THE UNIVERSITY of York

Discussion Papers in Economics

No. 2003/05

Hospital Regulation in the Presence of Performance Variation

by

Michael Kuhn, Diane Dawson, Peter C Smith and Andrew Street

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Hospital regulation in the presence of performance variation[+]

by Michael Kuhn[*,a,b], Diane Dawson[b], Peter C. Smith[a,b], Andrew Street[b]

[a]Department of Economics and Related Studies, University of York, England

[b]Centre for Health Economics, University of York, England

21 February 2003

Abstract

Performance standards are advocated for reduction of service variations in the provision of health care. This paper provides an analysis of standard and price regulation of hospitals that differ in the productivity relating to volume and quality of care. A purchaser is concerned with efficiency and reducing variations in outcome across hospitals and sets a fixed budget and a case-rate, quality being non-contractible. A second-best solution entails a provider specific case-rate. Real-world purchasers are often constrained to uniform case-rates but may prescribe minimum standards on activity. We derive, for this setting, conditions under which a second-best can be implemented by combined use of the instruments, and a rule for the use of a standard or a case-rate in third-best situations.

Key words: factor heterogeneity, hospital regulation, performance variation, prices vs. standards.

JEL classification: I11, I18, L51.

[+] We are grateful to Jan-Eirik Askildsen, Richard Cookson and Hugh Gravelle as well as to participants of the 3rd Workshop on Health Economics, Marseille, 2002, and of the winter meeting of the HESG study group, Leeds, 2003, for their helpful comments. The responsibility for any errors lies entirely with us.

[*] Corresponding author: Centre for Health Economics, University of York, Heslington, York YO10 5DD, England; tel: +44 1904 43 3682; e-mail: mk19@york.ac.uk.

1 Introduction

Policy makers in most developed countries are interested in increasing activity rates and improving the quality of health care per unit of expenditure. Evidence on widespread variation in performance appears to suggest that it is possible to meet these objectives without making additional resources available. 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) (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 (Greene, 1993, Dopuch & Gupta, 1997, Hollingsworth, Dawson, & Maniadakis, 1999). In the UK this has led to the introduction of an array of performance standards, laid down in the National Service Frameworks (Department of Health 2000).

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 a belief 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.

Provider heterogeneity has important implications at least on two grounds: Firstly, an efficient regulation of providers requires the use of differentiated policy instruments that reflect the differences in providers. In a real-world context, however, the policy-maker is likely to be constrained to a limited number of instruments so that policies cannot be adjusted for all of the relevant dimensions of heterogeneity. Thus, policy-making takes place in a second-best context.

Secondly, where provider heterogeneity leads to variation in volume or quality of care, an equity issue arises if patients are restricted in their choice of providers. The policy-maker may care about equity among patients as well as efficiency. Variations in the quality and volume

of care, the latter in the form of waiting times and access to care, are a central policy concern in the UK (Department of Health 2000, HM Treasury, 2002). For a tax financed health service, such as the NHS, a dominant concern is that citizens pay taxes according to a uniform national schedule of taxation, and that they should therefore receive uniform services. These equity concerns suggest that some convergence of quality and volume of health care is valued, even at the expense of aggregate 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, Dranove, Kessler, McClellan, & Satterthwaite, 2002). Public reporting initiatives rely on an implicit market-based incentive for providers to improve quality. Yet most areas of health care remain resistant to satisfactory quality measurement, and reviews of the US experience suggest that – even when it is attempted – patients tend to take little notice of quality reports (Reilly & Meyer, 2002). Chalkley & Malcomson (1998) therefore note that there are strong grounds to believe that patient demand will not be effective by itself in maintaining health care quality. 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.

The Chalkley and Malcomson model implies a tailor-made reward schedule for each provider. Health care providers are likely to vary in their productivity 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 two main reasons. Firstly, the transaction costs of developing tailor-made contracts for each provider are likely to be prohibitive. The informational requirements of bespoke contracts with numerous providers are substantial, and give rise to serious opportunities for

gaming on the part of providers. Secondly, 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 an environment – such as health care – in which contracts are seriously incomplete, and there is considerable reliance on provider morale and benevolence, the level of fairness of the system, as perceived by providers may materially affect outcomes (Fehr & Schmidt, 1999). A purchaser may therefore secure considerable benefits from appearing to treat providers even-handedly (Milgrom & Roberts, 1990).

The purpose of this paper is to develop a model of contracting with heterogeneous providers, when the purchaser has available only a uniform price and a uniform volume standard as 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 developed, 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 disadvantage consumers as long as they 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 (Gabsewicz & Thisse, 1979, Shaked & Sutton, 1982). However, in European health care systems, where 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 choose a higher quality service. However, both asymmetric information and limited physical mobility render it less likely that patient choice reflects the quality of hospital services.

3 The Model

We extend a model developed by Chalkley & Malcomson (1998) in order to explore the consequences of variations in factor productivity. A hospital produces health benefit $b(x,q)$ when treating x patients with (single dimensional) quality q. We assume $b(x,q)$ to be increasing and concave in both arguments. Volume is measurable and contractible, quality above a certain minimum (normalised to zero) is not contractible. 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.

Let a hospital's total monetary and effort cost be given by

$$C^j = c(x,q,e) + v(x,q,e) + \gamma^j x + \delta^j q,$$

where monetary cost $c(x,q,e)$ is increasing in x and q and decreasing in cost-reducing effort e. Further, we follow Chalkley & Malcomson (1998) in assuming that $c(x,q,e)$ is differentiable and convex in all arguments and strictly convex in q. A hospital's non-monetary cost is given by $v(x,q,e) + \gamma^j x + \delta^j q$, where $v(x,q,e)$ is strictly increasing, differentiable and strictly convex in all arguments. Hospitals are heterogeneous in the part $\gamma^j x + \delta^j q$ of their non-monetary cost. Here, $\gamma^j x$ represents a volume-related effort, where differences in productivity are reflected in the parameter γ^j. Specifically, the greater γ^j the less productive is hospital j with regard to volume. Likewise, $\delta^j q$ is a quality-related effort cost where a greater δ^j implies a lower level of productivity.

We will subsequently focus on a scenario in which the purchaser is committed to (prospective) case-payments, exposing the hospital to the full monetary cost. In the absence of cost sharing the hospital will always choose effort $\hat{e}(x,q) = \arg\min_e [c(\cdot) + v(\cdot)]$. Furthermore, if the hospital's participation constraint binds rather than its budget constraint then $\hat{e}(x,q)$ corresponds to the first-best effort from a welfare point of view (Chalkley & Malcomson 1998). In order to facilitate notation, let $m(x,q) := c(x,q,\hat{e}(x,q)) + v(x,q,\hat{e}(x,q))$. Using the implicit function theorem, it is easy to check that $m(x,q)$ is increasing and strictly convex in x and q. The sign of the cross partial $m_{xq}(x,q)$ is undetermined (subscripts denote the derivatives with respect to the relevant argument).[1]

Given this cost structure, we can write the objective of hospital j as

[1] Chalkley & Malcomson (1998) also consider contracts that may include cost-reimbursement, such that $B[x, c(x,q,e)]$, with $B_c \geq 0$. Since quality is not verifiable, the purchaser cannot usually attain the first-best by adjusting the case-payment p^j alone. She will, therefore, 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

$$H^j = B^j(x) + \beta b(x,q) - m(x,q) - \gamma^j x - \delta^j q ,$$

where β, $\beta \in [0,1]$, represents the extent to which the hospital seeks to promote population health.

We assume that hospital heterogeneity arises with respect to one and only one of the parameters γ^j or δ^j. For analytical and expositional convenience we do not consider a joint distribution of γ^j and δ^j. Furthermore, as we will see, volume related heterogeneity (γ^j) has policy implications which are quite distinct from those for quality related heterogeneity (δ^j). We assume that hospitals are one of two types (labelled 0 and 1), where $\lambda \in [0,1]$ denotes the share of type 0 in the hospital population.

The purchaser's objective function is given by

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

which contains, for each hospital type, the population health benefit less the cost of production and the shadow cost of public funds, with marginal cost α. The final term captures the purchaser's loss of utility associated with variation in performance. The parameter ψ, $\psi \geq 0$, measures the marginal social cost of variation.

The hospital's problem is $\max_{x^j, q^j} H^j \quad s.t. \quad p^j \quad j = 0,1$ with first-order conditions

$$H_x^j = p^j + \beta b_x(x^j,q^j) - m_x(x^j,q^j) - \gamma^j = 0 \qquad \text{(1a)},$$

$$H_q^j = \beta b_q(x^j,q^j) - m_q(x^j,q^j) - \delta^j = 0 \qquad \text{(1b)}.^2$$

A unique optimum exists if the Hessian is positive $\mathrm{H} = H_{xx}^j H_{qq}^j - \left(H_{xq}^j\right)^2 > 0$, where $H_{xx}^j < 0$ and $H_{qq}^j < 0$. Note that the sign of the cross-derivative

---

more general setting $B_c \geq 0$.

$$H^j_{xq} = \beta b_{xq}\left(x^j, q^j\right) - m_{xq}\left(x^j, q^j\right)$$

is undetermined. It will be positive if volume and quality are complements and negative if they are substitutes. Let $\hat{x}^j\left(p^j, \gamma^j, \delta^j\right)$ and $\hat{q}^j\left(p^j, \gamma^j, \delta^j\right)$ denote hospital j's choice of volume and quality, respectively. Comparative static analysis provides the following relationships:

$$\hat{x}^j_p := \frac{dx^j}{dp^j} = \frac{-H^j_{qq}}{\mathrm{H}} > 0 \qquad \text{(2a)}; \qquad \hat{q}^j_p := \frac{dq^j}{dp^j} = \frac{H^j_{xq}}{\mathrm{H}} \qquad \text{(2b)};$$

$$\hat{x}^j_\gamma := \frac{dx^j}{d\gamma^j} = \frac{H^j_{qq}}{\mathrm{H}} < 0 \qquad \text{(2c)}; \qquad \hat{q}^j_\gamma := \frac{dq^j}{d\gamma^j} = \frac{-H^j_{xq}}{\mathrm{H}}; \qquad \text{(2d)}$$

$$\hat{x}^j_\delta := \frac{dx^j}{d\delta^j} = \frac{-H^j_{xq}}{\mathrm{H}} \qquad \text{(2e)}; \qquad \hat{q}^j_\delta := \frac{dq^j}{d\delta^j} = \frac{H^j_{xx}}{\mathrm{H}} < 0 \qquad \text{(2f)}.$$

The first two terms give hospital j's best-response in volume and quality to the payment rate. The remaining terms describe the effects on volume and quality of heterogeneity in γ and δ, respectively. Note that the cross effects of a higher marginal cost of volume, γ, on quality and of a higher marginal cost of quality, δ, on volume depend on the complementarity. Furthermore, the effects of heterogeneity in γ and δ on population health $b(\cdot)$ are unambiguous only if quality and volume are complements. In this case, less productive hospitals with a higher γ or δ produce a lower level of population health.

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 $H^j \geq 0$.[3] This implies

[2] We assume an interior solution exists.

[3] If the degree of benevolence β is large, a situation might arise in which the hospital incurs a financial loss, where $\overline{B}^j + p^j x^j < c\left(x^j, q^j\right)$. For the purpose of this paper, we rule out this case. Formally, this requires

$$B^j\left(x^j\right)=\overline{B}^j+p^jx^j=-\beta b\left(x^j,q^j\right)+m\left(x^j,q^j\right)+\gamma^jx^j+\delta^jq^j.$$

Thus, the purchaser's objective function becomes

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

Hence, $\max_{p^j} R \quad j=0,1$ subject to the hospitals' best-responses $\hat{x}^j\left(p^j\right)$ and $\hat{q}^j\left(p^j\right)$, as in (2a) and (2b), gives the first order conditions

$$R_{p^j}=R_{x^j}\hat{x}_p^j+R_{q^j}\hat{q}_p^j=\left(R_{x^j}+R_{q^j}\hat{q}_x^j\right)\hat{x}_p^j=0;\qquad j=0,1,$$

where the second equation follows under observation of $\hat{q}_p^j=\hat{q}_x^j\hat{x}_p^j$, as from (2a) and (2b), where $\hat{q}_x^j:=\frac{-H_{xq}^j}{H_{qq}^j}$. Observing $m_x\left(x^j,q^j\right)+\gamma^j=p^j+\beta b_x\left(x^j,q^j\right)$ from (1a); $m_q\left(x^j,q^j\right)+\delta^j=\beta b_q\left(x^j,q^j\right)$ from (1b) we can calculate the explicit first-order conditions

$$R_{p^0}=\lambda\left\langle\begin{array}{l}\left\{1-\beta-\psi(1-\lambda)\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\}\left[b_x\left(x^0,q^0\right)+b_q\left(x^0,q^0\right)\hat{q}_x^0\right]\\-(1+\alpha)p^0\end{array}\right\rangle\hat{x}_p^0=0 \qquad (3a);$$

$$R_{p^1}=(1-\lambda)\left\langle\begin{array}{l}\left\{1-\beta+\psi\lambda\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\}\left[b_x\left(x^1,q^1\right)+b_q\left(x^1,q^1\right)\hat{q}_x^1\right]\\-(1+\alpha)p^1\end{array}\right\rangle\hat{x}_p^1=0 \qquad (3b).$$

In order to simplify the subsequent analysis, we make the following assumptions

$$b_x\left(x^j,q^j\right)+b_q\left(x^j,q^j\right)\hat{q}_x^j\geq 0\quad j=1,2 \qquad (4a);$$

$$b_{xx}+2b_{xq}\hat{q}_x^j+b_{qq}\left(\hat{q}_x^j\right)^2+b_q\frac{d\hat{q}_x^j}{dx}<0 \qquad (4b);$$

$$1-\beta\geq\max\left\{\psi(1-\lambda)\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right];-\psi\lambda\left[b\left(x^0,q^0\right)-b\left(x^1,q^1\right)\right]\right\} \qquad (4c).$$

$\beta b\left(\hat{x}^j,\hat{q}^j\right)-v\left(\hat{x}^j,\hat{q}^j\right)-\gamma^j\hat{x}^j-\delta^j\hat{q}^j<0$, where $\left(\hat{x}^j,\hat{q}^j\right)$ gives the hospital's volume and quality response to the contract $\left(\overline{B}^j,p^j\right)$.

Lemma l. The unique set of optimal case-payments is given by

$$p^{0*} = \frac{\{1-\beta-\psi(1-\lambda)[b(x^0,q^0)-b(x^1,q^1)]\}}{1+\alpha}[b_x(x^0,q^0)+b_q(x^0,q^0)\hat{q}_x^0] \quad (5a);$$

$$p^{1*} = \frac{\{1-\beta+\psi\lambda[b(x^0,q^0)-b(x^1,q^1)]\}}{1+\alpha}[b_x(x^1,q^1)+b_q(x^1,q^1)\hat{q}_x^1] \quad (5b).$$

Proof: See Appendix.

Remark: Assumptions (4b) and (4c) are sufficient but not necessary for the existence of a unique set of optimal rates. Assumption (4b) is technical and relates to the concavity of $b(\cdot)$ when taking into account the best-quality-response to volume. Assumption (4c) implies that an adjustment of payment rates for the concern about variation should never reverse the sign of the optimal payment rate. Assumption (4a) has no bearing on the existence or uniqueness of the optimal payment rates but together with (4c) implies strictly positive payment rates. We employ (4a) to contain the analysis in the subsequent sections, where the possibility of negative payment rates would unduly expand the number of cases to consider.

The case-payments in (5a) and (5b) reflect in turn

- The degree to which the hospital internalises population health β, which has a negative impact on the case-payment.
- 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, $\hat{q}_x^j \geq 0$ and, thus, $b_x(x^j,q^j)+b_q(x^j,q^j)\hat{q}_x^j > 0$. Here, case-payment and volume are greater than in the first-best because of the purchaser's desire to stimulate quality. If quality and volume are substitutes, $\hat{q}_x^j < 0$, 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, $b_x\left(x^j,q^j\right)+b_q\left(x^j,q^j\right)\hat{q}_x^j>0$, 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.

4 The effect of a concern for variation

In this section, we analyse the effect of a regulatory concern about variation on the second-best rates of case-payment. This provides a benchmark for our subsequent analysis of the scope for attaining a second-best with the use of a uniform case-payment and a uniform standard on volume. In what follows, we use the following definitions:

Definitions: We define

- hospital 0 as the hospital that, for any given payment, produces the lower volume, hence $\hat{x}^0(p)\leq\hat{x}^1(p)$.
- $\tilde{\psi}$ as the value of ψ at which there is no variation between hospitals in their volume

$$\tilde{\psi}:=\psi\left|x^{0*}(\psi)=x^{1*}(\psi)\right. \tag{6a}$$

- ψ^* as the value of ψ at which hospitals receive the same optimal payment rate

$$\psi^*:=\psi\left|p^{0*}(\psi)=p^{1*}(\psi)\right. \tag{6b}$$

where $x^{j*}(\psi)=\hat{x}^j\left[p^{j*}(\psi)\right],\ j=0,1$ is the volume implemented by the optimal payment rates.

Also, let the differences between the optimal payment rates, volumes and contributions to population health for the two types of hospital be denoted by

$$\Delta^{p*}(\psi)=p^{0*}(\psi)-p^{*1}(\psi) \tag{7a},$$

$$\Delta^{x*}(\psi)=x^{0*}-x^{1*}=\hat{x}^0\left[p^{0*}(\psi)\right]-\hat{x}^1\left[p^{1*}(\psi)\right] \tag{7b},$$

$$\Delta^{b*}(\psi):=b\left[x^{0*}(\psi),q^{0*}(\psi)\right]-b\left[x^{1*}(\psi),q^{1*}(\psi)\right] \tag{7c},$$

where $q^{j*}(\psi)=\hat{q}^j\left[p^{j*}(\psi)\right];\ j=0,1$ denotes the quality implemented by the optimal payment

rates.

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: volume-related heterogeneity in γ (case 1); quality-related heterogeneity in δ when volume and quality are complements (case 2); and quality-related heterogeneity in δ when volume and quality are substitutes (case 3).

Case 1: Volume-related Heterogeneity in γ.

First we examine 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). From (2c), 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 in this case.

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

- $\Delta^{p*}(\psi) > 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) \leq 0; \Delta^{x*}_{\psi}(\psi) > 0$ for all $\psi \geq 0$.
- $\lim_{\psi \to \infty} \Delta^{p*}(\psi) = \overline{\Delta}^{p*} > 0; \lim_{\psi \to \infty} \Delta^{x*}(\psi) = 0$

Proof: See Appendix.

The Lemma can be understood with reference to figure 1 which plots $\Delta^{p*}(\psi)$, the difference in optimal payment rates for hospital 0 and hospital 1, as per definition (7a), and $\Delta^{x*}(\psi)$, the difference between the volume produced, as per definition (7b).

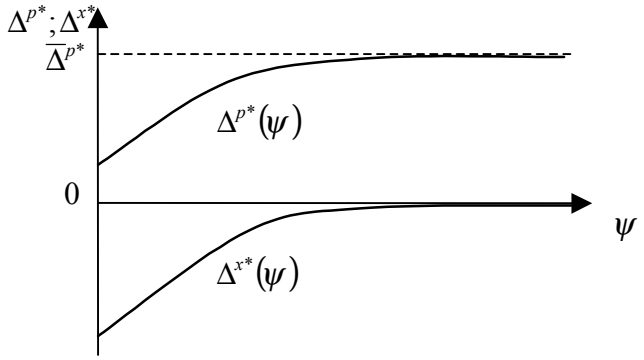

Figure 1[4]

In accordance with assumption (4a), 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 that 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 $x^0 < x^1$. It follows that the fee differential $\Delta^{p*}(\psi)$ should increase in the purchaser's disutility of variation as long as $\Delta^{x*}(\psi) < 0$.

Furthermore, the marginal impact of payment rate on population health (taking into account the quality response) is greater for the less productive hospital. Hence, even if the purchaser is unconcerned with variation ($\psi = 0$), the case-payment should be set at a higher level for hospital 0, implying $\Delta^{p*}(\psi) > 0$ for all $\psi \geq 0$. Yet, absent a concern of variation, 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 difference in volume $\Delta^{x*}(\psi)$ increases in ψ. Under assumption (4a), 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 difference in payments, $\overline{\Delta}^{p*}$, for which $\Delta^{x*} = 0$.

[4] In this figure and in the subsequent figures, Δ^{x*} and Δ^{p*} are not drawn to the same scale. Our use of a single vertical axis is for mere graphical convenience.

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.[5] However, variation becomes arbitrarily small for high values of ψ.[6]

Case 2: Quality-related heterogeneity in δ when 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.

- $\psi \in [0, \psi^*[\Leftrightarrow \left\{\Delta^{p*}(\psi) < 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) < 0; \Delta^{x*}_{\psi}(\psi) > 0\right\}$;
- $\psi \in [\psi^*, \tilde{\psi}] \Leftrightarrow \left\{\Delta^{p*}(\psi) \geq 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) \leq 0; \Delta^{x*}_{\psi}(\psi) > 0\right\}$;
- $\psi > \tilde{\psi} \Leftrightarrow \left\{\Delta^{p*}(\psi) > 0; \Delta^{p*}_{\psi}(\psi) > 0; \Delta^{x*}(\psi) > 0; \Delta^{x*}_{\psi}(\psi) > 0\right\}$;
- $\lim_{\psi \to \infty} \Delta^{p*}(\psi) = \overline{\Delta}^{p*} > 0; \lim_{\psi \to \infty} \Delta^{x*}(\psi) = \overline{\Delta}^{x*} > 0$

Proof: See Appendix.

---

[5] Lemma A1 in the Appendix proves formally that variation in population health is not eliminated for finite values of ψ.

[6] Note from $\lim_{\psi \to \infty} \Delta^{p*}(\psi) = \overline{\Delta}^{p*}$ that the difference between optimal payment rates is bounded from above. Using definition (7a), we can write $p^{1*}(\psi) = p^{0*}(\psi) - \Delta^{p*}(\psi)$. In the proof of Lemma A2 in the Appendix it is shown that $\operatorname{sgn} p^{0*}_{\psi} = -\operatorname{sgn} p^{1*}_{\psi} = -\operatorname{sgn} \Delta^{b*}$. But then, it follows for $\lim_{\psi \to \infty} \Delta^{b*}(\psi) = 0 > \Delta^{b*}(0)$, as in case 1, that $p^{0*}(\psi) > 0$ and $p^{1*}(\psi) \geq \lim_{\psi \to \infty} p^{1*}(\psi) = \lim_{\psi \to \infty} \left[p^{0*}(\psi) - \Delta^{p*}(\psi)\right] = \lim_{\psi \to \infty} p^{0*}(\psi) - \overline{\Delta}^{p*}$ for all $\psi \geq 0$. It follows that $p^{1*}(\psi) \geq 0$ for all $\psi \geq 0$ if the difference $\overline{\Delta}^{p*}$ is not too large. Recall from (5a) and (5b) that $p^{1*}(\psi) \geq 0$ and $p^{0*}(\psi) > 0$ imply the inequality in (4c). Hence, as long as variation can be eliminated at a not too large price difference $\overline{\Delta}^{p*}$ the assumption in (4c) is justified irrespective of the value ψ. A similar argument applies to cases 2 and 3 below.

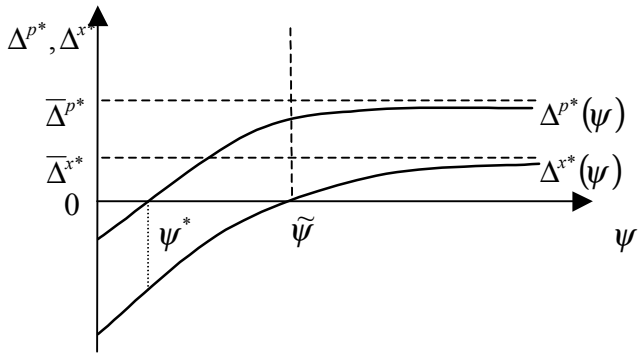

Figure 2

The Lemma can be understood with reference to figure 2. Taking into account assumption (4a), heterogeneity in δ implies for $H^0_{xq} \geq 0$ that the less productive hospital 0, with $\delta^0 > \delta^1$, produces a lower health benefit, $b^0 < b^1$ for any output $x^0 < \bar{x}$, where $\bar{x} > x^1$. 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 $\Delta^{x*}(\psi) = \overline{\Delta}^{x*} > 0$. 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 $\Delta^{x*}(\psi) < \overline{\Delta}^{x*}$.

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 the fee difference $\overline{\Delta}^{p^*} > 0$ induces the quality difference $\overline{\Delta}^{x^*} > 0$. Again, while elimination of variation is feasible, it is never optimal.

Case 3: Quality-related heterogeneity in δ when 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 $\hat{x}^0\left(p, \delta^0\right) \leq \hat{x}^1\left(p, \delta^1\right)$. Hence, this scenario remains consistent with our definition of hospital 0.

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

- $\psi \in [0, \widetilde{\psi}[\Leftrightarrow \left\{\Delta^{p^*}(\psi) > 0; \Delta^{p^*}_{\psi}(\psi) < 0; \Delta^{x^*}(\psi) > 0; \Delta^{x^*}_{\psi}(\psi) < 0\right\}$;
- $\psi \in \left[\widetilde{\psi}, \psi^*\right] \Leftrightarrow \left\{\Delta^{p^*}(\psi) \geq 0; \Delta^{p^*}_{\psi}(\psi) < 0; \Delta^{x^*}(\psi) \leq 0; \Delta^{x^*}_{\psi}(\psi) < 0\right\}$;
- $\psi > \psi^* \Leftrightarrow \left\{\Delta^{p^*}(\psi) < 0; \Delta^{p^*}_{\psi}(\psi) < 0; \Delta^{x^*}(\psi) < 0; \Delta^{x^*}_{\psi}(\psi) < 0\right\}$.
- $\lim_{\psi \to \infty} \Delta^{p^*}(\psi) = \underline{\Delta}^{p^*} < 0; \lim_{\psi \to \infty} \Delta^{x^*}(\psi) = \underline{\Delta}^{x^*} < 0$

Proof: See Appendix.

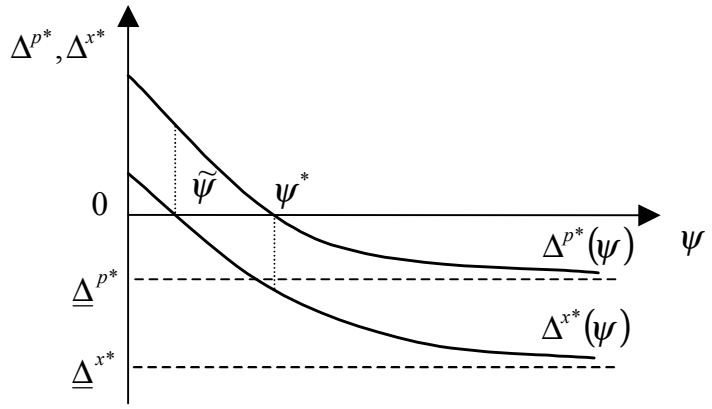

Figure 3

The Lemma can be understood with reference to figure 3. Taking into account assumption (4a), heterogeneity in δ when volume and quality are substitutes implies that hospital 0, which for $\delta^0 < \delta^1$ is more productive in quality, produces greater population health, $b^0 > b^1$ for any volume $x^0 > \underline{x}$, where $\underline{x} < x^1$. Variation in population health is now eliminated only if hospital 0 is induced to produce a volume that is strictly less than that of hospital 1, such that $\Delta^{x*} = \underline{\Delta}^{x*}(\psi) < 0$. This implies that the fee differential $\Delta^{p*}(\psi)$ should decrease in the disutility of variation ψ as long as $\Delta^{x*}(\psi) > \underline{\Delta}^{x*}$.

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

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 $\psi > \tilde{\psi}$. For $\psi > \psi^*$, the differential in payment rates is now negative, $\Delta^{p*}(\psi) < 0$, with hospital 0 still attaining the greater population health as long as $\Delta^{p*}(\psi) > \underline{\Delta}^{p*}$. Again, variation becomes arbitrarily small as ψ tends towards infinity

5 Use of standards

Let us, for the moment, consider the use of standards in the benchmark case in which

provider-specific case-rates can be used. Specifically, 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

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

where the shadow price of the constraint satisfies $\varphi^j > 0 \Leftrightarrow \underline{x}^j = x^j$. The hospital chooses volume and quality according to the first-order conditions

$$H'^j_x = p^j + \beta b_x\left(x^j, q^j\right) - m_x\left(x^j, q^j\right) - \gamma^j + \varphi^j = 0 \qquad \text{(1a')}$$

and (1b). Consider the purchaser's choice of a standard $\underline{x}^j$ given the payment p^j. The best responses to the standard are given by

$$\hat{x}^j_{\underline{x}} := \begin{cases} \dfrac{dx^j}{d\underline{x}^j} = 1 \Leftrightarrow x^j = \underline{x}^j \\ 0 \qquad\qquad \Leftrightarrow x^j > \underline{x}^j \end{cases} \qquad \text{(2a')} \qquad \hat{q}^j_{\underline{x}} := \begin{cases} \dfrac{dq^j}{d\underline{x}^j} = \hat{q}^j_x \Leftrightarrow x^j = \underline{x}^j \\ 0 \qquad\qquad \Leftrightarrow x^j > \underline{x}^j \end{cases} \qquad \text{(2b')}$$

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

$$R_{\underline{x}^j} = R_{x^j}\hat{x}^j_{\underline{x}} + R_{q^j}\hat{q}^j_{\underline{x}} = \left(R_{x^j} + R_{q^j}\hat{q}^j_x\right)\hat{x}^j_{\underline{x}}; \qquad j = 0,1.$$

Observing $m_x\left(x^j, q^j\right) + \gamma^j = p^j + \beta b_x\left(x^j, q^j\right) + \varphi^j$ from (1a') and $m_q\left(x^j, q^j\right) + \delta^j = \beta b_q\left(x^j, q^j\right)$ from (1b), we can write

$$R_{\underline{x}^0} = \lambda\begin{Bmatrix} \left\langle \left\{1 - \beta - \psi(1-\lambda)\left[b\left(x^0, q^0\right) - b\left(x^1, q^1\right)\right]\right\}b_x\left(x^0, q^0\right) - (1+\alpha)\left(p^0 + \varphi^0\right)\right\rangle \\ + \left\{1 - \beta - \psi(1-\lambda)\left[b\left(x^0, q^0\right) - b\left(x^1, q^1\right)\right]\right\}b_q\left(x^0, q^0\right)\hat{q}^0_x \end{Bmatrix}\hat{x}^0_{\underline{x}} \qquad \text{(8a)}$$

$$R_{\underline{x}^1} = (1-\lambda)\begin{Bmatrix} \left\langle \left\{1 - \beta + \psi\lambda\left[b\left(x^0, q^0\right) - b\left(x^1, q^1\right)\right]\right\}b_x\left(x^1, q^1\right) - (1+\alpha)\left(p^1 + \varphi^1\right)\right\rangle \\ + \left\{1 - \beta + \psi\lambda\left[b\left(x^0, q^0\right) - b\left(x^1, q^1\right)\right]\right\}b_q\left(x^1, q^1\right)\hat{q}^1_x \end{Bmatrix}\hat{x}^1_{\underline{x}} \qquad \text{(8b).}$$

Using (3a) and (3b) and observing the best-responses (2a) and (2b) as well as (2b'), 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 second-best payment rates.

Only if for some reason the purchaser cannot set the optimal case-payment is there a role for a standard. Inserting $\varphi^j = -p^j + m_x\left(x^j, q^j\right) + v_x\left(x^j, q^j\right) - \beta b_x\left(x^j, q^j\right) + \gamma^j$ from (1a') into (8a) and (8b) and comparing with (3a) and (3b), we can verify for $p^j \leq p^{j*}$ that $R_{\underline{x}^j} = 0 \Leftrightarrow \underline{x}^j = x^{j*}$. The standard should then be set at the level of volume that would be realised under an optimal payment rate.

Uniform payment and uniform standard

We now consider the more realistic case in which the purchaser is constrained to use a uniform payment scheme $B^o(x) = B^1(x) = B(x) = \overline{B} + px$ and a uniform standard $\underline{x}^0 = \underline{x}^1 = \underline{x}$. Participation requires that the uniform payment satisfies $\min\left\{H^0\left(B\left(x^0\right), x^0, q^0\right); H^1\left(B\left(x^1\right), x^1, q^1\right)\right\} \geq 0$, where $\min\left\{x^0; x^1\right\} \geq \underline{x}$. By mimicking the decisions of the less productive hospital, the more productive one can attain at least the same level of utility, so that when optimising according to its own technology it can be expected to attain a greater utility. It follows that under a uniform payment scheme, the participation constraint has to bind for the less productive type. Drawing on our previous definition of type 0 and the three cases derived from this, we obtain that the participation constraint binds for type 0 in case 1 ($\gamma^0 > \gamma^1$) and case 2 ($\delta^0 > \delta^1$ and $H^0_{xq} > 0$) and for type 1 in case 3 ($\delta^0 < \delta^1$ and $H^0_{xq} < 0$).

Without loss of generality, consider cases 1 or 2, where the participation constraint binds for the less productive type 0 such that

$$\overline{B} \geq -px^0 - \beta b\left(x^0, q^0\right) + m\left(x^0, q^0\right) + \gamma^0 x^0 + \delta^0 q^0$$

where $\max R \Leftrightarrow \min \overline{B}$ implies that the above holds with equality. Taking into account that type 0 and 1 receive payments $B\left(x^o\right) = \overline{B} + px^o$ and $B\left(x^1\right) = \overline{B} + px^1$, respectively, the purchaser's objective is then given by

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

The effect of the payment rate on social welfare can then be determined as

$$R_p=R_{p^0}+R_{p^1}-(1-\lambda)\alpha(x_1-x_0) \tag{9}$$

with R_{P^0} and R_{P^1} given by (3a) and (3b), respectively. The last term reflects an adjustment that reduces the rent accruing to the productive type 1 who receives a lump sum payment that is larger than the one necessary to guarantee participation. Here, rent extraction implies a reduction in the case rate below the level that would otherwise be optimal.

Likewise, the effect of a standard on welfare is given by

$$R_{\underline{x}}=R_{\underline{x}^0}+R_{\underline{x}^1}-(1-\lambda)\alpha\left(\varphi^0\hat{x}^0_{\underline{x}}-\varphi^1\hat{x}^1_{\underline{x}}\right) \tag{10}$$

with $R_{\underline{x}^0}$ and $R_{\underline{x}^1}$ as given by (8a) and (8b). Again, the last term reflects an adjustment aimed at curbing type 1's rent. Assuming that the standard only binds for type 0 such that $\varphi^0\hat{x}^0_{\underline{x}}=\varphi^0$ and $\varphi^1\hat{x}^1_{\underline{x}}=0$, rent extraction implies a reduction in the standard below the level that would otherwise be optimal.

In order to facilitate the exposition let us assume in the following that the rent adjustments in (9) and (10) are negligible. Specifically, assume $\alpha\to 0$ such that $R_p\to R_{p^0}+R_{p^1}$ and $R_{\underline{x}}\to R_{\underline{x}^0}+R_{\underline{x}^1}$. Recalling $\hat{x}^0(p)\le\hat{x}^1(p)$, 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*}.

Proposition 1 (i) 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 $x^{0*} \leq x^{1*}$. (ii) If $x^{0*} > x^{1*}$ 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. (iii) If $p^{0*} < p^{1*}$ the purchaser sets $p^* = \lambda p^{0*} + (1-\lambda)p^{1*}$. There is no role for the standard.

Proof: See Appendix.

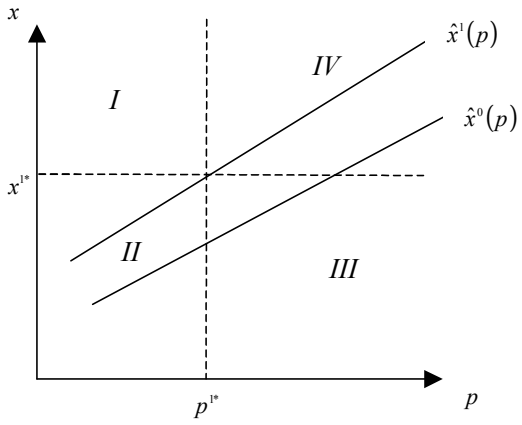

Figure 4.

The intuition of the Proposition can be understood with reference to figure 4, which depicts the supply function $\hat{x}^j(p)$ for each type $j = 0,1$. A second-best is attainable, in principle, as the purchaser can use the two instruments $\underline{x}$ and p to induce the two production targets x^{0*} and x^{1*}. The problem is that the instruments are not hospital specific and, therefore, apply to both types. For $x^{1*} \neq x^{0*}$ this implies that in a second-best the standard must be set at $\underline{x} = x^{0*}$. To understand this consider $\underline{x} = x^{1*}$. If $x^{0*} < x^{1*}$, the standard forces hospital 0 to 'over-produce' at $\hat{x}^0 = \underline{x} > x^{0*}$. If $x^{0*} > x^{1*}$ the price has to be used to raise hospital 0's output to the optimum, $\hat{x}^0(p) = x^{0*} > \underline{x}$. However, since $\hat{x}^0(p) < \hat{x}^1(p)$, this, in turn, would induce over-production by type 1, where $\hat{x}^1(p) > \hat{x}^0(p) > x^{1*}$.

Once the standard has been fixed at $\underline{x} = x^{0*}$, the purchaser can use the price to induce the optimal output by type 1, where $p = p^{1*}$. However, it is now easy to see that this is possible only if the two conditions $x^{1*} > x^{0*}$ and $p^{1*} < p^{0*}$ are satisfied. If $x^{1*} < x^{0*}$ the standard forces type 1 to over-produce, where $\hat{x}^1 = \underline{x} > x^{1*}$. Price has no longer a role in this scenario

and the regulator sets the standard at the weighted average $\underline{x} = \lambda x^{0*} + (1-\lambda)x^{1*}$. If $p^{1*} > p^{0*}$ the price induces type 0 to over-produce, where $\hat{x}^0(p^{1*}) > \underline{x} = x^{0*}$. In this case, the standard is slack; and the regulator sets the price at the weighted average $p = \lambda p^{0*} + (1-\lambda)p^{1*}$. In figure 4, we can see that a second-best is feasible only if (p^{1*}, x^{1*}) lies on the segment of $\hat{x}^0(p)$ in segment III. If (p^{1*}, x^{1*}) lies on the segment of $\hat{x}^0(p)$ in segment II and IV, respectively, a third-best is attained with a price and a standard, respectively.

We can now analyse which of the cases described in the above proposition arises depending on the type of heterogeneity (i.e. heterogeneity in γ or δ); the hospitals' 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 reconsider the three cases we have characterised before.

Case 1: Volume-related heterogeneity in γ.

Recall our assumptions $\gamma^0 > \gamma^1$ and $\delta^0 = \delta^1$ implying that the less productive hospital 0 produces lower volume and a lower health benefit (from assumption (4a)) for any given payment rate.

Proposition 2: The second-best can always be implemented if hospitals are heterogeneous in γ but not in δ.

Proof: Immediate from Lemma 2 and Proposition 1.

Inspection of figure 1 shows that the conditions $p^{0*} \geq p^{1*}$ and $x^{0*} \leq x^{1*}$, or equivalently $\Delta^{p*} \geq 0$ and $\Delta^{x*} \leq 0$, are satisfied for all possible $\psi \geq 0$. Thus, irrespective of the disutility 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: Quality-related heterogeneity in δ when volume and quality are complements.

Recall our assumptions $\gamma^0 = \gamma^1$ and $\delta^0 > \delta^1$ and $H^j_{xq} \geq 0$ implying that the less productive hospital 0 tends to produce lower quality, lower volume and a lower health benefit for any

given payment rate.

Proposition 3: The second-best is unattainable in either of two cases: (i) The marginal disutility from variation is 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 marginal disutility from variation is 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: Immediate from Lemma 3 and Proposition 1.

Inspection of figure 2 shows that the condition $\Delta^{p*} \geq 0$ is not satisfied if $\psi < \psi^*$, and that the condition $\Delta^{x*} \leq 0$ is not satisfied if $\psi > \tilde{\psi}$. 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) volume. 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. 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 of variation is high, the purchaser would like to induce the unproductive hospital 0 to produce a greater volume than hospital 1. Such an allocation is not feasible as a uniform standard would bind for both hospitals. The purchaser can merely achieve a third-best with a standard set at $\underline{x}^* = \lambda x^{0*}(\psi) + (1-\lambda)x^{1*}(\psi)$.

Case 3: Quality-related heterogeneity in δ when volume and quality are substitutes.

Recall our assumptions $\gamma^0 = \gamma^1$, $\delta^0 < \delta^1$ and $H^j_{xq} < 0$ implying that the less productive hospital 1 tends to produce lower quality, greater volume and a lower health benefit (from assumption (4a)) for any given payment rate.

Proposition 4: The second-best is unattainable in either of two cases: (i) The marginal disutility from variation is 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) The marginal disutility of variation is 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: Immediate from Lemma 4 and Proposition 1.

Inspection of figure 3 shows that the condition $\Delta^{x^*} \leq 0$ is violated for $\psi < \tilde{\psi}$, whereas $\Delta^{p^*} \geq 0$ fails to hold for $\psi > \psi^*$. If the disutility of variation is low it would be efficient for the purchaser to induce the more productive hospital 0 to produce 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 $\underline{x}^* = \lambda x^{0*}(\psi) + (1-\lambda) x^{1*}(\psi)$. Attainment of the second-best is also ruled out if the disutility of variation is large. Again, the purchaser can only attain a third-best allocation, here by setting $p^* = \lambda p^{0*}(\psi) + (1-\lambda) p^{1*}(\psi)$.

Corollary 1: When hospital heterogeneity is quality-related the second-best is unfeasible for weak complementarity between volume and quality, i.e. for $\left| H^0_{xq} \right| \to 0$.

Proof: See Appendix.

If quality and volume are poor substitutes or complements then quality-related heterogeneity has little impact on volume, and the two hospitals produce (almost) the same volume for any given payment rate. In this case, any difference in the optimal payment rate will induce a difference in volume in the same direction, i.e. $\operatorname{sgn} \Delta^{p^*} = \operatorname{sgn} \Delta^{x^*}$. This implies $\tilde{\psi} \approx \psi^*$ (in figures 2 and 3) so that there is no scope for a second-best. The problem arises from the purchaser's inability to contract on quality, the variable that is directly affected by heterogeneity. As both instruments price and standard relate to the contractible variable, volume, a second-best can only be achieved if the (positive or negative) correlation between quality and volume is sufficiently strong. If, in contrast heterogeneity relates to the contractible variable (volume), the purchaser can attain the second-best as the instruments apply directly.

6 Conclusions

Provider heterogeneity within health care systems is widely recognised. In particular within public health care systems such as the UK's NHS, the resulting variation in volume and/or quality of care is the focus of policy initiatives. This has led to the advocacy of benchmarking and the use of minimum performance standards. We analyse these policies by extending a model by Chalkley & Malcomson (1998) into which we introduce provider heterogeneity, and a concern about variation on behalf of the purchaser.

Our results can be summarised as follows. When hospitals are heterogeneous and quality is non-contractible, the purchaser can implement a second-best solution by appropriately setting differentiated case-payments. In many practical situations the purchaser may be restricted to a uniform case-payment and/or a uniform standard on volume. Instrument choice will then depend on the production technology and the purchaser's concern for variation. If factor heterogeneity relates to the production of volume, the second-best solution can be obtained by the combined use of a uniform case-payment and standard. In contrast, where factor heterogeneity affects the production of quality a second-best solution is unattainable if the degree of complementarity between volume and quality is too low. In this case, only a third-best can be attained which involves the exclusive use of either a standard or a case-payment. Which one of the instruments is optimal depends on the form of complementarity and on the concern for variation.

Our model is related to the growing literature on the impact of equity concerns on contractual relationships (Meyer & Mookherjee, 1987, Milgrom & Roberts, 1990, von Siemens, 2001, Englmaier & Wambach, 2002). Most of this literature deals with a principal facing a single type of agent, where the principal is concerned about an equitable allocation either for altruistic or for motivational reasons. Whereas these models are mainly concerned about the effects of fairness concerns between principal and agent on the efficiency of contracts, in our model the issue rather lies with the principal's concern about the inequality in outcomes delivered by heterogeneous agents.[7]

Our model re-appraises the use of quantities (activity standards) and/or prices (case-

[7] Von Siemens (2001) deals with agent heterogeneity in an adverse selection context, whereas our interest does not lie with informational concerns but rather with the problems arising under a lack of instruments.

payments) in regulation. Weitzman (1974) analyses the role of standards and prices under asymmetric information about the underlying benefits and costs. We obtain a rule about the optimality of a price versus a standard when producers (hospitals) are heterogeneous and the regulator is constrained to a single price and/or a single standard. In this regard, our work is related to Heyes (2001) who compares an input standard against an emission tax in a setting of environmental regulation. With taxes being based on firms' non-verifiable reports of their emissions, heterogeneity of firms in their propensity to report truthfully can lead to the standard outperforming the tax, in spite of the latter being superior in balancing marginal damage cost against marginal abatement cost. However, the two instruments are never applied simultaneously. In our model, there are circumstances under which the purchaser can target case-payments at high volume providers and standards at low volume providers. Hence, while in the case of homogeneous hospitals standard and case-payment are perfect substitutes, the simultaneous use of standards and prices may be optimal from a second-best point of view when hospitals are heterogeneous.[8]

Our findings, therefore, provide a rationale for the simultaneous use of performance standards and case-payments by purchasers of health care. For example, the UK's Department of Health plans to introduce at national level payments according to health care resource groups (HRG) (Department of Health, 2002).[9] HRG payments reflect cost-differences owing to the nature of the clinical condition, but their implementation as uniform payments at national level also implies that they do not reflect differences in provider productivity. Variation across hospitals in the volume and/or quality of care provided within a HRG is likely to be present. Our analysis shows that a purchaser can then improve the allocation by complementing the HRG payment with a volume standard if heterogeneity is predominantly volume-related. If heterogeneity is quality-related, there is scope for the use of both instruments if and only if the correlation between volume and quality is sufficiently strong. In all other cases the optimal solution entails either that a standard is redundant or that it should effectively replace

[8] A standard (on volume, say) may also be understood as a benchmark against which the total payment to the provider is gauged such that the participation and/or budget constraint is satisfied if and only if volume corresponds to the standard (e.g. Holmström, 1979, Baker, 1992). However, in contrast to our finding incentives arise only from the price (the case rate) and not the standard. Sherstyuk (2000) considers incentives for an agent who faces a performance risk and is subject to limited liability. A standard can then complement a payment, and is effective in bad states of nature where limited liability leads to slack in the monetary incentive. The issue of heterogeneity is not addressed, and a situation in which incentives arise simultaneously from the standard and a payment cannot arise.

[9] These are roughly equivalent to the diagnostic related groups as used in the US and other countries.

the HRG payment as the main source of incentives.

7 References

Baker, G. (1992), 'Incentive Contracts and Performance Measurement.' *Journal of Political Economy*, vol. 100, 598-614.

Chalkley, M. and Malcomson, J.M. (1998), 'Contracting for health services when patient demand does not reflect quality.' *Journal of Health Economics*, vol. 17,1-19.

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

Department of Health (2000), *The NHS Plan. A Plan for Investment. A Plan for Reform*. London: Stationary Office.

Department of Health (2002), *Reforming NHS financial flows. Payments by results*. London: Stationary Office.

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

Dranove, D., Kessler, D., McClellan, M. and Satterthwaite, M. (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.

Englmaier, F. and Wambach, A. (2002), Contracts and Inequity Aversion, mimeo, University of Munich.

Fehr, E. and Schmidt, K.M. (1999), 'A Theory of Fairness, Competition, and Cooperation.' *Quarterly Journal of Economics*, vol. 114, 817-68.

Gabsewicz, J.J. and Thisse, J.-F. (1979), 'Price Competition, Quality and Income Disparities.' *Journal of Economic Theory*, vol. 20, 340-359.

Greene, W.H. (1993), 'The econometric approach to efficiency analysis.' In Fried, H.O., Lovell, C.A.K. and Schmidt, S.S. (eds.), *The measurement of productive efficiency:*

Techniques and applications. New York: Oxford University Press.

Heyes, A. (2001), 'Honesty in a Regulatory Context - Good Thing or Bad?' *European Economic Review*, vol. 45, 215-232.

HM Treasury (2002), *Public spending review 2002: public service agreements*. London: HMSO.

Holmström, B. (1979), 'Moral Hazard and Observability.' *Bell Journal of Economics*, vol. 10, 74-91.

Hollingsworth, B., Dawson, P.J. and Maniadakis, N. (1999), 'Efficiency measurement of health care: a review of non-parametric methods and applications.' *Health Care Management Science*, vol. 2, 161-72.

Meyer, M.A. and Mookherjee, D. (1987), 'Incentives, Compensation, and Social Welfare.' *Review of Economic Studies*, vol. 54, 209-226.

Milgrom, P. and Roberts, J. (1990), 'The efficiency of equity in organizational decision-processes.' *American Economic Review*, vol. 80(2), 154-59.

NHS Executive (1997), *The New NHS: modern, dependable*. Leeds: NHS Executive.

Reilly, T. and Meyer, G. (2002), 'Providing performance information for consumers: experience from the United States.' In Smith, P. (ed.), *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*, vol. 39, 432-78.

Rosen, S. (1981), 'The economics of superstars.' *American Economic Review*, vol. 71, 845-58.

Von Siemens, F. (2001), Adverse Selection and Inequity Aversion, mimeo, University of Munich.

Shaked, A. and Sutton, J. (1982), 'Relaxing Price Competition through Product Differentiation.' *Review of Economic Studies*, vol. 49, 3-13.

Sherstyuk, K. (2000), 'Performance Standards and Incentive Pay in Agency Contracts.' *Scandinavian Journal of Economics*, vol. 102, 725-36.

Weitzman, M. (1974), 'Prices vs. Quantities.' *Review of Economic Studies*, vol. 41, 477-91.

Appendix

Proof of Lemma 1: To ease further exposition, we reformulate (3a) and (3b) to obtain:

$$\Pi^0\left(p^0, p^1, \psi, \gamma^0, \delta^0\right) - p^0 = 0 \qquad \text{(A1a)} \qquad \Pi^1\left(p^0, p^1, \psi, \gamma^0, \delta^0\right) - p^1 = 0 \qquad \text{(A1b)}$$

where

$$\Pi^0(\cdot) = \frac{\left\{1 - \beta - \psi(1-\lambda)\left[b\left(x^0, q^0\right) - b\left(x^1, q^1\right)\right]\right\}}{1+\alpha}\left[b_x\left(x^0, q^0\right) + b_q\left(x^0, q^0\right)\hat{q}_x^0\right] \qquad \text{(A1c)}$$

$$\Pi^1(\cdot) = \frac{\left\{1 - \beta + \psi\lambda\left[b\left(x^0, q^0\right) - b\left(x^1, q^1\right)\right]\right\}}{1+\alpha}\left[b_x\left(x^1, q^1\right) + b_q\left(x^1, q^1\right)\hat{q}_x^1\right] \qquad \text{(A1d).}$$

Note that a unique and stable solution $\left\{p^{0*}; p^{1*}\right\}$ exists if and only if the Hessian of (A1a) and (A1b), satisfies $Z = \Pi^0_{p^0}\Pi^1_{p^1} - \Pi^0_{p^1}\Pi^1_{p^0} > 0$.

Defining $\vartheta^j =: b_x\left(x^{j*}, q^{j*}\right) + b_q\left(x^{j*}, q^{j*}\right)\hat{q}_x^j$, we obtain from (A1c) and (A1d)

$$\Pi^0_{p^0} = \frac{\left\langle -\psi(1-\lambda)\left(\vartheta^0\right)^2 + \left[1 - \beta - \psi(1-\lambda)\Delta^{b*}\right]\frac{d\vartheta^0}{dx}\right\rangle \hat{x}_p^0}{1+\alpha} < 0 \qquad \text{(A1e)}$$

$$\Pi^1_{p^1} = \frac{\left\langle -\psi\lambda\left(\vartheta^1\right)^2 + \left[1 - \beta + \psi\lambda\Delta^{b*}\right]\frac{d\vartheta^1}{dx}\right\rangle \hat{x}_p^1}{1+\alpha} < 0 \qquad \text{(A1f)}$$

$$\Pi^0_{p^1} = \frac{\psi(1-\lambda)\vartheta^0\vartheta^1\hat{x}_p^1}{1+\alpha} > 0 \qquad \text{(A1g)} \qquad \Pi^1_{p^0} = \frac{\psi\lambda\vartheta^0\vartheta^1\hat{x}_p^0}{1+\alpha} > 0 \qquad \text{(A1h)}$$

where $\frac{d\vartheta^j}{dx} = b_{xx} + 2b_{xq}\hat{q}_x^j + b_{qq}\left(\hat{q}_x^j\right)^2 + b_q\frac{d\hat{q}_x^j}{dx}$.

The inequalities in (A1e)-(A1h) then follow under assumption of (4a)-(4c). Using (A1e)-(A1h) one can verify $Z > 0$. ■

In the following we will refer repeatedly to the following Lemmas A1 and A2, which we, therefore, prove in separate.

Lemma A1: If $\Delta^{b*}(0) \neq 0$, there exists no finite value $\overline{\psi} > 0$ that satisfies $\Delta^{b*}(\overline{\psi}) = 0$.

Proof: Suppose that $\Delta^{b*}(0) \neq 0$ and, by contradiction, suppose a finite value $\overline{\psi} > 0$ exists such that $\Delta^{b*}(\overline{\psi}) = 0$. The first-order conditions for the optimal payment rates then require

$$R_{p^0} = \lambda\left\langle\left\{1-\beta-\overline{\psi}(1-\lambda)\Delta^{b*}(\overline{\psi})\right\}\vartheta^0 - (1+\alpha)p^0\right\rangle\hat{x}_p^0 = \lambda\left\langle\{1-\beta\}\vartheta^0 - (1+\alpha)p^0\right\rangle\hat{x}_p^0 = 0;$$

$$R_{p^1} = (1-\lambda)\left\{\left\{1-\beta+\overline{\psi}\lambda\Delta^{b*}(\overline{\psi})\right\}\vartheta^1 - (1+\alpha)p^1\right\}\hat{x}_p^1 = (1-\lambda)\left\{\{1-\beta\}\vartheta^1 - (1+\alpha)p^1\right\}\hat{x}_p^1 = 0;$$

where the second equality in each expression follows from $\Delta^{b*}(\overline{\psi}) = 0$. But then, the last equalities together with the uniqueness of the pair $\{p^{0*}(\psi), p^{1*}(\psi)\}$ imply $p^{j*}(\overline{\psi}) = p^{j*}(0)$ and from the uniqueness of the hospital's optimum $x^{j*}(\overline{\psi}) = x^{j*}(0)$, $q^{j*}(\overline{\psi}) = q^{j*}(0)$ and $b^{j*}(\overline{\psi}) = b^{j*}(0)$. Consequently, $\Delta^{b*}(\overline{\psi}) = \Delta^{b*}(0) \neq 0$, a contradiction. Thus, unless $\Delta^{b*}(0) = 0$, there exists no $\overline{\psi} \in]0;\infty[$ such that $\Delta^{b*}(\overline{\psi}) = 0$. ■

Lemma A2. (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: Comparative static analysis yields

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

where $\Pi_\psi^0 = \frac{-(1-\lambda)\Delta^{b*}\vartheta^0}{1+\alpha}$, $\Pi_\psi^1 = \frac{\lambda\Delta^{b*}\vartheta^1}{1+\alpha}$; and $Z = \Pi_{p^0}^0\Pi_{p^1}^1 - \Pi_{p^1}^0\Pi_{p^0}^1 > 0$. Observing $\vartheta^j > 0;\ j = 0,1$, and (A1a)-(A1d) one can verify that $\operatorname{sgn} p_\psi^{0*} = -\operatorname{sgn} p_\psi^{1*} = -\operatorname{sgn}\Delta^{b*}$. But then, using definition (6d)

$$\operatorname{sgn}\Delta_\psi^{p*} = \operatorname{sgn}\left(p_\psi^{0*} - p_\psi^{1*}\right) = -\operatorname{sgn}\Delta^{b*} \tag{A2}$$

From $x_\psi^{j*} = \hat{x}_p^j p_\psi^{j*}$ for $j = 0,1$ and under observation of (2a) it follows that $\operatorname{sgn} x_\psi^{j*} = \operatorname{sgn} p_\psi^{j*}$ and, thus,

$$\operatorname{sgn} x_{\psi}^{0*} = -\operatorname{sgn} x_{\psi}^{1*} = -\operatorname{sgn} \Delta^{b*} \tag{A3}$$

Using definition (7b),

$$\operatorname{sgn} \Delta_{\psi}^{x*} = \operatorname{sgn}\left(x_{\psi}^{0*} - x_{\psi}^{1*}\right) = -\operatorname{sgn} \Delta^{b*} \tag{A4}$$

Using definition (7c), we find that

$$\operatorname{sgn} \Delta_{\psi}^{b*} = \operatorname{sgn}\left(\vartheta^{0} x_{\psi}^{0*} - \vartheta^{1} x_{\psi}^{1*}\right) = -\operatorname{sgn} \Delta^{b*}(\psi) \tag{A5},$$

where the last equality follows from (A3). Finally, we note that

$$\Delta^{b*}(\psi) = 0 \Leftrightarrow \Delta_{\psi}^{b*} = \Delta_{\psi}^{x*} = \Delta_{\psi}^{p*} = 0 \tag{A6}.$$

Suppose that $\Delta^{b*}(0) \neq 0$. It then follows from Lemma A1 that there exists no finite value $\overline{\psi} > 0$ such that $\Delta^{b*}(\overline{\psi}) = 0$. But then, it follows from (A5) and (A6) that $\lim_{\psi \to \infty} \Delta^{b*}(\psi) = 0$ and $\lim_{\psi \to \infty} \Delta_{\psi}^{b*}(\psi) = \lim_{\psi \to \infty} \Delta_{\psi}^{p*} = \lim_{\psi \to \infty} \Delta_{\psi}^{x*} = 0$, which proves part (ii). Furthermore, since $\Delta^{b*}(\psi) \neq 0 \quad \forall \psi < \infty$ it follows that $\operatorname{sgn} \Delta^{b*}(\psi) = \operatorname{sgn} \Delta^{b*}(0) \quad \forall \psi \geq 0$. Together with (A2) and (A4) this implies part (i) of the Lemma.

Finally, consider $\Delta^{b*}(0) = 0$. In this case, (A6) implies $\Delta^{b*}(\psi) = 0 \quad \forall \psi \geq 0$, a trivial case that is embraced by parts (i) and (ii) of the Lemma. This completes the proof.■

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. From part (ii) of Lemma A2, $\lim_{\psi \to \infty} \Delta^{b*}(\psi) = 0$, which together with (b) above implies $\lim_{\psi \to \infty} \Delta^{x*}(\psi) = 0$. Together with (a) above this implies $\lim_{\psi \to \infty} \Delta^{p*}(\psi) = \overline{\Delta}^{p*}$. Furthermore, (a) and (b) imply $\gamma^{0} > \gamma^{1} \Rightarrow \Delta^{b*}(0) < 0$ so that from part (i) of

Lemma A2 $\operatorname{sgn}\Delta^{p^*}_{\psi}(\psi)=\operatorname{sgn}\Delta^{x^*}_{\psi}(\psi)=-\operatorname{sgn}\Delta^{b^*}(0)=1$ for all $\psi\geq 0$ as indicated in the Lemma. But then, it follows immediately from (a) that $\Delta^{p^*}(\psi)>0$ for all $\psi\geq 0$. Recall $\lim_{\psi\to\infty}\Delta^{x^*}(\psi)=0$, which together with (a) implies $\Delta^{x^*}(\psi)\leq 0$ for all $\psi\geq 0$. Hence, (a) and (b) imply the properties presented in the Lemma. We now turn to prove (a) and (b).

(a) Consider the comparative static properties

$$p^{0^*}_{\gamma} := \frac{dp^{0^*}}{d\gamma^0}=\frac{\Pi^0_{\gamma^0}-\left(\Pi^0_{\gamma^0}\Pi^1_{p^1}-\Pi^1_{\gamma^0}\Pi^0_{p^1}\right)}{Z} \tag{A7a}$$

$$p^{1^*}_{\gamma} := \frac{dp^{1^*}}{d\gamma^0}=\frac{\Pi^1_{\gamma^0}-\left(\Pi^1_{\gamma^0}\Pi^0_{p^0}-\Pi^0_{\gamma^0}\Pi^1_{p^0}\right)}{Z} \tag{A7b}$$

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

Using (A7a)-(A7b) together with (A1e)-(A1h) one can show that

$$\Delta^{x^*}_{\gamma}=\hat{x}^0_{\gamma}+\hat{x}^0_p p^{0^*}_{\gamma}-\hat{x}^1_p p^{1^*}_{\gamma}$$

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

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

(b) Observing from (1b) 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 (4a) that $\Delta^{b^*}\leq 0\Leftrightarrow b\left(x^{0^*},q^{0^*}\right)=b\left[x^{0^*},\hat{q}^0\left(x^{0^*}\right)\right]\leq b\left[x^{1^*},\hat{q}^1\left(x^{1^*}\right)\right]=b\left(x^{1^*},q^{1^*}\right)\Leftrightarrow x^{0^*}\leq x^{1^*}\Leftrightarrow\Delta^{x^*}\leq 0$, which

proves (b).■

Proof of Lemma 3: We prove in turn

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

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

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

Suppose that (a)-(c) hold. Together, (a) and (b) imply $\left\{\delta^0 > \delta^1; H^0_{xq} \in [0, k^+[\right\} \Rightarrow \Delta^{b*}(0) < 0$ so that from part (i) of Lemma A2 $\operatorname{sgn} \Delta^{p*}_{\psi}(\psi) = \operatorname{sgn} \Delta^{x*}_{\psi}(\psi) = -\operatorname{sgn} \Delta^{b*}(0) = 1$ for all $\psi \geq 0$.

But then, it follows immediately from (c) together with the definitions (6a) and (6c) that $\psi \in [0, \psi^*[\Leftrightarrow \left\{\Delta^{p*}(\psi) < 0; \Delta^{x*}(\psi) < 0\right\}$; $\psi \in [\psi^*, \widetilde{\psi}] \Leftrightarrow \left\{\Delta^{p*}(\psi) \geq 0; \Delta^{x*}(\psi) \leq 0\right\}$; and $\psi > \widetilde{\psi} \Leftrightarrow \left\{\Delta^{p*}(\psi) > 0; \Delta^{x*}(\psi) > 0\right\}$. The limits $\lim_{\psi \to \infty} \Delta^{p*}(\psi) = \overline{\Delta}^{p*} > 0$ and $\lim_{\psi \to \infty} \Delta^{x*}(\psi) = \overline{\Delta}^{x*} > 0$ follow from (a) and (b) in conjunction with part (ii) of Lemma A2. Hence, (a)-(c) are sufficient for the properties listed in part (ii) of the present Lemma. We now turn to proving (a)-(c).

(a) Consider the comparative static properties

$$\left(p^{0*}_{\delta} := \frac{dp^{0*}}{d\delta^0}\right)_{\psi=0} = \frac{\Pi^0_{\delta^0} - \left(\Pi^0_{\delta^0}\Pi^1_{p^1} - \Pi^1_{\delta^0}\Pi^0_{p^1}\right)}{Z}\Big|_{\psi=0} \tag{A8a}$$

$$\left(p^{1*}_{\delta} := \frac{dp^{1*}}{d\delta^0}\right)_{\psi=0} = \frac{\Pi^1_{\delta^0} - \left(\Pi^1_{\delta^0}\Pi^0_{p^0} - \Pi^0_{\delta^0}\Pi^1_{p^0}\right)}{Z}\Big|_{\psi=0} \tag{A8b}$$

with $Z > 0$, where

$$\Pi^j_{p^j}\Big|_{\psi=0} = \left(\tfrac{1-\beta}{1+\alpha}\right)\tfrac{d\vartheta^j}{dx}\hat{x}^j_p < 0; \quad j = 0,1 \tag{A9}$$

and $\Pi^0_{p^1}\Big|_{\psi=0} = \Pi^1_{p^0}\Big|_{\psi=0} = 0$ from (A1e)-(A1h); where

$$\Pi^0_{\delta^0}\Big|_{\psi=0} = \frac{(1-\beta)\Omega\hat{q}^0_\delta}{1+\alpha} \tag{A10}$$

with

$$\Omega := \left(-b_{xx} + b_{xq}\frac{H^0_{xq}}{H^0_{qq}}\right)\frac{H^0_{xq}}{H^0_{xx}} + b_{xq} - b_{qq}\frac{H^0_{xq}}{H^0_{qq}} \tag{A11}$$

and where $\Pi^1_{\delta^0}\Big|_{\psi=0} = 0$. Inserting these into (A8a) and (A8b), it is easy to check that

$$\Delta^{p^*}_\delta\Big|_{\psi=0} = p^{0^*}_\delta\Big|_{\psi=0} - p^{1^*}_\delta\Big|_{\psi=0} = p^{0^*}_\delta\Big|_{\psi=0} = \frac{\Pi^0_{\delta^0}\left(1-\Pi^1_{p^1}\right)}{Z}\Big|_{\psi=0} \tag{A12}$$

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

$$\Delta^{p^*}_\delta\Big|_{\psi=0} < 0 \Leftrightarrow p^{0^*}_\delta\Big|_{\psi=0} < 0 \Leftrightarrow \Pi^0_{\delta^0}\Big|_{\psi=0} < 0 \Leftrightarrow \Omega > 0 \tag{A13}$$

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

$$\frac{d\Omega}{dH^0_{xq}} = \left(-b_{xx} + 2b_{xq}\frac{H^0_{xq}}{H^0_{qq}}\right)\left(\frac{1}{H^0_{xx}}\right) - b_{qq}\left(\frac{1}{H^0_{qq}}\right) < \frac{-\left[\left(b_{xq}H^0_{qq}\right)^2 - 2b_{xq}H^0_{qq}b_{qq}H^0_{xq} + \left(b_{qq}H^0_{xq}\right)^2\right]}{b_{qq}H^0_{qq}} < 0 \tag{A14}$$

Observing $\Omega\Big|_{H^0_{xq}=0} = b_{xq} > 0$ it then follows that there exists a $k^+ > 0$ such that $\Omega > 0 \Leftrightarrow H^0_{xq} < k^+$. But then, from (A13), $\left\{\delta^0 > \delta^1; H^0_{xq} \in [0, k^+[\right\} \Rightarrow \Delta^{p^*}(0) < 0$.

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

(b) Observing from (1b) 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 (4a) it follows that $\Delta^{x^*} \le 0 \Leftrightarrow x^{0^*} \le x^{1^*} \Rightarrow b\left(x^{0^*}, q^{0^*}\right) < b\left(x^{1^*}, q^{1^*}\right) \Leftrightarrow \Delta^{b^*} < 0$, which proves (b).

(c) Recall $\Delta^{x^*} \leq 0 \Rightarrow \Delta^{b^*} < 0$ from part (b) and $\Delta^{x^*}(0) < 0$ and $\Delta^{x^*}_{\psi}(0) > 0$ 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 $\{\delta^0 > \delta^1; H^0_{xq} \in [0, k^+[\} \Rightarrow \hat{x}^0(p) \leq \hat{x}^1(p)$. But then, $\Delta^{x^*}(\psi) = 0 \Rightarrow \Delta^{p^*}(\psi) \geq 0$. Observing $\{\Delta^{p^*}_{\psi}(\psi), \Delta^{x^*}_{\psi}(\psi)\} > 0$ together with $\Delta^{p^*}(0) < 0$ this implies $\tilde{\psi} \geq \psi^* > 0$. This completes the proof of (c).■

Proof of Lemma 4: We prove in turn

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

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

(c) $H^0_{xq} \geq k^- \Rightarrow \infty > \psi^* \geq \tilde{\psi} > 0$, 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 A2 $\operatorname{sgn} \Delta^{p^*}_{\psi}(\psi) = \operatorname{sgn} \Delta^{x^*}_{\psi}(\psi) = -\operatorname{sgn} \Delta^{b^*}(0) = -1$ for all $\psi \geq 0$.

But then, it follows immediately from (c) together with the definitions (6a) and (6c) that $\psi \in [0, \tilde{\psi}[\Leftrightarrow \{\Delta^{p^*}(\psi) > 0; \Delta^{x^*}(\psi) > 0\}$; $\psi \in]\tilde{\psi}, \psi^*] \Leftrightarrow \{\Delta^{p^*}(\psi) \leq 0; \Delta^{x^*}(\psi) \geq 0\}$; and $\psi > \psi^* \Leftrightarrow \{\Delta^{p^*}(\psi) < 0; \Delta^{x^*}(\psi) < 0\}$ as stated in part (ii) of the present Lemma. The limits $\lim_{\psi \to \infty} \Delta^{p^*}(\psi) = \underline{\Delta}^{p^*} < 0; \lim_{\psi \to \infty} \Delta^{x^*}(\psi) = \underline{\Delta}^{x^*} < 0$ follow from (a) and (b) together with part (ii) of Lemma A2. Hence, (a)-(c) guarantee the pattern indicated in the Lemma. We now turn to prove (a)-(c).

(a) Recall from (A13) that $\Delta^{p^*}_{\delta}\big|_{\psi=0} < 0 \Leftrightarrow p^{0^*}_{\delta}\big|_{\psi=0} < 0 \Leftrightarrow \Omega > 0$, where Ω as defined in (A11). Observing $\Omega\big|_{H^0_{xq}=0} = b_{xq} > 0$ and $\frac{d\Omega}{dH^0_{xq}} < 0$, as from (A14), it follows that $H^0_{xq} \leq 0 \Rightarrow \Omega > 0$ and by implication $H^0_{xq} \leq 0 \Rightarrow \Delta^{p^*}_{\delta}\big|_{\psi=0} < 0$. But then, from (A13), $\{\delta^0 < \delta^1; H^0_{xq} \leq 0\} \Rightarrow \Delta^{p^*}(0) > 0$.

Since $H^0_{xq} \leq 0 \Rightarrow \{p^{0^*}_{\delta}\big|_{\psi=0} < 0; \hat{x}^0_{\delta} > 0\}$ 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 (2e) and from the RHS of (A12); then from (A9) and (A10); and finally from (A11); 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 $\left\{\delta^0 < \delta^1; H^0_{xq} \le 0\right\} \Rightarrow \Delta^{x^*}(0) > 0$, which completes the proof of (a).

(b) Observing from (1b) 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 (4a) it follows that $\Delta^{x^*} \ge 0 \Leftrightarrow x^{0^*} \ge x^{1^*} \Rightarrow b\left(x^{0^*}, q^{0^*}\right) > b\left(x^{1^*}, q^{1^*}\right) \Leftrightarrow \Delta^{b^*} > 0$, which proves (b).

(c) Recall $\left\{\delta^0 < \delta^1; H^0_{xq} \le 0\right\} \Rightarrow \hat{x}^0(p) \le \hat{x}^1(p)$ such that $\Delta^{p^*}(\psi) = 0 \Rightarrow \Delta^{x^*}(\psi) \le 0$. Observing $\left\{\Delta^{p^*}_{\psi}(\psi), \Delta^{p^*}_{\psi}(\psi)\right\} < 0$ 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 (6a) and (6c), $H^0_{xq} = 0 \Rightarrow \tilde{\psi} = \psi^* < \infty$. It is easily checked that $\frac{d\left(\tilde{\psi} - \psi^*\right)}{dH^0_{xq}} < 0$ implying that there exists a $k^- < 0$ such that $H^0_{xq} \in \left]k^-, 0\right] \Leftrightarrow \tilde{\psi} \le \psi^* < \infty$. This proves (c).■

Proof of Proposition 1:

Part (i): For $x^{0^*} \ne x^{1^*}$ 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 $\hat{x}^1\left(p^{0^*}\right) \le x^{1^*} < x^{0^*}$. 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 $\hat{x}^0\left(p^{1^*}\right) \le x^{0^*} \le x^{1^*}$. This yields the conditions and

$\hat{x}^0\left(p^{1^*}\right) \le x^{0^*} \Leftrightarrow p^{0^*} \ge p^{1^*}$ $\qquad\qquad$ $x^{0^*} \le x^{1^*}$

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

Part (ii): $x^{0*} > x^{1*}$ obviously violates feasibility. 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 $\underline{x} = x^{0*} > x^{1*} = \hat{x}^1(p^{1*})$ binds for both types and is overly restrictive for type 1, forcing over-production. Formally, $R_{\underline{x}}[x^{0*}; x^{0*}] = \overbrace{R_{\underline{x}^0}[x^{0*}]}^{=0} + \overbrace{R_{\underline{x}^1}[x^{0*}]}^{<0} < 0$. Thus, it is optimal to set the standard at a level $\underline{x} \in]x^{1*}; x^{0*}[$ such that $R_{\underline{x}}[\underline{x}; \underline{x}] = \overbrace{R_{\underline{x}^0}[\underline{x}]}^{>0} + \overbrace{R_{\underline{x}^1}[\underline{x}]}^{<0} = 0$. Using (8a) and (8b) one can verify that this implies $\underline{x}^* = \lambda x^{0*} + (1-\lambda)x^{1*}$. Since $\underline{x}^* > x^{1*} \geq \hat{x}^0(p^{1*})$, both hospitals choose their output independent of the payment rate $p = p^{1*}$. Consider now alternative values of the case-rate. Let $\overline{p} := p | \hat{x}_1(p) = \underline{x}^*$. 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 $R_{\underline{x}}[\underline{x}^*; \underline{x}^*] = 0$ it follows that $R_{\underline{x}}[\max\{\hat{x}^0(p); \underline{x}^*\}; \hat{x}^1(p)] < 0$ for all $p > \overline{p}$, where $\hat{x}_1(p) > \underline{x}^*$. Finally, consider $p = \overline{p}$. While this implements $R_{\underline{x}}[\underline{x}^*; \underline{x}^*] = R_{\underline{x}}[\underline{x}^*; \hat{x}^{1*}(p)] = 0$ in the presence of a standard, $\hat{x}_0(\overline{p}) \leq \hat{x}_1(\overline{p}) = \underline{x}^*$ implies $R_{\underline{x}}[\hat{x}^{0*}(\overline{p}); \hat{x}^{1*}(\overline{p})] = R_{\underline{x}}[\hat{x}^{0*}(\overline{p}); \underline{x}^*] > 0$ in the absence of a standard. Thus, only a standard can implement the third-best. A case-payment is redundant.

Part (iii): $p^{0*} < p^{1*}$ implies $x^{0*} < \hat{x}^0(p^{1*}) \leq x^{1*}$. But then, a standard set at $\underline{x} = x^{0*} < \hat{x}^0(p^{1*}) \leq x^{1*}$ does not bind for type 0, which over-produces relatively to the first-best. Formally, $R_p[\hat{x}^0(p^{1*}); x^{1*}] = \overbrace{R_{p^0}[\hat{x}^0(p^{1*})]}^{<0} + \overbrace{R_{p^1}[x^{1*}]}^{=0} < 0$. Thus, it is optimal to set the case rate at $p \in]p^{1*}; p^{0*}[$ such that $R_p[\hat{x}^0(p); \hat{x}^1(p)] = \overbrace{R_{p^0}[\hat{x}^0(p)]}^{>0} + \overbrace{R_{p^1}[\hat{x}^1(p)]}^{<0} = 0$. Using (3a) and (3b) one can verify that this implies $p^* = \lambda p^{0*} + (1-\lambda)p^{1*}$. Since $p^* \geq p^{0*}$ and, thus, $\underline{x}^* = x^{0*} \leq \hat{x}^0(p^*) \leq \hat{x}^1(p^*)$, both hospitals choose their output above the standard. Consider now alternative values of the standard. Obviously, any $\underline{x} \leq \hat{x}^0(p^*)$ has no effect. From the concavity of $R(\cdot)$ and from $R_p[\hat{x}^0(p^*); \hat{x}^1(p^*)] = 0$ it follows that $\operatorname{sgn} R_{\underline{x}}[\underline{x}^*; \max\{\hat{x}^1(p); \underline{x}^*\}] = \operatorname{sgn} R_p[\underline{x}^*; \max\{\hat{x}^1(p); \underline{x}^*\}] = -1$ for all $\underline{x} > \hat{x}^0(p^*)$. Finally,

consider $\underline{x} = \hat{x}^0(p^*)$. While this implies $R_p[\underline{x}; \hat{x}^1(p^*)] = 0$ in the presence of a case-based payment, $\max\{\hat{x}_1(0), \underline{x}\} \leq \hat{x}_1(p^*)$ implies $\operatorname{sgn} R_{\underline{x}}[\underline{x}; \max\{\hat{x}_1(0), \underline{x}\}] = \operatorname{sgn} R_p[\underline{x}; \max\{\hat{x}_1(0), \underline{x}\}] > 0$ in the absence of a case-payment. Thus, only a case-payment can implement the third-best. A standard is redundant.■

Proof of Corollary 1: $H^0_{xq} = 0$ implies $\hat{x}^0(p, \delta^0) = \hat{x}^1(p, \delta^1)$ irrespective of $\delta^0 \neq \delta^1$. But then from definitions (6a) and (6b), $H^0_{xq} = 0 \Rightarrow \tilde{\psi} = \psi^*$. It is then easy to check from Lemma 3 and Lemma 4, that $\Delta^{p^*}(\psi) \geq 0 \Leftrightarrow \Delta^{x^*}(\psi) \geq 0$ so that a first-best is unattainable if and only if $\psi = \tilde{\psi}$. For a continuous distribution of ψ on $[0, \infty[$, this is a zero probability event. By continuity, $\tilde{\psi} \to \psi^*$ for $\left|H^0_{xq}\right| \to 0$ implying that $\psi \in [\tilde{\psi}, \psi^*]$ or $\psi \in [\psi^*, \tilde{\psi}]$ is an almost zero probability event. ■