m)$, where $B(q)$ is increasing and strictly concave in q. For each observed type the relative qualities offered by the two insurers will influence their market shares. These qualities are costly to insurers (see below).

Let us thus consider the individual's choice of insurer. Individuals are assumed to have idiosyncratic preferences for insurers. Let x denote the additional utility that in period 1 an individual gets from being a client of insurer B rather than of A, in short his preference for insurer B. By assumption, x is distributed across individuals according to a uniform distribution over interval $[-1/2, +1/2]$. In period 1 there cannot be any differentiation of quality across insurers. Therefore, in this period insurer A is chosen by all individuals with $x \leq 0$ and insurer B by all those with $x > 0$, which means that clients are equally shared between the two insurers.

The idiosyncratic preferences of individuals for insurers are, however, assumed to be subject to some random shocks between period 1 and 2. Although this feature is not unrealistic, the main reason for introducing it is to avoid some nondifferentiability at the symmetric equilibrium later on. Let y denote an individual's preference for insurer B in the second period. It is related to his preference in the first period by $y = x + \varepsilon$, where ε is a white noise that is i.i.d across individuals. Its distribution is uniform over interval $[-\alpha, +\alpha]$, with $0 < \alpha < 1$.

We are now ready to determine how the market shares of insurers in the second period are influenced by the quality of service they offer their clients (q_k^m, $m \in M$, $k = A, B$). For reasons that will become clear below, we also need to know whether the clients of an insurer in the second period have switched insurers between the two periods. Thus let $N_{i,j}^m$ be the number of individuals of observed type m who were clients of insurer i in period 1 and are clients of insurer j in period 2 ($i, j = A, B$).

In period 2, an individual with observed type m and idiosyncratic preference y for insurer B chooses insurer B if $y \geq \hat{y}^m \equiv \delta[B(q_A^m) - B(q_B^m)]$, and insurer A otherwise, where parameter $\delta(> 0)$ can be seen as a measure of competitiveness in the insurance market. Focusing on the clients of insurer A in period 2, we show in Appendix A that

$$N_{A,A}^m = \begin{cases} N^m \left[\dfrac{1}{2} - \alpha + \dfrac{1}{4\alpha}(\alpha + \hat{y}^m)(3\alpha - \hat{y}^m)\right] & \text{if } \delta \Delta B^m \leq \alpha \\ \dfrac{N^m}{2} & \text{if } \delta \Delta B^m > \alpha \end{cases} \tag{7a}$$

and

$$N_{B,A}^m = N^m \left(\frac{1}{2} + \delta \Delta B^m\right) - N_{A,A}^m, \tag{7b}$$

where N^m is the number of insurees with observed type m in the population[5] and $\Delta B^m \equiv B(q_A^m) - B(q_B^m)$.

Let us finally turn to the government instruments. The regulator first has to fix the risk-adjusted capitation payments for the different risk-adjustment groups. Given that we are focusing on one particular group, in our model this prospective risk-adjusted payment boils down to a per-capita lump-sum one. For reasons that will become clear later on, we will differentiate this lump-sum payment for the two periods: the capitation payments in the first and the second period will be denoted by R_1 and R_2 respectively. The regulator is aware of the problem of quality differentiation (or risk selection) in the second period, but as noted before, she is unable to observe and regulate

[5] Since the size of the group is 2, we have $N^m = 2\,\Pr(m)$, $m \in M$, where $\Pr(m)$ is given in (4).

quality directly. To mitigate risk selection, she will therefore resort to second-best instruments in addition to the prospective payments. One such instrument is a cost-based payment: a proportion of second-period expenditures, denoted by r, is reimbursed to insurers. If $r \neq 0$, we get the mixed reimbursement scheme, which has been analyzed quite extensively in the literature. In our setting, however, the regulator can use an additional instrument. Since she also has information on individuals' past expenditures, she can use that public information to further reduce the incentives for quality differentiation by reimbursing in the second period a fraction t of prior (or first-period) expenditures of an insurer's second-period clients. This basically boils down to the use of past expenditures as a risk adjuster. Our purpose in the following is to analyze the conditions under which the use of this instrument leads to a welfare improvement and to what extent it should be used.

In the next section, we describe the symmetric equilibria of this regulated insurance market for given values of R_1, R_2, r, and t. In Section 4 we will draw some comparative statics results related to these instruments. These will then be used in Section 5 to derive the optimal regulatory policy.

3. The symmetric equilibrium in the insurance market

■ Insurers are interested in maximizing their (discounted) profits net of the disutility of the cost-reducing effort. Concentrating on insurer A, his net profits are given by

$$\pi_A^1 + \beta \sum_m N_A^m \pi_A^m - (1+\beta) D(e_A), \tag{8}$$

where β is the discount factor and $D(e)$ is the disutility of cost-reducing effort expressed in monetary units. In this expression, π_A^1 refers to the expected profit per client in the first period:

$$\pi_A^1 = R_1 - E[C(e_A)] = R_1 - \frac{\theta_h + \theta_\ell}{2} + e_A \tag{9}$$

(recall that the population size is 2 and is equally shared between the two insurers in the first period). π_A^m in (8) stands for the expected profit per client of observed type $m = (s, \hat{s})$ in the second period:

$$N_A^m \pi_A^m = N_A^m \left\{R_2 - (1-r) E\left[C(e_A)/m\right]\right\} + t(N_{AA}^m C_s(e_A) + N_{BA}^m C_s(e_B)) - \phi N_A^m q_A^m, \quad m = (s, \hat{s}) \in M, \tag{10}$$

where $N_A^m = N_{AA}^m + N_{BA}^m$ and where ϕ is the cost per client and per unit of the quality offered. The second term on the right-hand side of the above expression accounts for the fact that the insurer is reimbursed, for each second-period client, a fraction t of the client's past (or first-period) expenditures.

As explained earlier, insurers have at their disposal two decision variables for maximizing their net profits. First, they must, at the start of the first period, choose their cost-reducing effort, e_k, which cannot be differentiated across periods. As indicated earlier, it is indeed a long-run decision. Second, in the second period they must decide upon the quality levels q_k^m to offer to the different observed types ($m \in M$). It is through these quality choices that the two insurers compete with each other in attracting clients. Since insurers expect the choice of effort in the first period to affect the Nash equilibrium in quality levels in the second period, a subgame-perfect equilibrium emerges that must be solved by backward induction. We will thus first investigate the choice of qualities in the second period, and then turn to the effort decision in the first period. In the following, we mainly concentrate on the symmetric equilibrium.

Let us then in a first stage characterize the Nash equilibrium in second-period qualities. This can be done separately for each of the four observed types. For a given m, insurer A takes q_B^m

as a parameter and maximizes with respect to q_A^m his objective (8) in which π_A^1 and $N_A^m \pi_A^m$ are substituted from (9) and (10). Using (7a) and (7b) yields the following first-order condition:

$$\delta B'(q_A^m)\left\{R_2 - (1-r)E[C(e_A)/m] + \frac{t}{2}\left(\left(1 - \frac{\hat{y}^m}{\alpha}\right)C_s(e_A) + \left(1 + \frac{\hat{y}^m}{\alpha}\right)C_s(e_B)\right) - \phi q_A^m\right\}$$
$$= \phi\left[\frac{1}{2} + \delta(B(q_A^m) - B(q_B^m))\right], \qquad m = (s, \hat{s}) \in M, \tag{11}$$

which is the standard rule for the optimal choice of quality by a firm.[6] From this condition we derive the reaction function $q_A^m(e_A, e_B, q_B^m)$, $m \in M$. Note that the expression in curly brackets is just π_A^m. Proceeding in the same way for the other insurer, we obtain a first-order condition equivalent to (11), from which we infer insurer B's reaction function $q_B^m(e_A, e_B, q_A^m)$, $m \in M$. Differentiating (11) with respect to q_A^m and q_B^m enables us to obtain the slope of insurer A's reaction curve at the symmetric Nash equilibrium where $q_A^m = q_B^m = q^m$ and $e_A = e_B = e$:

$$\frac{dq_A^m}{dq_B^m} = \left[-\frac{B''(q^m)}{(B'(q^m))^2}\frac{1}{2\delta} + 2\right]^{-1},$$

which is positive and smaller than one. A similar expression can be obtained for the slope of insurer B's reaction curve. Therefore, the quality choices of the two insurers are strategic complements.

From the insurers' two reaction functions, the Nash equilibrium in qualities can be obtained as functions of the cost-reducing efforts chosen in the first period:

$$Q_A^m(e_A, e_B) \quad \text{and} \quad Q_B^m(e_A, e_B), \qquad m \in M.$$

When setting their cost-reducing effort in the first period, the insurers will anticipate how these effort levels will shift the second-period Nash equilibrium. In Appendix B we show that

$$\frac{dq_A^m}{de_A} = (K(q^m)+1)^{-1}(K(q^m)+3)^{-1}\left[(1-r)(K(q^m)+2) - \frac{t}{2}(K(q^m)+3)\right]\frac{1}{\phi}, \tag{12}$$
$$\frac{dq_B^m}{de_A} = (K(q^m)+1)^{-1}(K(q^m)+3)^{-1}\left[(1-r) - \frac{t}{2}(K(q^m)+3)\right]\frac{1}{\phi} \tag{13}$$

at the symmetric Nash equilibrium, where we have $e_A = e_B = e$ and $q_A^m = q_B^m = q^m$. In these expressions, $K(q^m)$ is defined by

$$K(q^m) \equiv -\frac{B''(q^m)}{B'(q^m)}\frac{1}{2\delta B'(q^m)} > 0. \tag{14}$$

At first sight we would expect that an increase in an insurer's cost-reducing effort will have a positive effect on his own quality choice in the second period, since the reduction in his per-client expenditures makes it more profitable to attract all types of clients. Moreover, we would expect that the quality of service provided by the other insurer in the Nash equilibrium will also rise because the two insurers' qualities are strategic complements. However, as equations (12) and (13) show, the opposite is true if t is very large.[7] This is easily understood: an increase in e_A lowers both $C_h(e_A)$ and $C_\ell(e_A)$ in the first period. At the symmetric Nash equilibrium, half of any clients that insurer B could attract in the second period by enhancing its quality of service were originally clients of insurer A in the first period. This, plus the fact that the payment related to

[6] Condition (11) is similar to: $(P - C)\partial N/\partial q = N\partial C/\partial q$, where P, C, and N denote the price, the average cost, and the number of clients respectively.

[7] This interpretation might change if $r > 1$. However, we will show later on that this can never be the case in the regulator's optimum.

the past expenditures of these clients, $(1/2)tC_s(e_A)$, is lower, means that insurer B's profits from attracting new clients decrease if e_A increases. If t is large enough, this discourages the provision of quality by insurer B. The same is true for insurer A.

Let us now turn to the decisions on cost-reducing efforts in the first period and concentrate on insurer A. As mentioned above, each insurer chooses his effort level anticipating its effect on the Nash equilibrium in qualities that will emerge in the second period. Therefore, insurer A maximizes objective (8) after having substituted N_A^m, π_A^1, and π_A^m from (7a), (9), and (10) as well as q_A^m and q_B^m by their Nash equilibrium values. Using the envelope theorem, this yields the following first-order condition:

$$(1+\beta)D'(e_A) = 1 + \beta \sum_m \left[N_A^m(1-r) - tN_{AA}^m\right] + \beta \sum_m \pi_A^m \frac{\partial N_A^m}{\partial q_B^m}\frac{dq_B^m}{de_A}, \tag{15}$$

where

$$\frac{\partial N_A^m}{\partial q_B^m} = -N^m \delta B'(q_B^m). \tag{16}$$

A condition similar to (15) holds with respect to e_B. At the symmetric subgame-perfect equilibrium, equalities $q_A^m = q_B^m = q^m$ and $e_A = e_B = e$ are satisfied, which we use to infer from (10), (16), and (11) that

$$\pi_A^m \frac{\partial N_A^m}{\partial q_B^m} = -\frac{1}{2}\phi N^m. \tag{17}$$

Then, at this symmetric equilibrium, condition (15) can be rewritten as

$$\begin{aligned}(1+\beta)D'(e) = 1 &+ \beta\left(1 - r - t\left(1 - \frac{\alpha}{2}\right)\right)\\ &- \frac{\beta}{2}\sum_m N^m(K(q^m)+1)^{-1}(K(q^m)+3)^{-1}\left[(1-r) - \frac{t}{2}\big(K(q^m)+3\big)\right],\end{aligned} \tag{18}$$

where we have used (7a), (13), $\sum_m N^m = 2$, and $\sum_m N_A^m = 1$. From (11) we can also infer the condition that q^m satisfies at the symmetric equilibrium:

$$\delta B'(q^m)[R_2 - (1-r)E[C(e)/m] + tC_s(e) - \phi q^m] = \frac{\phi}{2}, \quad m = (s, \hat{s}) \in M. \tag{19}$$

Expressions (18) and (19) together characterize the values of e and $q^m, m \in M$, at the symmetric subgame-perfect equilibrium. These are the key expressions that we will use in the next section.

4. Comparative statics of the symmetric equilibrium

■ The comparative statics of e and $q^m s$ with respect to the regulator's policy instruments are complex because as shown by (18) and (19), the choices of e and $q^m s$ generally interact with each other. We move this comparative static analysis to the Appendix for the general case, and we concentrate here on the case where $B(q) = \log(q)$. With this specification, we have $K(q^m) = 1/2\delta$, which does not depend any more upon q^m. Accordingly, (18) and (19) can be solved in a recursive manner: the equilibrum value of e is first obtained from (18) and then plugged into (19) to determine $q^m, m \in M$ at equilibrium. This simplifies considerably the comparative static analysis of e and q^m. We adopt the logarithmic specification of $B(q)$ in the rest of this article.

With this specification we first infer from (18) the comparative statics of e at the symmetric equilibrium:

$$\frac{de}{dr} = \frac{1}{(1+\beta)D''(e)}\beta\left[\left(\frac{1}{2\delta}+1\right)^{-1}\left(\frac{1}{2\delta}+3\right)^{-1} - 1\right] < 0, \tag{20}$$

$$\frac{de}{dt} = \frac{1}{(1+\beta)D''(e)}\beta\left[\frac{1}{2}\left(\frac{1}{2\delta}+1\right)^{-1} - \left(1-\frac{\alpha}{2}\right)\right] < 0, \tag{21}$$

$$\frac{de}{dR_1} = \frac{de}{dR_2} = 0. \tag{22}$$

The inequalities in (20) and (21) agree with intuition: increases in the cost-sharing parameters, r and t, lower the effort level at the symmetric equilibrium. More important, we have $-de/dr > -de/dt$.

Lemma 1. If $B(q) = \log q$, increases in r and t cause the cost-reducing effort at the symmetric equilibrium to decrease. Furthermore, we have

$$-\frac{de}{dr} > -\frac{de}{dt} > 0.$$

The disincentive effect on e of an increase in t is therefore lower than that of an increase in r. This result can be explained as follows. The first-period expenditures of an insurer's second-period clients will be reimbursed at rate t, while their second-period expenditures will be at rate r. However, some of those second-period clients will not have been with this insurer in the first period, since some insurees will have changed insurer in the meantime. So the first-period expenditures of these clients will not have been affected by the cost-reducing effort of the insurer. Accordingly, the disincentive effect of t on e is mitigated relative to that of r.

Turning now to the comparative statics of equilibrium qualities, we have

$$\frac{dq^m}{dx} = \frac{\partial q^m}{\partial x} + \frac{\partial q^m}{\partial e}\frac{de}{dx}, \qquad x = R_2, r, t; \;\; m \in M,$$

where

$$\frac{\partial q^m}{\partial R_2} = \frac{1}{\phi}\frac{1}{(1/2\delta)+1} > 0, \qquad m \in M \tag{23}$$

$$\frac{\partial q^m}{\partial r} = E\big[C(e)/m\big]\frac{\partial q^m}{\partial R_2} > 0, \qquad m \in M \tag{24}$$

$$\frac{\partial q^m}{\partial t} = C_s(e)\frac{\partial q^m}{\partial R_2} > 0, \qquad m \in M \tag{25}$$

and

$$\frac{\partial q^m}{\partial e} = (1-r-t)\frac{\partial q^m}{\partial R_2}, \qquad m \in M. \tag{26}$$

Therefore, concentrating on the partial derivatives of q^m i.e., keeping e constant, an increase of the capitation payment in the second period leads to a rise in the quality offered to clients of all observed types. The same is true for increases in r and t. These results are not very surprising because all the policy changes make it more profitable to attract further clients. More interestingly, since for any $\hat{s}$ we have

$$E\big[C(e)/(h,\hat{s})\big] < C_h(e), \qquad \forall\, \hat{s} \in (\hat{h}, \hat{\ell})$$
$$C_\ell(e) < E\big[C(e)/(\ell,\hat{s})\big], \qquad \forall\, \hat{s} \in (\hat{h}, \hat{\ell}),$$

we can also state the following lemma.

Lemma 2. For any given effort level, we have $0 < \partial q^m/\partial r < \partial q^m/\partial t$ for $m = (h, \hat{s})$ and any $\hat{s} \in (\hat{h}, \hat{\ell})$. Symmetrically, we also have $\partial q^m/\partial r > \partial q^m/\partial t > 0$ for $m = (\ell, \hat{s})$ and any $\hat{s} \in (\hat{h}, \hat{\ell})$.

This lemma has an obvious interpretation. The proportional reimbursement parameter t refers to past expenditures; these are known with certainty at the beginning of the quality competition process in the second period. Actual expenditures in the second period reimbursed through r, however, are uncertain. This explains why an increase in t has a stronger positive effect on quality for the observed types with high expenditures in the first period and a weaker effect on quality for the observed types with low expenditures in the first period. Therefore, we conclude from Lemma 2 that *ceteris paribus*, t performs better than r at reducing quality discrimination.

Lemmas 1 and 2 together give the basic reason for the introduction of past expenditures in the risk-adjustment scheme. Increasing t has a less negative effect on the incentives for efficiency (Lemma 1) and is more effective at reducing quality discrimination (Lemma 2) than increasing r. We will now analyze more formally the optimal government policy.

5. Optimal government policy

■ With the logarithmic specification of the benefit function $B(q)$, the expressions characterizing the cost-reducing effort and qualities at the symmetric subgame-perfect equilibrium simplify to

$$D'(e) = 1 - \frac{\beta}{1+\beta}\left[r + \left(1 - \frac{\alpha}{2}\right)t + \frac{2\delta}{1+2\delta}\left(\frac{2\delta}{1+6\delta}(1-r) - \frac{t}{2}\right)\right], \tag{27}$$

$$q^m = \frac{2\delta}{1+2\delta}\frac{1}{\phi}\big(R_2 + tC_s(e) - (1-r)E[C(e)/m]\big), \qquad m = (s, \hat{s}) \in M. \tag{28}$$

The regulator must fix the four policy instruments (R_1, R_2, r, and t) so as to maximize her objective function. This is assumed to be quadratic in the qualities offered at equilibrium:

$$W = E\left[q^m - \frac{\rho}{2}(q^m)^2\right] - \lambda G, \tag{29}$$

where E is the expectation operator taken with respect to the observed types, $m \in M$, and G stands for the regulator's overall payment to insurers (or public spending):

$$G = 2\left[R_1 + \beta R_2 + \beta(r+t)\frac{1}{2}(C_h(e) + C_\ell(e))\right]. \tag{30}$$

With the quadratic part of W being expressed in money equivalent, parameter λ can be interpreted as the marginal cost of public funds. As usual, we posit $\lambda > 1$, because public funds are raised by means of distortive taxes. The above objective function can also be written as

$$W = \overline{q} - \frac{\rho}{2}\overline{q}^2 - \frac{\rho}{2}\mathrm{var}\,(q^m) - \lambda G, \tag{31}$$

where $\overline{q} \equiv E[q^m]$ is the average quality offered to insurees. Through the presence of $\mathrm{var}(q^m)$, this expression makes it explicit that, as stated in the Introduction, the regulator wants to keep quality discrimination across observed types as low as possible. In other words, the regulator wants to minimize risk selection by insurers.

When optimizing W, the regulator faces the requirement that the insurers' discounted expected profits be nonnegative. Using (3) and (8)–(10), this participation constraint can be written as

$$R_1 + \beta R_2 - \frac{1}{2}\left[1 + \beta(1-r-t)\right](C_h(e) + C_\ell(e)) - \beta\phi\overline{q} - (1+\beta)D(e) \geq 0. \tag{32}$$

Since R_1 does not affect any of the insurers' decisions, it is immediate that we can eliminate it from further consideration by choosing its value so as to make the participation constraint

binding.[8] Therefore, substituting R_1 from (32) with equality into (31), the regulator's optimization can then be written as

$$\max_{R_2,r,t} W = (\overline{q} - \frac{\rho}{2}\overline{q}^2) - \frac{\rho}{2}\text{var}(q^m) - 2\lambda\left[\frac{1+\beta}{2}(\theta_h + \theta_\ell - 2e) + \beta\phi\overline{q} + (1+\beta)D(e)\right], \quad (33)$$

where $e = e(r, t)$ and $q^m = q^m(R_2, r, t, e)$, $m \in M$, are obtained from (27) and (28).

□ **The first-best solution as a benchmark.** Before solving the above second-best problem, we investigate as a benchmark the solution to the first-best problem, in which the regulator is assumed to have direct control over the qualities offered to the observed types of insurees. Maximizing W in (33) with respect to e and q^m yields the following first-order conditions:

$$q^m = \overline{q} = \frac{1 - 2\lambda\beta\phi}{\rho}, \qquad m \in M, \quad (34)$$

$$D'(e) = 1. \quad (35)$$

These results agree with intuition. In the first best there is no quality discrimination, and the uniform level of quality is set so as to equate at the margin its social benefit to its social cost. As for e, it is pushed up to the level where at the margin the reduction in health expenditures is equated to the additional disutility of effort to insurers.

□ **The second-best solution.** Let us now return to the second-best problem in which the regulator can monitor effort and qualities only indirectly by means of the policy instruments at her disposal.

The first-order condition with respect to R_2 for a maximum of (33) is given by

$$\frac{dW}{dR_2} = [(1 - \rho\overline{q}) - 2\lambda\beta\phi]\frac{d\overline{q}}{dR_2} = 0, \quad (36)$$

which in our setting yields for $\overline{q}$ the same level as in the first-best solution (see (34)). For given values of r and t, R_2 must be adjusted so that the average quality reaches its first-best level.[9] Using (36), the first-order conditions with respect to r and t simplify to

$$\frac{dW}{dr} = -\frac{\rho}{2}\frac{d}{dr}\text{var}(q^m) + 2(1+\beta)\lambda(1 - D'(e))\frac{de}{dr} = 0, \quad (37)$$

$$\frac{dW}{dt} = -\frac{\rho}{2}\frac{d}{dt}\text{var}(q^m) + 2(1+\beta)\lambda(1 - D'(e))\frac{de}{dt} = 0. \quad (38)$$

Let us observe that the gain from raising r and t is to reduce $\text{var}(q^m)$ and thus risk selection. Together with Lemma 1, this enables us to state the following lemma.

Lemma 3. In the second-best solution, $D'(e) < 1$, i.e., the cost-reducing effort is lower than in the first best. Furthermore, the average quality is identical to its first-best level.

In the Appendix we show by substituting q^m from (28) into (37) and (38) that these first-order conditions yield a system of two linear equations in $1 - r$ and t:

$$t\,\text{cov}(C_s(e), E[C(e)/m]) - (1 - r)\,\text{var}(E[C(e)/m]) = \frac{2(1+\beta)}{\rho}\left(\frac{dq}{dR_2}\right)^{-2}\lambda(1 - D'(e))\frac{de}{dr} \quad (39)$$

[8] Remember that we focus here on one risk-adjustment group. Imposing a binding participation constraint separately for each such group implies that R_1 acts like a conventional risk adjuster in the sense that it removes all incentives for risk selection between the risk-adjustment groups.

[9] This finding is similar to the one in Frank, Glazer, and McGuire (2000) and Glazer and McGuire (2000). In their models as well, the capitation payment plays a crucial role in the determination of the optimal quality levels.

$$t\ \mathrm{var}(C_s(e)) - (1-r)\,\mathrm{cov}(C_s(e), E[C(e)/m]) = \frac{2(1+\beta)}{\rho}\left(\frac{dq}{dR_2}\right)^{-2}\lambda(1-D'(e))\frac{de}{dt}, \tag{40}$$

where $dq/dR_2 \equiv dq^m/dR_2 = 2\delta(1+2\delta)^{-1}\phi^{-1}$, $m \in M$, is independent of m as a consequence of the logarithmic specification of $B(q)$.[10]

Starting from the above system of equations, we prove the following proposition in the Appendix.

Proposition 1. If $B(q) = \log q$ and the private signal, $\hat{s}$, is informative ($\hat{p} > 1/2$), we have

(i) $t > 0$,

(ii) $r < 1$

at the second-best solution.

This proposition enables us to answer positively the question raised in the Introduction: Is it optimal for the regulator to use past expenditures as a risk adjuster? The reason for this result is that an increase in t reduces the incentive for risk selection (or quality differentiation). Indeed, this increase makes the payment to insurers rise more for those clients who have experienced the highest cost in the first period and are therefore the bad risks prior to the second period. One could correctly argue that an increase in r also discourages risk selection, since it reduces the gap between the expected profits on good and bad risks. However, increasing t turns out to be more effective at giving the right incentives than does increasing r. Equally important, in our setting a rise in t has a lower negative impact on the cost-reduction effort than does a rise in r (see Lemma 1).

Our purpose in the next subsection is to analyze how the second-best values of t and r are influenced by changes in some parameters such as the accuracy of the private signal $\hat{s}$. However, it will be worthwhile to first investigate the case where the private signal is not informative ($\hat{p} = 1/2$), which is equivalent to having no private signal at all.

Without private signal, the key consideration is that the number of observed types reduces from four to two: $s = h, \ell$. The qualities and cost-reducing effort at the subgame-perfect equilibrium remain characterized by expressions (27) and (28), except that the second one simplifies to

$$q^s = \frac{2\delta}{1+2\delta}\frac{1}{\phi}\big(R_2 + tC_s(e) - (1-r)E[C(e)/s]\big), \qquad s = h, \ell. \tag{41}$$

The above consideration means that the regulator has at her disposal three instruments (R_2, r, and t) to control the values of three variables (q^h, q^ℓ, and e), which implies that the first best can be achieved by an appropriate choice of these instruments.

In the Appendix, we show that without private information the optimal values of r and t are given by

$$r = -\frac{A}{1-A} < 0, \tag{42}$$

$$t = (2p-1)^2(1-r) > 0, \tag{43}$$

with

$$0 < A = (2p-1)^2\left(\frac{\delta+1}{2\delta+1} - \frac{\alpha}{2}\right) + \frac{2\delta}{2\delta+1}\,\frac{2\delta}{6\delta+1} < 1.$$

Those values of r and t allow the regulator to eliminate any discrimination in the qualities offered to the clients of the various observed types and to achieve the first-best level of cost-reducing

[10] See the comparative statics of q^m in Section 4. If $B(q) = \log(q)$, $K(q^m) = 1/2\delta$.

effort as given by (35). The prospective payment R_2 must be set so that the uniform level of qualities satisfies the first-best level as given by (34). The result that $r < 0$ may be surprising. The justification is that in the case without private information, t is an ideal instrument to counteract quality discrimination. Increasing t will have efficiency effects, however. These are compensated for by decreasing r sufficiently.

□ **Comparative static analysis of t and r with respect to $\hat{p}$ and var($C_s(e)$).** In the Appendix, we prove the comparative static results stated in the following proposition.

Proposition 2. Assuming $B(q) = \log q$ and $D(e) = (1/2)e^2$,

(i) $\dfrac{d}{d\hat{p}}r > 0$ and $\dfrac{d}{d\hat{p}}t < 0$,

(ii) $\dfrac{d}{d\sigma^2}r > 0$ and $\dfrac{d}{d\sigma^2}t < 0$,

with $\sigma^2 \equiv \text{var}(C_s(e)) = (1/4)(\theta_h - \theta_\ell)^2$.

An increase in $\hat{p}$ means that insurers get better informed about the risks of their clients, and this induces them to enhance risk selection through larger quality differences. The only way to mitigate this negative effect is to increase the cost-sharing rate r. However, this has the negative consequence of diluting the cost-reducing effort. To counteract this behavioral response, t is decreased at the same time that r is increased. The choice of t and r reflects some optimal tradeoff between encouraging cost-reducing effort and avoiding quality differentiation. Rises in either t or r cause the cost-reducing effort to diminish (negative effect) and quality differences to fade away (positive effects). When private information improves, r becomes more efficient than t at mitigating quality differentiation.

When $\sigma^2 = (1/4)(\theta_h - \theta_\ell)^2$ rises, differences in expected profits across observed types get larger. This induces insurers to increase risk selection through quality differentiation. Again, r turns out to be more effective than t at counteracting this effect.[11]

As stated in Proposition 2, a rise in the accuracy of the private information, $\hat{p}$, causes the optimal value of r to increase and that of t to decrease, with opposite effects on the cost-reducing effort. We show in the Appendix that the overall effect of these changes is to decrease the cost-reducing effort: $de/d\hat{p} < 0$. The regulator reacts to an increase in $\hat{p}$ pushing up risk selection by giving less weight to the objective of cost reduction.

Proposition 1 provides no information about the sign of r. However, in the case where the private signal, $\hat{s}$, is not informative, we have shown r to be negative. This calls for the following question: Under which circumstances is r positive? To answer this, let us rewrite the system of equations (39) and (40) assuming that $D(e) = e^2/2$:

$$t - f(\hat{p})(1 - r) - F\frac{1 - e}{(2p - 1)^2\sigma^2}\frac{de}{dr} = 0, \tag{44}$$

$$t - (2p - 1)^2(1 - r) - F\frac{1 - e}{\sigma^2}\frac{de}{dt} = 0, \tag{45}$$

where $f(\hat{p})$ is defined by (A18) in the Appendix, $F = 2(1 + \beta)\lambda\rho^{-1}(dq/dR_2)^{-2}$, and (to recall) $\sigma^2 = (1/4)(\theta_h - \theta_\ell)^2$. From these expressions we infer that as $\sigma^2 \to \infty$, $r \to 1$ and $t \to 0$ at the regulator's optimum. Together with $dr/d\sigma^2 > 0$ in Proposition 2, this enables us to state the following proposition.

Proposition 3. Assuming $B(q) = \log q$ and $D(e) = (1/2)e^2$, given $\hat{p} > 1/2$, there exists some

[11] Proposition 2 is in line with the claim of Newhouse (1986, 1994, 1998) that actual expenditures yield a more sensitive measure of predictable variation in expected cost. As shown in Proposition 1, this does not imply that $t = 0$ in the regulator's optimum.

critical value $\hat{\sigma}^2$ such that at the regulator's optimum,

$$r > 0 \qquad \text{for } \sigma^2 > \hat{\sigma}^2.$$

Therefore, when $\theta_h - \theta_\ell$ is large enough, the cost-sharing rate r at the regulator's optimum is positive. The value of t was shown in Proposition 1 to be always positive.

6. Concluding remarks

■ We have argued that the dominant position not to include past expenditures as a risk adjuster in a risk-adjustment scheme should be reconsidered once we accept the desirability of partial capitation (or mixed reimbursement) to curb the incentives for risk selection. As we have shown, the use of prior expenditures may be more efficient than the use of actual expenditures to curb risk selection under the highly realistic assumption that information about these past expenditures is used by insurers as a signal to distinguish between different risk groups. If past expenditures are the only signal available to insurers, the regulator can even reach the first-best solution by making use of that same information. In general, if insurers have additional private information about the insurees, the regulator should implement a reimbursement scheme in which the capitation is mixed with both actual and past expenditures. Their weights will depend on the specific informational features of the problem and the value judgments concerning quality discrimination.

Some assumptions have been made to obtain meaningful results, while others have been for the sake of simplicity. It would be worth investigating how our results are modified when the probabilities that individuals of risk-type H have health expenditures $C_h(e)$ and $C_\ell(e)$ are p_H and $1 - p_H$ respectively while those probabilities for risk-type L are $1 - p_L$ and p_L. We have considered here the symmetric case where $p_H = p_L \equiv p$ (see the third paragraph of Section 2).

We have derived these results in a two-period model of quality competition between two insurers. In our model the first period only sets the stage for the real competition, which takes place in the second period. An obvious extension would be to formulate an overlapping-generations model in which "first period" and "second period" are overlapping for different cohorts of insurees. This extension would require us to think carefully about the exact timing of the cost-reducing effort, which here as been held constant for the two periods. It would also allow a richer dynamic description of the use of information by both the insurers and the regulator.

Another line of research would recognize that insurees live for more than two periods. Public and private information about the risk type of insurees would then accumulate all along their life cycle, implying that instruments r and t should possibly be made contingent on the age of insurees. A multiperiod setting would also offer the possibility of introducing chronic illnesses in a natural way.

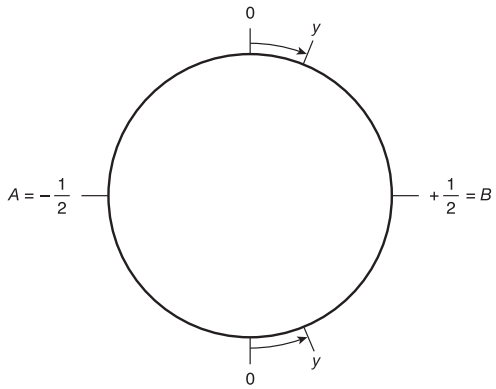
It would also be worth investigating how our setting is modified if different insurers do not share the same private information about insurees. Insurers may have more information on their past and present clients than on other individuals. Moreover, contrary to most of the existing literature, we have concentrated on the selection decisions made on the supply side of the market. In a more general model it would be necessary to integrate the phenomenon of adverse selection if insurees have superior information about their own type and insurers can offer different benefit packages. Even in countries with a broad compulsory insurance package, the latter possibility can arise if insurers get the opportunity to make selective contracts with providers or can offer differentiated supplementary insurance.

Appendix

■ In this first subsection we derive the formulas expressing $N^m_{A,A}$ and $N^m_{B,A}$ in terms of q^m_A and q^m_B. To do so we use Hotelling's standard model, identifying an individual's preference for insurer B with his location on a circle.[12] This is

[12] If individuals were distributed along a line segment in the first period, some intricacies would result when deriving the distribution of individuals at the lower and upper ends of this segment in the second period. Whether a linear or circular model is used does not at all affect the properties of the symmetric subgame-perfect equilibrium.

FIGURE A1



depicted in Figure A1, with insurers A and B being located at points $-1/2$ and $+1/2$ respectively and an individual with idiosyncratic preference y for insurer B in period 2 being located at point y. There is a cutoff value of y,

$$\hat{y}^m = \delta[B(q_A^m) - B(q_B^m)],$$

such that all individuals of observed type m with $y \leq \hat{y}^m$ choose insurer A in period 2, and all those with $y > \hat{y}^m$ choose insurer B. In the location analogy, the per-unit-of-distance travel cost is the inverse of δ.

As explained in the main text, each individual is subjected to some random shock between the two periods. If he is located at point y in period 2, he was located at point $x = y - \varepsilon$ in period 1, where ε is a random variable uniformly distributed on interval $[-\alpha, +\alpha]$. Therefore, the probability that an individual located at y in period 2 had been a client of insurer A in period 1 is given by

$$\begin{aligned}
\Pr(x = y - \varepsilon \leq 0) &= \Pr(y \leq \varepsilon), \\
&= 1 && \text{if } y \leq -\alpha, \\
&= \alpha - y/2a && \text{if } -\alpha \leq y \leq +\alpha \\
&= 0 && \text{if } +\alpha \leq y.
\end{aligned}$$

We then have

$$N_{AA}^m = \begin{cases} \left[\displaystyle\int_{-1/2}^{-\alpha} dy + \int_{-\alpha}^{\hat{y}^m} \frac{\alpha - y}{2\alpha} dy \right] N^m & \text{if } \hat{y}^m \leq \alpha, \\ \left[\displaystyle\int_{-1/2}^{-\alpha} dy + \int_{-\alpha}^{+\alpha} \frac{\alpha - y}{2\alpha} dy \right] N^m & \text{if } \hat{y}^m > \alpha, \end{cases}$$

which yields

$$N_{AA}^m = \begin{cases} \left[\dfrac{1}{2} - \alpha + \dfrac{1}{4\alpha}(\hat{y}^m + \alpha)(3\alpha - \hat{y}^m) \right] N^m & \text{if } \hat{y}^m \leq \alpha, \\ \dfrac{1}{2} N^m & \text{if } \hat{y}^m > \alpha. \end{cases}$$

We also have

$$N_A^m \equiv N_{AA}^m + N_{BA}^m = \left(\frac{1}{2} + \hat{y}^m \right) N^m,$$

from which we can infer the value of N_{BA}^m.

□ **Proof of (12) and (13).** Starting from (11), let us define

$$\begin{aligned}
\Delta_A^m \equiv \delta B'(q_A^m) &\left\{ R_2 - (1 - r)E[C(e_A)/m] + \frac{t}{2}\left((1 - \frac{\hat{y}_m}{\alpha}) C_s(e_A) + (1 + \frac{\hat{y}_m}{\alpha}) C_s(e_B) \right) - \phi q_A^m \right\} \\
&- \phi \left[\frac{1}{2} + \delta(B(q_A^m) - B(q_B^m)) \right],
\end{aligned}$$

and likewise for Δ_B^m. The qualities q_A^m and q_B^m at the Nash equilibrium satisfy $\Delta_A^m = 0$ and $\Delta_B^m = 0$. Evaluating the derivatives of Δ_A^m at the symmetric equilibrium ($q_A^m = q_B^m = q^m$) gives

$$\begin{aligned}
\frac{\partial \Delta_A^m}{\partial q_A^m} &= \delta B''(q^m)\left\{R_2 - (1-r)E[C(e_A)/m] + \frac{t}{2}(C_s(e_A) + C_s(e_B)) - \phi q^m\right\} - 2\phi\delta B'(q^m) \\
&= \frac{\phi}{2}\frac{B''(q^m)}{B'(q^m)} - 2\phi\delta B'(q^m), \\
\frac{\partial \Delta_A^m}{\partial q_B^m} &= \phi\delta B'(q^m), \\
\frac{\partial \Delta_A^m}{\partial e_A} &= \delta B'(q^m)(1 - r - \frac{t}{2}), \\
\frac{\partial \Delta_A^m}{\partial e_B} &= -\delta B'(q^m)\frac{t}{2},
\end{aligned}$$

and likewise for the derivatives of Δ_B^m. Differentiating equations $\Delta_A^m = 0$ and $\Delta_B^m = 0$ and using the above derivatives provides us with the following system of equations in dq_A^m/de_A and dq_B^m/de_A:

$$\begin{vmatrix} -(K(q^m)+2) & 1 \\ 1 & -(K(q^m)+2) \end{vmatrix} \begin{vmatrix} \dfrac{dq_A^m}{de_A} \\ \dfrac{dq_B^m}{de_A} \end{vmatrix} = \frac{1}{\phi} \begin{vmatrix} -\left(1 - r - \dfrac{t}{2}\right) \\ \dfrac{t}{2} \end{vmatrix},$$

where first we have everywhere divided by $\delta\phi B'(q^m)$ the results of the differentiation of the two equations and then used $K(q^m)$ as defined in (14). The determinant of the matrix on the left-hand side is

$$J = (K(q^m)+2)^2 - 1 = (K(q^m)+3)(K(q^m)+1) > 0.$$

Solving the system of linear equations yields

$$\begin{vmatrix} \dfrac{dq_A^m}{de_A} \\ \dfrac{dq_B^m}{de_A} \end{vmatrix} = \frac{1}{\phi J} \begin{vmatrix} -(K(q^m)+2) & -1 \\ -1 & -(K(q^m)+2) \end{vmatrix} \begin{vmatrix} -(1-r) + \dfrac{t}{2} \\ \dfrac{t}{2} \end{vmatrix},$$

from which we obtain (12) and (13).

□ **The general case where $B(q)$ can be any increasing and strictly concave function.** In this case, we have, from (18),

$$\frac{de}{dx} = \frac{\partial e}{\partial x} + \sum_m \frac{\partial e}{\partial q^m}\frac{dq^m}{dx}, \qquad x = R_2, r, t.$$

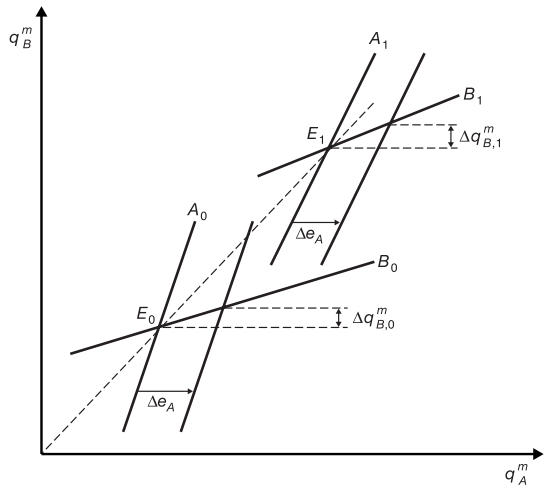
According to this, the total effects (or total derivatives) can be decomposed into direct effects (the partial derivatives on the right-hand side) and indirect effects operating through changes in q^m (the second terms on the right-hand side). These indirect effects were nil with the logarithmic specification of $B(q)$, and therefore the partial derivatives were also the total derivatives. From (18) we infer the partial derivatives

$$\begin{aligned}
\frac{\partial e}{\partial r} &= \frac{1}{(1+\beta)D''(e)}\beta\left\{\frac{1}{2}\sum_m \frac{N^m}{1+K(q^m)}\frac{1}{3+K(q^m)} - 1\right\} < 0, \\
\frac{\partial e}{\partial t} &= \frac{1}{(1+\beta)D''(e)}\beta\left\{\frac{1}{4}\sum_m \frac{N^m}{1+K(q^m)} - (1 - \frac{\alpha}{2})\right\} < 0, \quad \text{and} \\
\frac{\partial e}{\partial R_1} &= \frac{\partial e}{\partial R_2} = 0
\end{aligned}$$

that generalize (20), (21), and (22). To understand the meaning of the indirect effects, let us observe that it is through the presence of $K(q^m)$ in the expression multiplying $(\beta/2)N^m$ on the right-hand side of (18) that these indirect effects work. We can also observe from (13) that this expression is equal to dq_B^m/de_A up to the multiplicative factor $-1/\phi$. Suppose that t is not too large, so that $dq_B^m/de_A > 0$, and consider an increase in r that causes q^m to rise. As can be inferred from (18), whether the cost-reducing effort at the symmetric equilibrium is affected positively or negatively by this rise of q^m depends

FIGURE A2

INDIRECT EFFECT



on whether $K'(q^m)$ is positive or negative. Since the slope of the reaction curves is given by $dq_A^m/dq_B^m = (K(q^m)+2)^{-1}$, a negative (positive) $K'(q^m)$ means that as q^m rises, insurers become more (less) agressive.[13] The effect of the rise of r on dq_B^m/de_A is depicted in Figure A2, where the equilibria before and after the increase in r are represented by E_0 and E_1 respectively. A rise in e_A will induce in equilibria E_0 and E_1 an increase in the quality provided by insurer B, $\Delta q_{B,0}^m$ and $\Delta q_{B,1}^m$ respectively. Whether $\Delta q_{B,1}^m$ is larger or smaller than $\Delta q_{B,0}^m$ depends on whether the slope of insurer B's reaction curve increases or decreases respectively in q^m, that is, whether insurers become more or less agressive.

We therefore conclude that when t is not too large and a rise in q^m makes insurers more agressive (i.e., their reaction curves become steeper), the direct and indirect effects on e of an increase in r go in the same direction: the effort level at the symmetric equilibrium diminishes. Indeed, if insurer B is more agressive at E_1 than at E_0, the cost to insurer A of increasing e_A marginally rises. In the opposite case where the insurers become less agressive when q^m rises, we cannot rule out *a priori* that the positive indirect effect of an increase in r through the lower intensity of quality competition offsets the direct negative effect.

□ **How to obtain (39) and (40) from (37) and (38).** We have, from (28),

$$q^m - \overline{q} = \frac{dq}{dR_2}\Big\{tC_s(e) - (1-r)E[C(e)/m] + (1-r-t)E[C(e)]\Big\}, \qquad m \in M, \tag{A1}$$

in which $dq/dR_2 \equiv dq^m/dR_2 = 2\delta(1+2\delta)^{-1}\phi^{-1}$. Using this relation allows us to write

$$\begin{aligned}\frac{d}{dr}\mathrm{var}(q^m) &= \frac{d}{dr}E_m[(q^m-\overline{q})^2]\\ &= 2\frac{dq}{dR_2}E_m\Big[(q^m-\overline{q})(E[C(e)/m]-E[C(e)]\Big]\\ &= 2\frac{dq}{dR_2}\mathrm{cov}(q^m, E[C(e)/m])\end{aligned} \tag{A2}$$

and

$$\begin{aligned}\frac{d}{dt}\mathrm{var}(q^m) &= \frac{d}{dt}E_m[(q^m-\overline{q})^2]\\ &= 2\frac{dq}{dR_2}E_m\Big[(q^m-\overline{q})(C_s(e)-E[C(e)])\Big]\\ &= 2\frac{dq}{dR_2}\mathrm{cov}(q^m, C_s(e)).\end{aligned} \tag{A3}$$

[13] If $B(q^m)$ is iso-elastic, i.e., if $B(q^m)$ is equal to $(1-\theta)^{-1}(q^m)^{1-\theta}$ for $\theta \neq 1$ and $\log(q^m)$ for $\theta = 1$, we have $K(q^m) = \theta(q^m)^{\theta-1}/2\delta$. Therefore, as q^m rises, the slope of reaction curves rises or falls according to whether θ is lower or higher than one.

Substituting q^m from (28) into the covariance terms appearing in (A2) and (A3) yields

$$\mathrm{cov}(q^m, E[C(e)/m]) = \frac{dq}{dR_2}\left\{t\,\mathrm{cov}(C_s(e), E[C(e)/m]) - (1-r)\mathrm{var}(E[C(e)/m])\right\} \tag{A4}$$

and

$$\mathrm{cov}(q^m, C_s(e)) = \frac{dq}{dR_2}\left\{t\,\mathrm{var}(C_s(e)) - (1-r)\mathrm{cov}(C_s(e), E[C(e)/m])\right\}. \tag{A5}$$

Using (A2)–(A5) enables us to infer (39) and (40) from (37) and (38).

□ **Proof of Proposition 1.** We first solve the system of linear equations (39) and (40).

$$t = \frac{1}{S}\frac{2(1+\beta)\lambda}{\rho}\left(\frac{dq}{dR_2}\right)^{-2}(1-D'(e))\left[\mathrm{cov}(C_s(e), E[C(e)/m])\left(-\frac{de}{dr}\right) - \mathrm{var}(E[C(e)/m])\left(-\frac{de}{dt}\right)\right], \tag{A6}$$

$$1-r = \frac{1}{S}\frac{2(1+\beta)\lambda}{\rho}\left(\frac{dq}{dR_2}\right)^{-2}(1-D'(e))\left[\mathrm{var}(C_s(e))\left(-\frac{de}{dr}\right) - \mathrm{cov}(C_s(e), E[C(e)/m])\left(-\frac{de}{dt}\right)\right], \tag{A7}$$

in which S denotes the determinant of the matrix of the coefficients of t and $(1-r)$ on the left-hand side of (39) and (40):

$$S = \mathrm{var}(C_s(e))\mathrm{var}(E[(C(e)/m]) - \mathrm{cov}^2(C_s(e), E[C(e)/m]).$$

To sign t and $1-r$ at the second-best solution, we will later prove the following relationships:

$$\mathrm{cov}(C_s(e), E[C(e)/m]) = (2p-1)^2\mathrm{var}(C_s(e)) \tag{A8}$$

$$\mathrm{var}(E[C(e)/m]) = f(\hat{p})\mathrm{cov}(C_s(e), E[C(e)/m]), \tag{A9}$$

where $f(\hat{p})$ is positive and increasing in $\hat{p}$ (except at $\hat{p} = 1/2$, where $f'(1/2) = 0$), $f(1/2) = (2p-1)^2$, and $f(1) = 1$. These results imply that if $\hat{p} > 1/2$, $S > 0$. Together with Lemma 1 they also imply that the expressions in brackets in (A6) and (A7) are both positive. Since $D'(e) < 1$ according to Lemma 3, we obtain Proposition 1.

We need, however, to prove (A8) and (A9) as well as the properties of $f(\hat{p})$. We proceed in several steps.

(i) We first obtain

$$\mathrm{var}(C_s) = \frac{1}{2}\sum_{s=\ell,h}(C_s - E[C])^2 = \frac{1}{4}(C_h - C_\ell)^2 = \frac{1}{4}(\theta_h - \theta_\ell)^2, \tag{A10}$$

where we have used $E[C] = \frac{1}{2}\sum_{s=h,\ell} C_s$.

(ii) To prove equality (A8), we start from

$$\begin{aligned}\mathrm{cov}(C_s, E[C/m]) &= \sum_{m\in M}\Pr(m)(C_s - E[C])(E[C/m] - E[C])\\ &= (C_h - E[C])\left\{\Pr(h,\hat{h})(E[C/h,\hat{h}] - E[C]) + \Pr(h,\hat{\ell})(E[C/h,\hat{\ell}] - E[C])\right\}\\ &\quad + (C_\ell - E[C])\left\{\Pr(\ell,\hat{h})(E[C/\ell,\hat{h}] - E[C]) + \Pr(\ell,\hat{\ell})(E[C/\ell,\hat{\ell}] - E[C])\right\}.\end{aligned} \tag{A11}$$

Let us first look at the expression in the first curly brackets in (A11). Using relations (4) and (5) in the main text, $E[C/h,\hat{h}]$ can be written as

$$E[C/h,\hat{h}] = \frac{1}{\Pr(h,\hat{h})}\left\{\frac{1}{2}p\hat{p}E[C/H] + \frac{1}{2}(1-p)(1-\hat{p})E[C/L]\right\}. \tag{A12}$$

Likewise, we have

$$E[C/h,\hat{\ell}] = \frac{1}{\Pr(h,\hat{\ell})}\left\{\frac{1}{2}p(1-\hat{p})E[C/H] + \frac{1}{2}(1-p)\hat{p}E[C/L]\right\}. \tag{A13}$$

Substituting from (A12) and (A13) and recognizing from (4) that $\Pr(h,\hat{h}) + \Pr(h,\hat{\ell}) = 1/2$, the expression in the first curly brackets of (A11) simplifies to

$$\frac{1}{2}pE[C/H] + \frac{1}{2}(1-p)E[C/L] - \frac{1}{2}E[C].$$

Using $E[C] = (1/2)(E[C/H] + E[C/L])$ and relations (1) in the text, it can be further simplified to

$$\left(p - \frac{1}{2}\right)^2 (C_h - C_\ell).$$

Since $C_h - E[C] = (1/2)(C_h - C_\ell)$, the first of the two terms on the right-hand side of (A11) is therefore equal to

$$\frac{1}{2}\left(p - \frac{1}{2}\right)^2 (C_h - C_\ell)^2 = 2\left(p - \frac{1}{2}\right)^2 \mathrm{var}(C_s),$$

where the equality is obtained using (A10). Proceeding in the same way, the second term on the right-hand side of (A11) can be shown to be equal to the same value as the first term. Therefore we conclude that

$$\mathrm{cov}(C_s, E[C/m]) = (2p - 1)^2 \mathrm{var}(C_s),$$

i.e., equality (A8).

(iii) To prove equality (A9), we begin with

$$\mathrm{var}(E[C/m]) = \sum_{m \in M} \Pr(m)(E[C/m] - E[C])^2. \tag{A14}$$

Since $E[C/m] = \Pr(H/m)E[C/H] + (1 - \Pr(H/m)E[C/L]$ and $E[C] = (1/2)(E[C/H] + E[C/L])$, we have

$$E[C/m] - E[C] = \left(\Pr(H/m) - \frac{1}{2}\right)(E[C/H] - E[C/L]),$$

which we substitute into (A14) to obtain

$$\mathrm{var}(E[C/m]) = (E[C/H] - E[C/L])^2 \sum_{m \in M} \Pr(m)\left(\Pr(H/m) - \frac{1}{2}\right)^2. \tag{A15}$$

From relations (5) we infer

$$\Pr(H/h, \hat{h}) = 1 - \Pr(H/\ell, \hat{\ell}) \quad \text{and} \quad \Pr(H/h, \hat{\ell}) = 1 - \Pr(H/\ell, \hat{h}).$$

Therefore, we have

$$\Pr(H/\ell, \hat{\ell}) - \frac{1}{2} = \frac{1}{2} - \Pr(H/h, \hat{h}) \quad \text{and} \quad \Pr(H/\ell, \hat{h}) - \frac{1}{2} = \frac{1}{2} - \Pr(H/h, \hat{\ell}).$$

On the other hand, from relations (4) we infer

$$\Pr(h, \hat{h}) = \Pr(\ell, \hat{\ell}) \quad \text{and} \quad \Pr(h, \hat{\ell}) = \Pr(\ell, \hat{h}).$$

Using these equalities in (A15), we obtain

$$\mathrm{var}(E[C/m]) = 2(E[C/H] - E[C/L])^2 \left\{\Pr(h, \hat{h})\left(\Pr(H/h, \hat{h}) - \frac{1}{2}\right)^2 + \Pr(h, \hat{\ell})\left(Pr(H/h, \hat{\ell}) - \frac{1}{2}\right)^2\right\}. \tag{A16}$$

From relations (5) we also obtain

$$\Pr(H/h, \hat{h}) - \frac{1}{2} = \frac{1}{2}\frac{p + \hat{p} - 1}{p\hat{p} + (1 - p)(1 - \hat{p})}$$

and

$$\Pr(H/h, \hat{\ell}) - \frac{1}{2} = \frac{1}{2}\frac{p - \hat{p}}{p(1 - \hat{p}) + (1 - p)\hat{p}}.$$

Using these relations, relations (4), and equality $(E[C/H] - E[C/L])^2 = (2p-1)^2(C_h - C_\ell)^2 = 4(2p-1)^2\text{var}(C_s)$ enables us to rewrite (A16) as follows:

$$\text{var}(E[C/m]) = (2p-1)^2\text{var}(C_s)f(\hat{p}), \tag{A17}$$

with

$$f(\hat{p}) = \frac{(p+\hat{p}-1)^2}{p\hat{p}+(1-p)(1-\hat{p})} + \frac{(p-\hat{p})^2}{p(1-\hat{p})+(1-p)\hat{p}}. \tag{A18}$$

Using (A8), (A17) is seen to be equivalent to (A9).

(iv) It remains to investigate the properties of $f(\hat{p})$ given in (A18). It is straightforward to show that $f(1/2) = (2p-1)^2$ and $f(1) = 1$. After some manipulations, the derivative of $f(\hat{p})$ can be written as

$$\begin{aligned} f'(\hat{p}) &= 2\left[\frac{p+\hat{p}-1}{p\hat{p}+(1-p)(1-\hat{p})} - \frac{p-\hat{p}}{p(1-\hat{p})+(1-p)\hat{p}}\right] \\ &\quad - (2p-1)\left[\left(\frac{p+\hat{p}-1}{p\hat{p}+(1-p)(1-\hat{p})}\right)^2 - \left(\frac{p-\hat{p}}{p(1-\hat{p})+(1-p)\hat{p}}\right)^2\right] \\ &= \left\{\frac{p+\hat{p}-1}{p\hat{p}+(1-p)(1-\hat{p})} - \frac{p-\hat{p}}{p(1-\hat{p})+(1-p)\hat{p}}\right\} \\ &\quad \cdot \left\{2-(2p-1)\left(\frac{p+\hat{p}-1}{p\hat{p}+(1-p)(1-\hat{p})} + \frac{p-\hat{p}}{p(1-\hat{p})+(1-p)\hat{p}}\right)\right\}. \end{aligned}$$

In the last expression, the first and second factors can be shown to be equal to $2p(2\hat{p}-1)(1-p)$ and $4p(1-p)$ respectively, both divided by $(p\hat{p}+(1-p)(1-\hat{p}))(p(1-\hat{p})+(1-p)\hat{p})$. Therefore, $f'(\hat{p}) > 0$ for $\hat{p} > 1/2$ and $f'(1/2) = 0$.

□ **Proof of relations (42) and (43).** Both (42) and (43) hold in the absence of private information (i.e., $\hat{p} = 1/2$). In this case the regulator has enough instruments to achieve the first best. In particular, these instruments are chosen so that the qualities offered to insurees are not differentiated: $q^h = q^\ell$. Using (28), the equality of qualities implies

$$t(C_h - C_\ell) = (1-r)(E[C/h] - E[C/\ell]), \tag{A19}$$

where

$$\begin{aligned} E[C/h] - E[C/\ell] &= (pE[C/H] + (1-p)E[C/L]) - ((1-p)E[C/H] + pE[C/L]) \\ &= (2p-1)(E[C/H] - E[C/L]) = (2p-1)^2(C_h - C_\ell). \end{aligned}$$

Therefore, (A19) is equivalent to $t = (2p-1)^2(1-r)$, i.e., relation (43).

To obtain (42) we start from (27), taking first into account that at the first best $D'(e) = 1$. Then, substituting t from (43) into (27) yields

$$r + \left\{(2p-1)^2\left(\frac{1+\delta}{1+2\delta} - \frac{\alpha}{2}\right) + \frac{2\delta}{1+2\delta}\frac{2\delta}{1+6\delta}\right\}(1-r) = 0, \tag{A20}$$

where the expression in curly brackets is equal to A, as defined in the text, which can be shown to be less than one. The value of r given in (42) follows from (A20).

□ **Proof of Proposition 2.** Making use of (A8) and (A9), (39) and (40) can be expressed as follows:

$$\Delta \equiv t - f(\hat{p})(1-r) - F\frac{1-e}{(2p-1)^2\sigma^2}\frac{de}{dr} = 0, \tag{A21}$$

$$\Omega \equiv t - (2p-1)^2(1-r) - F\frac{1-e}{\sigma^2}\frac{de}{dt} = 0, \tag{A22}$$

in which (to recall) $\sigma^2 \equiv \text{var}(C_s)$ and

$$F \equiv \frac{2(1+\beta)}{\rho}\left(\frac{dq}{dR_2}\right)^{-2}\lambda.$$

Assuming $D(e) = e^2/2$ and using relations (A8) and (A9), first-order conditions (39) and (40) correspond to $\Delta = 0$ and $\Omega = 0$. It is worth noting that under assumption $D(e) = (1/2)e^2$, according to (27), e is linear in r and t. Therefore, the second derivatives of e with respect to r and t are nil.

To analyze the comparative static properties of the optimal values of r and t with respect to $\hat{p}$ and σ^2, we shall need to use the following derivatives:

$$\begin{aligned}
\Delta_t &= 1 + \frac{F}{(2p-1)^2\sigma^2}\frac{de}{dt}\frac{de}{dr} > 0,\\
\Delta_r &= f(\hat{p}) + \frac{F}{(2p-1)^2\sigma^2}\left(\frac{de}{dr}\right)^2 > 0,\\
\Delta_{\hat{p}} &= -f'(\hat{p})(1-r) < 0,\\
\Delta_{\sigma^2} &= \frac{F}{(2p-1)^2\sigma^4}(1-e)\frac{de}{dr} < 0,\\
\Omega_t &= 1 + \frac{F}{\sigma^2}\left(\frac{de}{dt}\right)^2 > 0,\\
\Omega_r &= (2p-1)^2 + \frac{F}{\sigma^2}\frac{de}{dt}\frac{de}{dr} > 0,\\
\Omega_{\hat{p}} &= 0,\\
\Omega_{\sigma^2} &= \frac{F}{\sigma^4}(1-e)\frac{de}{dt}.
\end{aligned}$$

Differentiating Δ and Ω gives

$$\begin{vmatrix} \Delta_r & \Delta_t \\ \Omega_r & \Omega_t \end{vmatrix}\begin{vmatrix} dr \\ dt \end{vmatrix} = -\begin{vmatrix} \Delta_{\hat{p}} \\ 0 \end{vmatrix} d\hat{p} - \begin{vmatrix} \Delta_{\sigma^2} \\ \Omega_{\sigma^2} \end{vmatrix} d\sigma^2. \tag{A23}$$

The determinant of the matrix on the left-hand side is

$$H = \Delta_r\Omega_t - \Delta_t\Omega_r,$$

the sign of which we have to determine. Substituting the above derivatives of Δ and Ω and rearranging yields

$$\begin{aligned}
H &= \left[f(\hat{p}) - (2p-1)^2\right] + \frac{F}{\sigma^2}\left[\frac{1}{(2p-1)^2}\left(\frac{de}{dr}\right)^2 - 2\frac{de}{dt}\frac{de}{dr} + f(\hat{p})\left(\frac{de}{dt}\right)^2\right]\\
&\geq \left[f(\hat{p}) - (2p-1)^2\right] + \frac{F}{\sigma^2}\left[\frac{1}{(2p-1)^2}\left(\frac{de}{dr}\right)^2 - 2\frac{de}{dt}\frac{de}{dr} + (2p-1)^2\left(\frac{de}{dt}\right)^2\right]\\
&= \left[f(\hat{p}) - (2p-1)^2\right] + \frac{F}{\sigma^2}\left[\frac{1}{2p-1}\frac{de}{dr} - (2p-1)\frac{de}{dt}\right]^2 > 0,
\end{aligned}$$

where we have used $f(\hat{p}) \geq (2p-1)^2$.

From (A23) we obtain

$$\begin{vmatrix} \dfrac{dr}{d\hat{p}} \\ \dfrac{dt}{d\hat{p}} \end{vmatrix} = -H^{-1}\begin{vmatrix} \Omega_t & -\Delta_t \\ -\Omega_r & \Delta_r \end{vmatrix}\begin{vmatrix} \Delta_{\hat{p}} \\ 0 \end{vmatrix}$$

and so

$$\frac{dr}{d\hat{p}} = -H^{-1}\Omega_t\Delta_{\hat{p}} > 0, \quad \frac{dt}{d\hat{p}} = H^{-1}\Omega_r\Delta_{\hat{p}} < 0, \tag{A24}$$

as stated in Proposition 2. Also we have

$$\begin{vmatrix} \dfrac{dr}{d\sigma^2} \\ \dfrac{dt}{d\sigma^2} \end{vmatrix} = -H^{-1}\begin{vmatrix} \Omega_t & -\Delta_t \\ -\Omega_r & \Delta_r \end{vmatrix}\begin{vmatrix} \Delta_{\sigma^2} \\ \Omega_{\sigma^2} \end{vmatrix} = H^{-1}\frac{F(1-e)}{\sigma^4}\begin{vmatrix} \Delta_t\dfrac{de}{dt} - \Omega_t\dfrac{1}{(2p-1)^2}\dfrac{de}{dr} \\ \Omega_r\dfrac{1}{(2p-1)^2}\dfrac{de}{dr} - \Delta_r\dfrac{de}{dt} \end{vmatrix}.$$

Since

$$\Delta_t\frac{de}{dt} - \Omega_t\frac{1}{(2p-1)^2}\frac{de}{dr} = \frac{de}{dt} - \frac{1}{(2p-1)^2}\frac{de}{dr} > \frac{de}{dt} - \frac{de}{dr} > 0,$$

we conclude that $dr/d\sigma^2 > 0$. Likewise, since

$$\Omega_r \frac{1}{(2p-1)^2}\frac{de}{dr} - \Delta_r \frac{de}{dt} = \frac{de}{dr} - f(\hat{p})\frac{de}{dt} \leq \frac{de}{dr} - \frac{de}{dt} < 0,$$

we have $dt/d\sigma^2 < 0$. This ends the proof of Proposition 2. *Q.E.D.*

□ **Proof that $de/d\hat{p} < 0$.** First we differentiate $\Omega = 0$, where Ω is given in (A22):

$$\frac{dt}{d\hat{p}} + (2p-1)^2 \frac{dr}{d\hat{p}} = -\frac{F}{\sigma^2}\frac{de}{dt}\frac{de}{d\hat{p}}$$

or, using (A24),

$$\frac{\Delta_{\hat{p}}}{H}(\Omega_r - (2p-1)^2\Omega_t) = -\frac{F}{\sigma^2}\frac{de}{dt}\frac{de}{d\hat{p}},$$

which yields, after having substituted for Ω_r and Ω_t,

$$\frac{\Delta_{\hat{p}}}{H}\left[\left(-\frac{de}{dr}\right) - (2p-1)^2\left(-\frac{de}{dt}\right)\right] = \frac{de}{d\hat{p}}.$$

Since $\Delta_{\hat{p}} < 0$, $H > 0$, and the expression in brackets can be shown to be positive (using Lemma 1 and $(2p-1)^2 < 1$), we conclude that $de/d\hat{p} < 0$.

References

Beebe, J. "An Outlier Pool for Medicare HMO Payments." *Health Care Financing Review*, Vol. 14 (1992), pp. 17–29.

Ellis, R.P. "Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins." *Journal of Health Economics*, Vol. 17 (1998), pp. 537–555.

Encinosa, W.E. and Sappington, D.E.M. "Competition Among Health Maintenance Organizations." *Journal of Economics and Management Strategy*, Vol. 6 (1997), pp. 129–150.

Frank, R., Glazer, J., and McGuire, T. "Measuring Adverse Selection in Managed Health Care." *Journal of Health Economics*, Vol. 19 (2000), pp. 829–854.

Glazer, J. and McGuire, T.G. "Optimal Risk Adjustment in Markets with Adverse Selection: An Application to Managed Care." *American Economic Review*, Vol. 90 (2000), pp. 1055–1071.

Jack, W. "Controlling Selection Incentives When Health Insurance Contracts Are Endogenous." *Journal of Public Economics*, Vol. 80 (2001), pp. 25–48.

Ma, C.T.-A. "Health Care Payment Systems: Cost and Quality Incentives." *Journal of Economics and Management Strategy*, Vol. 3 (1994), pp. 93–112.

Newhouse, J.P. "Rate Adjusters for Medicare Under Capitation." *Health Care Financing Review*, Vol. 7 (1986), pp. 45–55.

———. "Patients at Risk: Health Reform and Risk Adjustment." *Health Affairs*, Vol. 13 (1994), pp. 132–146.

———. "Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection." *Journal of Economic Literature*, Vol. 34 (1996), pp. 1236–1263.

———. "Risk Adjustment: Where Are We Now?" *Inquiry*, Vol. 35 (1998), pp. 122–131.

van Barneveld, E., Lamers, L., van Vliet, R.C.J.A., and van de Ven, W. "Risk Sharing as a Supplement to Imperfect Capitation: A Tradeoff Between Selection and Efficiency." *Journal of Health Economics*, Vol. 20 (2001), pp. 147–168.

van de Ven, W. and Ellis, R. "Risk Adjustment in Competitive Health Plan Markets." In A. Culyer et al., eds., *Handbook of Health Economics, Vol. 1*. Amsterdam: North-Holland, 2000.

van Vliet, R.C.J.A. "Predictability of Individual Health Care Expenditures." *Journal of Risk and Insurance*, Vol. 59 (1992), pp. 443–460.