Regulating health service standards to reduce variations: an economic analysis

Diane Dawson, Michael Kuhn, Peter C. Smith, Andrew Street
Centre for Health Economics, University of York

Work in Progress; please do not quote without authors' permission

Regulators advocate imposing uniform standards in order to reduce service variations across public health systems. Such policies may be counterproductive in the presence of factor heterogeneity. We extend the model developed by Chalkley and Malcomson (1998) in which hospital output comprises patients treated and the quality of care, the latter being non-contractible. The CM model implies a tailor-made optimal reward schedule for each provider. In practice, health care providers appear to vary substantially in both their productive capacity and preferences. We introduce heterogeneity in factor productivity and assume this to be specific to each output. The purchaser is concerned with improving efficiency and reducing variations. Reflecting commonly observed practice we restrict the purchaser to two instruments, setting a uniform case payment and imposing a uniform standard on activity. We derive the conditions under which, in the presence of factor heterogeneity, the same (second-best) allocation can be obtained as in the CM model. This solution depends upon the complementarity between activity and quality and upon the strength of the purchaser's preference for reducing variations. For those cases where the CM allocation is not achieved, either only a case payment or a standard is appropriate. The conditions for exclusive use of one or other of the instruments are derived and the policy implications are detailed.

Regulating health service standards to reduce variations: an economic analysis

1 Introduction

Policy makers in most developed countries are interested in increasing activity rates and improving the quality of health care provision. Many policy makers appear to believe it is possible to meet these objectives without making additional resources available. The belief is based on evidence suggesting that there is widespread variation in performance. For example, in the UK, the government highlighted 'unacceptable variations' in survival rates, rates of treatment and costs as indicative of inefficiency among hospitals in the National Health Service (NHS Executive, 1997). More generally, the growth of benchmarking activities and the use of techniques such as data envelopment analysis or stochastic frontier analysis often are predicated on such a belief (Dopuch & Gupta, 1997, Greene, 1993, Hollingsworth, Dawson, & Maniadakis, 1999).

An implicit assumption of policies to reduce variations in performance is that relatively poor performers can, with the right incentives, produce the levels of activity and quality achieved by the best. This implies that providers have identical production functions. Yet, if there are inherent differences in the capabilities of the workforce, production functions will differ. In the health sector, some doctors will always provide higher levels of activity or superior health outcomes than others. Surgeons labelled 'fast cutters' by their peers are able to treat more patients than others without compromising quality. Some clinicians are better at reaching correct diagnoses, resulting in higher quality outcomes. Factor heterogeneity leads to a complex trade-off between quality and activity as purchasers try to reduce variations in performance.

Variation in activity levels is readily observable. Measurement of variations in health care quality is more problematic. Numerous experiments with quality measurement are underway, but thus far few of these have been used in earnest to affect health care purchasing. An exception is the public reporting of outcomes from cardiac surgery for individual physicians implemented in the states of New York and Pennsylvania (Chassin, 2002). Evidence from such initiatives suggests that quality information can influence provider behaviour, although

there is a continuing debate as to whether the benefits (improved outcomes of care) outweigh unsought for behavioural responses (eg high risk patients failing to secure treatment) (Dranove, Kessler, McClellan, & Satterthwaite, 2002).

Public reporting initiatives such as the New York and Pennsylvania schemes rely on an implicit market-based incentive for providers to improve quality. Reviews of the US experience suggest that patients tend to take little notice of quality reports (Reilly & Meyer, 2002). Chalkley and Malcomson note that there are strong grounds to believe that patient demand will not be effective by itself in maintaining health care quality (Chalkley & Malcomson, 1998). They develop a model in which an institutional purchaser, concerned with costs, volume, and quality of care, contracts with a provider who is partially motivated by a concern for quality. Quality cannot be monitored effectively, and so the contract must be based on volume and costs only. The purchaser then selects an optimal contract, the nature of which is determined in part by the degree of 'benevolence' of the provider with respect to quality.

The Chalkley and Malcomson model implies a tailor-made optimal reward schedule for each provider. In practice, health care providers are likely to vary in their productivity. This can be due to inherent difference in the productivity of labour or to differences in preferences with respect to income, effort and altruism. Under these circumstances, a health care system is likely to require a great variety of Chalkley and Malcomson contracts.

In practice purchasers rarely offer more than two or three types of contract. This is likely to be for three main reasons. First, the transaction costs of developing tailor-made contracts for each provider are likely to be prohibitive. Second, variations in the terms of contracts between providers may be seen as creating a perception of unfairness within the provider market, leading to the potential for adverse provider responses. In short, a purchaser may secure considerable benefits from appearing to treat providers even-handedly (Milgrom & Roberts, 1990). And third, purchasers may care about equity among patients as well as efficiency, suggesting that some convergence of the quality of outcomes may be valued even at the expense of volume.

The purpose of this paper is to develop a model of optimal contracting with heterogeneous providers when the purchaser has available only two contract instruments, and may care about the equity of outcomes. In particular, we are interested in systems where the heterogeneity of labour inputs results in a distribution of production possibilities over the set of providers and in the regulatory implications that arise from this distribution. The paper is arranged as follows. Section 2 sketches the small economic literature that has examined the implications of heterogeneity in labour inputs to production. In section 3 the analytical framework is established, with factor heterogeneity allowed to effect volume and quality independently and with hospitals differing in their levels of productivity. A benchmark case is described, in which the purchaser is able to make differentiated case-payments. The effect of a concern for variation on the case-payment is considered in section 4, under scenarios based on the form of factor heterogeneity and the complementarity between volume and quality. In section 5, the role of uniform case-payment and a uniform standard on activity is considered under each scenario. Concluding comments are offered in section 6.

2 Heterogeneous Inputs

It is a convention of neoclassical models to treat all units of an input as homogeneous. If there are differences in the quality or characteristics of inputs, these are treated as different inputs in order to preserve the useful assumption of homogeneity of units within each input type.

There are important exceptions to this approach. In education and labour market analyses, screening models seek to throw light on how organisations attempt to sort labour by expected productivity when they know the units are not identical but lack information on which individuals are of higher quality (Riley, 2001). The literature on 'superstars' is particularly interesting (Rosen, 1981). Basketball players, opera singers and university researchers differ in inherent ability and the best will be able to command differential rents, the amount of which will be influenced by the consumers' willingness to pay for a higher quality product and the size of the market.

Variation in the productivity of inputs will affect final users if neither input nor output prices

adjust to differences in productivity. In most health care systems, input prices (wages or fees) do not fully reflect differences in productivity because, to a great extent, wages are set (or negotiated) centrally. In sectors where relative earnings are constrained, institutions such as universities or hospitals rely on prestige, working conditions and career development to attract the higher quality labour. A 'pecking order' of institutions emerges. Where location-specific, non-monetary rewards replace cash rent for differential factor productivity, there can be important implications for final consumers. Specifically this is the case if the differential in input productivity maps into a quality differential in output. Differences in the quality of final output do not necessarily disadvantage consumers as long as these differences are reflected in relative output prices. Many industries are characterised by vertical product differentiation, where higher quality variants of a product sell at a premium. However, in European health care systems, where health care services are provided at regulated or zero prices, there is very limited scope for quality adjustments in the prices of these services. Even then consumers would not be affected if they were mobile and able to move to a higher quality service. But if the consumer is a hospital patient, it is less likely that she is able to select an institution that attracts the higher quality medical labour, primarily due to limited mobility.

Heterogeneity of inputs creates two problems for a third party purchaser. First, when observing differences between hospitals in activity and quality, the purchaser cannot distinguish between differences that are due to inherent differences in productivity and those that are due to differences in effort. Purchasers commonly seek to create incentives to increase effort in low productivity hospitals as a way of increasing overall efficiency. The greater the importance of factor heterogeneity, the more limited the scope for reducing variation through contract incentives directed at effort. Second, given that hospitals are location specific, differences in performance due to factor heterogeneity can have important equity implications for access and quality. The purchaser may need to address explicitly the resulting trade-off between efficiency and equity and seek new instruments to reduce the impact on equity of variations in activity and quality.

3 The Model

We extend the model developed by Chalkley and Malcomson in order to explore the consequences of variations in factor productivity (Chalkley & Malcomson, 1998). The purchaser is concerned with maximising its perception of population health, *b*, this being a function of the volume of patients treated (*x*) and a single dimensional measure of quality (*q*). Volume is measurable and contractible, quality is not – at least, not in a cost-effective manner. The budget set by the purchaser takes the form $B(x) = \overline{B} + px$ where $\overline{B}$ represents a fixed payment and *p* the price per patient.

The objective of hospital *j* is

$$H^j = B^j(x) - c(x,q) - v(x,q) + \beta b(x,q) - \gamma^j x - \delta^j q \qquad (1)$$

Following Chalkley and Malcomson, but omitting the argument for effort [1], we denote $c(x,q)$ as the monetary cost of treating *x* patients of *q* quality, with $v(x,q)$ capturing non-monetary cost. We assume that costs and effort are increasing at an increasing rate in both volume and quality. The direction of the cross partials $c_{xq}(x,q)$ and $v_{xq}(x,q)$ are undetermined (subscripts denote the derivatives with respect to the relevant argument). β represents the extent to which the hospital seeks to promote population health, this 'benevolence' being distributed as $\beta \in [0,1]$. Factor heterogeneity is introduced by the two parameters γ and δ. γ identifies relative productivity in volume, the number of patients treated by given labour inputs. δ identifies relative quality, the quality of patient outcomes for patients treated by given labour inputs. For expediency we assume that hospitals are one of two types, hospitals with high productivity and hospitals with low productivity, but that productivity in volume does not necessarily imply productivity in quality. For expediency, we assume that hospitals are distributed into two types in terms of both γ and δ, with $\lambda \in [0,1]$, but information about type is not known to either party at the contracting stage.

The purchaser's objective function is written as

$$R = \begin{Bmatrix} \lambda\left[b\left(x^0,q^0\right)-c\left(x^0,q^0\right)-v\left(x^0,q^0\right)-\gamma^0 x^0-\delta^0 q^0-\alpha B^0\left(x^0\right)\right] \\ +\left(1-\lambda\right)\left[b\left(x^1,q^1\right)-c\left(x^1,q^1\right)-v\left(x^1,q^1\right)-\gamma^1 x^1-\delta^1 q^1-\alpha B^1\left(x^1\right)\right] \\ -\frac{\psi}{2}\lambda\left(1-\lambda\right)\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]^2 \end{Bmatrix} \qquad (2)$$

which contains, for each hospital type, arguments for the population health benefit, the private utility of the hospital (we omit any benefit the hospital may derive from health production in order to avoid double counting), and the disutility associated with the budget allocation. The final term captures the purchaser's loss of utility associated with variation in performance. For $\psi = 0$, variation is not of concern. For $\psi > 0$, it is.

The hospital's maximisation problem is:

$$\max_{x^j,q^j} H^j \quad s.t. \quad p^j \quad j = 0,1 \qquad (3)$$

with first-order conditions:

$$H_x^j = p^j - c_x\left(x^j,q^j\right)-v_x\left(x^j,q^j\right)+\beta b_x\left(x^j,q^j\right)-\gamma^j = 0 \qquad (3a)$$

$$H_q^j = -c_q\left(x^j,q^j\right)-v_q\left(x^j,q^j\right)+\beta b_q\left(x^j,q^j\right)-\delta^j = 0 \qquad (3b)$$

A unique optimum exists if the Hessian is positive $\mathrm{H} = H_{xx}^j H_{qq}^j - \left(H_{xq}^j\right)^2 > 0$, with second-order derivatives:

$$H_{xx}^j = -c_{xx}\left(x^j,q^j\right)-v_{xx}\left(x^j,q^j\right)+\beta b_{xx}\left(x^j,q^j\right) < 0; \qquad (3c)$$

$$H_{qq}^j = -c_{qq}\left(x^j,q^j\right)-v_{qq}\left(x^j,q^j\right)+\beta b_{qq}\left(x^j,q^j\right) < 0; \qquad (3d)$$

$$H_{xq}^j = -c_{xq}\left(x^j,q^j\right)-v_{xq}\left(x^j,q^j\right)+\beta b_{xq}\left(x^j,q^j\right). \qquad (3e)$$

The cross-derivative will be positive if volume and quality are complements and negative if they are substitutes.

An analysis of the comparative statics provides the following relationships with respect to the case payment, with the quality effect depending on the nature of the complementarity between volume and quality:

$$\hat{x}_p^j := \frac{dx^j}{dp^j} = \frac{-H_{qq}^j}{\mathrm{H}} > 0 \qquad (4a); \qquad \hat{q}_p^j := \frac{dq^j}{dp^j} = \frac{H_{xq}^j}{\mathrm{H}} \geq 0 \Leftrightarrow H_{xq}^j \geq 0 \qquad (4b)$$

With respect to the heterogeneity parameters, for γ we have:

$$\hat{x}_{\gamma}^{j} \coloneqq \frac{dx^{j}}{d\gamma^{j}} = \frac{H_{qq}^{j}}{\mathrm{H}} < 0 \qquad (4c) \qquad\qquad \hat{q}_{\gamma}^{j} \coloneqq \frac{dq^{j}}{d\gamma^{j}} = \frac{-H_{xq}^{j}}{\mathrm{H}} \leq 0 \Leftrightarrow H_{xq}^{j} \geq 0 \quad (4d)$$

$$\frac{db^{j}}{d\gamma^{j}} = b_{x}\hat{x}_{\gamma}^{j} + b_{q}\hat{q}_{\gamma}^{j} = \frac{b_{x}H_{qq}^{j} - b_{q}H_{xq}^{j}}{\mathrm{H}} < 0 \Leftarrow H_{xq}^{j} \geq 0 \qquad (4e)$$

These imply that heterogeneity in γ leads to variation in volume, quality and population health. For example, if volume and quality are complements, a low level of γ implies higher volume, higher quality, and higher health benefits. The equivalent comparative statics for δ are:

$$\hat{x}_{\delta}^{j} \coloneqq \frac{dx^{j}}{d\delta^{j}} = \frac{-H_{xq}^{j}}{\mathrm{H}} > 0 \Leftrightarrow H_{xq}^{j} < 0 \quad (4f) \qquad\qquad \hat{q}_{\delta}^{j} \coloneqq \frac{dq^{j}}{d\delta^{j}} = \frac{H_{xx}^{j}}{\mathrm{H}} < 0 \qquad (4g)$$

Benchmark case

As a benchmark for subsequent analysis, we examine the case where the purchaser is able to make differentiated payments across hospitals. The purchaser sets the budget for each hospital $\left\{\overline{B}^{j}; p^{j}\right\}$ $j = 0,1$ in order to maximise its objective function subject to the hospital's participation constraint $H^{j} \geq 0$. This implies $B^{j}\left(x^{j}\right) = \overline{B}^{j} + p^{j}x^{j} = c\left(x^{j}, q^{j}\right) + v\left(x^{j}, q^{j}\right) - \beta b\left(x^{j}, q^{j}\right) + \gamma^{j}x^{j} + \delta^{j}q^{j}$.

Thus, the purchaser's objective function becomes:

$$R = \left\langle \begin{array}{l} \lambda\left\{\left(1+\alpha\beta_{b}^{0}\right)b\left(x^{0}, q^{0}\right) - (1+\alpha)\left[c\left(x^{0}, q^{0}\right) + v\left(x^{0}, q^{0}\right) + \gamma^{0}x^{0} + \delta^{0}q^{0}\right]\right\} \\ +(1-\lambda)\left\{\left(1+\alpha\beta_{b}^{1}\right)b\left(x^{1}, q^{1}\right) - (1+\alpha)\left[c\left(x^{1}, q^{1}\right) + v\left(x^{1}, q^{1}\right) + \gamma^{1}x^{1} + \delta^{1}q^{1}\right]\right\} \\ -\frac{\psi}{2}\lambda(1-\lambda)\left[b\left(x^{0}, q^{0}\right) - b\left(x^{1}, q^{1}\right)\right]^{2} \end{array} \right\rangle \qquad (5)$$

Hence, $\max_{p^{j}} R \quad j = 0,1$ subject to hospital's best-responses $\hat{x}^{j}\left(p^{j}\right)$ and $\hat{q}^{j}\left(p^{j}\right)$, as in (4a) and (4b), gives the first order conditions $R_{p^{j}} = R_{x^{j}}\hat{x}_{p}^{j} + R_{q^{j}}\hat{q}_{p}^{j} = 0$; $j = 0,1$. The first-order derivatives are:

$$R_{t^{0}} = \lambda\left\{ \begin{array}{l} \left(1+\alpha\beta_{b}^{0}\right)b_{t}\left(x^{0}, q^{0}\right) - (1+\alpha)\left[c_{t}\left(x^{0}, q^{0}\right) + v_{t}\left(x^{0}, q^{0}\right) + \tau_{t}^{0}\right] \\ -\psi(1-\lambda)\left[b\left(x^{0}, q^{0}\right) - b\left(x^{1}, q^{1}\right)\right]b_{t}\left(x^{0}, q^{0}\right) \end{array} \right\}; \quad t = x, q \qquad (5a)$$

$$R_{t^1} = (1-\lambda)\begin{Bmatrix}(1+\alpha\beta_b^1)b_t(x^1,q^1)-(1+\alpha)\left[c_t(x^1,q^1)+v_t(x^1,q^1)+\tau_t^1\right]\\ +\psi\lambda\left[b(x^0,q^0)-b(x^1,q^1)\right]b_t(x^1,q^1)\end{Bmatrix};\quad t=x,q \qquad (5b)$$

where $\tau_x^j=\gamma^j;\quad \tau_q^j=\delta^j$. Observing $c_x(x^j,q^j)+v_x(x^j,q^j)+\gamma^j=p^j+\beta_b^j b_x(x^j,q^j)$ from (3a) and $c_q(x^j,q^j)+v_q(x^j,q^j)+\delta^J=\beta b_q(x^j,q^j)$ from (3b), we obtain the explicit first-order conditions:

$$R_{p^0}=\lambda\begin{Bmatrix}\left\langle\left\{1-\beta-\psi(1-\lambda)\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}b_x(x^0,q^0)-(1+\alpha)p^0\right\rangle\hat{x}_p^0\\ +\left\{1-\beta-\psi(1-\lambda)\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}b_q(x^0,q^0)\hat{q}_p^0\end{Bmatrix}=0 \qquad (5c)$$

$$R_{p^1}=(1-\lambda)\begin{Bmatrix}\left\langle\left\{1-\beta+\psi\lambda\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}b_x(x^1,q^1)-(1+\alpha)p^1\right\rangle\hat{x}_p^1\\ +\left\{1-\beta+\psi\lambda\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}b_q(x^1,q^1)\hat{q}_p^1\end{Bmatrix}=0 \qquad (5d)$$

To ease future exposition, we reformulate these two conditions to obtain:

$$\Pi^0(p^0,p^1,\psi,\gamma^0,\delta^0)-p^0=0 \qquad (6a) \qquad \Pi^1(p^0,p^1,\psi,\gamma^0,\delta^0)-p^1=0 \qquad (6b)$$

where

$$\Pi^0(\cdot)=\frac{\left\{1-\beta-\psi(1-\lambda)\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}}{1+\alpha}\left[b_x(x^0,q^0)+b_q(x^0,q^0)\xi^0\right] \qquad (6c)$$

$$\Pi^1(\cdot)=\frac{\left\{1-\beta+\psi\lambda\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}}{1+\alpha}\left[b_x(x^1,q^1)+b_q(x^1,q^1)\xi^1\right] \qquad (6d)$$

with

$$\xi^j:=\frac{\hat{q}_p^j}{\hat{x}_p^j}=\frac{-H_{xq}^j}{H_{qq}^j};\quad j=0,1.$$

Note that a unique and stable solution $\{p^{0*};p^{1*}\}$ exists if and only if the Hessian of (6a) and (6b), satisfies $Z=\Pi_{p^0}^0\Pi_{p^1}^1-\Pi_{p^1}^0\Pi_{p^0}^1>0$. This is verified in the appendix to Lemma 1.

From (6a) and (6b) the optimum case-payments follow as

$$p^{0*}=\frac{\left\{1-\beta-\psi(1-\lambda)\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}}{1+\alpha}\left[b_x(x^0,q^0)+b_q(x^0,q^0)\xi^0\right] \qquad (7a)$$

$$p^{1*}=\frac{\left\{1-\beta+\psi\lambda\left[b(x^0,q^0)-b(x^1,q^1)\right]\right\}}{1+\alpha}\left[b_x(x^1,q^1)+b_q(x^1,q^1)\xi^1\right] \qquad (7b)$$

The case-payments reflect in turn

- The degree to which the hospital is interested in improving population health β. The value of the reimbursement rate falls β. If the purchaser is not concerned about variation, $\psi = 0$, reimbursement rates are set to zero if hospitals fully internalise population health, i.e. if $\beta = 1$.
- The cost of funds α, which bears negatively on the case-payment.
- The complementarity between quality and volume in health production. An increase in the case-payment rate has a non-negative impact on quality if and only if quality and volume are complements. In this case, $\xi^j \geq 0$ and, thus, $b_x\left(x^j, q^j\right) + b_q\left(x^j, q^j\right)\xi^j > 0$. In this case, case-payment and volume are greater than would be observed in the first-best situation, because of a desire stimulate quality. If quality and volume are substitutes, $\xi^j < 0$, the case-payment is adjusted downward depending on the effect of quality on health benefit b_q. Volume is lower than in the first-best situation.
- The cost of variation $\psi > 0$. Given a positive net effect of reimbursement on population health, $b_x\left(x^j, q^j\right) + b_q\left(x^j, q^j\right)\xi^j > 0$, the case-payment is adjusted downwards (upwards) for the hospital which attains the greater (smaller) population health. The extent of this adjustment increases with the other type's share in the population.

In the next section, we analyse in greater depth the implications of a regulatory concern about variation on the second-best rates of case-payment. This provides the basis for our subsequent analysis of the scope for attaining a Chalkley and Malcomson second-best with the use of a uniform case-payment and a uniform standard on volume.

4 The effect of a concern for variation

In order to simplify our analysis and in line with empirical relevance we exclude the case of negative case-payments by assuming

$$b_x\left(x^j, q^j\right) + b_q\left(x^j, q^j\right)\xi^j \geq 0 \quad j = 1,2 \tag{8a}$$

$$\psi \leq \psi^{\max}, \text{ with } \psi^{\max} := \psi \left| \min\left\{p^{0*}(\psi), p^{1*}(\psi)\right\} = 0 \right. \tag{8b}$$

In what follows, we use the following definitions:

Definitions.

- We define hospital 0 as being the hospital that, for any given payment, will produce the lower volume, hence $\hat{x}^0(p) \leq \hat{x}^1(p)$.
- $\widetilde{\psi}$ the value of ψ at which there is no variation between hospitals in their volume
$$\widetilde{\psi} := \psi \left| x^{0*}(\psi) = x^{1*}(\psi) \right. \quad (9a)$$
- ψ^* the value of ψ at which there is no variation between hospitals in their optimal payment rate $\psi^* := \psi \left| p^{0*}(\psi) = p^{1*}(\psi) \right.$ (9b)
- $\overline{\psi}$ the value of ψ at which there is no variation between hospitals in their contribution to population health $\overline{\psi} := \psi \left| b\left[x^{0*}(\psi), q^{0*}(\psi)\right] = b\left[x^{1*}(\psi), q^{1*}(\psi)\right] \right.$ (9c)

where $x^{j*}(\psi) = \hat{x}^j\left[p^{j*}(\psi)\right]$ and $q^{j*}(\psi) = \hat{q}^j\left[p^{j*}(\psi)\right]$; $j = 0,1$.

We assume that $\overline{\psi} \leq \psi^{\max}$ throughout. This guarantees that $\min\left\{p^{0*}(\psi), p^{1*}(\psi)\right\} \geq 0$ for all $\psi \geq 0$.

Also, let

- $\Delta^{p*}(\psi) = p^{0*}(\psi) - p^{*1}(\psi)$ (9d)
- $\Delta^{x*}(\psi) = x^{0*} - x^{1*} = \hat{x}^0\left[p^{0*}(\psi)\right] - \hat{x}^1\left[p^{1*}(\psi)\right]$ (9e)
- $\Delta^{b*}(\psi) := b\left[x^{0*}(\psi), q^{0*}(\psi)\right] - b\left[x^{1*}(\psi), q^{1*}(\psi)\right]$ (9f)

These denote the difference between the optimal payment rates, volumes and contributions to population health for the two types of hospital.

The following Lemma 1 characterises the effect of the cost of variation on the optimal case-payments, $\Delta^{p*}(\psi)$, $\Delta^{x*}(\psi)$ and $\Delta^{b*}(\psi)$.

Lemma 1: (i) $\operatorname{sgn}\Delta^{p*}_{\psi} = \operatorname{sgn}\Delta^{x*}_{\psi} = -\operatorname{sgn}\Delta^{b*}(0)$ for all $\psi \geq 0$, and (ii) $\lim_{\psi\to\infty}\Delta^{p*}_{\psi} = \lim_{\psi\to\infty}\Delta^{x*}_{\psi} = \lim_{\psi\to\infty}\Delta^{b*}(\psi) = 0$.

Proof: See appendix.

Using the definitions above, we examine the pattern of optimal payment rates, conditional upon the purchaser's valuation of ψ, in the presence of factor heterogeneity under three scenarios:

- heterogeneity in γ;
- heterogeneity in δ when volume and quality are complements;
- heterogeneity in δ when volume and quality are substitutes.

Case 1: Heterogeneity in γ.

First we examine heterogeneity in γ, with $\gamma^0 > \gamma^1$ and $\delta^0 = \delta^1$. Hence, hospital 0 suffers lower productivity in volume but we assume that heterogeneity in the productivity of labour has no direct impact on the production of quality (although an indirect effect is likely). This implies that $\hat{x}^0(p) < \hat{x}^1(p)$, consistent with our earlier definition of hospital 0.

The following Lemma characterises the pattern of optimal payments:

Lemma 2: The optimal payment rates and volumes depend on ψ as follows.

$\Delta^{p*}(\psi) > 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) < 0; \Delta^{x*}_{\psi}(\psi) > 0$ for all $\psi \geq 0$.

Proof: See Appendix.

The Lemma can be understood with reference to figure 1.

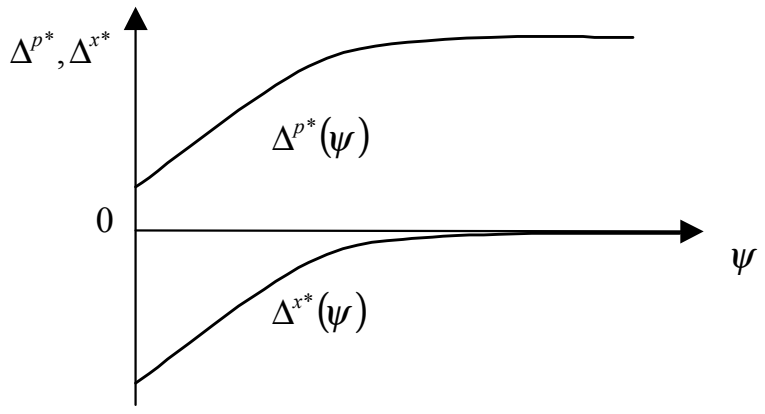

Figure 1

Figure 1 plots $\Delta^{p^*}(\psi)$, the difference in optimal payment rates for hospital 0 and hospital 1, as per definition (9d), and $\Delta^{x^*}(\psi)$, the difference between the volume produced, as per definition (9e).

In accordance with assumption (8a), population health never decreases in the payment rate even when allowing for a possible reduction in quality under a substitutive relationship between volume and quality. Thus, with reference to figure 1, heterogeneity in γ implies the following:

- The less productive hospital 0, with $\gamma^0 > \gamma^1$, produces less population health, $b^0 < b^1$ as long as it produces a lower output $x^0 < x^1$. This in turn, implies that the fee differential $\Delta^{p^*}(\psi)$ should increase in the purchaser's disutility from the existence of variation as long as $\Delta^{x^*}(\psi) < 0$.
- The marginal impact of payment rate on population health (taking into account the quality response) is greater for the less productive hospital so that, if the purchaser is not concerned with variation (i.e. $\psi = 0$), the case-payment should be set at a higher level for hospital 0, such that $\Delta^{p^*}(\psi) > 0$ for all $\psi \geq 0$.
- Again, if variation is not of concern, hospital 0 should produce a lower output, such that $\Delta^{x^*}(0) < 0$.

As the rate differential Δ^{p^*} increases in ψ, this, in turn, implies that the differential in volume $\Delta^{x^*}(\psi)$ increases in ψ. Under assumption (8a), this reduces not only variation in volume but also in population health.

Since, by assumption, hospitals do not differ in their productivity with regard to quality, $\delta^0 = \delta^1$, this implies that if hospitals 0 and 1 produce the same volume $x^0 = x^1$, they choose the same quality and attain the same population health, $x^0 = x^1 \Rightarrow \hat{q}^0(x^0) = \hat{q}^1(x^1) \Rightarrow b^0 = b^1$. Consequently, variation in volume and population health could be eliminated by inducing a payment differential for which $\Delta^{x^*} = 0$. While this is feasible, it is never optimal because the marginal gain from complete elimination of variation is zero while the marginal cost in

inefficiency terms is large. However, variation becomes arbitrarily small for high values of ψ.

Case 2: Heterogeneity in δ and volume and quality are complements.

Second we examine the case where there is heterogeneity in δ, assuming that volume and quality are complementary outputs. Specifically, assume $\gamma^0 = \gamma^1$, $\delta^0 > \delta^1$ and $H^0_{xq} \geq 0$. Under this scenario, hospital 0 suffers lower productivity in quality. While we assume that heterogeneity has no direct impact on the production of volume, because volume and quality are complements, hospital 0 will produce lower volume also, implying $\hat{x}^0\left(p,\delta^0\right) \leq \hat{x}^1\left(p,\delta^1\right)$. Again this scenario is consistent with our definition of hospital 0.

Lemma 3: (i) There exists $k^+ > 0$ such that $H^0_{xq} \in [0,k^+[\Leftrightarrow \infty > \tilde{\psi} \geq \psi^* > 0$. (ii) The optimal payment rates and volumes then depend on ψ as follows.

$\psi \in [0,\psi^*[\Leftrightarrow \left\{\Delta^{p*}(\psi) < 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) < 0; \Delta^{x*}_{\psi}(\psi) > 0\right\};$

$\psi \in [\psi^*,\tilde{\psi}] \Leftrightarrow \left\{\Delta^{p*}(\psi) \geq 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) \leq 0; \Delta^{x*}_{\psi}(\psi) > 0\right\};$

$\psi > \tilde{\psi} \Leftrightarrow \left\{\Delta^{p*}(\psi) > 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) > 0; \Delta^{x*}_{\psi}(\psi) > 0\right\};$

Proof: See appendix.

The Lemma can be understood with reference to figure 2.

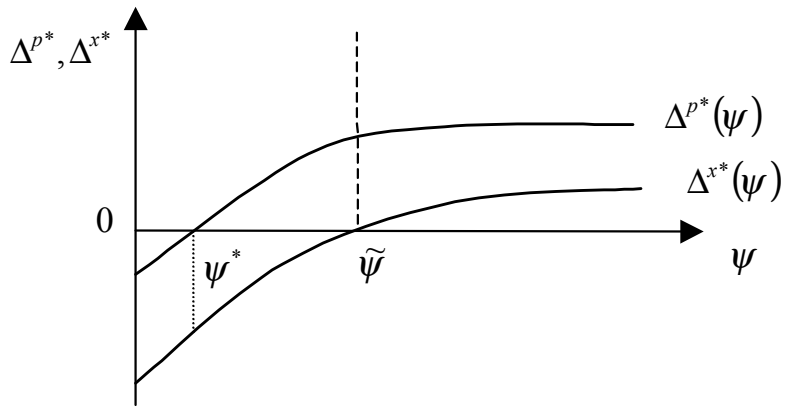

Figure 2

Taking into account assumption (8a), heterogeneity in δ implies for $H^0_{xq} \geq 0$:

- The less productive hospital 0, with $\delta^0 > \delta^1$, produces a lower health benefit, $b^0 < b^1$ for any output $x^0 < \hat{x}$, where $\hat{x} > x^1$. In contrast to the previous case (of heterogeneity in γ) variation in population health is eliminated entirely only if the unproductive hospital is induced to produce a volume strictly in excess of that of the productive hospital, such that $\Delta^{x^*}(\psi) = \hat{x} - x^{1*} > 0$. This is because, when quality is complementary to volume, only 'over-production' can induce hospital 0 to generate the increase in quality that is necessary to attain the same health output as hospital 1. Hence the fee differential $\Delta^{p^*}(\psi)$ should increase in the disutility of variation ψ as long as $\Delta^{x^*}(\psi) < \hat{x} - x^{1*}$.
- The marginal impact of payment rate on population health (taking into account the quality response) is lower for the less productive hospital so that, if the purchaser is not concerned with variation i.e. $\psi = 0$, the payment rate should be set at a higher level for hospital 1, such that $\Delta^{p^*}(0) < 0$. This also implies that hospital 0 will produce a lower volume, such that $\Delta^{x^*}(0) < 0$.

If the disutility of variation is sufficiently great the payment for hospital 0 should exceed that for hospital 1, i.e, $\Delta^{p^*}(\psi) \geq 0$ if $\psi \geq \psi^*$. As long as $\psi \in [\psi^*, \tilde{\psi}]$, the difference in payment rates is insufficient to induce hospital 0 to produce greater volume than hospital 1, implying that $\Delta^{x^*}(\psi) \leq 0$. However, for $\psi > \tilde{\psi}$, the fee differential is sufficient for hospital 0 to produce the greater volume, $\Delta^{x^*}(\psi) > 0$. Note, however, that since hospital 1 produces higher quality, the variance in population health is not eliminated unless $\Delta^{x^*}(\psi) = \hat{x} - x^{1*} > 0$. Again, while elimination of variation is feasible, it is not optimal.

Case 3: Heterogeneity in δ and volume and quality are substitutes.

Finally, we examine the case where there is heterogeneity in δ, assuming that volume and quality are substitutes. Specifically, we assume $\gamma^0 = \gamma^1$, $\delta^0 < \delta^1$ and $H^0_{xq} \leq 0$. Unlike under the previous scenarios, we now assume that hospital 1 is less productive (in quality terms), i.e. $\delta^0 < \delta^1$. Because volume and quality are substitutes, however, hospital 0 will produce

lower volume for a given payment rate, implying $\hat{x}^0(p,\delta^0) \leq \hat{x}^1(p,\delta^1)$. Hence, ultimately, this scenario remains consistent with our definition of hospital 0.

Lemma 4: Let $\{\gamma^0 = \gamma^1; \delta^0 < \delta^1; H^0_{xq} \leq 0\}$. (i) There exists $k^- < 0$ such that $H^0_{xq} \in]k^-, 0] \Leftrightarrow \infty > \psi^* \geq \tilde{\psi} > 0$. (ii) The second-best payment rates and volumes then depend on ψ as follows.

$\psi \in [0, \tilde{\psi}[\Leftrightarrow \{\Delta^{p*}(\psi) > 0; \Delta^{p*}_{\psi}(\psi) < 0; \Delta^{x*}(\psi) > 0; \Delta^{x*}_{\psi}(\psi) < 0\}$;

$\psi \in [\tilde{\psi}, \psi^*] \Leftrightarrow \{\Delta^{p*}(\psi) \geq 0; \Delta^{p*}_{\psi}(\psi) < 0; \Delta^{x*}(\psi) \leq 0; \Delta^{x*}_{\psi}(\psi) < 0\}$;

$\psi > \psi^* \Leftrightarrow \{\Delta^{p*}(\psi) < 0; \Delta^{p*}_{\psi}(\psi) < 0; \Delta^{x*}(\psi) < 0; \Delta^{x*}_{\psi}(\psi) < 0\}$.

Proof: See appendix.

The Lemma can be understood with reference to figure 3.

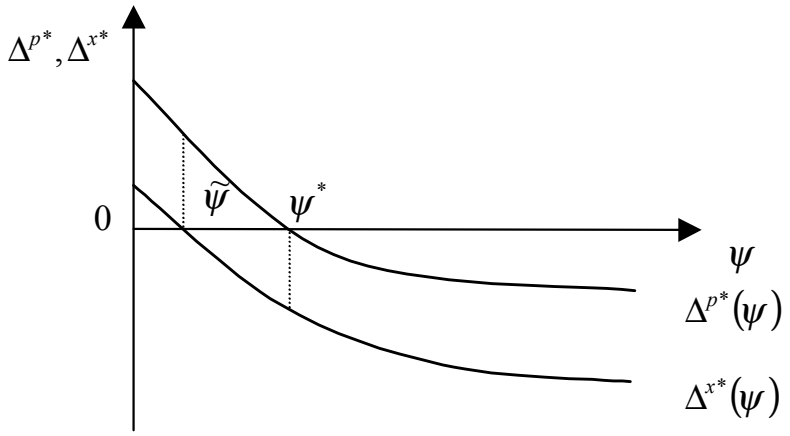

Figure 3

Taking into account assumption (8a), heterogeneity in δ when volume and quality are substitutes implies that:

- Hospital 0, which is more productive in quality given that $\delta^0 < \delta^1$, produces greater population health, $b^0 > b^1$ for any volume $x^0 > \hat{x}$, where $\hat{x} < x^1$. As when volume and quality were complements, variation in volume is now eliminated entirely only if the hospital 0 is induced to produce an output of population health that is strictly less than that of the hospital 1, $\Delta^{x*}(\psi) = \hat{x} - x^{1*} < 0$. Although quality trades-off against volume,

assumption (8a) implies that only 'under-production' can induce hospital 0 to generate the decrease in population health that is necessary to attain the same volume as hospital 1. Hence, this implies that the fee differential $\Delta^{p*}(\psi)$ should decrease in the disutility of variation ψ as long as $\Delta^{x*}(\psi) > \hat{x} - x^{1*}$.

- The marginal impact of payment rate on population health (taking into account the quality response) is lower for the hospital 0, so that, if the purchaser is not concerned with variation ($\psi = 0$), the payment rate should be set at a higher level for hospital 0, and, thus, $\Delta^{p*}(0) > 0$. This implies that hospital 1 should produce a lower volume, such that $\Delta^{x*}(0) > 0$.

If the disutility of variation is sufficiently great the purchaser will induce hospital 0 to produce a lower volume than hospital 1 such that $\Delta^{x*}(\psi) < 0$. This is the case for $\psi > \tilde{\psi}$. Compared to the complementary scenario, for $\psi > \psi^*$, the differential in payment rates is now negative, $\Delta^{p*}(\psi) < 0$, with hospital 0 still attaining the greater population health. Variation in population health becomes arbitrarily small for high ψ.

5 Use of standards

5.1 No role for standards in benchmark case

Consider the effect of a binding minimum standard on volume that restricts the hospital to $x^j \geq \underline{x}^j$. The hospital's objective is now

$$H^j{}' = B^j(x) - c(x,q) - v(x,q) + \beta b(x,q) - \gamma^j x - \delta^j q - \varphi^j\left(\underline{x}^j - x\right) \qquad (1')$$

where the shadow price of the constraint satisfies $\varphi^j > 0 \Leftrightarrow \underline{x}^j = x^j$. First-order conditions are now

$$H'^j_x = p^j - c_x\left(x^j, q^j\right) - v_x\left(x^j, q^j\right) + \beta b_x\left(x^j, q^j\right) - \gamma^j + \varphi^j = 0 \qquad (3a')$$

and (3b). Consider now the purchaser's choice of a standard $\underline{x}^j$ given the payment p^j. The best responses to the standard are given by

$$\hat{x}_x^j := \begin{cases} \dfrac{dx^j}{d\underline{x}^j} = 1 \Leftrightarrow x^j = \underline{x}^j \\ 0 \qquad\qquad\quad \Leftrightarrow x^j > \underline{x}^j \end{cases} \qquad (4a') \qquad \hat{q}_x^j := \begin{cases} \dfrac{dq^j}{d\underline{x}^j} = \dfrac{-H_{xq}^j}{H_{qq}^j} \Leftrightarrow x^j = \underline{x}^j \\ 0 \qquad\qquad\quad\;\; \Leftrightarrow x^j > \underline{x}^j \end{cases} \qquad (4b')$$

Thus the effect of the standard on the purchaser's objective is given by:

$R_{\underline{x}^j} = R_{x^j}\hat{x}_x^j + R_{q^j}\hat{q}_x^j;\ j = 0,1.$

Inserting (5a) and (5b), respectively; and observing $c_x\left(x^j,q^j\right)+v_x\left(x^j,q^j\right)+\gamma^j = p^j+\beta b_x\left(x^j,q^j\right)+\varphi^j$ from (3a') and $c_q\left(x^j,q^j\right)+v_q\left(x^j,q^j\right)+\delta^j = \beta b_q\left(x^j,q^j\right)$ from (3b), we obtain

$$R_{\underline{x}^0} = \lambda\begin{Bmatrix} \left\langle \left\{1-\beta-\psi(1-\lambda)\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\} b_x\left(x^0,q^0\right)-(1+\alpha)\left(p^0+\varphi^0\right)\right\rangle \hat{x}_x^0 \\ +\left\{1-\beta-\psi(1-\lambda)\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\} b_q\left(x^0,q^0\right)\hat{q}_x^0 \end{Bmatrix} \qquad (10a)$$

$$R_{\underline{x}^1} = (1-\lambda)\begin{Bmatrix} \left\langle \left\{1-\beta+\psi\lambda\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\} b_x\left(x^1,q^1\right)-(1+\alpha)\left(p^1+\varphi^1\right)\right\rangle \hat{x}_x^1 \\ +\left\{1-\beta+\psi\lambda\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\} b_q\left(x^1,q^1\right)\hat{q}_x^1 \end{Bmatrix} \qquad (10b).$$

Using (5c) and (5d) and observing the best-responses (4a) and (4b) as well as (4b'), it is readily verified that $p^j \geq p^{j*} \Rightarrow R_{\underline{x}^j} \leq 0$. Since the standard and case-payment are perfect substitutes under perfect information there is no role for a standard if the purchaser is able to set the first-best payment rates.

However, if the purchaser cannot set the optimal case-payment there may be a role for a standard. Inserting $\varphi^j = -p^j + c_x\left(x^j,q^j\right)+v_x\left(x^j,q^j\right)-\beta b_x\left(x^j,q^j\right)+\gamma^j$ from (3a') into (10a) and (10b) and comparing with (5c) and (5d), we can verify for $p^j \leq p^{j*}$ that $R_{\underline{x}^j} = 0 \Leftrightarrow \underline{x}^j = x^{j*}$. The standard should be set at the level that would be realised under an optimal payment rate.

5.2 Uniform payment and uniform standard

We now consider the more realistic case in which the purchaser is constrained to use a uniform payment scheme $B^o(x) = B^1(x) = B(x)$ and a uniform standard $\underline{x}^0 = \underline{x}^1 = \underline{x}$. This

case would arise if, at the point of contracting, neither the purchaser nor hospitals themselves know the hospital type (0 or 1) and contingent contracts cannot be written.

For the following exposition, we make the following assumptions:

- hospitals learn their type after contracting but before production, while the purchaser is informed throughout only about the distribution parameter λ.
- ex-post participation is satisfied (justified perhaps if contractual fines can be imposed if a hospital withdraws from a contract).
- hospitals are risk-neutral.

Ex-ante participation requires, for $B(x) = \overline{B} + px$,

$$\overline{B} \geq \left\{ \begin{array}{l} \lambda\left[c\left(x^0, q^0\right) + v\left(x^0, q^0\right) - \beta b\left(x^0, q^0\right) - \left(p - \gamma^0\right)x^0 + \delta^0 q^0\right] \\ + (1-\lambda)\left[c\left(x^1, q^1\right) + v\left(x^1, q^1\right) - \beta b\left(x^1, q^1\right) - \left(p - \gamma^1\right)x^1 + \delta^1 q^1\right] \end{array} \right\} \qquad (11)$$

where $\max_{\overline{B}} R \Leftrightarrow \min \overline{B}$ implies that the above holds with equality. Taking the binding participation constraint into account, the purchaser's objective is identical to (2). The effect of the payment rate on social welfare is given by $R_p = R_{p^0} + R_{p^1}$ with R_{P^0} and R_{P^1} given by (5c) and (5d), respectively. Likewise, the effect of a standard on welfare is given by $R_{\underline{x}} = R_{\underline{x}^0} + R_{\underline{x}^1}$ with $R_{\underline{x}^0}$ and $R_{\underline{x}^1}$ as given by (10a) and (10b).

We can now demonstrate under which conditions the purchaser can choose a combination of payment rate and standard $\left(p^*; \underline{x}^*\right)$ that implements the second-best, i.e. the allocation that would be realised by using type specific payment rates p^{j*}.

Recall that $\hat{x}^0(p) \leq \hat{x}^1(p)$.

Proposition 1.1: The second-best allocation $\left(p^* = p^{1*}; \underline{x}^* = x^{0*}\right)$ is attainable if and only if $p^{0*} \geq p^{1*}$ and $x^{0*} \leq x^{1*}$.

Proof: See appendix.

The aim is to induce hospital 1, which is more responsive to the payment rate, to produce an optimal volume. This can be achieved by setting $p^* = p^{1*}$. At this payment rate, hospital 0

will produce a lower volume but, with a standard set at $\underline{x}^* = x^{0*}$, can be encouraged to expand volume to its optimal level. This applies only when $p^{0*} \geq p^{1*}$ and $x^{0*} \leq x^{1*}$.

Proposition 1.2: If $x^{0*} > x^{1*}$ the second-best is unattainable and the purchaser sets $\underline{x}^* = \lambda x^{0*} + (1-\lambda)x^{1*}$. The standard binds for both types, whereas there is no role for the payment rate.

Proof: See appendix.

If the optimum requires over-production by hospital 0, this will always imply over-production by hospital 1. The implication is that the standard should be set with reference to a weighted mean of the optimal volume produced by each hospital.

Proposition 1.3: If $p^{0*} < p^{1*}$ the second-best is unattainable and the purchaser sets $p^* = \lambda p^{0*} + (1-\lambda)p^{1*}$. There is no role for the standard.

Proof: See appendix.

If the optimum requires that the payment rate for hospital 1 is greater than for hospital 0, this will always imply that hospital 0 produces a greater volume than is optimal. The implication is that the payment rate should be set with reference to a weighted mean of the optimal payment rates for each hospital.

We can now characterise which of the cases described in the above propositions is likely to arise depending on:

- the type of heterogeneity (i.e. heterogeneity in γ or δ);
- the hospital's technology and preferences, or more specifically, the complementarity or substitutability of quality and volume in production (i.e. sgn H_{xq}); and
- the purchaser's preference for eliminating variation (i.e. ψ).

In order to do so, we can go back to the three cases we have characterised before.

Case 1: Heterogeneity in γ.

Proposition 2: The second-best can always be implemented if hospitals are heterogeneous in

γ but not in δ.

Proof: The second-best obtains if and only if $p^{0*} \geq p^{1*}$ and $x^{0*} \leq x^{1*}$ or equivalently if and only if $\Delta^{p*}(\psi) \geq 0$ and $\Delta^{x*}(\psi) \leq 0$. But from Lemma 2, it follows that this is satisfied for all $\psi \geq 0$.■

Inspection of figure 1 shows that the conditions $p^{0*} \geq p^{1*}$ and $x^{0*} \leq x^{1*}$ are satisfied for all possible $\psi \geq 0$. Thus, irrespective of the disutility arising from variation, the purchaser is able to implement the second-best solution by setting the uniform payment rate and uniform standard at $\{p^* = p^{1*}(\psi), \underline{x}^* = x^{0*}(\psi)\}$.

Case 2: Heterogeneity in δ and volume and quality are complements.

Proposition 3: The second-best is unattainable in either of two cases: (i) The disutility arising from variation ψ is sufficiently high such that $\psi > \tilde{\psi}$. Then, a standard $\underline{x}^* = \lambda x^{0*}(\psi) + (1-\lambda)x^{1*}(\psi)$ implements the third-best, and there is no role for a case payment. (ii) The disutility of variation ψ is sufficiently low such that $\psi < \psi^*$. Then, a case payment $p^* = \lambda p^{0*}(\psi) + (1-\lambda)p^{1*}(\psi)$ implements the third best, and there is no role for a standard.

Proof: The second-best obtains if and only if $\Delta^{p*}(\psi) \geq 0$ and $\Delta^{x*}(\psi) \leq 0$. From Lemma 3, it follows that $\Delta^{p*}(\psi) \geq 0$ is violated for $\psi < \psi^*$ and $\Delta^{x*}(\psi) \leq 0$ is violated for $\psi > \tilde{\psi}$. The use of the instruments then follows from Propositions 1.2 and 1.3.■

Inspection of figure 2 shows that the condition $p^{0*} \geq p^{1*}$ and $x^{0*} \leq x^{1*}$ are not satisfied simultaneously if either $\psi < \psi^*$ or $\psi > \tilde{\psi}$. If the disutility of variation is low, it would be efficient for the purchaser to implement an allocation with significant variation at which the productive (unproductive) hospital chooses a high (low) case load. However, this implies that the productive hospital 1 receives a greater case payment, $p^{1*} > p^{0*}$. We have argued above that such an allocation cannot be implemented with a uniform rate-cum-standard. At a

uniform rate $p = p^{1*}$, the unproductive hospital would over-produce and a minimum standard cannot mitigate this situation as it would not bind. The best the purchaser can hope for is a third-best allocation with $p^* = \lambda p^{0*}(\psi) + (1-\lambda)p^{1*}(\psi)$.

If, in contrast, the disutility arising from variation is significant, the purchaser would like to raise the volume of the unproductive hospital 0 over and above the one for hospital 1. However, such an allocation is not feasible, as a uniform standard would have to bind for both hospitals and a second-best is again not attained. The purchaser can achieve merely a third-best with a standard set at $\underline{x}^* = \lambda x^{0*}(\psi) + (1-\lambda)x^{1*}(\psi)$.

Case 3: Heterogeneity in δ and volume and quality are substitutes.

Proposition 4: The second-best is unattainable in either of two cases: (i) The disutility arising from variation ψ is sufficiently low such that $\psi < \tilde{\psi}$. Then, a standard $\underline{x}^* = \lambda x^{0*}(\psi) + (1-\lambda)x^{1*}(\psi)$ implements the third best, and there is no role for a payment rate. (ii) $\psi^* < \overline{\psi}$ and the disutility of variation ψ is sufficiently high such that $\psi > \psi^*$. Then, a case payment $p^* = \lambda p^{0*}(\psi) + (1-\lambda)p^{1*}(\psi)$ implements the third-best, and there is no role for a standard.

Proof: The second-best obtains if and only if $\Delta^{p*}(\psi) \geq 0$ and $\Delta^{x*}(\psi) \leq 0$. From Lemma 5, it follows that $\Delta^{x*}(\psi) \leq 0$ is violated for $\psi < \tilde{\psi}$. $\Delta^{p*}(\psi) \geq 0$ is violated if and only if $\psi^* < \min\{\overline{\psi}, \psi\}$. The use of the instruments then follows from Propositions 1.2 and 1.3.■

Inspection of figure 3, which is drawn for the case $\psi^* < \overline{\psi}$, shows that the condition $x^{0*} \leq x^{1*}$ is violated for $\psi < \tilde{\psi}$, whereas $p^{0*} \geq p^{1*}$ fails to hold for $\psi > \psi^*$. If the disutility of variation is low, it would be efficient for the purchaser to implement an allocation at which the productive hospital 0 chooses a higher volume. We have argued that such an allocation cannot be implemented with a uniform case-rate-cum-standard. Indeed, the best the purchaser can do is to set a uniform standard at $\underline{x}^* = \lambda x^{0*}(\psi) + (1-\lambda)x^{1*}(\psi)$. Recall from Lemma 5 that

$\psi^* < \overline{\psi}$ if quality is important in generating population health. In this case, attainment of the second-best is ruled out if the disutility of variation is $\psi > \psi^*$. Here, the purchaser can only attain a third-best allocation by setting $p^* = \lambda p^{0*}(\psi) + (1-\lambda)p^{1*}(\psi)$.

6 Conclusions

Regulators advocate imposing uniform standards in order to reduce service variations across public health systems. We have argued that such policies may be counterproductive in the presence of factor heterogeneity. We extend the model developed by Chalkley and Malcomson by introducing heterogeneity in factor productivity and assuming this to be specific to volume and quality.

When the purchaser is able to make differentiated payments, a second-best solution is always attainable. The differential in the case payment between hospitals depends on the source of factor heterogeneity, on the degree of complementarity between volume and quality, and on the strength of the purchaser's preference to eliminate variation.

In many practical situations the purchaser may be restricted to a uniform case-payment and/or a uniform standard on volume. The choice of instrument will depend on specific contextual characteristics.

- If there is factor heterogeneity effecting the production of volume, the second-best solution can be obtained by the combined use of a uniform case-payment and standard.
- If factor heterogeneity effects the production of quality, and volume and quality are complements, a second-best solution is unattainable if concern for eliminating variation is either very high or very low. If there is limited concern for eliminating variation, the purchaser would be able to achieve a third-best using the co-payment, recognising that the unproductive hospital would over-produce. If there is a strong concern for eliminating variation, the purchaser should set a standard on volume to achieve a third-best solution.
- When volume and quality are substitutes the circumstances under which each instrument is to be preferred are the reverse of when the outputs are complements. If volume and quality are substitutes, a uniform standard should be chosen if the concern to eliminate

variation is low, while a uniform case-payment is preferable if concern is high.

Footnote

1 Chalkley and Malcomson consider effort in cost-reductions, e, as a further non-contractible instrument, where $c(x,q,e)$ and $v(x,q,e)$. Contracts may include cost-reimbursement, such that $B[x,c(x,q,e)]$, with $B_c \geq 0$. It is easily checked that for $B_c = 0$, i.e.no cost-reimbursement, the hospital chooses the optimal level of effort $e^*(x,q)$ for each volume-quality pairing. This corresponds to our set-up in which we have not explicitly modelled effort provision. As the purchaser cannot attain the first-best by adjusting the case-payment p^j alone, she will usually distort B_c away from zero in order to improve the allocation. We ignore this case for the sake of tractability and merely note that our main points carry over to the more general setting with effort and $B_c \geq 0$.

References

Chalkley, M. & J.M. Malcomson. 1998. Contracting for health services when patient demand does not reflect quality. *Journal of Health Economics*, 17: 1-19.

Chassin, M. 2002. Achieving and sustaining improved quality: lessons for New York State and cardiac surgery. *Health Affairs*, 21(4): 40-51.

Dopuch, N. & M. Gupta. 1997. Estimation of benchmark performance standards: an application to public school expenditures. *Journal of Accounting and Economics*, 23: 147-61.

Dranove, D., D. Kessler, M. McClellan, & M. Satterthwaite. 2002. Is more information better? The effoect of 'report cards' on health care providers, *NBER working paper 8697*. Cambridge MA: National Bureau of Economic Research.

Greene, W.H. 1993. The econometric approach to efficiency analysis.In Fried, H.O., C.A.K. Lovell, & S.S. Schmidt, editors, *The measurement of productive efficiency: Techniques and applications.* New York: Oxford University Press.

Hollingsworth, B., P.J. Dawson, & N. Maniadakis. 1999. Efficiency measurement of health

care: a review of non-parametric methods and applications. *Health Care Management Science*, 2: 161-72.

Milgrom, P. & J. Roberts. 1990. The efficiency of equity in organizational decion-processes. *American Economic Review*, 80(2): 154-59.

NHS Executive. 1997. The New NHS: modern, dependable. Leeds: NHS Executive.

Reilly, T. & G. Meyer. 2002. Providing performance information for consumers: experience from the United States.In Smith, P., editor, *Measuring Up: improving health systems performance in OECD countries*. Paris: OECD.

Riley, J.G. 2001. Silver signals: twenty-five years of screening and signalling. *Journal of Economic Literature*, 39: 432-78.

Rosen, S. 1981. The economics of superstars. *American Economic Review*, 71: 845-58.

Appendix

Lemma 1. (i) $\operatorname{sgn}\Delta_{\psi}^{p^*} = \operatorname{sgn}\Delta_{\psi}^{x^*} = -\operatorname{sgn}\Delta^{b^*}(0)$ for all $\psi \geq 0$, and (ii) $\lim_{\psi\to\infty}\Delta_{\psi}^{p^*} = \lim_{\psi\to\infty}\Delta_{\psi}^{x^*} = \lim_{\psi\to\infty}\Delta^{b^*}(\psi) = 0$.

Proof of Lemma 1:

The Hessian for the system (6a) and (6b) is given by $Z = \Pi^0_{p^0}\Pi^1_{p^1} - \Pi^0_{p^1}\Pi^1_{p^0}$. Defining $\vartheta^j =: b_x\left(x^{j^*}, q^{j^*}\right) + b_q\left(x^{j^*}, q^{j^*}\right)\xi^j$, we obtain from (6c) and (6d)

$$\Pi^0_{p^0} = \frac{\left\langle -\psi(1-\lambda)\left(\vartheta^0\right)^2 + \left[1-\beta-\psi(1-\lambda)\Delta^{b^*}\right]\frac{d\vartheta^0}{dx}\right\rangle \hat{x}^0_p}{1+\alpha} < 0 \tag{A1a}$$

$$\Pi^0_{p^1} = \frac{\psi(1-\lambda)\vartheta^0\vartheta^1\hat{x}^1_p}{1+\alpha} > 0 \tag{A1b}$$

$$\Pi^1_{p^1} = \frac{\left\langle -\psi\lambda\left(\vartheta^1\right)^2 + \left[1-\beta+\psi\lambda\Delta^{b^*}\right]\frac{d\vartheta^1}{dx}\right\rangle \hat{x}^1_p}{1+\alpha} < 0 \tag{A1c}$$

$$\Pi^1_{p^0} = \frac{\psi\lambda\vartheta^0\vartheta^1\hat{x}^0_p}{1+\alpha} > 0 \tag{A1d}$$

where

$$\frac{d\vartheta^j}{dx} = b_{xx} + 2b_{xq}\xi^j + b_{qq}\left(\xi^j\right)^2 + b_q\frac{d\xi^j}{dx} < 0.$$

The inequalities follow under assumptions (8a) and (8b), under the assumption that $\frac{d\xi^j}{dx} \approx 0$, and under the assumption $b_{xx} \leq \frac{b^2_{xq}}{b_{qq}}$, with the latter two assumptions implying $\frac{d\upsilon^j}{dx} \leq \frac{1}{b_{qq}H^2_{qq}}\left(b^2_{xq}H^2_{qq} - b^2_{qq}H^2_{xq}\right)^2 < 0$.

Using (A1a)-(A1d) one can then easily verify that $Z > 0$. Comparative static analysis is then feasible and yields

$$p^{0^*}_{\psi} := \frac{dp^{0^*}}{d\psi} = \frac{\Pi^0_{\psi} - \left(\Pi^0_{\psi}\Pi^1_{p^1} - \Pi^1_{\psi}\Pi^0_{p^1}\right)}{Z},$$

$$p^{1^*}_{\psi} := \frac{dp^{1^*}}{d\psi} = \frac{\Pi^1_{\psi} - \left(\Pi^1_{\psi}\Pi^0_{p^0} - \Pi^0_{\psi}\Pi^1_{p^0}\right)}{Z},$$

where

$$\Pi_{\psi}^{0} = \frac{-(1-\lambda)\Delta^{b*}\vartheta^{0}}{1+\alpha}, \qquad \Pi_{\psi}^{1} = \frac{\lambda\Delta^{b*}\vartheta^{1}}{1+\alpha}.$$

Observing $\vartheta^{j} > 0;\ j = 0,1$, and (A1a)-(A1d) one can verify that

$$\operatorname{sgn} p_{\psi}^{0*} = -\operatorname{sgn} p_{\psi}^{1*} = -\operatorname{sgn}\Delta^{b*} \tag{A10}$$

But then, using definition (9d)

$$\operatorname{sgn}\Delta_{\psi}^{p*} = \operatorname{sgn}\left(p_{\psi}^{0*} - p_{\psi}^{1*}\right) = -\operatorname{sgn}\Delta^{b*} \tag{A10b}$$

From $x_{\psi}^{j*} = \hat{x}_{p}^{j} p_{\psi}^{j*}$ for $j = 0,1$ and under observation of (4a) it follows that $\operatorname{sgn} x_{\psi}^{j*} = \operatorname{sgn} p_{\psi}^{j*}$ and, thus,

$$\operatorname{sgn} x_{\psi}^{0*} = -\operatorname{sgn} x_{\psi}^{1*} = -\operatorname{sgn}\Delta^{b*} \tag{A11}$$

Using definition (9e),

$$\operatorname{sgn}\Delta_{\psi}^{x*} = \operatorname{sgn}\left(x_{\psi}^{0*} - x_{\psi}^{1*}\right) = -\operatorname{sgn}\Delta^{b*} \tag{A11b}$$

Finally, using definition (9f), we find that $\Delta_{\psi}^{b*} = \vartheta^{0} x_{\psi}^{0*} - \vartheta^{1} x_{\psi}^{1*} \geq 0 \Leftrightarrow \Delta^{b*}(\psi) \leq 0$, where the last equality follows from (A11). This implies $\lim_{\psi\to\infty}\Delta^{b*}(\psi) = \lim_{\psi\to\infty}\Delta_{\psi}^{b*}(\psi) = 0$ and, likewise, $\lim_{\psi\to\infty}\Delta_{\psi}^{p*} = \lim_{\psi\to\infty}\Delta_{\psi}^{x*} = 0$, which proves part (ii).

We conclude the proof by showing that $\operatorname{sgn}\Delta^{b*}(\psi) = \operatorname{sgn}\Delta^{b*}(0)$ for all $\psi \geq 0$, which together with (A10b) and (A11b) implies part (i) of the Lemma. Suppose there exists a finite $\overline{\psi} \in [0,\infty[$ such that $\Delta^{b*}(\overline{\psi}) = 0$. Then, it follows from $\Delta^{b*} = 0 \Leftrightarrow x_{\psi}^{j*} = 0 \Leftrightarrow \Delta_{\psi}^{b*} = 0$ that $\Delta^{b*}(\psi) = \Delta^{b*}(\overline{\psi}) = 0$ for all $\psi \geq 0$. But this is (trivially) satisfied if and only if $\Delta^{b*}(0) = 0$. By contradiction, if $\Delta^{b*}(0) \neq 0$ there exists no $\overline{\psi} \in [0,\infty[$ such that $\Delta^{b*}(\overline{\psi}) = 0$. But then, $\operatorname{sgn}\Delta^{b*}(\psi) = \operatorname{sgn}\Delta^{b*}(0)$ for all $\psi \geq 0$, which completes the proof.■

Lemma 2: The optimal payment rates and volumes depend on ψ as follows.

$\Delta^{p*}(\psi) > 0; \Delta_{\psi}^{p*}(\psi) > 0; \Delta^{x*}(\psi) < 0; \Delta_{\psi}^{x*}(\psi) > 0$ for all $\psi \geq 0$.

Proof of Lemma 2: We prove in turn

(a) $\gamma^{0} > \gamma^{1} \Rightarrow \Delta^{p*}(0) > 0 > \Delta^{x*}(0)$;

(b) $\Delta^{b*}(\psi) \leq 0 \Leftrightarrow \Delta^{x*}(\psi) \leq 0$

Suppose for the moment that (a) and (b) are true. Together, they imply $\gamma^{0} > \gamma^{1} \Rightarrow \Delta^{b*}(0) < 0$

so that from part (i) of Lemma 1 $\operatorname{sgn}\Delta_{\psi}^{p^*}(\psi)=\operatorname{sgn}\Delta_{\psi}^{x^*}(\psi)=-\operatorname{sgn}\Delta^{b^*}(0)=1$ for all $\psi\geq 0$ as indicated in the Lemma. But then, it follows immediately from (a) that $\Delta^{p^*}(\psi)>0$ for all $\psi\geq 0$.

From part (ii) of Lemma 1, $\lim_{\psi\to\infty}\Delta^{b^*}(\psi)=0$ which together with (b) above implies $\lim_{\psi\to\infty}\Delta^{x^*}(\psi)=0$ and, thus, together with (a), $\Delta^{x^*}(\psi)\leq 0$ for all $\psi\geq 0$.

(a) Consider the comparative static properties

$$p_{\gamma}^{0*}:=\frac{dp^{0*}}{d\gamma^{0}}=\frac{\Pi_{\gamma^0}^{0}-\left(\Pi_{\gamma^0}^{0}\Pi_{p^1}^{1}-\Pi_{\gamma^0}^{1}\Pi_{p^1}^{0}\right)}{Z} \tag{A2a}$$

$$p_{\gamma}^{1*}:=\frac{dp^{1*}}{d\gamma^{0}}=\frac{\Pi_{\gamma^0}^{1}-\left(\Pi_{\gamma^0}^{1}\Pi_{p^0}^{0}-\Pi_{\gamma^0}^{0}\Pi_{p^0}^{1}\right)}{Z} \tag{A2b}$$

where $\Pi_{\gamma^0}^{j}=\Pi_{p^0}^{j}\frac{\hat{x}_{\gamma}^{0}}{\hat{x}_{p}^{0}}$; $j=0,1$. Using (A1a) and (A1d) and observing $\hat{x}_{\gamma}^{0}<0$, it is then readily verified that $p_{\gamma}^{0*}>0$ and $p_{\gamma}^{1*}<0$. But then, $\Delta_{\gamma}^{p^*}=p_{\gamma}^{0*}-p_{\gamma}^{1*}>0$ and, thus, $\gamma^0>\gamma^1\Rightarrow\Delta^{p^*}(\psi)>0$ for all $\psi\geq 0$.

Using (A2a)-(A2b) together with (A1a)-(A1d) one can show that

$$\begin{aligned}\Delta_{\gamma}^{x^*}&=\hat{x}_{\gamma}^{0}+\hat{x}_{p}^{0}p_{\gamma}^{0*}-\hat{x}_{p}^{1}p_{\gamma}^{1*}\\&=\frac{\hat{x}_{\gamma}^{0}\left\langle 1-\hat{x}_{p}^{1}\left\{\left[1-\beta+\psi\lambda\left[b\left(x^{o*},q^{0*}\right)-b\left(x^{1*},q^{1*}\right)\right]\right]\frac{d\vartheta^1}{dx}+\psi\lambda\left(\vartheta^0-\vartheta^1\right)\vartheta^0\right\}\right\rangle}{Z}.\end{aligned}$$

Evaluating the RHS expression at $\psi=0$ gives $\Delta_{\gamma}^{x^*}\big|_{\psi=0}=\frac{\hat{x}_{\gamma}^{0}\left\langle 1-\hat{x}_{p}^{1}(1-\beta)\right\rangle}{Z}<0$ implying that $\gamma^0>\gamma^1\Rightarrow\Delta^{x^*}(0)<0$. This proves (a).

(b) Observing from (3b) that $\left\{\delta^0=\delta^1\right\}\Rightarrow\hat{q}^0(x)=\hat{q}^1(x)$, it follows that $x^{0*}=x^{1*}\Leftrightarrow\hat{q}^0\left(x^{0*}\right)=\hat{q}^1\left(x^{1*}\right)$ and, thus, $b\left(x^{0*},q^{0*}\right)=b\left(x^{1*},q^{1*}\right)\Leftrightarrow x^{0*}=x^{1*}$. But then, it follows from assumption (8a) that $\Delta^{b^*}\leq 0\Leftrightarrow b\left(x^{0*},q^{0*}\right)=b\left[x^{0*},\hat{q}^0\left(x^{0*}\right)\right]\leq b\left[x^{1*},\hat{q}^1\left(x^{1*}\right)\right]=b\left(x^{1*},q^{1*}\right)\Leftrightarrow x^{0*}\leq x^{1*}\Leftrightarrow\Delta^{x^*}\leq 0$.■

Lemma 3: (i) There exists $k^+>0$ such that $H_{xq}^{0}\in[0,k^+[\Leftrightarrow\infty>\tilde{\psi}\geq\psi^*>0$. (ii) The optimal payment rates and volumes then depend on ψ as follows.

$\psi\in[0,\psi^*[\Leftrightarrow\left\{\Delta^{p^*}(\psi)<0;\Delta_{\psi}^{p^*}(\psi)>0;\Delta^{x^*}(\psi)<0;\Delta_{\psi}^{x^*}(\psi)>0\right\}$;

$\psi \in [\psi^*, \tilde{\psi}] \Leftrightarrow \{\Delta^{p^*}(\psi) \geq 0; \Delta^{p^*}_{\psi}(\psi) > 0; \Delta^{x^*}(\psi) \leq 0; \Delta^{x^*}_{\psi}(\psi) > 0\};$

$\psi > \tilde{\psi} \Leftrightarrow \{\Delta^{p^*}(\psi) > 0; \Delta^{p^*}_{\psi}(\psi) > 0; \Delta^{x^*}(\psi) > 0; \Delta^{x^*}_{\psi}(\psi) > 0\};$

Proof of Lemma 3: We prove in turn

(a) $\{\delta^0 > \delta^1; H^0_{xq} \in [0, k^+[\} \Rightarrow 0 > \{\Delta^{p^*}(0); \Delta^{x^*}(0)\};$

(b) $\Delta^{x^*}(\psi) \leq 0 \Rightarrow \Delta^{b^*}(\psi) < 0$

(c) $\infty > \tilde{\psi} \geq \psi^* > 0$, which proves part (i) of the Lemma.

Suppose that (a)-(c) hold. Together, (a) and (b) imply $\{\delta^0 > \delta^1; H^0_{xq} \in [0, k^+[\} \Rightarrow \Delta^{b^*}(0) < 0$ so that from part (i) of Lemma 1 $\operatorname{sgn} \Delta^{p^*}_{\psi}(\psi) = \operatorname{sgn} \Delta^{x^*}_{\psi}(\psi) = -\operatorname{sgn} \Delta^{b^*}(0) = 1$ for all $\psi \geq 0$.

But then, it follows immediately from (c) together with the definitions (9a) and (9c) that $\psi \in [0, \psi^*[\Leftrightarrow \{\Delta^{p^*}(\psi) < 0; \Delta^{x^*}(\psi) < 0\}$; $\psi \in [\psi^*, \tilde{\psi}] \Leftrightarrow \{\Delta^{p^*}(\psi) \geq 0; \Delta^{x^*}(\psi) \leq 0\}$; and $\psi > \tilde{\psi} \Leftrightarrow \{\Delta^{p^*}(\psi) > 0; \Delta^{x^*}(\psi) > 0\}$ as stated in the Lemma.

(a) Consider the comparative static properties

$$\left(p^{0*}_{\delta} := \frac{dp^{0*}}{d\delta^0} \right)_{\psi=0} = \frac{\Pi^0_{\delta^0} - \left(\Pi^0_{\delta^0} \Pi^1_{p^1} - \Pi^1_{\delta^0} \Pi^0_{p^1} \right)}{Z} \Big|_{\psi=0} \tag{A3a}$$

$$\left(p^{1*}_{\delta} := \frac{dp^{1*}}{d\delta^0} \right)_{\psi=0} = \frac{\Pi^1_{\delta^0} - \left(\Pi^1_{\delta^0} \Pi^0_{p^0} - \Pi^0_{\delta^0} \Pi^1_{p^0} \right)}{Z} \Big|_{\psi=0} \tag{A3b}$$

where

$$\Pi^j_{p^j} \Big|_{\psi=0} = \left(\frac{1-\beta}{1+\alpha} \right) \frac{d\vartheta^j}{dx} \hat{x}^j_p < 0; \quad j = 0,1 \tag{A31}$$

and $\Pi^0_{p^1} \Big|_{\psi=0} = \Pi^1_{p^0} \Big|_{\psi=0} = 0$ from (A1a)-(A1d); where

$$\Pi^0_{\delta^0} \Big|_{\psi=0} = \frac{(1-\beta)\Omega \hat{q}^0_{\delta}}{1+\alpha} \tag{A1a'}$$

with

$$\Omega := \left(-b_{xx} + b_{xq} \frac{H^0_{xq}}{H^0_{qq}} \right) \frac{H^0_{xq}}{H^0_{xx}} + b_{xq} - b_{qq} \frac{H^0_{xq}}{H^0_{qq}} \tag{A32}$$

and where $\Pi^1_{\delta^0} \Big|_{\psi=0} = 0$. Inserting these into (A3a) and (A3b), it is easy to check that

$$\Delta^{p^*}_{\delta} \Big|_{\psi=0} = p^{0*}_{\delta} \Big|_{\psi=0} - p^{1*}_{\delta} \Big|_{\psi=0} = p^{0*}_{\delta} \Big|_{\psi=0} = \frac{\Pi^0_{\delta^0} \left(1 - \Pi^1_{p^1} \right)}{Z} \Big|_{\psi=0} \tag{A33}$$

Since $\frac{\left(1 - \Pi^1_{p^1} \right)}{Z} \Big|_{\psi=0} > 0$, it follows that

$$\Delta_{\delta}^{p^*}\Big|_{\psi=0} < 0 \Leftrightarrow p_{\delta}^{0*}\Big|_{\psi=0} < 0 \Leftrightarrow \Pi_{\delta^0}^{0}\Big|_{\psi=0} < 0 \Leftrightarrow \Omega > 0 \qquad \text{(A4)}$$

For $b_{xx} \le \frac{b_{xq}^2}{b_{qq}}$ and $H_{xx}^0 \le \frac{\left(H_{xq}^0\right)^2}{H_{qq}^0}$ we obtain

$$\frac{d\Omega}{dH_{xq}^0} = \left(-b_{xx} + 2b_{xq}\frac{H_{xq}^0}{H_{qq}^0}\right)\left(\frac{1}{H_{xx}^0}\right) - b_{qq}\left(\frac{1}{H_{qq}^0}\right) < \frac{-\left[\left(b_{xq}H_{qq}^0\right)^2 - 2b_{xq}H_{qq}^0 b_{qq}H_{xq}^0 + \left(b_{qq}H_{xq}^0\right)^2\right]}{b_{qq}H_{qq}^0} < 0 \qquad \text{(A5)}$$

Observing $\Omega\Big|_{H_{xq}^0=0} = b_{xq} > 0$ it then follows that there exists a $k^+ > 0$ such that $\Omega > 0 \Leftrightarrow H_{xq}^0 < k^+$. But then, from (A4), $\left\{\delta^0 > \delta^1; H_{xq}^0 \in [0, k^+[\right\} \Rightarrow \Delta^{p^*}(0) < 0$.

Furthermore, $\Delta_{\delta}^{x^*}\Big|_{\psi=0} = \hat{x}_{\delta}^0 + \hat{x}_p^0 p_{\delta}^{0*}\Big|_{\psi=0} - \hat{x}_p^1 p_{\delta}^{1*}\Big|_{\psi=0} = \hat{x}_{\delta}^0 + \hat{x}_p^0 p_{\delta}^{0*}\Big|_{\psi=0}$. Recalling $H_{xq}^0 \ge 0 \Leftrightarrow \hat{x}_{\delta}^0 \le 0$, it follows from the condition $\Delta_{\delta}^{p^*}\Big|_{\psi=0} < 0 \Leftrightarrow p_{\delta}^{0*}\Big|_{\psi=0} < 0$ that $\Delta_{\delta}^{p^*}\Big|_{\psi=0} \le 0 \Rightarrow \Delta_{\delta}^{x^*}\Big|_{\psi=0} < 0$ and, thus, $\left\{\delta^0 > \delta^1; H_{xq}^0 \in [0, k^+[\right\} \Rightarrow \Delta^{x^*}(0) < 0$. This proves (a).

(b) Observing from (3b) that $\left\{\delta^0 > \delta^1\right\} \Rightarrow \hat{q}^0(x) < \hat{q}^1(x)$, it follows that $x^{0*} = x^{1*} \Leftrightarrow \hat{q}^0\left(x^{0*}\right) < \hat{q}^1\left(x^{1*}\right)$. Under assumption (8a) it follows that $\Delta^{x^*} \le 0 \Leftrightarrow x^{0*} \le x^{1*} \Rightarrow b\left(x^{0*}, q^{0*}\right) < b\left(x^{1*}, q^{1*}\right) \Leftrightarrow \Delta^{b^*} < 0$.

(c) Recall $\Delta^{x^*} \le 0 \Rightarrow \Delta^{b^*} < 0$ from part (b) and $\Delta^{x^*}(0) < 0$ and $\Delta_{\psi}^{x^*}(0) > 0$ from part (a). Since $\lim_{\psi \to \infty} \Delta^{b^*}(\psi) = 0$, it then follows from the monotony of $\Delta^{b^*}(\psi)$ that $\infty > \tilde{\psi} > 0$. Furthermore, recall $\left\{\delta^0 > \delta^1; H_{xq}^0 \in [0, k^+[\right\} \Rightarrow \hat{x}^0(p) \le \hat{x}^1(p)$. But then, $\Delta^{x^*}(\psi) = 0 \Rightarrow \Delta^{p^*}(\psi) \ge 0$. Observing $\left\{\Delta_{\psi}^{p^*}(\psi), \Delta_{\psi}^{x^*}(\psi)\right\} > 0$ together with $\Delta^{p^*}(0) < 0$ this implies $\tilde{\psi} \ge \psi^* > 0$. This completes the proof of part (c).■

Lemma 4: Let $\left\{\gamma^0 = \gamma^1; \delta^0 < \delta^1; H_{xq}^0 \le 0\right\}$. (i) There exists $k^- < 0$ such that $H_{xq}^0 \in]k^-, 0] \Leftrightarrow \infty > \psi^* \ge \tilde{\psi} > 0$. (ii) The second-best payment rates and volumes then depend on ψ as follows.

$\psi \in [0, \tilde{\psi}[\Leftrightarrow \left\{\Delta^{p^*}(\psi) > 0; \Delta_{\psi}^{p^*}(\psi) < 0; \Delta^{x^*}(\psi) > 0; \Delta_{\psi}^{x^*}(\psi) < 0\right\}$;

$\psi \in [\tilde{\psi}, \psi^*] \Leftrightarrow \left\{\Delta^{p^*}(\psi) \ge 0; \Delta_{\psi}^{p^*}(\psi) < 0; \Delta^{x^*}(\psi) \le 0; \Delta_{\psi}^{x^*}(\psi) < 0\right\}$;

$\psi > \psi^* \Leftrightarrow \left\{\Delta^{p^*}(\psi) < 0; \Delta_{\psi}^{p^*}(\psi) < 0; \Delta^{x^*}(\psi) < 0; \Delta_{\psi}^{x^*}(\psi) < 0\right\}$;

Proof of Lemma 4: We prove in turn

(a) $\{\delta^0 < \delta^1; H^0_{xq} \le 0\} \Rightarrow \{\Delta^{p^*}(0); \Delta^{x^*}(0)\} > 0$;

(b) $\Delta^{x^*}(\psi) \ge 0 \Rightarrow \Delta^{b^*}(\psi) > 0$;

(c) $H^0_{xq} \ge k^- \Rightarrow \infty > \psi^* \ge \tilde{\psi} > 0$, which proves part (i) of the Lemma.

Suppose that (a)-(c) hold. Together, (a) and (b) imply $\{\delta^0 < \delta^1; H^0_{xq} \in]k^-, 0]\} \Rightarrow \Delta^{b^*}(0) > 0$ so that from part (i) of Lemma 1 $\operatorname{sgn} \Delta^{p^*}_{\psi}(\psi) = \operatorname{sgn} \Delta^{x^*}_{\psi}(\psi) = -\operatorname{sgn} \Delta^{b^*}(0) = -1$ for all $\psi \ge 0$.

But then, it follows immediately from (c) together with the definitions (9a) and (9c) that $\psi \in [0, \tilde{\psi}[\Leftrightarrow \{\Delta^{p^*}(\psi) > 0; \Delta^{x^*}(\psi) > 0\}$; $\psi \in [\tilde{\psi}, \psi^*] \Leftrightarrow \{\Delta^{p^*}(\psi) \le 0; \Delta^{x^*}(\psi) \ge 0\}$; and $\psi > \psi^* \Leftrightarrow \{\Delta^{p^*}(\psi) < 0; \Delta^{x^*}(\psi) < 0\}$ as stated in the Lemma.

(a) Recall from (A4) that $\Delta^{p^*}_{\delta}\big|_{\psi=0} < 0 \Leftrightarrow p^{0^*}_{\delta}\big|_{\psi=0} < 0 \Leftrightarrow \Omega > 0$, where Ω as defined in (A52). Observing $\Omega\Big|_{H^0_{xq}=0} = b_{xq} > 0$ and $\frac{d\Omega}{dH^0_{xq}} < 0$, as from (A5), it follows that $H^0_{xq} \le 0 \Rightarrow \Omega > 0$ and by implication $H^0_{xq} \le 0 \Rightarrow \Delta^{p^*}_{\delta}\big|_{\psi=0} < 0$. But then, from (A4), $\{\delta^0 < \delta^1; H^0_{xq} \le 0\} \Rightarrow \Delta^{p^*}(0) > 0$.

Since $H^0_{xq} \le 0 \Rightarrow \{p^{0^*}_{\delta}\big|_{\psi=0} < 0; \hat{x}^0_{\delta} > 0\}$ the sign of $\Delta^{x^*}_{\delta}\big|_{\psi=0} = \hat{x}^0_{\delta} + \hat{x}^0_{p} p^{0^*}_{\delta}\big|_{\psi=0}$ is not immediate. However, inserting successively from (4f) and from the RHS of (A33); then from (A31) and (A1a'); and finally from (A32); while observing $\frac{d\vartheta^j}{dx} = b_{xx} - 2b_{xq}\frac{H^j_{xq}}{H^j_{qq}} + b_{qq}\left(\frac{H^j_{xq}}{H^j_{qq}}\right)^2$ one can verify after tedious but straightforward calculations that $H^0_{xq} \le 0 \Rightarrow \Delta^{x^*}_{\delta}\big|_{\psi=0} < 0$. It follows that $\{\delta^0 < \delta^1; H^0_{xq} \le 0\} \Rightarrow \Delta^{x^*}(0) > 0$, which completes the proof of part (a).

(b) Observing from (3b) that $\{\delta^0 < \delta^1\} \Rightarrow \hat{q}^0(x) > \hat{q}^1(x)$, it follows that $x^{0^*} = x^{1^*} \Leftrightarrow \hat{q}^0(x^{0^*}) > \hat{q}^1(x^{1^*})$. Under assumption (8a) it follows that $\Delta^{x^*} \ge 0 \Leftrightarrow x^{0^*} \ge x^{1^*} \Rightarrow b(x^{0^*}, q^{0^*}) > b(x^{1^*}, q^{1^*}) \Leftrightarrow \Delta^{b^*} > 0$.

Recall $\{\delta^0 < \delta^1; H^0_{xq} \le 0\} \Rightarrow \hat{x}^0(p) \le \hat{x}^1(p)$ such that $\Delta^{p^*}(\psi) = 0 \Rightarrow \Delta^{x^*}(\psi) \le 0$. Observing $\{\Delta^{p^*}_{\psi}(\psi), \Delta^{p^*}_{\psi}(\psi)\} < 0$ together with $\Delta^{x^*}(0) > 0$ this implies $\psi^* \ge \tilde{\psi} > 0$. Furthermore, since $\Delta^{x^*} \ge 0 \Rightarrow \Delta^{b^*} > 0$ from part (b) it follows from $\lim_{\psi \to \infty} \Delta^{b^*}(\psi) = 0$ and from the monotony of $\Delta^{b^*}(\psi)$ that $\tilde{\psi} < \infty$.

Now, observe $H^0_{xq} = 0 \Rightarrow \hat{x}^0(p) = \hat{x}^1(p)$. But then, $\Delta^{p^*}(\psi) = 0 \Leftrightarrow \Delta^{x^*}(\psi) = 0$ so that under the

definitions (9a) and (9c), $H_{xq}^{0}=0\Rightarrow\tilde{\psi}=\psi^{*}<\infty$. It is easily checked that $\frac{d\left(\tilde{\psi}-\psi^{*}\right)}{dH_{xq}^{0}}<0$ implying that there exists a $k^{-}<0$ such that $H_{xq}^{0}\in\left]k^{-},0\right]\Leftrightarrow\tilde{\psi}\leq\psi^{*}<\infty$. This completes the proof of part (c).■

Proof of Proposition 1.1: For $x^{0*}\neq x^{1*}$ the second-best allocation is either given by $\left(p=p^{1*};\underline{x}=x^{0*}\right)$ or by $\left(p=p^{0*};\underline{x}=x^{1*}\right)$. The latter gives rise to a second-best allocation if and only if $\hat{x}^{1}\left(p^{0*}\right)\leq x^{1*}<x^{0*}$. But $\hat{x}^{1}\left(p^{0*}\right)<x^{0*}=\hat{x}^{0}\left(p^{0*}\right)$ is a contradiction. Thus, other than for the trivial case $x^{0*}=x^{1*}$, $\left(p=p^{0*};\underline{x}=x^{1*}\right)$ cannot be a second-best allocation. Consider now $\left(p=p^{1*};\underline{x}=x^{0*}\right)$. This is a second-best allocation if and only if $\hat{x}^{0}\left(p^{1*}\right)\leq x^{0*}\leq x^{1*}$. This yields the conditions and

$\hat{x}^{0}\left(p^{1*}\right)\leq x^{0*}\Leftrightarrow p^{0*}\geq p^{1*}$ (A6a) $x^{0*}\leq x^{1*}$ (A6b)

which are necessary and sufficient for the feasibility of a second-best.■

Proof of Proposition 1.2: $x^{0*}>x^{1*}$ obviously violates a feasibility condition. Note from our assumption that $\hat{x}^{0}(p)\leq\hat{x}^{1}(p)$, it follows that $x^{0*}>x^{1*}\Rightarrow p^{0*}>p^{1*}$. If $x^{0*}>x^{1*}$ a standard set at $\underline{x}=x^{0*}>x^{1*}=\hat{x}^{1}\left(p^{1*}\right)$ binds for both types and is overly restrictive for type 1, forcing over-production. Formally, $R_{\underline{x}}\left[x^{0*};x^{0*}\right]=\overbrace{R_{\underline{x}^{0}}\left[x^{0*}\right]}^{=0}+\overbrace{R_{\underline{x}^{1}}\left[x^{0*}\right]}^{<0}<0$. Thus, it is optimal to set the standard at a level $\underline{x}\in\left]x^{1*};x^{0*}\right[$ such that $R_{\underline{x}}\left[\underline{x};\underline{x}\right]=\overbrace{R_{\underline{x}^{0}}\left[\underline{x}\right]}^{>0}+\overbrace{R_{\underline{x}^{1}}\left[\underline{x}\right]}^{<0}=0$. Using (10a) and (10b) one can verify that this implies $\underline{x}^{*}=\lambda x^{0*}+(1-\lambda)x^{1*}$. Since $\underline{x}^{*}>x^{1*}\geq\hat{x}^{0}\left(p^{1*}\right)$, both hospitals choose their output independent of the payment rate $p=p^{1*}$. Consider now alternative values of the case-rate. Let $\overline{p}:=p\left|\hat{x}_{1}(p)=\underline{x}^{*}\right.$. Obviously, $p<\overline{p}$ has no effect as $\hat{x}^{0}(p)\leq\hat{x}^{1}(p)<\underline{x}^{*}$. From the concavity of $R(\cdot)$ and from $R_{\underline{x}}\left[\underline{x}^{*};\underline{x}^{*}\right]=0$ it follows that $R_{\underline{x}}\left[\max\left\{\hat{x}^{0}(p);\underline{x}^{*}\right\};\hat{x}^{1}(p)\right]<0$ for all $p>\overline{p}$, where $\hat{x}_{1}(p)>\underline{x}^{*}$. Finally, consider $p=\overline{p}$. While this implements $R_{\underline{x}}\left[\underline{x}^{*};\underline{x}^{*}\right]=R_{\underline{x}}\left[\underline{x}^{*};\hat{x}^{1^{*}}(p)\right]=0$ in the presence of a standard, $\hat{x}_{0}(\overline{p})\leq\hat{x}_{1}(\overline{p})=\underline{x}^{*}$ implies $R_{\underline{x}}\left[\hat{x}^{0^{*}}(\overline{p});\hat{x}^{1^{*}}(\overline{p})\right]=R_{\underline{x}}\left[\hat{x}^{0^{*}}(\overline{p});\underline{x}^{*}\right]>0$ in the absence of a

standard. Thus, only a standard can implement the third-best. A case payment is redundant.■

Proof of Proposition 1.3: $p^{0*} < p^{1*}$ implies $x^{0*} < \hat{x}^0\left(p^{1*}\right) \leq x^{1*}$. But then, a standard set at $\underline{x} = x^{0*} < \hat{x}^0\left(p^{1*}\right) \leq x^{1*}$ does not bind for type 0, which over-produces relatively to the first-best. Formally, $R_p\left[\hat{x}^0\left(p^{1*}\right); x^{1*}\right] = \overbrace{R_{p^0}\left[\hat{x}^0\left(p^{1*}\right)\right]}^{<0} + \overbrace{R_{p^1}\left[x^{1*}\right]}^{=0} < 0$. Thus, it is optimal to set the case rate at $p \in \left]p^{1*}; p^{0*}\right[$ such that $R_p\left[\hat{x}^0(p); \hat{x}^1(p)\right] = \overbrace{R_{p^0}\left[\hat{x}^0(p)\right]}^{>0} + \overbrace{R_{p^1}\left[\hat{x}^1(p)\right]}^{<0} = 0$. Using (5c) and (5d) one can verify that this implies $p^* = \lambda p^{0*} + (1-\lambda)p^{1*}$. Since $p^* \geq p^{0*}$ and, thus, $\underline{x}^* = x^{0*} \leq \hat{x}^0\left(p^*\right) \leq \hat{x}^1\left(p^*\right)$, both hospitals choose their output above the standard. Consider now alternative values of the standard. Obviously, any $\underline{x} \leq \hat{x}^0\left(p^*\right)$ has no effect. From the concavity of $R(\cdot)$ and from $R_p\left[\hat{x}^0\left(p^*\right); \hat{x}^1\left(p^*\right)\right] = 0$ it follows that $\operatorname{sgn} R_{\underline{x}}\left[\underline{x}^*; \max\left\{\hat{x}^1(p); \underline{x}^*\right\}\right] = \operatorname{sgn} R_p\left[\underline{x}^*; \max\left\{\hat{x}^1(p); \underline{x}^*\right\}\right] = -1$ for all $\underline{x} > \hat{x}^0\left(p^*\right)$. Finally, consider $\underline{x} = \hat{x}^0\left(p^*\right)$. While this implies $R_p\left[\underline{x}; \hat{x}^1\left(p^*\right)\right] = 0$ in the presence of a case-based payment, $\max\left\{\hat{x}_1(0), \underline{x}\right\} \leq \hat{x}_1\left(p^*\right)$ implies $\operatorname{sgn} R_{\underline{x}}\left[\underline{x}; \max\left\{\hat{x}_1(0), \underline{x}\right\}\right] = \operatorname{sgn} R_p\left[\underline{x}; \max\left\{\hat{x}_1(0), \underline{x}\right\}\right] > 0$ in the absence of a case payment. Thus, only a case payment can implement the third-best. A standard is redundant.■