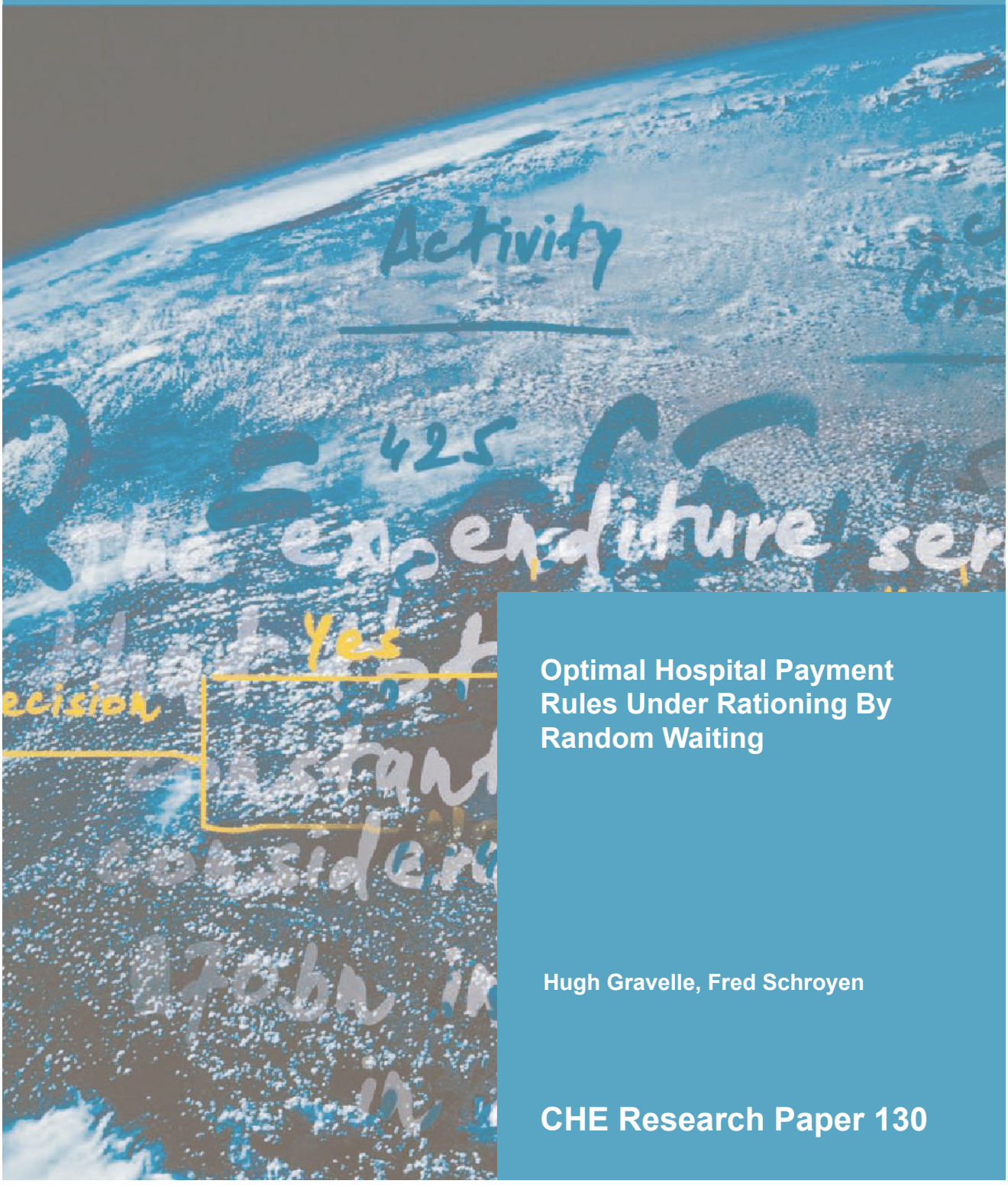
Monograph:
Gravelle, Hugh Stanley Emrys and Schroyen, Fred (2016) Optimal hospital payment rules under rationing by random waiting. Discussion Paper. CHE Research Paper . Centre for Health Economics, University of York , York, UK.

CHE
Centre For Health Economics

UNIVERSITY of York



Optimal Hospital Payment Rules Under Rationing By Random Waiting

Hugh Gravelle, Fred Schroyen

CHE Research Paper 130

Optimal hospital payment rules under rationing by random waiting

[1]Hugh Gravelle
[2]Fred Schroyen

[1]Centre for Health Economics, University of York, York, UK
[2]Department of Economics, Norwegian School of Economics, Bergen, Norway

May 2016

Background to series

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Acknowledgements

We received helpful comments from Gjermund Grimsby, Philippe Bontem and other participants at the 2015 Annual Conference of the Norwegian Economic Association (Bergen) and at the 2015 European Health Economics Workshop (Toulouse).

Further copies

Copies of this paper are freely available to download from the CHE website www.york.ac.uk/che/publications/ Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321405 for further details.

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/che

Optimal hospital payment rules under rationing by random waiting*

Hugh Gravelle† Fred Schroyen‡

26 April 2016

Abstract We derive optimal rules for paying hospitals in a public health care system in which providers can choose quality and random patient demand is rationed by waiting time. Since waiting time imposes real costs on patients hospital payment rules should take account of their effect on waiting time as well as on quality and the number of patients treated. We develop a general stochastic model of rationing by waiting and use it to derive welfare maximising payment to hospitals linked to output, expected waiting times, quality, hospital capacity and length of stay. We show that, although prospective output pricing gives hospitals an incentive to attract patients by raising quality and reducing waiting times, it must be supplemented by prices attached to other hospital decisions and outcomes except under very strong assumptions about the welfare function, patient preferences, and whether patients lose income whilst waiting.

Keywords: Rationing. Waiting times. Queues. Prospective payment. Hospitals.

JEL code: I11, I13, I18, L51, D81

*We received helpful comments from Gjermund Grimsby, Philippe Bontem and other participants at the 2015 Annual Conference of the Norwegian Economic Association (Bergen) and at the 2015 European Health Economics Workshop (Toulouse).

†Centre for Health Economics, University of York. Email: hugh.gravelle@york.ac.uk.

‡Department of Economics, Norwegian School of Economics. Email: fred.schroyen@nhh.no.

1 Introduction

Public hospital systems, like those in Scandinavia, the UK, and other OECD countries, are mainly financed through general taxation or compulsory social insurance. Patients face zero or very low money prices and consequently non-emergency treatment is rationed by waiting time (Cullis *et al.*, 2000; Siciliani and Iversen, 2012). Such waiting times are often long and a source of concern to both patients and policy makers.[1] There are also considerable variations in quality and costs amongst hospitals. Hospitals in these systems are increasingly paid prospectively for each case treated (Paris *et al.*, 2010) and in some countries there are attempts to improve hospital quality by linking payment directly to quality as well as to output (Jha *et al.*, 2012; Sutton *et al.*, 2012).

In this paper we derive optimal rules for paying hospitals in a public health care system in which patient demand is rationed by waiting time and hospitals can choose quality and make supply decisions that change the distribution of waiting times facing patients.

Although both quality and waiting time affect patient demand and both can be influenced by hospital decisions, quality is determined solely by hospital decisions whereas waiting time is determined by both patient demand and hospital supply. Waiting time is not just another type of quality and cannot be treated as just another quality dimension when analysing payment rules: it is necessary to allow for the fact that waiting time is also a rationing device which equates patient demand and hospital supply and does so by imposing costs on patients which are not offset by gains to providers. Payment rules have to be designed in the light of their effect on these rationing costs as well as on hospital quality and the number of patients treated.

The literature on the welfare implications of alternative hospital payment systems is reviewed in Chalkley and Malcomson (2000). The aim of the payment system is to induce welfare maximising hospital behaviour: treatment of an optimal number of patients with optimal quality at minimum cost. In this literature it is assumed that payment cannot be linked directly to unverifiable or unobserved quality and cost reducing effort. Policy makers are restricted to setting a price for output and to reimbursing hospitals for their costs. In Ellis and McGuire (1986) the number of patients requiring treatment is not affected by hospital decisions and first best quality and cost reducing effort are not achievable. They show that the second best welfare maximising reimbursement rule combines a prospective price per patient treated and partial reimbursement of hospital costs. With only a prospective price and an exogenously determined number of patients requiring treatment the hospital will skimp on quality, unless it is perfectly altruistic, because quality is costly and has no effect on its revenue. Partial reimbursement of costs reduces the marginal cost of quality and so induces the hospital to increase quality. But partial cost reimbursement also reduces the incentive for cost reducing effort so that the second best mixed reimbursement scheme trades off quality and cost reducing effort. Despite having two policy targets (quality and cost reducing effort) and two policy instruments the first best is not achievable because, with a fixed number of patients, the prospective price is equivalent to a lump sum payment with no incentive properties: the only instrument which affects the

[1] For example, the median waiting time from being placed on the waiting list for hip replacement to treatment in 2011 was 108 days in Australia, 113 in Finland, 87 in Portugal, and 82 in England (Siciliani *et al.*, 2014). See Cullis *et al.* (2000), Iversen and Siciliani (2011) and Siciliani and Iversen (2012) for surveys of the health economics waiting time literature.

hospital quality decision is cost reimbursement.

Chalkley and Malcomson (1998) consider payments regimes when patient demand varies with quality.[2] They show that, if (a) there is only one dimension of quality and (b) it is optimal to treat all patients who demand care at the optimal quality, then first best quality and output can be achieved at minimum cost with a single instrument: a prospective output price. Because higher quality attracts more patients, thus increasing revenue, hospitals respond to a higher price by increasing quality. It is thus possible to set the price so that the hospital chooses the optimal quality and this results in the optimal number of patients being treated. And with no cost reimbursement the hospital bears all the costs of producing care and so has the appropriate incentive for cost reducing effort. Remarkably, this result does not depend on the policy maker and patients having the same valuation of quality and the benefits of treatment. It is not even necessary that patients correctly perceive quality when demanding care, only that their demand is increasing in quality as perceived by the policy maker. But Chalkley and Malcomson (1998) also show that if quality is multi-dimensional, the first best is implementable via the output price only if the policy maker and patients have the same relative marginal valuations of the different quality dimensions.

The insights from this literature are obtained from models which do not take account of rationing by waiting time which is a salient feature of many public health care systems. It is implicitly assumed that demand is not affected by waiting time.[3] But there is considerable evidence that waiting times do affect demand for elective care. Higher waiting times lead patients to switch within the public sector to hospitals with lower waiting times (Sivey, 2012), to opt for private hospitals (Besley *et al.* 1999, Aarbu, 2010) or to forgo care entirely (Martin and Smith, 1999; Gravelle *et al.*, 2002; Windmeijer *et al.*, 2005).

Following Lindsay and Feigenbaum (1984) most formal models of rationing by waiting in health care assume that demand and supply, and hence waiting time, are deterministic.[4] The certain waiting time adjusts, like the money price in standard markets, to ensure that the certain demand equals the certain supply. Such deterministic waiting time models are useful for some purposes but are flawed as a basis for modelling hospital behaviour and the welfare implications of regulation and pricing regimes. With a positive waiting time a hospital can reduce quality whilst holding supply constant. Waiting time will fall to equate demand and the unchanged supply and so hospital revenue is unchanged. Since cost is reduced because quality is lower, profit is increased. Hence, a profit maximising hospital whose revenue varies with the volume of patients treated will never choose to have both positive waiting time and positive quality. Thus in deterministic models of rationing by waiting, the only way to explain the coexistence of positive waiting times and positive quality, is by assuming sufficiently great direct provider concern with quality.

In stochastic queueing models waiting times are determined by the random demand for treatment (conditional on quality) and random length of treatment. Waiting times are

[2] The Ellis and McGuire (1986) setting is akin to emergency treatment and in Chalkley and Malcomson (1998) it is akin to elective treatment where patients choose amongst alternatives (including no treatment).

[3] When demand exceeds supply Chalkley and Malcomson (1998) assume that there is perfect rationing (all patients treated have higher benefits that those who are not selected for treatment) or random rationing. But neither method of rationing is assumed to impose any direct costs on patients (other than not being treated if not selected) and patient demand is assumed unaffected the probability of treatment.

[4] See, for example, Marchand and Schroyen (2005), and Gravelle and Siciliani (2008).

therefore also random. In equilibrium there is a steady state distribution of waiting times determining the mean waiting time and the mean number of patients treated per period. The mean number treated is equal to the mean demand and less than the capacity of the hospital. The mean waiting time is positive since only an infinitely large service capacity can result in all patients having a zero realised waiting time. A reduction in quality will reduce costs but it will also reduce expected demand and thus expected output and revenue and so may not increase expected profit. Thus the equilibrium of the system will always have positive expected waiting time and may, if quality is not too costly, have a positive quality.

Two papers in the health economics literature have considered stochastic waiting time models with demand depending on the distribution of waiting times.[5] In Goddard *et al.* (1995) it is assumed that a patient observes the length of the waiting list before deciding whether to join the list or not. The resulting complicated expressions for the steady state probabilities on the number of people in the system and expected waiting time are used to derive comparative static predictions about the effects of patient income and the price of private care. Iversen and Lurås (2002) use a much simpler queueing model to examine competition between GPs via their choice of quality and expected waiting time.

Because we use our model of rationing with random waits for normative rather than positive analysis we derive demand functions for treatment from patient preferences over income, quality and waiting times, rather than making plausible but *ad hoc* assumptions about the demand functions. We derive a welfare function based on these preferences to examine policy options. Like Iversen and Lurås (2002) we take an *ex ante*, or rational expectations, approach, though we have a much more general specification of the queueing model and of individual preferences.[6] In the rational expectations equilibrium individuals decide whether to seek public treatment on the basis of an anticipated waiting time distribution and their decisions generate the anticipated distribution. By contrast to Goddard *et al.* (1995), this approach has the advantage of yielding an analytically tractable equilibrium steady state distribution of waiting times for the public system which can be used to examine the welfare properties of payment schemes.[7]

We derive first and second-best payment schemes for a public hospital taking into account their effects on the hospital's choice of quality, the number of beds and its service rate, and their impacts on the equilibrium waiting time distribution. In the first best, when a prospective price is combined with payments related to any two of quality, beds, and service rate, the first best price per patient treated is less than the marginal social

[5] There are stochastic waiting time models of hospitals in the operations research literature (see, e.g., Worthington (1987, 1991) and the survey by Fomundam and Herrmann (2007)). But none of these allow for balking, i.e., for patient decisions to join the waiting list being affected by the distribution of waiting times.

Some of the queuing literature does consider endogenous arrival processes or balking (Hassin and Haviv, 2003). Analyses of pricing have focussed on the use of user charges to influence demand and curb congestion, rather than on provider prices to encourage supply and quality. For economic analyses of user charges in stochastic queueing models see Edelson and Hildenbrand (1975) and Naor (1969).

[6] For example, we allow demand to depend on the distribution of waiting times not just on the mean waiting time.

[7] The assumption also explains the purchase of supplementary insurance against the cost of private treatment by individuals before they fall ill. This decision must be made ex ante and so be based on unconditional expectations about the distribution waiting times, not the distribution conditional on the number waiting at the date the individual falls ill.

benefit per additional patient. This difference is larger the stronger the hospital's degree of altruism, the greater the marginal cost of public funds, and the greater is the effect of waiting lists and waiting times on lost earnings due to waiting for care. With a prospective price and one other instrument the optimal second best price per patient treated is higher than the first best price to compensate for the fact that only one of the hospital decisions (quality, service rate, number of beds) is directly incentivised. Only if the welfare function respects patient preferences over waiting time and quality, all patients are willing to trade off waiting time and quality at the same rate, and if there are no lost earnings from waiting for treatment, can a prospective output price yield the first best quality and service rates in the absence of other policy instruments.

In the Section 2 we describe the stochastic queueing process and patient choices between public and private treatment, examine the effects of hospital choice of quality and supply decisions on the equilibrium demand and waiting time distribution, and set out the welfare function. In Section 3 we derive the first best hospital financing scheme when the regulator has a sufficient set of instruments. In Section 4, we limit this set and derive and discuss second and third best pricing rules. Section 5 compares stochastic and deterministic waiting time models and shows that the deterministic specification is unsatisfactory as a model of hospital behaviour and hence as a the basis for a welfare analysis of optimal hospital payment systems. Section 6 concludes.

2 Model

2.1 Queueing model

We use a general model of the queueing process which includes some of the standard stochastic queueing models as special cases.[8] However, unlike most queueing models we allow for the fact that demand (the arrival process) depends on the distribution of waiting times. Our focus is on obtaining a tractable analytical model of the resulting market equilibrium as a basis for deriving first and second-best payment schemes for public hospitals.

We assume that patients require elective treatment for a non-epidemic condition with probability σ. All patients have the same severity and health gain from treatment and there is no prioritisation of patients who are treated in order of arrival on the waiting list: the queue discipline is "first come, first served". The mean rate of arrivals (patients joining the waiting list) per unit of time is λ.

The hospital has k beds allowing it to treat k patients simultaneously. A patient's length of stay once admitted is uncertain though the hospital can influence it by varying staffing levels, theatre hours, and better coordination between departments. We summarise these supply decisions by μ. Both k and μ will affect the distribution of the length of time a patient waits on the list to be treated. μ will also effect the distribution of length of stay for patients when admitted. Note that although for definiteness we interpret k,μ as hospital decisions on beds and service rate they could be any hospital decisions shifting the distribution of waiting times.

[8] See Taylor and Karlin (1998, ch 9) or Gross *et al.* (2008, ch 2) for an introduction to queueing theory.

Because the arrival rate and length of stay are random, the time between being placed on the waiting list and admission to the hospital is also random. We assume that the stochastic processes governing additions to the list and length of stay imply that the total time w between referral to the hospital and completed treatment has a steady state distribution function

$$H(w;\lambda,k,\mu), \qquad H_\lambda < 0, H_s > 0 \ (s = k,\mu). \tag{1}$$

Thus increases in λ and reductions in k and μ produce first degree stochastic dominating changes in the distribution of waiting times, implying that the mean wait $\bar{w}$ is increasing in the referral rate and decreasing in (k,μ):

$$\bar{w} = \int_0^\infty w \mathrm{d}H(w;\lambda,k,\mu) = \bar{w}(\lambda,k,\mu), \qquad \bar{w}_\lambda > 0, \bar{w}_s < 0 \ (s = k,\mu). \tag{2}$$

More importantly for our purposes, (1) implies that under the assumption that patients prefer shorter waiting times, increases in λ and reductions in k or μ reduce the expected utility of patients who decide to join the queue.

Table 1. Symbols and definitions

Symbol	Definition
σ	probability of ill health
λ	demand for public hospital (referral rate)
μ, k	public hospital supply decisions: service rate, number of beds
$w, \bar{w}$	random, mean wait for public hospital
$H(w;\lambda,k,\mu)$	waiting time distribution function
q	quality in public hospital
y	income
$F(y)$	income distribution function
$u(y,q,w)$	utility if ill and treatment in public hospital after wait of w
$\bar{u}(y,q,\lambda,k,\mu)$	expected utility if ill and treatment in public hospital
$u^N(y)$	utility if not ill
$v = \sigma\bar{u} + (1-\sigma)\,u^N$	expected utility if treatment in public hospital when ill
$v^o(y-\gamma)$	expected utility with private insurance at premium γ
$\hat{y}$	threshold income: choice of public hospital if $y \leq \hat{y}$
$B(q,\lambda,k,\mu)$	aggregate patient welfare
$c^H(q,\lambda,k,\mu)$	public hospital expected cost
$c^I(q,\lambda,k,\mu)$	expected total earnings loss due to waiting

Our results hold for all queuing systems which yield a waiting time distribution satisfying (1). One specific, and relatively straightforward, example is the $M/M/k$ system[9] in which the number of arrivals has a Poisson distribution with mean rate λ, the length of stay has a negative exponential distribution with parameter μ, and there are k servers (beds), with $\lambda < k\mu$, so that the mean number of patients joining the waiting list is less than the mean number who are treated per unit of time. When $k\mu$ exceeds the arrival

[9] $M/M/k$ is the Kendall notation for a queueing system with Markov (memoryless) arrivals, Markov service time and k servers.

rate λ, there is a steady state for the queueing system such that the probabilities for a given number of patients in the system (either waiting or being treated) are well-defined. From these, one can derive the probability of a patient being admitted without waiting, $\pi_0(\lambda, k, \mu)$ ($\pi_{0\lambda} < 0$, $\pi_{0k} > 0$, $\pi_{0\mu} > 0$) as well as the distribution for the waiting time when all beds are occupied. The distribution of the total waiting time (time on the list plus and time under treatment) is the convolution of two negative exponential distributions and the mean wait is $\bar{w}(\lambda, k, \mu) = \frac{1}{\mu} + \frac{1-\pi_0}{k\mu-\lambda}$. In the simple but instructive $M/M/1$ system the expected wait is $\bar{w}(\lambda, \mu) = \frac{1}{\mu-\lambda}$. In the Appendix (Theorem A.1) we prove that the $M/M/k$ system has a distribution of waiting times satisfying (1).

2.2 Patients

A compulsory public health insurance system covers the costs of treatment in the public hospital. Income y per unit of time is distributed over $[y_{\min}, y_{\max}]$ with distribution function $F(y)$. We assume that a patient waiting for treatment is unable to work but is fully reimbursed by the social insurance system for foregone earnings of $w \cdot y$.[10]

Utility when ill and treated in the public hospital with quality q after a time w is

$$u = u(y, q, w), \qquad u_y > 0, u_q > 0, u_w < 0.$$

$u(\cdot)$ is a cardinal function which is increasing and concave in y and q and decreasing in w.[11] We assume usually that there is a single dimension of quality but also consider the implications of multiple quality dimensions. We normalize the quality variable so that the minimum quality is $q = 0$. Quality q reflects aspects of the hospital stay and treatment that alter utility but do not affect length of stay. Examples might be the extent to which patients receive adequate pain management, are informed about the diagnosis, treated with respect by staff, and aspects of hotel services, such as privacy, visiting hours, and quality of food. The adoption of minimally invasive surgery techniques, nursing intensity or effective hygiene which reduce the risk of acquiring hospital infections, are interpreted as efforts to reduce average length of stay and are captured by the treatment intensity μ.

Waiting time in the public hospital is uncertain and expected utility when ill for a patient who decides not to take out private health care insurance and to be treated in the public hospital is

$$\bar{u}(y, q, \lambda, k, \mu) = \int u(y, q, w) \mathrm{d}H(w; \lambda, k, \mu), \qquad \bar{u}_\lambda < 0, \bar{u}_s > 0 \ (s = k, \mu).$$

The first order stochastic dominance properties of H and the assumption that patients dislike waiting ($u_w < 0$) imply that expected utility is decreasing in the arrival rate and increasing in k and μ.[12]

[10] Allowing the proportion of income lost whilst waiting to be less than one or to vary with income or to be jointly distributed with income would not change the results substantively.

[11] In general, patients may distinguish between time spent on the waiting list and time spent in the hospital til discharge. We ignore this distinction because it would unnecessarily complicate the model without affecting its main results.

[12] In their seminal paper, Lindsay and Feigenbaum (1984) assume that the effect of waiting time is captured by exponential discounting: $u(y, q)e^{-\phi w}$. With these preferences and uncertain w expected utility is $u(y, q)Ee^{-\phi w} = u(y, q)J^H(-\phi)$, where $J^H(-\phi)$ is the moment generating function for distribution

We sometimes consider two benchmark cases of patient preferences over public treatment. In the first case preferences are quasi-separable (QS) so that the marginal rate of substitution between quality and waiting time is independent of income. Equivalently, the marginal rates of substitution between any pair of q, λ, μ, k are independent of income

$$\bar{u} = \int [a^1(y) + a^2(y) r(q,w)] \mathrm{d}H(w;\lambda,k,\mu) = a^1(y) + a^2(y) R(q,\lambda,k,\mu). \tag{3}$$

with $R(q,\lambda,k,\mu) \stackrel{\text{def}}{=} \int r(q,w)] \mathrm{d}H(w;\lambda,k,\mu)$.

In the second special case (LN preferences) u is linear in w so that citizens only care about the expected wait:

$$\bar{u} = \int [t^1(y,q) + t^2(y,q) w] \mathrm{d}H(w;\lambda,k,\mu) = t^1(y,q) + t^2(y,q) \bar{w}(\lambda,k,\mu). \tag{4}$$

With LN preferences the marginal rates of substitution amongst λ, μ, k are independent of income and quality.

Utility when in good health and not requiring hospital treatment $u^N(y)$ is an increasing concave function of income with $u^N(y) > u(y,q,w)$ (all y,q,w), so that immediate treatment never makes a patient better off than if healthy. Expected utility from not taking out private health care insurance and being treated in the public hospital when ill is

$$v(y,q,\lambda,k,\mu) \stackrel{\text{def}}{=} \sigma \bar{u}(y,q,\lambda,k,\mu) + (1-\sigma) u^N(y), \quad v_y > 0, v_\lambda < 0, v_z > 0 \ \ (z = q,k,\mu).$$

There is also a private hospital sector which provides care with a low certain wait w^o and higher quality q^o than the public hospital. Individuals who know they will prefer to use the private sector when ill buy full cover supplementary private insurance at an actuarially fair price γ. Thus their utility when ill is $u^o = u(y-\gamma, q^o, w^o) = u^o(y-\gamma)$ and utility when in good health is $u^N(y-\gamma)$. Expected utility from taking out private insurance and being treated in the private hospital is

$$v^o(y-\gamma) \stackrel{\text{def}}{=} \sigma u^o(y-\gamma) + (1-\sigma) u^N(y-\gamma).$$

We assume that $v_y(y,q,\lambda,k,\mu) < v^o_y(y-\gamma)$, which holds if there is declining marginal utility of income, weak Edgeworth complementarity between income and quality of treatment ($u_{yq} \geq 0$) and $u_{yw} \leq 0$. Hence there is a threshold income level $\hat{y}$ (assumed to be in the interior of $[y_{\min}, y_{\max}]$) defined by

$$v(\hat{y},q,\lambda,k,\mu) - v^o(\hat{y}-\gamma) = 0, \tag{5}$$

such that all individuals with $y \leq \hat{y}$ choose the option of no private insurance and treatment

$H(w)$. This has the analytical advantage of yielding tractable expressions for expected utility with some distributions. For example, in the $M/M/k$ system w has a negative exponential distribution and $J^H(-\phi) = \frac{k\mu-\lambda+\phi\pi_0}{k\mu-\lambda+\phi}$. However, the utility function is convex in w implying that a mean preserving spread in the distribution of w would increase expected utility: the patient would be better off with a riskier waiting time distribution.

in the public hospital when ill. The threshold income $\hat{y}(q,\lambda,k,\mu)$ has derivatives

$$\hat{y}_z(q,\lambda,k,\mu) = -\frac{v_z(\hat{y},q,\lambda,k,\mu)}{v_y(\hat{y},q,\lambda,k,\mu)-v_y^o(\hat{y}-\gamma)} > 0, \qquad (z=q,k,\mu) \tag{6}$$

$$\hat{y}_\lambda(q,\lambda,k,\mu) = -\frac{v_\lambda(\hat{y},q,\lambda,k,\mu)}{v_y(\hat{y},q,\lambda,k,\mu)-v_y^o(\hat{y}-\gamma)} < 0. \tag{7}$$

2.3 Rational expectations equilibrium

Since individuals fall ill at a rate σ and choose the public hospital if and only if they have $y \leq \hat{y}(q,\lambda,k,\mu)$, they join the waiting list at the rate $\sigma F(\hat{y}(q,\lambda,k,\mu))$. Hence the equilibrium arrival rate (patient demand for public hospital care), $\lambda(q,k,\mu)$, is implicitly defined by

$$\lambda - \sigma F(\hat{y}(q,\lambda,k,\mu)) = 0.$$

This embodies the rational expectations assumption: the distribution $H(w;\lambda,k,\mu)$ upon which decisions about joining the waiting list for the public hospital are based coincides with the distribution $H(w;\lambda(q,k,\mu),k,\mu)$ that these decisions give rise to. Demand is increasing in quality and supply since both increase the utility of the marginal patient:

$$\lambda_z(q,k,\mu) = \frac{\sigma f(\hat{y})\hat{y}_z}{1-\sigma f(\hat{y})\hat{y}_\lambda} > 0, \qquad z=q,k,\mu. \tag{8}$$

We use ^{e} to denote equilibrium values of variables and functions. At the equilibrium the threshold income level depends on hospital quality and supply decisions

$$\hat{y}^e(q,k,\mu) = \hat{y}(q,\lambda(q,k,\mu),k,\mu),$$

with

$$\hat{y}_z^e = \hat{y}_z + \hat{y}_\lambda\lambda_z = \hat{y}_z + \hat{y}_\lambda\frac{\sigma f(\hat{y})\hat{y}_z}{1-\sigma f(\hat{y})\hat{y}_\lambda} = \frac{\hat{y}_z}{1-\sigma f(\hat{y})\hat{y}_\lambda} \in (0,\hat{y}_z), \qquad (z=q,k,\mu). \tag{9}$$

Using these results we have

Lemma 1 *Hospital attributes z,x $(z,x=q,k,\mu,\ z\neq x)$ have the same relative marginal effects on demand, threshold income and expected utility :*

$$\frac{\lambda_z}{\lambda_x} = \frac{\widehat{y}_z(q,\lambda(q,k,\mu),k,\mu)}{\widehat{y}_x(q,\lambda(q,k,\mu),k,\mu)} = \frac{\widehat{y}_z^e(q,k,\mu)}{\widehat{y}_x^e(q,k,\mu)} = \frac{v_z^e(\widehat{y},q,k,\mu)}{v_x^e(\widehat{y},q,k,\mu)} = \frac{v_z(\widehat{y},\lambda(q,k,\mu),q,k,\mu)}{v_x(\widehat{y},\lambda(q,k,\mu),q,k,\mu)}. \tag{10}$$

If utility is linear in the waiting time then it also true that

$$\frac{v_z}{v_x} = \frac{\overline{w}_z}{\overline{w}_x} \qquad (z,x=\lambda,k,\mu). \tag{11}$$

The direct effects of q and k,μ on the expected utility v of those choosing the public hospital are positive. But they all also have an indirect effect in increasing λ and this reduces v. Defining $v^e(y,q,k,\mu) = v(y,q,\lambda(q,k,\mu),k,\mu)$ and using (6) and (7), the effect

of $z = q, k, \mu$ on the expected utility of those choosing the public hospital is

$$\begin{aligned} v_z^e(y,q,k,\mu) &= v_z(y,q,\lambda,k,\mu) + v_\lambda(y,q,\lambda,k,\mu)\lambda_z = v_z + v_\lambda \frac{\sigma f(\hat{y})\hat{y}_z}{1-\sigma f(\hat{y})\hat{y}_\lambda}, \\ &= v_z \frac{1+\sigma f \hat{y}_z \left(\frac{v_\lambda(y,q,\lambda,k,\mu)}{v_z(y,q,\lambda,k,\mu)} - \frac{v_\lambda(\hat{y},q,\lambda,k,\mu)}{v_z(\hat{y},q,\lambda,k,\mu)}\right)}{1-\sigma f(\hat{y})\hat{y}_\lambda}. \end{aligned} \tag{12}$$

Thus we have

Proposition 1 *In equilibrium an increase in $z = q, k, \mu$ will make all users of the public hospital better off if the marginal rate of substitution of z for λ is constant or increasing with income.*

The marginal rate of substitution of z for λ ($-v_\lambda/v_z$) is the increase in z required to compensate the individual for an increase in λ. For the marginal individual with income $\hat{y}$ we know that $v_z^e(\hat{y}, q, k, \mu) > 0$ ($z = q, k, \mu$) and so, with an increasing marginal rate of substitution, all individuals with $y < \hat{y}$ are also made better off by an increase in z despite the induced increase in demand. The condition in Proposition 1 is satisfied for QS preferences (3). From Lemma 1 (11), the condition in Proposition 1 is also satisfied for $z = k, \mu$ (but not for q) for LN preferences (4).

2.3.1 Equilibrium mean wait

Denote the equilibrium mean waiting time as

$$w^e(q,k,\mu) = \bar{w}(\lambda(q,k,\mu), k, \mu).$$

An increase in quality increases demand and so always increases the equilibrium mean waiting time

$$w_q^e(q,k,\mu) = \bar{w}_\lambda(\lambda(q,k,\mu),k,\mu)\lambda_q > 0. \tag{13}$$

However, it is possible that the expected waiting time increases after an increase in supply because it also induces a change in demand[13]

$$w_z^e(q,k,\mu) = \bar{w}_z(\lambda(q,k,\mu),k,\mu) + \bar{w}_\lambda(\lambda(q,k,\mu),k,\mu)\lambda_z, \qquad (z = k, \mu) \tag{14}$$

This possibility would be worrying if we were considering a market in which the money price varies to equate demand and an exogenous supply: the only way in which price could increase following an increase in exogenous supply would be if demand was increasing in the price.

In the market for the public sector hospital, demand depends on the *distribution* of waiting times. In general we cannot interpret the mean wait as the price which adjusts to clear the market and hence need not be concerned about whether an increase in supply reduces or increases the expected wait. What matters is the demand response to an exogenous supply increase and, as we showed above (see (8) and (9)), an increase in

[13] Braess (1968) demonstrated that adding an additional connection in a road network can increase the journey time of all users. Cohen and Kelly (1990) provide an example of a stochastic queuing network in which adding an additional route increases the mean wait of all users.

supply induces a change in the distribution of waiting times which increases the expected utility for the marginal patient choosing public hospital and so increases demand.

To ensure that $w_z^e(q,k,\mu) < 0$ $(z = k, \mu)$ requires further restrictions on preferences or on the distribution of waiting times generated by the queueing system.

Proposition 2 *The equilibrium expected waiting time is decreasing in $z = \mu, k$ if (a) preferences are linear in waiting time (4) or (b) equal increases in the arrival rate λ and in supply z leave the mean waiting time unchanged ($\bar{w}_z = -\bar{w}_\lambda$).*

The proofs are in the Appendix. One example of a queueing process satisfying the second condition is $M/M/1$.[14] We stress that none of the subsequent analysis requires the assumption that an increase in supply reduces the equilibrium mean waiting time, though it is sometimes useful in interpreting some of the results on optimal pricing rules.

2.3.2 Equilibria in deterministic and stochastic models.

Almost all waiting time models in the health economics literature assume that both demand and supply and hence the waiting time are certain. The purpose of this section is to highlight some qualitative differences with our approach by comparing simple versions of the stochastic and deterministic waiting time models.

Suppose that the stochastic queueing system is $M/M/1$ with a single server and an exponential distribution of waiting times and that patients have linear preferences (4) and so are concerned only with the expected waiting time $\bar{w}$ which in the $M/M/1$ system is just $\bar{w} = 1/(\mu - \lambda)$. Hence demand is $\lambda(q, \bar{w})$, with $\lambda_q > 0$, $\lambda_{\bar{w}} < 0$. In Figure 1 , with the service rate set at μ^0 the expected waiting time $\bar{w} = 1/(\mu^0 - \lambda)$ increases with λ from $(1/\mu^0)$ at $\lambda = 0$ and tends to infinity as $\lambda \to \mu$. Because demand depends on the mean waiting time, the equilibrium is determined by the intersection of the downward sloping demand curve $\lambda(q^0, \bar{w})$ and the upward sloping expected waiting time locus $\bar{w} = 1/(\mu^0 - \lambda)$. The equilibrium mean wait $\bar{w}^0$ and the equilibrium expected number of patients treated per period is $\lambda^0 = \lambda(q^0, \bar{w}^0)$.[15] Note that the expected output (number of patients treated per period) is equal to expected demand $\lambda(q^0, \bar{w}^0)$ and strictly less than the service rate μ^0 which is the maximum possible expected output.

Suppose that a reduction in quality from q^0 to q^1 induces a parallel downward shift in the demand curve to $\lambda(q^1, \bar{w})$. The new equilibrium is $\bar{w}^1$ with lower mean waiting time and a smaller expected number of patients $\lambda^1 = \lambda(q^1, \bar{w}^1)$ treated.

Now consider a deterministic waiting time process. Since both demand and service times are certain so is the waiting time. In each period nature randomly picks a proportion σ of patients to become ill, so that the number falling ill each period is certain but each patient faces the probability σ of falling ill. Expected utility for a patient who will choose the public hospital when ill is

$$v^D(y, q, w^D) = \sigma u(y, q, w^D) + (1 - \sigma)\, u^N(y),$$

[14] In a "one bed"-hospital, $\overline{w} = \frac{1}{\mu - \lambda}$ so that $\overline{w}_\mu = -\overline{w}_\lambda$.

[15] In the usual stochastic queueing model demand is exogenous–there is no balking by patients, and the equilibrium mean wait is determined by the intersection of the $\bar{w} = 1/(\mu^0 - \lambda)$ locus and the vertical line at the exogenous arrival rate.

where w^D is the certain wait in the deterministic system. To ensure that the demand curve is the same as in the stochastic case shown in Figure 1 we assume that patient preferences are also given by (4). The certain demand for treatment in the public sector is

$$\lambda^D(q, w^D) = \sigma F(\hat{y}^D(q, w^D)),$$

where the threshold income $\hat{y}^D(q, w)$ is defined by

$$v^D(y, q, w^D) - V^o(y - \gamma) = 0.$$

Hence in Figure 1 LN preferences ensure that the deterministic and stochastic demand curves are identical: $\lambda^D = \lambda(q, w^D) = \lambda(q, \bar{w})$.

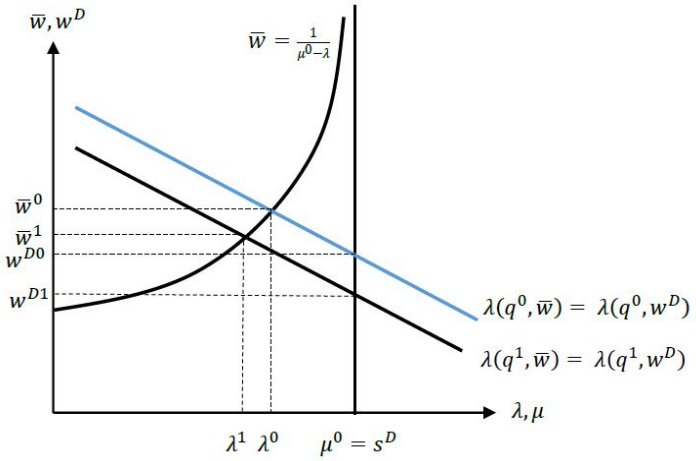


Figure 1. Equilibria in stochastic and deterministic waiting time models. $\overline{w}$ expected waiting time in stochastic model, w^D certain waiting time in deterministic model, $\lambda(q, \overline{w})$, expected demand in stochastic model, $\lambda(q, w^D)$ certain demand in deterministic model.

We denote the certain supply of treatments per unit of time as s^D. This is a function of hospital decisions affecting length of stay, number of beds and so on. To compare the stochastic and deterministic models we assume that the certain supply is equal to the number of beds times the service rate: $s^D = k\mu$, which in the special $M/M/1$ case in Figure 1 is just $s^D = \mu^0$.

With $v_{w^D} < 0$ the waiting time adjusts to clear the market: if λ^D exceeds s^D the number waiting to be treated will increase and so will the waiting time, reducing the inflow of new patients until $\lambda(q, w^D) = s^D$. Conversely if λ^D is less than s^D the waiting time will fall until either $\lambda(q, w^D) = s^D$ or $\lambda(q, 0) \leq s^D$. Since supply has a positive marginal cost the hospital will never choose to have $s^D > \lambda(q, 0)$.

In Figure 1 the initial equilibrium waiting time w^{D0} in the deterministic case is given by the intersection of $\lambda(q^0, w^D)$ and the vertical supply curve at $s^D = \mu^0$. The certain equilibrium waiting time is less than the mean wait in the stochastic equilibrium because with certain demand and supply there will never be unused capacity: effective supply is larger in the deterministic case and equal to the certain demand.

Increases in supply have the same qualitative effects in the deterministic model and the stochastic models: equilibrium waiting time falls and the (expected) number of patients treated increases.[16] But the implications of a demand shift are different. A reduction in quality from q^0 to q^1 with unchanged supply will reduce the equilibrium certain waiting time to $w^{D1}$and the equilibrium expected waiting time to $\bar{w}^1$. In the stochastic case the equilibrium expected output (patients treated) will also decrease to $\lambda^1 = \lambda(q^1, \bar{w}^1)$. But in the deterministic case there is no reduction in equilibrium output which, by assumption is equal to the unchanged supply. The deterministic equilibrium is re-established solely by a reduction in waiting time so that demand is equated to unchanged supply. Thus in the deterministic case a hospital whose revenue varies only with the number of patients treated could increase its profit by reducing quality, shifting the certain demand curve downwards, and allowing the certain waiting time to fall to keep demand and output unchanged. In the stochastic waiting time specification the hospital would lose revenue if it reduced quality. We examine the implications of this crucial difference between stochastic and deterministic waiting time models for deriving optimal hospital payment rules in Section 5.

2.4 Welfare function

2.4.1 Patient welfare

The patient welfare function is additive over individuals, with the benefit to an individual being $b^e(y, q, k, \mu) = b(y, q, \lambda(q, k, \mu), k, \mu)$, $(y \in [y_{\min}, y_{\max}])$ and total patient welfare

$$B^e(q, k, \mu) = \int_{y^{\min}}^{\hat{y}^e(q,k,\mu)} b^e(y, q, s)\mathrm{d}F(y) + \int_{\hat{y}^e(q,k,\mu)}^{y^{\max}} b^o(y, q^o)\mathrm{d}F(y),$$

with, for $z = q, k, \mu$,

$$B^e_z(q, k, \mu) = \int_{y^{\min}}^{\hat{y}^e(q,k,\mu)} b^e_z(y, q, k, \mu)\mathrm{d}F(y) + \left[b^e(\hat{y}, q, k, \mu) - b^o(\hat{y} - \gamma, q^o)\right] f\left(\hat{y}^e\right) \hat{y}^e_z. \quad (15)$$

We allow for the fact that the individual benefit $b(y, \cdot)$ which the regulator takes into account, may not coincide with the expected utility $v^e(y, \cdot)$ on which citizens with income y base their decision. Hence, the welfare of the marginal patient may not be the same in the public and private hospital.[17] The specification also reflects the assumption that it is not possible to directly affect the decision to seek public treatment except via q, k or μ, so that $\hat{y}$ is determined by patient decisions, not by the regulator. In addition, we assume that the level of q, k and μ in the public sector does not affect the insurance premium, quality, or the waiting time in the private sector. With a utilitarian welfare function respecting patient preferences $b^e(y, q, k, \mu) = v^e(y, q, k, \mu)$, $b^o = v^o(y - \gamma)$ and the

[16] Although by less than the capacity increase in the stochastic model since $\lambda_\mu < 1$ (cf (8)).

[17] For example, the regulator may be of the opinion that the individual benefit should not vary with income ($b(q, k, \lambda, \mu)$), or that the marginal benefit of attributes should be income independent ($b_z(q, \lambda, k, \mu)$), or that the marginal willingness to pay should be independent of income: $\frac{b_z(y,q,k,\lambda,\mu)}{b_y(y,q,k,\lambda,\mu)} = \frac{b_z(q,k,\lambda,\mu)}{b_y(q,k,\lambda,\mu)}$. This last is similar to what Tobin (1970: 264) called *specific egalitarianism* ("the view that certain specific scarce commodities should be distributed less unequally than the ability to pay for them.")

last term in (15) vanishes because the marginal patient is indifferent between the sectors (see (5)).

In one specification of the welfare function, with implications which we discuss in Section 3, patients have quasi separable preferences (3) which are partly respected by the welfare function. The welfare of a patient choosing the public sector is

$$b^e(y,q,k,\mu) = \sigma[m^0(y) + m^1(y)m^2(R^e(q,k,\mu))] + (1-\sigma)m^N(y), \qquad m^2_R > 0. \tag{16}$$

In this case the welfare function respects patient preferences $R^e(q,k,\mu) = R(\lambda(q,k,\mu),q,k,\mu)$ over characteristics of hospital treatment in the sense that the regulator's marginal rate of substitution of q for k or μ is the same as that of the patient. But the monetary valuation of hospital treatment characteristics, and so the willingness to pay, may differ.

2.4.2 Costs

The second component of the welfare function is the public hospital's expected cost

$$c^{He}(q,k,\mu) = c^H(q,\lambda(q,k,\mu),k,\mu).$$

We assume that increasing quality is costly ($c^H_q > 0$) as are supply decisions ($c^H_z > 0,\ z = k,\mu$) which induce a more favourable distribution of waiting times. We also allow for the possibility that expected hospital cost depend on the expected number of patients treated (λ), for example because each patient treated requires drugs and other consumables.[18] The marginal cost of expected output is $c^H_\lambda > 0$ and so $c^{He}_z(q,k,\mu) = c^H_z + c^H_\lambda \lambda_z > 0$, ($z = q,k,\mu$). We ignore, until section 4, the possibility that the cost of producing treatments of given quality can be affected by cost reducing effort.

We assume that patients do not work whilst waiting and are fully compensated for lost earnings with the cost of lost output borne by a social insurance fund. We normalise labour supply when healthy to 1. Since the mean wait is $w^e(q,k,\mu)$, the expected total compensation payment from the insurance fund (c^{Ie}) in equilibrium is

$$c^{Ie}(q,k,\mu) = \sigma w^e(q,k,\mu) \int_{y_{\min}}^{\hat{y}^e(q,k,\mu)} y \mathrm{d}F(y).$$

An increase in attribute $z(= q,k,\mu)$ alters expected insurance cost both by changing the expected waiting time (waiting time effect) and by changing the number of individuals waiting (waiting list effect):

$$c^{Ie}_z = \sigma w^e_z \int_{y_{\min}}^{\hat{y}^e(q,s)} y \mathrm{d}F(y) + \sigma w^e(q,s)\,\hat{y}^e f(\hat{y}^e)\hat{y}^e_z. \tag{17}$$

Since $\hat{y}^e_z > 0$ ($z = q,k,\mu$), increases in quality and supply conditions induce richer patients to join the waiting list and so increase the income loss at a given mean wait. An increase in quality also increases the waiting time ($w^e_q = \bar{w}_\lambda(\lambda,k,\mu)\lambda_q > 0$) and hence always increase

[18] We assume that in the steady state equilibrium the expected output rate equals the expected arrival rate λ. In queueing theory, this is property is known as *Burke's Theorem* (Burke, 1956) and holds for the $M/M/k$ system.

the compensation payment: $c_q^{Ie} > 0$. As we noted in section 2.3.1 in general the effect of an increase in supply on the mean waiting time is ambiguous because whilst the increase in supply reduces the mean wait it also induces a partially offsetting increase in demand which increases the mean wait: $w_z^e = \bar{w}_z + \bar{w}_\lambda \lambda_z$ $(z = k, \mu)$. If we make the intuitive assumption that $w_z^e < 0$ the sign of c_z^{Ie} is ambiguous.

The regulator's objective function is[19,20]

$$A^e(q, k, \mu) \stackrel{\text{def}}{=} B^e(q, k, \mu) - (1 + \theta)\, C^e(q, k, \mu),$$

where θ is the marginal cost of public funds and

$$C^e(q, k, \mu) \stackrel{\text{def}}{=} C(q, \lambda(q, k, \mu), k, \mu) = c^{He}(q, k, \mu) + c^{Ie}(q, k, \mu). \tag{18}$$

In the next sections, we inquire about the optimal hospital payment schemes under different assumptions about which hospital decisions and outcomes can be observed.

3 Optimal payment schemes

We first derive first best levels of hospital quality and supply and then examine how they can be implemented with payment schemes.[21]

3.1 First best regulation

Social welfare depends on hospital decisions about treatment quality (q), treatment intensity (μ) and the number of beds (k). The first best levels of these attributes satisfy the first order conditions

$$A_s^e = B_s^e - (1 + \theta)C_s^e = B_q - (1 + \theta)C_s + [B_\lambda - (1 + \theta)C_\lambda]\lambda_s = 0, \quad (s = \mu, k), \tag{19}$$

$$A_q^e = B_q^e - (1 + \theta)C_q^e = B_q - (1 + \theta)C_q + [B_\lambda - (1 + \theta)C_\lambda]\lambda_q \leq 0, \quad q \geq 0. \tag{20}$$

The condition on q holds with complementary slackness: we allow for the possibility that the first best quality is minimal ($q = 0$) but ignore the trivial solution where no patients

[19] One set of assumptions which yields this form is that the regulator is only concerned with patient welfare and tax financed public expenditure, and sets a lump sum tax or subsidy so that provider just breaks even financially after any incentive payments. Or welfare is the sum of patient benefit and the hospital utility and the lump sum tax or subsidy drives hospital utility to zero.

[20] Implicitly, we renormalise patient benefit to make it commensurable with the currency that costs are measured in. Since both patient utility (v) or the regulator's perception of that utility (b) are cardinal functions, these can be rescaled by a positive constant.

Note that we could reformulate our model in terms of the willingness to pay for public treatment, $P(y, \cdot)$, defined as $v(y - P(y, \cdot), \cdot) = v^0(y - \gamma)$ and measure social surplus as $\int_{y^{\min}}^{\hat{y}^e(q,\cdot)} P^e(y, \cdot)\mathrm{d}F(y)$. The critical citizen would have a willingness to pay $P(\widehat{y}, \cdot) = 0$. Paternalistic social preferences would replace $P^e(y, \cdot)$ by some other WTP function, also measured in the same currency as income and costs.

[21] Strictly, we are considering the second best because we are not giving the regulator the means to directly control patient demand. Such control would correct for the externality that arises because the marginal patient ignores the effect of her decision to join the waiting list on the average waiting time. See Noar (1969), Littlechild (1974), and Edelson and Hildebrand (1975) on policies to control decisions to join the queue.

are treated in the public sector ($\mu = 0$ and $k = 0$).

To examine the circumstances in which first best quality is positive we use (19) to substitute $[B_s - (1+\theta)C_s]/\lambda_s$ for $[B_\lambda - (1+\theta)C_\lambda]$ in (20) to get

$$A_q^e = \left(B_q - B_s\frac{\lambda_q}{\lambda_s}\right) - (1+\theta)\left(C_q - C_s\frac{\lambda_q}{\lambda_s}\right) \ (s = \mu, k). \tag{21}$$

The rate at which s must be reduced to keep demand constant after an increase in q is $(-\lambda_q/\lambda_s)$ so that first best quality is positive if, starting from $q = 0$, the net increase in patient welfare from such a reform exceeds the net increase in production and insurance cost.[22]

From Theorem A.3 in the appendix

$$B_q^e - B_s^e\frac{\lambda_q}{\lambda_s} = B_q - B_s\frac{\lambda_s}{\lambda_q} = \int_{y_{\min}}^{\hat{y}} b_s(y,\cdot)\left[\frac{b_q(y,\cdot)}{b_s(y,\cdot)} - \frac{v_q(\hat{y},\cdot)}{v_s(\hat{y},\cdot)}\right]\mathrm{d}F(y). \tag{22}$$

We can sign the first term in (21) under some assumptions about welfare and patient preferences. If the welfare function respects patient preferences in the sense that the regulator's marginal rates of substitution between quality and supply variables are equal to those of patients then (22) is positive if the average marginal welfare valuation of q in terms of s is greater than the marginal valuation revealed by the demand responses of the marginal patient who is indifferent between the public and private hospital. If patient preferences are quasi-separable and are respected in the welfare function then (22) is zero: the reform does not increase patient benefit.

Next consider the net cost implications of increasing q and adjusting s ($s = \mu, k$) to keep demand constant. From (18) and Lemma A5 (A.8), we have

$$\begin{aligned} C_q - C_s\frac{\lambda_q}{\lambda_s} &= \left(c_q^H - c_s^H\frac{\lambda_q}{\lambda_s}\right) + \left(c_q^I - c_s^I\frac{\lambda_q}{\lambda_s}\right), \\ &= \left(c_q^H - c_s^H\frac{\lambda_q}{\lambda_s}\right) - \sigma\bar{w}_s\int_{y^{\min}}^{\hat{y}} y\mathrm{d}F\frac{\lambda_q}{\lambda_s} \ (s = \mu, k). \end{aligned}$$

If the marginal hospital cost of quality at zero quality is small then the overall effect of the reform is to reduce hospital cost. Since the reform keeps demand constant the number of people who require income compensation is unchanged and the effect of the reform is to increase expected insurance cost because the reduction in attribute s increases the average wait. Thus while the overall effect of the reform on cost is ambiguous, we may expect it to be negative if minimal quality in the public hospital encourages most citizens to take out private insurance. So we conclude that first best quality is more likely to be above its minimal level when (i) patients with lower incomes have a higher marginal valuation of quality in terms of s, (ii) the marginal hospital cost of quality is low when quality is minimal, and (iii) the effect of increases in the equilibrium wait on the total income lost due to waiting is small at minimal quality.

[22] Equivalently, multiplying (19) through by $\frac{\lambda_q}{\lambda_s}$ and subtracting from (20) yields (21). By construction, this reform gets rid of the indirect effects on B and C due to changes in demand so that $B_q^e - B_s^e\frac{\lambda_q}{\lambda_s} = B_q - B_s\frac{\lambda_q}{\lambda_s}$ and $C_q^e - C_s^e\frac{\lambda_q}{\lambda_s} = C_q - C_s\frac{\lambda_q}{\lambda_s}$.

In what follows, we will assume that first best quality always exceeds the minimal level.

3.2 First best payment schemes

In general, we require as many policy instruments with linearly independent effects on hospital decisions as the hospital has decision variables. The hospital makes decisions on q, μ and k which result in an expected number of treatments, $\lambda(q, \mu, k)$, as well as an expected waiting time $\overline{w}^e(q, \mu, k)$. We will assume that it is always possible to observe output λ and so to set a prospective price per completed treatment p_λ. It may also be possible to attach a price to quality p_q (a pay for performance scheme), a (possibly negative) price to the average wait ($p_{\bar{w}}$), and to reward supply decisions via p_μ and to have a beds subsidy p_k.[23] With five instruments available to influence three provider decisions there are 10 possible first best schemes. Given the increased use of prospective output pricing, we examine three of the six pricing schemes which include a prospective output price p_λ. In Section 4 we consider some examples of second best pricing schemes in which there is only one other instrument available in addition to the prospective price. We then discuss the third best prospective price when there are no other instruments. Finally, we allow for the possibility that unobserved provider effort affects cost and consider cost reimbursement rules.

3.2.1 Payment for output, beds and average length of stay

We first assume that the risk neutral public hospital receives a payment per patient treated, p_λ, per bed installed, p_k, and per unit of service rate, p_μ – recall that average length of stay is $\frac{1}{\mu}$. The risk neutral public hospital chooses q, k and μ to maximise a weighted sum of expected profit and patient welfare with $\alpha \geq 0$ reflecting the hospital's degree of concern for patients. It also receives a lump sum transfer T (possibly negative) to ensure that it breaks even:[24]

$$\max_{q,k,\mu} p_\lambda \lambda(q, k, \mu) + p_k k + p_\mu \mu - c^{He}(q, k, \mu) + \alpha B^e(q, k, \mu) + T \tag{23}$$

First order conditions for an interior solution are

$$p_\lambda \lambda_z + p_k \mathbf{1}_{(z=k)} + p_\mu \mathbf{1}_{(z=\mu)} + \alpha B^e_z = c^{He}_z \quad (z = q, k, \mu) \tag{24}$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function equal to 1 if the condition $(\cdot)$ is true and equal to zero otherwise. In the appendix (section A.5) we use a general approach to derive first, second and third best prices. Here, we will focus on the main results and their interpretation. To shorten notation, we will define the *residual marginal social benefit* (RMSB) of decision z as $S^e_z \stackrel{\text{def}}{=} \beta B^e_z - c^{Ie}_z$, where $\beta \stackrel{\text{def}}{=} \frac{1-\alpha(1+\theta)}{1+\theta}$. S_z denotes that part of the social welfare effect of decision z which is not internalised by the hospital. Recall that the hospital

[23] In addition to the prospective output price Chalkley and Malcomson (1998) also consider linking payment to the number of patients not treated but added to a waiting list. However, the costs of deferred treatment for these patients is implicitly assumed to be zero and so both the first best and the payment mechanism required to achieve it take no account of the costs of rationing demand.

[24] Hospital profit is $\pi = p_\lambda \lambda(q, k, \mu) + p_k k + p_\mu \mu - c^{He}(q, k, \mu) + T$. For π to be zero, $T = -p_\lambda \lambda - p_k k - p_\mu \mu + c(q, k, \mu)$. The hospital perceives T as lump sum.

takes into account a fraction α of patient benefit, as well as the entire hospital cost (see (23)). Hence, S_z^e, is the remaining part, 'discounted' by the marginal cost of public funds. The conditions for first best imply that $S_z^e > 0$ $(z = q, k, \mu)$. First best prices on quality, beds and service rate would be set at S_z^e $(z = q, k, \mu)$, the superscript e indicating that demand responses are taken into account–a result of the fact that demand is not directly controlled by the planner: thus $S_z^e = S_z + S_\lambda \lambda_z$ $(z = q, k, \mu)$. However, we are assuming that quality is not observed and that instead a price is attached to output, λ, which is a good substitute because demand (and therefore output) is sensitive to quality of treatment. In the Appendix, we show

Proposition 3 *The first best prices per treated patient, bed and unit of service rate are*

$$p_\lambda^{FB*} = \frac{S_q^e}{\lambda_q}, \tag{25}$$

$$p_k^{FB*} = S_k^e - p_\lambda^{FB*}\lambda_k \tag{26}$$

$$p_\mu^{FB*} = S_\mu^e - p_\lambda^{FB*}\lambda_\mu \tag{27}$$

where all terms on the right hand sides are evaluated at the first best quality, service rate and number of beds.

The first expression reflects the fact that rewarding output incentivises quality choice. Since $\frac{S_q^e}{\lambda_q} = \beta\frac{B_q^e}{\lambda_q} - \frac{c_q^{Ie}}{\lambda_q}$, the first best output price p_λ^{FB*} is less than the marginal social benefit per patient attracted by higher quality $\left(\frac{B_q^e}{\lambda_q}\right)$ to the extent that (i) hospitals are intrinsically motivated, (ii) raising public funds is costly, and (iii) a quality increase results in larger social insurance expenditure because it attracts more to public treatment if ill and therefore increases the waiting time and the waiting list. To bring this out starkly, suppose that the provider is not altruistic ($\alpha = 0$) and that there is no loss of earnings whilst waiting for treatment. Then (25) can be written as $p_\lambda^{FB}\lambda_q = B_q^e/(1+\theta)$, so that the provider's marginal revenue from increasing quality should be less than the marginal patient welfare from higher quality only because of the marginal deadweight cost of public funds.

The remaining prices are adjusted for the fact that choice of beds and service rate are also rewarded through output, and should therefore be rewarded at a lower rate than their RMSB would require. Substituting for p_λ^{FB*}, the prices for z $(z = k, \mu)$ are $p_z^{FB*} = S_z^e - S_q^e\frac{\lambda_z}{\lambda_q}$ and making use of (A.3) and (A.7), they can be written as

$$p_z^{FB*} = \beta\int_{y^{\min}}^{\hat{y}^e(q,k,\mu)} b_q(y,\cdot)\left[\frac{b_z(y,\cdot)}{b_q(y,\cdot)} - \frac{v_z(\widehat{y},\cdot)}{v_q(\widehat{y},\cdot)}\right]\mathrm{d}F(y)$$
$$-\sigma w_z(q,k,\mu)\int_{y^{\min}}^{\hat{y}^e(q,k,\mu)} y\mathrm{d}F(y) \quad (z = k, \mu). \tag{28}$$

The first term is the effect of an increase in hospital decision z $(z = k, \mu)$ on expected patient benefit when quality is simultaneously reduced in order to keep demand, λ, constant. $\frac{b_z}{b_q}$ is the regulator's perception of citizen y's marginal valuation for quality in terms of hospital attribute z, while $\frac{\widehat{v}_z}{\widehat{v}_q}$ is the corresponding valuation for individual $\hat{y}$. If the

regulator respects individual preferences, $\frac{b_z(y,\cdot)}{b_q(y,\cdot)} = \frac{v_z(y,\cdot)}{v_q(y,\cdot)}$ and the first term will be positive (negative) if the marginal valuation for quality in terms of income increases uniformly faster (slower) with income than that for attribute z (cf (A.6)). If both increase equally fast, which is the case of quasi-separable preferences, the first term vanishes. The first term also vanishes when the regulator's perception of the preferences of the person with income $\hat{y}$ respects their quasi-separable structure. The second term is the reduction in sickness leave compensation because of a reduction in average wait. Both effects call for a positive incentive for supply decision z ($z = k, \mu$).

3.2.2 Payment for patients, beds, and quality

Next suppose that the prospective output price is combined with prices linked to quality and the number of beds. The rationale given for the pricing structure stated in Proposition 3 immediately suggests

Proposition 4 *The first best prices for output, beds, and quality are*

$$\begin{aligned} p_\lambda^{FB**} &= \frac{S_\mu^e}{\lambda_\mu} \\ p_k^{FB**} &= S_k^e - p_\lambda^{FB**}\lambda_k \\ p_q^{FB**} &= S_q^e - p_\lambda^{FB**}\lambda_q \end{aligned}$$

where all terms on the right hand sides are evaluated at the first best quality, service rate and number of beds.

Since μ is not priced, rewarding treated patients becomes a substitute for rewarding the service rate. The rewards per bed and quality unit reflect the MRSB, marked-down to take into account that these attributes are indirectly rewarded through the prospective output price.

3.2.3 Payment for output, beds and waiting times

Now consider combining a prospective price for output with a beds subsidy and a price linked to the mean waiting time. In the hospital's objective function (23) $p_w w^e(q,k,\mu)$ is substituted for $p_\mu \mu$, while the generic first order condition for attribute z turns into

$$p_\lambda \lambda_z + p_w w_z^e + p_k \mathbf{1}_{(z=k)} + \alpha B_z^e = c_z^{He} \quad (z = q, k, \mu).$$

In the appendix, we prove

Proposition 5 *The first best prices for mean wait, treated patients, and beds are*

$$p_w^{FB+} = \frac{\left(-\frac{dS}{d\mu}\Big|_{d\lambda=0, dk=0} \right)}{(-w_\mu)}, \tag{29}$$

$$p_\lambda^{FB+} = \frac{S_q^e}{\lambda_q} - p_w^{FB+} w_\lambda, \tag{30}$$

$$p_k^{FB+} = S_k^e - p_\lambda^{FB+}\lambda_k - p_w^{FB+} w_k. \tag{31}$$

The substitution of w for μ in the reward scheme (27) intuitively replaces $p_{\mu}^{FB*} = \left(-\frac{\mathrm{d}S}{\mathrm{d}\mu}\Big|_{\mathrm{d}\lambda=0,\mathrm{d}k=0}\right)$ with (29). Since a reward for the waiting time now indirectly rewards output as well to the extent that the average wait is responsive to demand ($w_{\lambda} > 0$), this leads to a smaller first best reward per patient treated as given by (30). The reward per bed is adjusted accordingly (31).[25]

Since $w_{\mu} < 0$, it transpires from (29) that p_w^{FB+} should take the opposite sign of p_{μ}^{FB*}, given by (28). Thus if $\frac{b_{\mu}(y,\cdot)}{b_q(y,\cdot)}$ is sufficiently smaller than $\frac{\widehat{v}_{\mu}}{\widehat{v}_q}$ (e.g., when $\frac{b_{\mu}(y,\cdot)}{b_q(y,\cdot)} = \frac{v_{\mu}(y,\cdot)}{v_q(y,\cdot)}$ and the marginal valuation for the service rate in terms of income rises sufficiently faster with income than that for quality) low income people are regarded to be better off with higher quality and length of stay. It is then optimal to have a negative price on the service rate and therefore a positive price on the average wait. The waiting time is then a signal of quality and length of stay, and subsidising it will promote both attributes.

4 Second and third best prospective output pricing

4.1 Second best output and mean wait prices

We now assume that the regulator can observe only the output and the mean waiting time. The hospital's objective function is now $p_{\lambda}\lambda(q,k,\mu) + p_w w^e(q,k,\mu) + \alpha B^e(q,k,\mu) - c^{He}(q,k,\mu)$ and its choices satisfy the first order conditions

$$p_{\lambda}\lambda_z + p_w w_z^e + \alpha B_z^e = c_z^{He} \quad (z = q,k,\mu). \tag{32}$$

Before inquiring about the optimal second best price levels for p_{λ} and p_w, notice that the hospital is left with some discretion in the second best since it makes decisions on three attributes, q, k and μ, while incentives are provided on two outcomes, λ and w. In response to a marginal change in the output price, the three attribute levels are adjusted, $\frac{\partial z}{\partial p_{\lambda}}$ $(z = q,k,\mu)$, resulting in an increase in the expected number of patients treated, $\frac{\partial \lambda^*}{\partial p_{\lambda}} \overset{\text{def}}{=} \sum_z \lambda_z \frac{\partial z}{\partial p_{\lambda}}$ and a change in the expected wait $\frac{\partial w^*}{\partial p_{\lambda}} \overset{\text{def}}{=} \sum_z w_z^e \frac{\partial z}{\partial p_{\lambda}}$.[26] Likewise, a marginal change in p_w triggers responses $\frac{\partial z}{\partial p_w}$ $(z = q,k,\mu)$ that result in an increase in the

[25] It can be shown that

$$\begin{pmatrix} p_{\lambda}^{FB+} \\ p_k^{FB+} \\ p_w^{FB+} \end{pmatrix} = \begin{pmatrix} 1 & 0 & -\frac{w_{\lambda}}{w_{\mu}} \\ 0 & 1 & -\frac{w_k}{w_{\mu}} \\ 0 & 0 & \frac{1}{w_{\mu}} \end{pmatrix} \begin{pmatrix} p_{\lambda}^{FB*} \\ p_k^{FB*} \\ p_{\mu}^{FB*} \end{pmatrix},$$

indicating the recursive relationship between (25)-(27) and (29)-(31). The ratios $-\frac{w_z}{w_{\mu}}$ $(z = \lambda, k)$ adjust the first best rewards for output and bed in a way that keeps the waiting time constant, and the reason is that w is now separately incentivised. A third way of writing the first best price for output is by using the equilibrium waiting time effects w_z^e $(z = \mu, q)$:

$$p_{\lambda}^{FB+} = \frac{(-w_{\mu}^e \lambda_q)}{\lambda_{\mu} w_q^e + (-\lambda_q w_{\mu}^e)} \frac{S_q^e}{\lambda_q} + \frac{w_q^e \lambda_{\mu}}{\lambda_{\mu} w_q^e + (-\lambda_q w_{\mu}^e)} \frac{S_{\mu}^e}{\lambda_{\mu}},$$

i.e., a weighted average of the MRSB for quality and service rate, the two hospital attributes that are not directly rewarded by the present scheme.

[26] The equilibrium demand and expected waiting time are $\lambda^*(p_{\lambda}, p_w) = \lambda(q(p_{\lambda}, p_w), k(p_{\lambda}, p_w), \mu(p_{\lambda}, p_w))$ and $w^*(p_{\lambda}, p_w) = w^e(q(p_{\lambda}, p_w), k(p_{\lambda}, p_w), \mu(p_{\lambda}, p_w))$.

expected wait, $\frac{\partial w^*}{\partial p_w} \stackrel{\text{def}}{=} \sum_z w_z^e \frac{\partial z}{\partial p_w}$ and a change in expected output $\frac{\partial \lambda^*}{\partial p_w} \stackrel{\text{def}}{=} \sum_z \lambda_z \frac{\partial z}{\partial p_w}$.[27] Consider now the *induced response* change in output following an increase in p_λ *when* w^* *is kept fixed*. This hospital response is given by $\frac{\partial \lambda^*}{\partial p_\lambda} - \frac{\partial \lambda^*}{\partial p_w}\left(\frac{\partial w^*}{\partial p_w}\right)^{-1}\frac{\partial w^*}{\partial p_\lambda}$.[28] Likewise, the optimal change in bed use following an increase in p_λ *when* w^* *is kept fixed*, is given by $\frac{\partial k}{\partial p_\lambda} - \frac{\partial k}{\partial p_w}\left(\frac{\partial w^*}{\partial p_w}\right)^{-1}\frac{\partial w^*}{\partial p_\lambda}$. As a result, when w^* is kept constant, the induced response effect on outcome λ per unit change in decision k can be defined as

$$\lambda_k^{\text{ir}}|_{\text{d}w=0} \stackrel{\text{def}}{=} \frac{\frac{\partial \lambda^*}{\partial p_\lambda} - \frac{\partial \lambda^*}{\partial p_w}\left(\frac{\partial w^*}{\partial p_w}\right)^{-1}\frac{\partial w^*}{\partial p_\lambda}}{\frac{\partial k}{\partial p_\lambda} - \frac{\partial k}{\partial p_w}\left(\frac{\partial w^*}{\partial p_w}\right)^{-1}\frac{\partial w^*}{\partial p_\lambda}}. \tag{33}$$

In the same vein, we can define the induced response effect on outcome w^e per unit change in k *when keeping* λ *fixed*:

$$w_k^{\text{ir}}|_{\text{d}\lambda=0} \stackrel{\text{def}}{=} \frac{\frac{\partial w^*}{\partial p_w} - \frac{\partial w^*}{\partial p_\lambda}\left(\frac{\partial \lambda^*}{\partial p_\lambda}\right)^{-1}\frac{\partial \lambda^*}{\partial p_w}}{\frac{\partial k}{\partial p_w} - \frac{\partial k}{\partial p_\lambda}\left(\frac{\partial \lambda^*}{\partial p_\lambda}\right)^{-1}\frac{\partial \lambda^*}{\partial p_w}}. \tag{34}$$

We can now state our second main result.

Proposition 6 *The second best prices for output and waiting time that decentralise the second best allocation are respectively*

$$p_\lambda^{SB} = p_\lambda^{FB++} + \frac{p_k^{FB++}}{\lambda_k^{ir}|_{dw=0}}, \text{ and} \tag{35}$$

$$p_w^{SB} = p_w^{FB++} + \frac{p_k^{FB++}}{w_k^{ir}|_{d\lambda=0}}. \tag{36}$$

where p_λ^{FB++}, p_w^{FB++} *and* p_k^{FB++} *are the first best prices per treatment, waiting time and bed as specified by (30), (29) and (31), but evaluated at the second best allocation* (q, k, μ).

We interpret these pricing rules as follows. In second best, the number of beds can no longer be directly incentivised. Therefore the prices on output and average wait should take over the rôle that p_k plays in first best. The optimal SB price for output is the FB price *plus* the FB reward per patient attracted by an extra bed. However, since the waiting time is steered through p_w, the proper number of attracted patients to use is the one given by $\lambda_k^{\text{ir}}|_{\text{d}w=0}$. Similarly, the second best price per wait is the first best price *minus* the first best reward per week reduction in average wait by installing an extra bed and

[27] Both $\frac{\partial \lambda^*}{\partial p_\lambda} > 0$ and $\frac{\partial w^*}{\partial p_w} > 0$ follow immediately from the second order condition for an optimal choice of z $(= q, k, \mu)$.

[28] The direct best response is $\frac{\partial \lambda^*}{\partial p_\lambda}$. However, since $\text{d}p_\lambda$ also triggers a change in waiting time $\frac{\partial w^*}{\partial p_\lambda}$ and since w needs to be kept constant, it is as if the (virtual) price per wait is reduced with $\left(\frac{\partial w^*}{\partial p_w}\right)^{-1}\frac{\partial w^*}{\partial p_\lambda}$, which then triggers an indirect response of λ^* equal to $-\frac{\partial \lambda^*}{\partial p_w}\left(\frac{\partial w^*}{\partial p_w}\right)^{-1}\frac{\partial w^*}{\partial p_\lambda}$. The virtual price interpretation is due to Neary and Roberts (1980).

keeping output constant (assuming $w_k^{\text{ir}}|_{\text{d}\lambda=0} < 0$). Thus the fact that w (λ) is still priced calls for the use of induced responses in the pricing rule for λ (w) that are conditioned on the outcome w (λ).

4.2 Can a price on output and waiting time achieve the first best?

From Proposition 6 it transpires that the second best prices p_λ^{SB}, p_w^{SB} will coincide with their first best counterparts if $p_k^{FB++} = 0$. The price p_k enables the regulator to directly influence the hospital choice of (μ, k). Note first that if the regulator can use p_λ and p_w, he can control both demand (and therefore the length of the waiting list) and the average waiting time, and therefore the social insurance cost c^I. Thus from this perspective, there is no need to influence the choice of (μ, k). The other reason for doing so is that the hospital ignores a fraction $(1-\alpha)$ of the aggregate benefit B. This benefit, though, will be independent of the choice of (μ, k) if preferences are quasi-separable or if they are linear in the waiting time. In the former case, any change in either μ or k accompanied by an appropriate change in q to keep demand constant does not affect patient benefit. In the latter case, any change in either μ or k that leaves the expected wait constant will not affect patient benefit either. Thus there is no reason for the regulator to wish to directly influence the hospital choice of (μ, k). Formally

Proposition 7 *The first best allocation is achievable using only a prospective output price and a price on waiting time if (a) individual and social preferences are quasi-separable as in (3) and (16) or if (b) the expected waiting time is a sufficient statistic for the waiting time distribution, both w.r.t. individual preferences and w.r.t. social preferences, i.e.,* $\overline{u}(y, q, \lambda, k, \mu) = \overline{u}(y, q, \overline{w}(\lambda, k, \mu))$ *and* $b(y, q, \lambda, k, \mu) = b(y, q, \overline{w}(\lambda, k, \mu))$.

The regulator can achieve the first best without the use of p_k if the two instruments were p_λ and p_q only under the more stringent condition (b) of the proposition. In this case the (μ, k)-choice will affect c^I and B through its effect on the expected wait. But if preferences are linear in waiting time, the demand responses of the critical citizen depend on the expected wait. Hence, any choice of (μ, k) that leaves demand constant will have no external effects on B or c^I that require internalisation via an additional instrument.

In general, when the prospective output price p_λ is complemented with either p_k or p_μ, it will not be possible to achieve a first best allocation. For example, if p_k is used, the margins left for the hospital are μ and q. But since quality has no direct effect on the expected waiting time, any choice of (μ, q) by the hospital that keeps demand constant will always affect c^I. It will only be by accident that this effect is exactly offset by the effect on $(1-\alpha)B$ such that direct influence on the choice (μ, q) is required.

4.3 Third best prospective output price

If the hospital payment scheme can no longer be made contingent upon the expected waiting time either, the regulator has only one instrument left to incentivise the hospital: the prospective output price p_λ. Recall that $\frac{\partial \lambda^*}{\partial p_\lambda}$ and $\frac{\partial w^*}{\partial p_\lambda}$ are the increase in the number of expected treatments and the change in expected wait, following the hospital's responses $\frac{\partial z}{\partial p_\lambda}$ to an increase in the prospective price. Then we can define the *unconstrained* induced

responses in λ and w following the hospital responses to a marginal increase in p_λ as

$$\lambda_k^{\text{ir}} \overset{\text{def}}{=} \frac{\frac{\partial \lambda^*}{\partial p_\lambda}}{\frac{\partial k}{\partial p_\lambda}} \text{ and } w_\lambda^{\text{ir}} \overset{\text{def}}{=} \frac{\frac{\partial w^*}{\partial p_\lambda}}{\frac{\partial \lambda^*}{\partial p_\lambda}}. \tag{37}$$

Using these definitions we can characterise the third-best prospective output price:

Proposition 8 *The price for output that decentralises the third best allocation is given by*

$$p_\lambda^{TB} = p_\lambda^{FB+++} + w_\lambda^{ir} \times p_w^{FB+++} + \frac{p_k^{FB+++}}{\lambda_k^{ir}}. \tag{38}$$

where p_λ^{FB+++}, p_w^{FB+++} and p_k^{FB+++} are the first best prices per treatment and bed as specified by (30), (29) and (31) but evaluated at the third best allocation of q, μ and k.

Suppose $\frac{\partial w^*}{\partial p_\lambda} > 0$ and $\frac{\partial k}{\partial p_\lambda} > 0$. Then $w_\lambda^{\text{ir}} > 0$ and $\lambda_k^{\text{ir}} > 0$ and the third best output price is the first best price with two mark-ups: one equal to the first best reward for an extra bed per patient attracted and another equal to the extra reward that the first best price on waiting time (possible negative) would have triggered per attracted patient. Thus the third best price for output picks up the incentives on bed installment and service rate decisions (the latter through incentivising the waiting time). Unlike in (33) and (34) the induced responses defined in (37) are no longer conditioned on a fixed wait or demand. The reason is that the average wait is no longer separately incentivised. In this sense, the third best output price is less sophisticated than the second best price.

The prospective output price by itself yields the first best only under stringent assumptions:

Proposition 9 *The first best allocation is achievable using only a prospective output price if individual and social preferences are quasi-separable and sickness leave compensation is zero.*

With QS preferences, the only role for p_w is to make the hospital aware of the consequences of its choice of (μ, k) for social insurance outlays. Without any sickness leave compensation, e.g., because people can continue working while waiting for treatment, this role disappears. If people had LN preferences as in (4), additional instruments are required since (μ, k) will in general affect $\overline{w}$ and therefore patient benefit.

4.4 Cost reducing effort

So far we have ignored the possibility that the hospital can exert unobservable effort to reduce production cost c^H. If the hospital bears all the production cost it has the incentive to select efficient cost reducing effort. However, this presumes that the regulator has a sufficient set of instruments available in order to provide the hospital with the right incentives at the other margins – quality, service rate and number of beds.

If the regulator has insufficient instruments to control quality, service rate and number of beds, cost sharing will become optimal. For example, if the hospital is rewarded per treated patient and receives a price (possibly negative) per week of average wait,

but is not incentivised along the bed dimension, cost-sharing can be a useful second-best instrument because it indirectly subsidises the installation of extra beds. To make this claim more precise, suppose that the hospital is refunded a fraction φ of its production cost $c^{He}(q, \mu, k, t)$ where t is cost-reducing effort which imposes non-verifiable cost $g(t)$ on the hospital. Suppose also that it receives a lump sum to ensure that it breaks even financially[29] and in the second best faces a prices on output and waiting time and in the third best there is only a price on output. The cost sharing parameter and the prices determine its decisions on q, μ, k and t, which determine demand and the expected waiting time w^e. Define the induced response marginal costs of an extra bed as

$$MC_k^{\text{ir}}\big|_{\text{d}\lambda=0} \stackrel{\text{def}}{=} \frac{\left.\frac{\partial c^{H*}}{\partial \varphi}\right|_{\text{d}\lambda=0}}{\left.\frac{\partial k}{\partial \varphi}\right|_{\text{d}\lambda=0}}, \text{ and}$$

$$MC_k^{\text{ir}}\big|_{\text{d}\lambda=0,\text{d}w=0} \stackrel{\text{def}}{=} \frac{\left.\frac{\partial c^{H*}}{\partial \varphi}\right|_{\text{d}\lambda=0,\text{d}w=0}}{\left.\frac{\partial k}{\partial \varphi}\right|_{\text{d}\lambda=0,\text{d}w=0}},$$

where $\frac{\partial c^{H*}}{\partial \varphi}$ is the ensuing increase in hospital cost when the cost-share parameter increases and the hospital adjusts its decisions on q, μ, k and t to maximise its utility, $\frac{\partial k}{\partial \varphi}$ is the change in the number of beds and the vertical bars indicates that these induced responses are conditioned on the output rate and, in the second best, the waiting time remaining constant. Then we have (see Appendix):

Proposition 10 *Suppose that (i) the hospital can reduce its cost by exerting an effort which has an unverifiable cost, (ii) the regulator rewards per treated patient and possibly also per unit of waiting time, and (iii) the first-best reward per bed is p_k^{FB} and per unit of waiting time is p_w^{FB}. Then the second-best value for the cost-sharing parameter, φ^{SBcs}, and the third-best value of that parameter when waiting time cannot be rewarded, φ^{TBcs}, respectively satisfy*

$$\varphi^{SBcs} \times MC_k^{ir}|_{d\lambda=0}^{dw=0} = p_k^{FB++}, \text{ and} \tag{39}$$

$$\varphi^{TBcs} \times MC_k^{ir}|_{d\lambda=0} = p_k^{FB+++} + p_w^{FB+++} \times \widehat{w}_k^{ir}|_{d\lambda=0}. \tag{40}$$

where $^{++}$ ($^{+++}$) indicates that the first-best prices are evaluated at the optimal second (third) best values for the decision variables and $\widehat{w}_k^{ir}|_{d\lambda=0} \stackrel{def}{=} \frac{\frac{\partial w^{e*}}{\partial \varphi}|_{d\lambda=0}}{\frac{\partial k}{\partial \varphi}|_{d\lambda=0}}$, the induced response effect of beds on the expected wait.

In second-best, when the waiting time is observable and rewardable, the responses used to calculate the induced response marginal costs are conditional on λ and w being held constant (since both are directly rewarded). Expression (39) then shows that the optimal second-best refund of the marginal cost of installing an extra bed should coincide to the first-best bed reward; i.e., what is given as a direct subsidy in first-best should comes as a cost refund in second-best. When w can no longer be rewarded, we enter a third-best

[29] The break-even condition includes the non-verifiable cost of effort.

situation. The proper induced marginal cost of an extra bed is then based on hospital responses that keep λ constant (because λ is still directly rewarded). According to (40), the third-best refund of the induced marginal cost of an extra bed amounts to the first-best bed reward *plus* the extra revenue the hospital would receive through the first-best waiting time reward. Thus cost sharing act as a good (but not perfect) substitute for direct incentives at the cost of distorting the choice of cost-reducing effort.

5 Optimal pricing with certain waiting time

We now contrast our results with those from a deterministic waiting time model. With the same patient and regulator preferences over waiting time and quality as in the stochastic case, patient welfare is

$$B^D(q, w^D) = \int_{y_{\min}}^{\hat{y}^D} b(y, q, w^D) dF(y).$$

In addition to choosing quality, the hospital makes supply decisions which affect the maximum number of patients who can be treated each period (s^D). We do not need to specify these decisions in detail but to retain as many similarities with the stochastic specification as possible we could assume that hospital capacity is give by the number of beds multiplied by the number of patients treated in each bed per period: $s^D = k\mu$. Whether made by the regulator or the hospital, these decisions minimise the cost of treating any given volume of patients at a given quality. Hence we can write the hospital cost as $c^{HD}(\lambda^D, q, s^D)$ so that, as in the stochastic case we allow for direct treatment costs to increase with the number treated ($c^{HD}_\lambda > 0$).

The welfare function is then

$$A^D = B^D(q, w^D) - (1+\theta)\left[c^{HD}(\lambda^D, q, s^D) + c^{DI}(q, w^D)\right], \tag{41}$$

where

$$c^{DI} = \sigma w^D \int_{y_{\min}}^{\hat{y}^D(q, w^D)} y \mathrm{d}F(y)$$

is the insurance cost of lost working time.

The first best quality, waiting time, and supply are chosen to maximise (41) subject to non-negativity constraints on quality and the waiting time and the market clearing condition which determines the waiting time. Since supply is costly it will never be welfare maximising to have $\lambda^D > s^D$ and so we can substitute $\lambda^D(q, w^D)$ for s^D in the hospital cost function. The first order conditions for a welfare maximum are

$$B^D_q - (1+\theta)\left[c^{HD}_\lambda \lambda^D_q + c^{HD}_q + c^{HD}_s \lambda^D_q + c^{DI}_q\right] \leq 0, \qquad q \geq 0,$$

$$B^D_w - (1+\theta)\left[c^{HD}_\lambda \lambda^D_w + c^{HD}_s \lambda^D_w + c^{DI}_w\right] \leq 0, \qquad w^D \geq 0,$$

where the conditions on q and w^D hold with complementary slackness.

There are four possible first best configurations of quality and waiting time: (i) positive quality and waiting time; (ii) quality and waiting time both zero; (iii) positive quality and

zero waiting time; (iv) zero quality and positive waiting time. All four configurations are possible depending on the welfare function, patient preferences and the cost functions.

Consider the hospital's choice of quality and supply when it receives only a prospective price p_λ, with no direct reward for quality or penalty for waiting time. Its objective function is $\alpha B^D + p_\lambda \lambda^D - c^H(\lambda, q, s^D)$ which is maximised by choice of q, s^D and w^D subject to the market clearing constraint $s^D \geq \lambda^D(q, w^D)$ and to non-negativity constraints on q, w^D. Given that supply is costly we can substitute $\lambda^D(q, w^D)$ for s^D in the cost function and the first order conditions are, ignoring the case where the hospital produces no output,

$$\alpha B_q^D + (p_\lambda - c_\lambda^{HD} - c_s^{HD})\lambda_q^D - c_q^{HD} \leq 0, \qquad q \geq 0, \tag{42}$$

$$\alpha B_w^D + (p_\lambda - c_\lambda^{HD} - c_s^{HD})\lambda_w^D \leq 0, \qquad w^D \geq 0, \tag{43}$$

where the conditions on q and w^D hold with complementary slackness.

If the hospital's optimal waiting time is positive ($w^{D*} > 0$), then (43) implies that $-\alpha B_w^D/\lambda_w^D = (p_\lambda - c_\lambda^{HD} - c_s^{HD})$. Substituting $-\alpha B_w^D/\lambda_w^D$ for $(p_\lambda - c_\lambda^{HD} - c_s^{HD})$, the sign of the derivative of the hospital objective function with respect to q at $q = 0$, $w^{D*} > 0$ is

$$\text{sgn } \alpha \left(\frac{B_q^D}{\lambda_q^D} - \frac{B_q^D}{\lambda_w^D} \right) - \frac{c_q^{HD}}{\lambda_q^D}. \tag{44}$$

Conversely if the optimal quality is positive, analogous substitutions give the same expression for the sign of the derivative of the hospital objective function with respect to w^D at $q^{D*} > 0$, $w^D = 0$.

Hence we have

Proposition 11 *There is a level of altruism* $\alpha^o \equiv \frac{c_q^{HD}}{\lambda_q^D} \left(\left| \frac{B_q^D}{\lambda_q^D} - \frac{B_w^D}{\lambda_w^D} \right| \right)^{-1}$ *such that if* $\alpha \in [0, \alpha^o)$ *the hospital will never choose to have both positive quality and waiting time when rewarded only by a prospective output price.*

The intuition is stark if altruism is zero and hospital is a pure profit maximiser. Suppose that the hospital has a positive quality and waiting time. Then if it reduces quality and keeps supply constant, demand will be reduced and the equilibrium waiting time will fall to bring demand again in line with supply. Hence, hospital revenue is unaffected but production cost is reduced because quality has a positive marginal cost. Therefore it can never be profit maximising for a hospital to have both positive quality and waiting time in this deterministic setting.

Figure 2 illustrates. With a price less than p_λ^0 the hospital cannot cover its marginal cost of output at zero quality and so output is zero with demand entirely choked off by a waiting time $w \geq w^{D0}$. For a price in the range $(p_\lambda^0, p_\lambda^2)$ the hospital produces a positive output where the price equals its marginal production cost $c_\lambda^{HD}(\lambda^D, 0, \lambda^D) + c_s^{HD}(\lambda^D, 0, \lambda^D)$. Further increases in price increase output and reduce the waiting time. Thus with a price $p_\lambda^1 \in (p_\lambda^0, p_\lambda^2)$ output is s^1 and the waiting time w^{D1} clears the market so that $\lambda^D(0, w^{D1}) = s^1$. At a price p_λ^2 waiting time is zero and the hospital can get an additional patient, and hence revenue, only by raising quality by $1/\lambda_q^D$. Hence the full marginal cost of additional patients is $c_\lambda^{HD} + c_s^{HD} + c_q^{HD}/\lambda_q^D$. If the marginal

cost of quality at zero quality and output s^{D2} is positive there will be a range of prices $[p^2_\lambda, p^3_\lambda]$ over which output and quality do not vary with price. Once the price exceeds $p^3_\lambda = c^{HD}_\lambda(s^2, 0, s^2) + c^{HD}_s(s^2, 0, s^2) + c^{HD}_q(s^2, 0, s^2)/\lambda^D_q(0, 0)$ the hospital will increase quality, with zero waiting time, to attract additional patients as price increases. Thus at a price p^4_λ, there will be s^4 patients receiving immediate treatment.

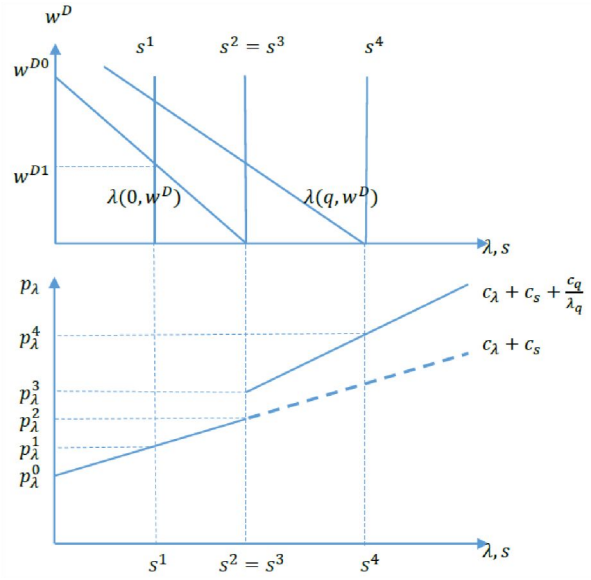


Figure 2. Effect of prospective output price on quality, output, and waiting time in deterministic waiting time model.

The implication of Proposition 11 is that with rationing by deterministic waiting time it is impossible to achieve a first best allocation which has positive quality and waiting time if the regulator is restricted to a prospective output price. If the demand for elective hospital care is rationed by waiting time then it is necessary to be specific about the rationing mechanism when modelling the behaviour of hospitals and examining the welfare implications of incentive schemes. But the apparently convenient simplification of assuming that demand and supply, and hence waiting time, are certain is not satisfactory in this context.

6 Conclusion

In this paper we have expanded previous analyses of optimal payment schemes for hospitals by taking explicit account of the facts that demand and treatment time are random and that patients care about waiting time. Queueing theory shows that under reasonable assumptions the stochastic processes generating demand and treatment time result in an equilibrium distribution of waiting times. However, unlike queueing theory which usually takes the demand (arrival) process as exogenous, we examine rational expectations equilibria in which patient decisions to join the waiting list depend on the equilibrium distribution of waiting times and their decisions give rise to the equilibrium distribution. Even when the mean waiting time is a sufficient statistic for patients when making their

provider choices, our stochastic model has properties that are qualitatively different from those that model the waiting time as a deterministic 'price' that balances demand and treatment capacity.

We develop a general queueing model and use it to derive the optimal first best and second best pricing rules for a hospital providing care to fully insured patients. In general three prices are required ensure that the hospital chooses the welfare maximising quality, the service intensity determining average length of stay, and the number of beds. The first best price per patient treated should be equal to the residual marginal social benefit of treating an extra patient. This is the part of the social benefit which the hospital ignores: some or all of the direct benefit of treatment plus the expected income loss to patients whilst they are on the waiting list. In addition to the first best prospective output price two other instruments are required. Candidates are a per bed subsidy, payments (possibly negative) linked to the average waiting time and average length of stay, and payment for quality. In a second best, when the regulator can only link payment to the number of patients treated and to the expected waiting time, we show that the optimal price per patient is the first-best price plus a mark-up that reflects the extra financing the hospital would have obtained in the first best by expanding the number of beds needed to attract an extra patient while at the same time keeping the expected wait constant (Proposition 6, (35)). And *vice versa*, the optimal reward per waiting time (possibly negative) should be set at the first best level minus the reward the hospital would have earned in first best by expanding the number of beds to reducing the expected wait while keeping (expected) demand constant ((36)). In other words, second-best prices should pick up the unavailable first best incentives, but the way they is done depends on the set of prices available in the second best.

When the hospital can exert effort to reduce its costs it is not necessary to directly incentivise such effort in the first best since it bears both production and effort costs. But in the second best it is optimal to refund a fraction of the total cost. This fraction should be set such that the refund of the marginal cost of installing an extra bed reflects the optimal first bed subsidy for such a bed (Proposition 10, (39)).

References

Aarbu, K. (2010). Demand patterns for treatment insurance in Norway. Ch 2 in *Empirical essays on risk taking and insurance*, PhD thesis, Norwegian School of Economics.

Besley, T., Hall, J., Preston I. (1999). The demand for private health insurance: Do waiting lists matter? *Journal of Public Economics* 72, 155–181.

Braess, D. (1968). Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung,* 12, 258–268. English translation: Braess, D., Nagurney, A., Wakolbinger, T. (2005). On a paradox of traffic planning. *Transportation Science*, 39, 446-450.

Burke, P., (1956). The output of a queuing system. *Operations Research* 4, 699-704.

Chalkley, M., Malcomson, J. (1998). Contracting for health services with unmonitored quality. *Economic Journal,* 108 (449), 1093-1110.

Chalkley, M., Malcomson, J. (2000). Government purchasing of health services. In Culyer, A. Newhouse, J. (eds.), *Handbook of Health Economics*, Volume 1. North Holland, Amsterdam.

Cohen, J., Kelly, F. (1990). A paradox of congestion in a queuing network. *Journal of Applied Probability*, 27, 730-734.

Cullis, J., Jones, P., Propper, C. (2000). Waiting lists and medical treatment: analysis and policies. In: Culyer, A.J., Newhouse, J.P. (Eds.), *Handbook of Health Economics*, North-Holland, Amsterdam

Edelson, N., Hildebrand, K. (1975). Congestion tolls for Poisson queueing processes. *Econometrica,* 43, 81-92.

Ellis, R., McGuire, T. (1986). Provider behavior under prospective reimbursement: cost sharing and supply, *Journal of Health Economics*, 5, 129-151.

Fomundam, S., Herrmann, J. (2007). A survey of queuing theory applications in healthcare. ISR report 2007-24, University of Maryland.

Goddard, J., Malek, M., Tavakoli, M. (1995). An economic model for hospital treamtent for non-urgent conditions, *Health Economics* 4, 41-55.

Gravelle, H., Dusheiko, M., Sutton, M. (2002).The demand for elective surgery in a public system: time and money prices in the UK National Health Service, *Journal of Health Economics*, 21, 423-449

Gravelle, H., Siciliani, L. (2008). Optimal quality, waits and charges in health insurance. *Journal of Health Economics*, 27, 663-674.

Gross, D. Shortle, J., Thompson, J., Harris, C. (2008). *Fundamentals of queueing theory* (3th ed.) (New York: Wiley).

Hassin, R., Haviv, M. (2003) *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Springer.

Iversen, T., Luras, H. (2002). Waiting time as a competitive device: and example from general medical practice. *International Journal of Health Care Finance and Economics*, 2, 189-204

Iversen, T., Siciliani, L. (2011). Non-price rationing and waiting times, in S. Glied and P. Smith (eds.) *The Oxford Handbook of Health Economics* (Oxford: OUP).

Jha, A., Joynt, K., Orav, E., Epstein, A. (2012). The long-term effect of Premier pay for performance on patient outcomes. *New England Journal of Medicine*, 366, 1606-1615.

Lindsay, C., Feigenbaum, B. (1984). Rationing by waiting lists, *Amercian Economic Review* 74, 404-417.

Littlechild, S. (1974). Optimal arrival rate in a simple queueing system. *International Journal of Production Research,* 12, 391-397

Marchand, M., Schroyen, F. (2005). Can a mixed health care system be desirable on equity grounds? *Scandinavian Journal of Economics* 107, 1-23.

Martin, S., Smith, P. (1999). Rationing by waiting lists: an empirical investigation, *Journal of Public Economics* 71, 141-164.

Naor, P. (1969). The regulation of queue size by levying tolls, *Econometrica*, 37, 15-24.

Neary, P., Roberts, K. (1980). The theory of household behaviour under rationing, *European Economic Review* 13, 25–42.

Paris, V., M. Devaux, M., Wei, L. (2010), Health Systems Institutional Characteristics: A Survey of 29 OECD Countries. OECD Health Working Papers, No. 50, OECD Publishing. http://dx.doi. org/ 10.1787/5kmfxfq9qbnr-en.

Siciliani, L., Moran, V., Borowitz, M. (2014). Measuring and comparing health care waiting times in OECD countries. *Health Policy*, 118, 292-303.

Siciliani, L, Iversen, T. (2012). Waiting times and waiting lists, in A. Jones (ed.) *The Elgar Companion To Health Economics.* 2nd ed. Edward Elgar Publishing.

Sivey, P. (2012). The effect of waiting time and distance on hospital choice for English cataract patients. *Health Economics*, 21, 444-456.

Sutton, M., Nikolova, S., Boaden, R., Lester, H., McDonald, R., Roland, M. (2012). Reduced mortality with pay for performance in England. *New England Journal of Medicine*, 379, 1821-1828.

Taylor, H., Karlin, S. (1998). *An Introduction to Stochastic Modelling* (3th edition) (New York: Academic Press).

Tobin, J. (1970). On Limiting the Domain of Inequality, *Journal of Law & Economics* 13, 263-277.

Windmeijer, F., Gravelle, H., Hoonhout, P. (2005). Waiting lists, waiting times and admissions: an empirical analysis. *Health Economics*, 2005, 14, 971-985

Worthington, D. (1987). Queueing models for hospital waiting lists, *Journal of the Operational Research Society* 38, 413-422

Worthington D. (1991) Hospital waiting list management models", *Journal of the Operational Research Society* 42, 833-843.

A Appendix

A.1 Waiting time distribution in the $M/M/k$ model

In the $M/M/k$ queueing model arrivals follow a Poisson process with mean arrival rate λ. The time it takes to treat a patient is negatively exponentially distributed with parameter μ. The mean treatment time is therefore $\frac{1}{\mu}$. If there are k servers (beds), the maximum number of patients that can be treated simultaneously is k. It can be shown that if $\lambda < k\mu$, there exists a steady state distribution for the system, in the sense that the probability p_n that there are n patients in the system (either waiting for treatment or under treatment) is well defined ($n = 0, 1, 2, ...$). From these state probabilities, one then can derive the distributions for the length of the waiting list, for the time spent waiting for treatment and for the total time spent in the system (w).

Let π_0 denote the probability that there is a vacant bed (i.e., $\pi_0 = \sum_{n=0}^{k-1} p_n$), and $\rho = \frac{\lambda}{\mu}$ then

$$\pi_0(k,\rho) = \left(\sum_{j=0}^{j=k-1} \frac{\rho^j}{j!}\right)\left[\sum_{j=0}^{j=k-1} \frac{\rho^j}{j!} + \frac{\rho^k}{k!\,(1-\rho/k)}\right]^{-1} = \frac{d^1(k,\rho)}{d^1(k,\rho)+d^2(k,\rho)}, \quad \text{(A.1)}$$

with obvious definitions of $d^1(\cdot)$ and $d^2(\cdot)$. If all beds are occupied, which happens with probability $1-\pi_0$, a patient joining the waiting list faces a time on the list which has a negative exponential distribution with mean $\frac{1}{k\mu-\lambda}$. In addition, this patient faces a time under treatment which is negatively exponentially distributed with mean $\frac{1}{\mu}$. Hence, the distribution of the total time in the system is the convolution of the distribution of the time on the waiting list given that this time is strictly positive, and the distribution of the treatment time is (see Gross *et al.* (2008)[30]

$$\begin{aligned} H(w;\pi_0,k,\mu,\lambda) &= \pi_0 \Pr[\tilde{w} \le w |\, n \le k-1] + (1-\pi_0)\Pr[\tilde{w} \le w |\, n \ge k] \\ &= \pi_0\left[1-e^{-\mu w}\right] + (1-\pi_0)\left[\frac{(k-\rho)\left(1-e^{-\mu w}\right) - \left(1-e^{-(k\mu-\lambda)w}\right)}{k-\rho-1}\right]. \end{aligned} \quad \text{(A.2)}$$

The expected total time in the system is.

$$\overline{w}(\pi_0,k,\mu,\lambda) = \int_0^\infty w dH(w;\pi_0,k,\mu,\lambda) = \frac{1}{\mu} + \frac{1-\pi_0}{k\mu-\lambda}.$$

Theorem A.1 *Consider an $M/M/k$ queueing system with mean referral rate λ and mean treatment time $\frac{1}{\mu}$, satisfying the stability condition $\frac{\lambda}{k\mu} < 1$. Reductions in the number of beds k, reductions in the treatment rate μ, and increases in the arrival rate λ all induce a stochastically dominating distribution for the total waiting time (time on the waiting list plus treatment time)*

Proof of Theorem A.1

[30] Two useful results are: (a) Let $a > 0$. Then $x(1-e^{-a}) - (1-e^{-xa}) \gtrless 0 \iff x \gtrless 1$. (b) Let $a > 0$. Then $\frac{x(1-e^{-a})-(1-e^{-ax})}{x-1} < 1$ for all x. Hence the second square bracketed term in in (A.2) is lies in $(0,1)$.

(a) We first show that

$$\frac{\partial H}{\partial k} = \underbrace{\frac{\partial H}{\partial \pi_0}}_{\text{(i)}} \underbrace{\frac{\partial \pi_0}{\partial k}}_{\text{(ii)}} + (1-\pi_0) \underbrace{\frac{\partial G(w;k,\mu,\lambda)}{\partial k}}_{\text{(iii)}} > 0,$$

where we treat π_0 as differentiable in k to reduce notational clutter.

(i) $\partial H/\partial \pi_0 > 0$.

$$\frac{\partial H}{\partial \pi_0} = \Pr[\tilde{w} \leq w | \, n \leq k-1] - \Pr[\tilde{w} \leq w | \, n \geq k] = \frac{e^{-\mu w} - e^{-(k-\rho)\mu w}}{k-\rho-1}.$$

If $k-\rho-1 > 0$ the numerator and denominator are positive and if $k-\rho-1 < 0$ they are both negative.

(ii) $\pi_0(k,\rho)$ is increasing in k.

Consider

$$\begin{aligned} \pi_0(k+1,\rho) - \pi_0(k,\rho) &= \frac{d^1(k+1,\rho)}{d^1(k+1,\rho)+d^2(k+1,\rho)} - \frac{d^1(k,\rho)}{d^1(k,\rho)+d^2(k,\rho)} \\ &= \frac{d^1(k+1,\rho)d^2(k,\rho) - d^1(k,\rho)d^2(k+1,\rho)}{[d^1(k+1,\rho)+d^2(k+1,\rho)]\,[d^1(k,\rho)d^2(k+1,\rho)]}, \end{aligned}$$

and so

$$\text{sign } [\pi_0(k+1,\rho) - \pi_0(k,\rho)] = \text{sign } \left[\frac{d^1(k+1,\rho)}{d^1(k,\rho)} - \frac{d^2(k+1,\rho)}{d^2(k,\rho)}\right].$$

Now

$$\frac{d^1(k+1,\rho)}{d^1(k,\rho)} = \frac{\sum_{j=0}^{j=k} \frac{\rho^j}{j!}}{\sum_{j=0}^{j=k-1} \frac{\rho^j}{j!}} > 1,$$

and

$$\begin{aligned} \frac{d^2(k+1,\rho)}{d^2(k,\rho)} &= \frac{\rho^{k+1}}{(k+1)!\,(1-\rho/(k+1))} \frac{k!\,(1-\rho/k)}{\rho^k} \\ &= \frac{\rho}{k+1} \frac{(1-\rho/k)}{(1-\rho/(k+1))} = \frac{\rho}{k} \frac{(k-\rho)}{(k+1-\rho)} < 1, \end{aligned}$$

since the stability of the queueing system requires $\rho < k$. Hence

$$\pi_0(k+1,\rho) - \pi_0(k,\rho) > 0,$$

and so $\pi_0(k,\rho)$ is increasing in k.

(iii) $G(w;k,\mu,\lambda) = \Pr[\tilde{w} \leq w | \, n \geq k]$ is increasing in k.

The random waiting time for a patient joining the queue, conditional on there being no empty bed, is the convolution of the distribution of two random variables: the wait for

a bed to become available w_1 and the service time w_2 whilst being treated:

$$\begin{aligned}\Pr[\tilde{w} \leq w | \, n \quad \geq \quad k] &= G(w; k, \mu, \lambda) \\ &= \int_0^w G_1(w - w_2; \delta_1) g_2(w_2; \delta_2) \mathrm{d}w_2 \\ &= \int_0^w \left[1 - e^{-\delta_1 (w - w_2)}\right] \delta_2 e^{-\delta_2 w_2} \mathrm{d}w_2,\end{aligned}$$

where G_1, g_2 are the distribution and density functions for the exponentially distributed wait for a bed (w_1) and the treatment time (w_2) which have means $\frac{1}{\delta_1} \stackrel{\text{def}}{=} \frac{1}{k\mu - \lambda}$ and $\frac{1}{\delta_2} \stackrel{\text{def}}{=} \frac{1}{\mu}$. Hence

$$\begin{aligned}\frac{\partial G(w; k, \mu, \lambda)}{\partial k} &= \frac{\partial}{\partial \delta_1} \left(\int_0^w \left[1 - e^{-\delta_1 (w - w_2)}\right] \delta_2 e^{-\delta_2 w_2} \mathrm{d}w_2 \right) \frac{\partial \delta_1}{\partial k} \\ &= \int_0^w (w - w_2) e^{-\delta_1 (w - w_2)} \delta_2 e^{-\delta_2 w_2} \mathrm{d}w_2 \mu > 0,\end{aligned}$$

which is positive since $w_2 \in [0, w]$.

(b) We establish

$$\frac{\partial H}{\partial \mu} = \underbrace{\frac{\partial H}{\partial \pi_0} \frac{\partial \pi_0}{\partial \rho} \frac{\partial \rho}{\partial \mu}}_{\text{(i) (iv) (v)}} + \pi_0 \underbrace{\frac{\partial \left(1 - e^{-\mu w}\right)}{\partial \mu}}_{\text{(vi)}} + (1 - \pi_0) \underbrace{\frac{\partial G(w; \delta_1, \delta_2)}{\partial \mu}}_{\text{(vi)}} > 0,$$

in similar fashion.

(iv) $\partial \pi_0(k, \rho) / \partial \rho < 0$.

From (A.1)

$$\text{sign } \frac{\partial \pi_0(k, \rho)}{\partial \rho} = \text{sign } \left[\frac{\partial d^1(k, \rho) / \partial \rho}{d^1(k, \rho)} - \frac{\partial d^2(k, \rho) / \partial \rho}{d^2(k, \rho)} \right] < 0,$$

since

$$\frac{\partial d^1(k, \rho) / \partial \rho}{d^1(k, \rho)} = \left[\sum_{j=0}^{j=k-2} \frac{\rho^j}{j!} \right] \left[\sum_{j=0}^{j=k-1} \frac{\rho^j}{j!} \right]^{-1} < 1,$$

and

$$\begin{aligned}\frac{\partial d^2(k, \rho) / \partial \rho}{d^2(k, \rho)} &= \frac{(1 - \rho/k) \, k \rho^{k-1} + \rho^k / k}{k! \, (1 - \rho/k)^2} \left[\frac{k! \, (1 - \rho/k)}{\rho^k} \right] \\ &= \frac{(1 - \rho/k) \, k \rho^{-1} + 1/k}{(1 - \rho/k)} \\ &= \frac{k}{\rho} + \frac{1}{k - \rho} > 1.\end{aligned}$$

(v) Since $\rho = \lambda / \mu$ we have $\partial \rho / \partial \mu < 0$ and so the product of (i), (iv) and (v) is positive.

(vi) The distribution function for the waiting time for patients who find an empty

queue is decreasing in μ

$$\frac{\partial(1-e^{-\mu w_1})}{\partial \mu} = w_1 e^{-\mu w_1} > 0.$$

(vi) $\partial G(w;k,\mu,\lambda)/\partial\mu > 0$.
μ affects both $\delta_1 = k\mu - \lambda$ and $\delta_2 = \mu$ and so

$$\begin{aligned}
\frac{\partial G(w;\delta_1,\delta_2)}{\partial \mu} &= \frac{\partial G(w;\delta_1,\delta_2)}{\partial \delta_1}\frac{\partial \delta_1}{\partial \mu} + \frac{\partial G(w;\delta_1,\delta_2)}{\partial \delta_2}\frac{\partial \delta_2}{\partial \mu} \\
&= \frac{\partial}{\partial \delta_1}\left(\int_0^w G_1(w-w_2;\delta_1)g_2(w_2;\delta_2)\mathrm{d}w_2\right)k \\
&\quad + \frac{\partial}{\partial \delta_2}\left(\int_0^w g_1(w_1;\delta_1)G_2(w-w_1;\delta_2)\mathrm{d}w_1\right) \\
&= \int_0^w (w-w^2)e^{-\lambda_1(w-w_2)}\lambda_2 e^{-\lambda_2 w_2}\mathrm{d}w_2 k \\
&\quad + \int_0^w \lambda_1 e^{-\lambda_1 w_1}(w-w_1)e^{-\lambda_2(w-w_1)}\mathrm{d}w_1 \; > \; 0 \; .
\end{aligned}$$

(c) Finally, we show

$$\frac{\partial H}{\partial \lambda} = \underbrace{\frac{\partial H}{\partial \pi_0}\frac{\partial \pi_0}{\partial \rho}\frac{\partial \rho}{\partial \lambda}}_{\text{(i) (iv) (vii)}} + (1-\pi_0)\underbrace{\frac{\partial G(w;\delta_1,\delta_2)}{\partial \lambda}}_{\text{(viii)}} < 0.$$

(vii) Since (i) is positive, (iv) is negative, and (viii) is positive because $\rho = \lambda/\mu$, the first term in $\partial H/\partial\lambda$ is negative. (viii) is

$$\begin{aligned}
\frac{\partial G(w;\delta_1,\delta_2)}{\partial \lambda} &= \frac{\partial}{\partial \delta_1}\left(\int_0^w \left[1-e^{-\delta_1(w-w_2)}\right]\delta_2 e^{-\delta_2 w_2}\mathrm{d}w_2\right)\frac{\partial \delta_1}{\partial \lambda} \\
&= -\int_0^w (w-w_2)e^{-\delta_1(w-w_2)}\delta_2 e^{-\delta_2 w_2}\mathrm{d}w_2 < 0,
\end{aligned}$$

which completes the proof of Proposition A1.■

A.2 Equilibrium mean waiting time

Lemma A.1 *Increases in supply conditions that are accompanied by a quality reduction to keep demand (expected wait) constant lower the expected wait (increase expected demand).*

Proof of Lemma A.1

$$\begin{aligned}
\left.\frac{\mathrm{d}w^e(q,k,\mu)}{\mathrm{d}s}\right|_{\mathrm{d}\lambda=0} &= \bar{w}_s + \bar{w}_\lambda\lambda_s - \bar{w}_\lambda\lambda_q\frac{\lambda_s}{\lambda_q} = \bar{w}_s < 0 \quad (s=k,\mu), \\
\left.\frac{\mathrm{d}\lambda}{\mathrm{d}s}\right|_{\mathrm{d}\bar{w}=0} &= \lambda_s - \lambda_q\frac{\bar{w}^e_s}{\bar{w}^e_q} = \lambda_s - \lambda_q\frac{(\bar{w}_s + \bar{w}_\lambda\lambda_s)}{\bar{w}_\lambda\lambda_q} = -\frac{\bar{w}_s}{\bar{w}_\lambda} > 0 \quad (s=k,\mu).
\end{aligned}$$

Proof of Proposition 2 (a) Linear (LN) preferences. Define utility if ill and treated in the public hospital after a wait of w as $u(y,q,w) = t^1(y,q) - t^2(y,q)w$ so that expected util-

ity from choosing the public hospital is $v^* = (1-\sigma)u^N(y)+\sigma\left[t^1(y,q)-t^2(y,q)\bar{w}(\lambda,k,\mu)\right] = v^*(y,q,\bar{w}(\lambda,k,\mu))$. The threshold income $\hat{y}(q,\lambda,k,\mu)$ is defined by $G = v^*(y,q,\bar{w}(\lambda,k,\mu))-v^o(y-\gamma)=0$. Equilibrium demand λ satisfies $\lambda-\sigma F(\hat{y}(q,\lambda,k,\mu))=0$. Then

$$\lambda_s = \frac{\sigma f(\hat{y})\hat{y}_s}{1-\sigma f(\hat{y})\hat{y}_\lambda} = \frac{\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)G_s}{1-\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)G_\lambda} = \frac{\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_s}{1-\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_\lambda} \quad (s=k,\mu),$$

and so

$$\begin{aligned} w^e_s &= \bar{w}_s(\lambda(q,s),s)+\bar{w}_\lambda(\lambda(q,s),s)\lambda_s = \bar{w}_s+\bar{w}_\lambda\left[\frac{\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_s}{1-\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_\lambda}\right] \\ &= \left[\frac{\bar{w}_s}{1-\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_\lambda}\right]\left[1-\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_\lambda+\bar{w}_\lambda\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\right] \\ &= \frac{\bar{w}_s}{1-\sigma f(\hat{y})\left(\frac{-1}{G_y}\right)v^*_{\bar{w}}\bar{w}_\lambda} < 0 \quad (s=k,\mu). \end{aligned}$$

(b) $M/M/1$. Letting s now be the service rate (inverse of expected length of stay) the density function for waiting time is $h(w;\lambda,s)=(s-\lambda)e^{-(s-\lambda)}$ and the expected waiting time is $\bar{w}=(s-\lambda)^{-1}$ (see Gross *et al.* (2008)). Hence $w^e_s=\bar{w}_s+\bar{w}_\lambda\lambda_s=\bar{w}_s(1-\lambda_s)$. Now

$$\begin{aligned} 1-\lambda_s &= 1-\frac{\sigma f(\hat{y})\hat{y}_s}{1-\sigma f(\hat{y})\hat{y}_\lambda} = \frac{1-\sigma f(\hat{y})\hat{y}_\lambda-\sigma f(\hat{y})\hat{y}_s}{1-\sigma f(\hat{y})\hat{y}_\lambda} \\ &= \frac{1-\sigma f(\hat{y})\left[\hat{y}_\lambda+\hat{y}_s\right]}{1-\sigma f(\hat{y})\hat{y}_\lambda}, \end{aligned}$$

where $\hat{y}(q,\lambda,s)$ is defined by

$$G=(1-\sigma)u^N(y)+\sigma\int_0^\infty u(y,q,w)h(w;\lambda,s)dw-v^o(y-\gamma)=0.$$

Hence

$$\begin{aligned} \hat{y}_\lambda+\hat{y}_s &= \frac{-1}{G_y}\left[G_\lambda+G_s\right] \\ &= \frac{-\sigma}{G_y}\left[\int_0^\infty u(y,q,w)[h_\lambda(w;\lambda,s)+h_s(w;\lambda,s)\right]dw=0, \end{aligned}$$

and so

$$1-\lambda_s=\frac{1}{1-\sigma f(\hat{y})\hat{y}_\lambda}>0,$$

implying that $w^e_s=\bar{w}_s+\bar{w}_\lambda\lambda_s=\bar{w}_s(1-\lambda_s)>0$. ■

A.3 Demand constant welfare changes

Theorem A.2 *In equilibrium an increase in hospital attribute z $(= q, \mu, k)$ accompanied by an adjustment in hospital attribute x $(x = q, \mu, k;\ x \neq z)$ to keep demand constant has no effect on patient welfare $B^e(q, \mu, k)$ if and only if*

(a) patient preferences are quasi-separable between quality and waiting time :

$$\begin{aligned} v &= \sigma \left\{ \int [a^1(y) + a^2(y) r(q, w)] \mathrm{d}H(w; \lambda, k, \mu) \right\} + (1 - \sigma) u^N(y) \\ &= \sigma \left[a^1(y) + a^2(y) R(q, \lambda, k, \mu) \right] + (1 - \sigma) u^N(y), \end{aligned}$$

and (b) the welfare function respects patient preferences over quality and waiting time

$$b^e(y, q, k, \mu) = \sigma[m^0(y) + m^1(y) m^2(R(q, k, \mu))] + (1 - \sigma) m^N(y), \qquad m^2_R > 0.$$

Proof of Theorem A.2. In equilibrium an increase in hospital attribute z accompanied by an adjustment in hospital attribute x to keep demand constant changes patient welfare at the rate

$$B^e_z(q, \mu, k) - B^e_x(q, \mu, k) \frac{\lambda_z(q, \mu, k)}{\lambda_x(q, \mu, k)}$$

From (15) and Lemma 1 in the main text:

$$\begin{aligned} B^e_z - B^e_x \frac{\lambda_z}{\lambda_x} &= \int_{y^{\min}}^{\hat{y}^e(q,,\mu,k)} \left[b^e_z - b^e_x \frac{\lambda_z}{\lambda_x} \right] \mathrm{d}F(y) \\ &\quad + \left[b^e(\hat{y}, q, k, \mu) - b^o(\hat{y} - \gamma, q^o) \right] \sigma f\left(\hat{y}^e\right) \left[\hat{y}^e_z - \hat{y}^e_x \frac{\lambda_z}{\lambda_x} \right] \\ &= \int_{y^{\min}}^{\hat{y}^e(q,\mu,k)} \left[b^e_z - b^e_x \frac{\lambda_z}{\lambda_x} \right] \mathrm{d}F(y) \\ &= \int_{y^{\min}}^{\hat{y}(q,\lambda,\mu,k)} \left[b_z(y, q, \lambda, k, \mu) + b_\lambda \lambda_z - b_x(y, q, \lambda, k, \mu) \frac{\lambda_z}{\lambda_x} - b_\lambda \lambda_x \frac{\lambda_z}{\lambda_x} \right] \mathrm{d}F(y) \\ &= \int_{y^{\min}}^{\hat{y}(q,\lambda,\mu,k)} \left[b_z(y, q, \lambda, k, \mu) - b_x(y, q, \lambda, k, \mu) \frac{\lambda_z}{\lambda_x} \right] \mathrm{d}F(y) \\ &= B_z(\hat{y}, q, \lambda, \mu, k) - B_x(\hat{y}, q, \lambda, \mu, k) \frac{v_z(\hat{y}, q, \lambda, \mu, k)}{v_x(\hat{y}, q, \lambda, \mu, k)} = T(\hat{y}, q, \lambda, \mu, k). \quad \text{(A.3)} \end{aligned}$$

For $T(\hat{y}, q, \lambda, \mu, k) = 0$ to hold for all the equilibria generated by the premium for private health care γ, which affects (A.3) only via the the threshold income $\hat{y}$, requires

$$\begin{aligned} \frac{\partial T}{\partial \hat{y}} &= \left[b_z(\hat{y}, q, \lambda, \mu, k) - b_x(\hat{y}, q, \lambda, \mu, k) \frac{v_z(\hat{y}, q, \lambda, \mu, k)}{v_x(\hat{y}, q, \lambda, \mu, k)} \right] f\left(\hat{y}\right) \\ &\quad - B_x \frac{1}{\hat{v}^2_x} \left[\hat{v}_{zy} \hat{v}_x - \hat{v}_z \hat{v}_{xy} \right] = 0 \end{aligned} \quad \text{(A.4)}$$

This holds in general (i.e., for all distributions of income) only if both square bracketed terms are zero. The first term is zero only if the marginal rate of substitution between

z and x for a patient of given income are the same for the regulator and the patient. The second is zero only if the patient marginal rate of substitution between z and x is unaffected by patient income. Sufficiency of (a) and (b) is easily established.■

Theorem A.3 *The effect on social benefit of an increase in hospital attribute z followed by an adjustment in attribute x to keep demand constant is given by*

$$B_z^e - B_x^e \frac{\lambda_z}{\lambda_x} = \int_{y_{\min}}^{\widehat{y}} b_x \left[\frac{b_z(y,\cdot)}{b_x(y,\cdot)} - \frac{v_z(\widehat{y},\cdot)}{v_x(\widehat{y},\cdot)} \right] dF(y) \tag{A.5}$$

$$= \int_{y_{\min}}^{\widehat{y}} b_z \left[\left(\frac{\partial \log \frac{b_z(y,\cdot)}{b_y(y,\cdot)}}{\partial y} - \frac{\partial \log \frac{b_x(y,\cdot)}{b_y(y,\cdot)}}{\partial y} \right)_{\widetilde{y} \in [y,\widehat{y}]} (y - \widehat{y}) + \left(\frac{b_z(\widehat{y},\cdot)}{b_x(\widehat{y},\cdot)} - \frac{v_z(\widehat{y},\cdot)}{v_x(\widehat{y},\cdot)} \right) \right] dF(y). \tag{A.6}$$

The theorem implies that when individual preferences are respected, so that the second round bracket term is zero, then $B_z^e - B_x^e \frac{\lambda_z}{\lambda_x} > 0 (< 0)$ if the marginal willingness to pay for attribute x increases faster (slower) with income than that for attribute z. Under quasi-separable individual and social preferences ((3) and (16)), $\frac{b_z(y,\cdot)}{b_x(y,\cdot)} = \frac{v_z(\widehat{y},\cdot)}{v_x(\widehat{y},\cdot)} = \frac{R_z(\cdot)}{R_x(\cdot)}$ and $B_z^e - B_x^e \frac{\lambda_z}{\lambda_x} = 0$. Under specific egalitarianism, $\frac{b_z(y,\cdot)}{b_x(y,\cdot)}$ is independent of income, and the sign of (A.5) is the sign of $\frac{b_z(\widehat{y},\cdot)}{b_x(\widehat{y},\cdot)} - \frac{v_z(\widehat{y},\cdot)}{v_x(\widehat{y},\cdot)}$.

Proof of Theorem A.3.

We have, making use of the mean value theorem in the third line

$$B_z^e - B_x^e \frac{\lambda_z}{\lambda_x} = \int_{y_{\min}}^{\widehat{y}} \left[b_z^e(y,q,k,\mu) - b_x^e(y,q,k,\mu) \frac{\lambda_z}{\lambda_x} \right] \mathrm{d}F(y)$$

$$= \int_{y_{\min}}^{\widehat{y}} b_x \left[\frac{b_z(y,q,\lambda,k,\mu)}{b_x(y,q,\lambda,k,\mu)} - \frac{b_z(\widehat{y},q,\lambda,k,\mu)}{b_x(\widehat{y},q,\lambda,k,\mu)} + \frac{b_z(\widehat{y},q,\lambda,k,\mu)}{b_x(\widehat{y},q,\lambda,k,\mu)} - \frac{\lambda_z}{\lambda_x} \right] \mathrm{d}F(y)$$

$$= \int_{y_{\min}}^{\widehat{y}} b_x(y,\cdot) \left[(y - \widehat{y}) \left. \frac{\partial \frac{b_z(\tilde{y},\cdot)}{b_x(\tilde{y},\cdot)}}{\partial y} \right|_{\widetilde{y} \in [y,\widehat{y}]} + \frac{b_z(\widehat{y},\cdot)}{b_x(\widehat{y},\cdot)} - \frac{v_z(\widehat{y},\cdot)}{v_x(\widehat{y},\cdot)} \right] \mathrm{d}F(y)$$

Define

$$\Delta(\widehat{y},\cdot) \stackrel{\text{def}}{=} \frac{b_z(\widehat{y},\cdot)}{b_x(\widehat{y},\cdot)} - \frac{v_z(\widehat{y},\cdot)}{v_x(\widehat{y},\cdot)}$$

$$\begin{aligned} B_z^e - B_x^e \frac{\lambda_z}{\lambda_x} &= \int_{y_{\min}}^{\widehat{y}} b_x(y,\cdot) \left[(y - \widehat{y}) \left. \frac{\partial \frac{b_z(\tilde{y},\cdot)}{b_x(\tilde{y},\cdot)}}{\partial y} \right|_{\widetilde{y} \in [y,\widehat{y}]} + \Delta(\widehat{y},\cdot) \right] \mathrm{d}F(y) \\ &= \int_{y_{\min}}^{\widehat{y}} b_x(y,\cdot) \left[(y - \widehat{y}) \left(\frac{b_z(\tilde{y},\cdot)}{b_x(\tilde{y},\cdot)} \left. \frac{\partial \log \frac{b_z(\tilde{y},\cdot)}{b_x(\tilde{y},\cdot)}}{\partial y} \right|_{\widetilde{y} \in [y,\widehat{y}]} \right) + \Delta(\widehat{y},\cdot) \right] \mathrm{d}F(y) \end{aligned}$$

If patient preferences are respected $\Delta(\widehat{y},\cdot)=0$ then

$$B_z^e - B_x^e\frac{\lambda_z}{\lambda_x} = \int_{y_{\min}}^{\widehat{y}} v_x(y,\cdot)\left[(y-\widehat{y})\left(\frac{b_z(\tilde{y},\cdot)}{b_x(\tilde{y},\cdot)}\left(\frac{\partial \log \frac{v_z(y,\cdot)}{v_y(y,\cdot)}}{\partial y} - \frac{\partial \log \frac{v_x(y,\cdot)}{v_y(y,\cdot)}}{\partial y}\right)\bigg|_{\tilde{y}\in[y,\widehat{y}]}\right)\right]\mathrm{d}F(y)$$

$$= \int_{y_{\min}}^{\widehat{y}} v_x(y,\cdot)\left[(y-\widehat{y})\left(\frac{b_z(\tilde{y},\cdot)}{b_x(\tilde{y},\cdot)}\left(\frac{\partial \log MWP_z(\tilde{y},\cdot)}{\partial y} - \frac{\partial \log MWP_x(\tilde{y},\cdot)}{\partial y}\right)\bigg|_{\tilde{y}\in[y,\widehat{y}]}\right)\right]\mathrm{d}F(y)$$

■

A.4 Demand constant change in insurance cost

Lemma A.2 *The change in equilibrium insurance costs when supply s ($s=k,\mu$) is increased and q is reduced to keep demand constant is*

$$\left.\frac{dc^{Ie}}{ds}\right|_{d\lambda=0} = c_s^{Ie} - c_q^{Ie}\frac{\lambda_s}{\lambda_q} = \sigma\bar{w}_s\int_{y_{\min}}^{\hat{y}^e} y dF(y) < 0, \ \ (s=k,\mu). \tag{A.7}$$

The change in equilibrium insurance costs when quality is increased and supply s ($s=k,\mu$) is reduced to keep demand constant is

$$\left.\frac{dc^{Ie}}{dq}\right|_{d\lambda=0} = c_q^{Ie} - c_s^{Ie}\frac{\lambda_q}{\lambda_s} = -\sigma\bar{w}_s\frac{\lambda_q}{\lambda_s}\int_{y_{\min}}^{\hat{y}^e} y dF(y) > 0 \quad (s=k,\mu). \tag{A.8}$$

Proof of Lemma A.2

$$\begin{aligned}
c_z^{Ie} - c_x^{Ie}\frac{\lambda_z}{\lambda_x} &= \sigma\left(w_z^e - w_x^e\frac{\lambda_z}{\lambda_x}\right)\int_{y_{\min}}^{\hat{y}} y\mathrm{d}F(y) + \sigma w\int_{y_{\min}}^{\hat{y}} y\mathrm{d}F(y)\hat{y}f(\hat{y})\left(\hat{y}_z^e - \hat{y}_x^e\frac{\lambda_z}{\lambda_x}\right)\\
&= \sigma\left(w_z - w_x\frac{\lambda_z}{\lambda_x}\right)\int_{y_{\min}}^{\hat{y}} y\mathrm{d}F(y) + \sigma w\int_{y_{\min}}^{\hat{y}} y\mathrm{d}F(y)\hat{y}f(\hat{y})\left(\hat{y}_z - \hat{y}_x\frac{\lambda_z}{\lambda_x}\right)\\
&= \sigma\left(w_z - w_x\frac{\lambda_z}{\lambda_x}\right)\int_{y_{\min}}^{\hat{y}} y\mathrm{d}F(y)
\end{aligned}$$

since the second round bracketed term in the second line is zero (cf (10)). ■

A.5 Derivation of optimal pricing rules

Proof of Proposition 3 The hospital maximises (23), leading to decisions $q(p_\lambda,p_k,p_\mu)$, $k(p_\lambda,p_k,p_\mu)$, $\mu(p_\lambda,p_k,p_\mu)$ satisfying (24). The regulation problem is

$$\max_{p_\lambda,p_k,p_\mu} B^e(q,k,\mu) - (1+\theta)[p_\lambda\lambda(q,k,\mu) + p_k k + p_\mu\mu + T + c^{Ie}(q,k,\mu)].$$

T is set to leave the hospital with no financial rent

$$T = -p_\lambda\lambda - p_k k - p_\mu\mu + c^{He}(q,k,\mu),$$

and so the regulator problem is equivalently

$$\begin{aligned}&\max_{p_\lambda,p_k,p_\mu} B^e(q(p_\lambda,p_k,p_\mu),k(p_\lambda,p_k,p_\mu),\mu(p_\lambda,p_k,p_\mu))\\&-(1+\theta)c^{He}(q(p_\lambda,p_k,p_\mu),k(p_\lambda,p_k,p_\mu),\mu(p_\lambda,p_k,p_\mu))\\&-(1+\theta)c^{Ie}(q(p_\lambda,p_k,p_\mu),k(p_\lambda,p_k,p_\mu),\mu(p_\lambda,p_k,p_\mu))\end{aligned}$$

The FOCs to this regulation problem can be written as

$$\begin{pmatrix}\frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda}\\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k}\\ \frac{\partial q}{\partial p_\mu} & \frac{\partial k}{\partial p_\mu} & \frac{\partial \mu}{\partial p_\mu}\end{pmatrix}\begin{pmatrix}B^e_q-(1+\theta)[c^{He}_q+c^{Ie}_q]\\ B^e_k-(1+\theta)[c^{He}_k+c^{Ie}_k]\\ B^e_\mu-(1+\theta)[c^{He}_\mu+c^{Ie}_\mu]\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

But making use of the hospital FOCs, the regulator's FOCs can be written as

$$\begin{pmatrix}\frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda}\\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k}\\ \frac{\partial q}{\partial p_\mu} & \frac{\partial k}{\partial p_\mu} & \frac{\partial \mu}{\partial p_\mu}\end{pmatrix}\left[\begin{pmatrix}S^e_q\\S^e_k\\S^e_\mu\end{pmatrix}-\begin{pmatrix}\lambda_q & 0 & 0\\ \lambda_k & 1 & 0\\ \lambda_\mu & 0 & 1\end{pmatrix}\begin{pmatrix}p^{FB*}_\lambda\\p^{FB*}_k\\p^{FB*}_\mu\end{pmatrix}\right]=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

where $S^e_z=\beta B^e_z-c^{Ie}_z$ where $\beta=[1-\alpha(1+\theta)]/(1+\theta)$.

The (3×3) matrix of price effects is invertible. Hence, the FB optimal prices are

$$\begin{aligned}\begin{pmatrix}p^{FB*}_\lambda\\p^{FB*}_k\\p^{FB*}_\mu\end{pmatrix}&=\begin{pmatrix}\lambda_q & 0 & 0\\ \lambda_k & 1 & 0\\ \lambda_\mu & 0 & 1\end{pmatrix}^{-1}\begin{pmatrix}S^e_q\\S^e_k\\S^e_\mu\end{pmatrix}=\begin{pmatrix}\frac{1}{\lambda_q} & 0 & 0\\ -\frac{\lambda_k}{\lambda_q} & 1 & 0\\ -\frac{\lambda_\mu}{\lambda_q} & 0 & 1\end{pmatrix}\begin{pmatrix}S^e_q\\S^e_k\\S^e_\mu\end{pmatrix}\\&=\begin{pmatrix}\frac{S^e_q}{\lambda_q}\\ S^e_k-\frac{\lambda_k}{\lambda_q}S^e_q\\ S^e_k-\frac{\lambda_\mu}{\lambda_q}S^e_q\end{pmatrix}.\blacksquare\end{aligned}$$

Proof of Proposition 4 The generic first order condition for the hospital facing prices p_λ, p_k and p_q is

$$p_\lambda\lambda_z+p_k\mathbf{1}_{(z=k)}+p_q\mathbf{1}_{(z=q)}+\alpha B^e_z=c^{He}_z\quad(z=q,k,\mu).$$

The first order conditions for the regulator's problem is

$$\begin{pmatrix}\frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda}\\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k}\\ \frac{\partial q}{\partial p_q} & \frac{\partial k}{\partial p_q} & \frac{\partial \mu}{\partial p_q}\end{pmatrix}\begin{pmatrix}B^e_q-(1+\theta)[c^{He}_q+c^{Ie}_q]\\ B^e_k-(1+\theta)[c^{He}_k+c^{Ie}_k]\\ B^e_\mu-(1+\theta)[c^{He}_\mu+c^{Ie}_\mu]\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

Replacing c^{He}_z by the LHS of the hospital's first order condition shows that

$$B^e_z-(1+\theta)[c^{He}_z+c^{Ie}_z]=(1+\theta)\left[S^e_z-p_\lambda\lambda_z-p_k\mathbf{1}_{(z=k)}-p_q\mathbf{1}_{(z=q)}\right],$$

which allows us to rewrite the regulator's FOCs as

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k} \\ \frac{\partial q}{\partial p_q} & \frac{\partial k}{\partial p_q} & \frac{\partial \mu}{\partial p_q} \end{pmatrix} \left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix} - \begin{pmatrix} \lambda_q & 0 & 1 \\ \lambda_k & 1 & 0 \\ \lambda_\mu & 0 & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB**} \\ p_k^{FB**} \\ p_\mu^{FB**} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Since the (3×3) response matrix is invertible, the FOCs are equivalent to the square bracket vector being zero. Hence

$$\begin{aligned} \begin{pmatrix} p_\lambda^{FB**} \\ p_k^{FB**} \\ p_\mu^{FB**} \end{pmatrix} &= \begin{pmatrix} \lambda_q & 0 & 1 \\ \lambda_k & 1 & 0 \\ \lambda_\mu & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix} = \begin{pmatrix} 0 & 0 & \frac{1}{\lambda_\mu} \\ 0 & 1 & -\frac{\lambda_k}{\lambda_\mu} \\ 1 & 0 & -\frac{\lambda_q}{\lambda_\mu} \end{pmatrix} \begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix}, \\ &= \begin{pmatrix} \frac{S_\mu^e}{\lambda_\mu} \\ S_k^e - \frac{\lambda_k}{\lambda_\mu} S_\mu^e \\ S_q^e - \frac{\lambda_q}{\lambda_\mu} S_\mu^e \end{pmatrix} = \begin{pmatrix} \frac{S_\mu^e}{\lambda_\mu} \\ S_k^e - \lambda_k p_\lambda^{FB**} \\ S_q^e - \lambda_q p_\lambda^{FB**} \end{pmatrix}. \blacksquare \end{aligned}$$

Proof of Proposition 5 The hospital choices $q(p_\lambda, p_k, p_w), k(p_\lambda, p_k, p_w), \mu(p_\lambda, p_k, p_w)$ satisfy (24). The regulation problem is then

$$\begin{aligned} &\max_{p_\lambda, p_k, p_w} B^e(q(p_\lambda, p_k, p_w), k(p_\lambda, p_k, p_w), \mu(p_\lambda, p_k, p_w)) \\ &-(1+\theta)c^{He}(q(p_\lambda, p_k, p_w), k(p_\lambda, p_k, p_w), \mu(p_\lambda, p_k, p_w)) \\ &-(1+\theta)c^{Ie}(q(p_\lambda, p_k, p_w), k(p_\lambda, p_k, p_w), \mu(p_\lambda, p_k, p_w)). \end{aligned}$$

The FOCs for welfare maximisation are

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} \end{pmatrix} \begin{pmatrix} B_q^e - (1+\theta)[c_q^{He} + c_q^{Ie}] \\ B_k^e - (1+\theta)[c_k^{He} + c_k^{Ie}] \\ B_\mu^e - (1+\theta)[c_\mu^{He} + c_\mu^{Ie}] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

By making use of the hospital FOCs, the regulator's FOCs can be written as

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} \end{pmatrix} \left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB+} \\ p_w^{FB+} \\ p_k^{FB+} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \quad \text{(A.9)}$$

The (3×3) matrix of price effects is invertible. Hence, under FB, the optimal prices are obtained by setting the square bracket term to zero and solving for the prices:

$$\begin{pmatrix} p_\lambda^{FB+} \\ p_w^{FB+} \\ p_k^{FB+} \end{pmatrix} = \begin{pmatrix} \frac{w_\mu^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} & 0 & -\frac{w_q^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} \\ -\frac{\lambda_\mu}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} & 0 & \frac{\lambda_q}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} \\ -\frac{\lambda_k w_\mu^e - \lambda_\mu w_k^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} & 1 & \frac{\lambda_k w_q^e - \lambda_q w_k^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} \end{pmatrix} \begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix}.$$

Using (13)-(14), and $S_z^e = S_z + S_\lambda \lambda_z$, this reduces to

$$\begin{pmatrix} p_\lambda^{FB+} \\ p_w^{FB+} \\ p_k^{FB+} \end{pmatrix} = \begin{pmatrix} \frac{S_q^e}{\lambda_q} - (S_\mu - S_q \frac{\lambda_\mu}{\lambda_q}) \frac{w_\lambda}{w_\mu} \\ (S_\mu - S_q \frac{\lambda_\mu}{\lambda_q}) \frac{1}{w_\mu} \\ (S_k - S_q \frac{\lambda_k}{\lambda_q}) - (S_\mu - S_q \frac{\lambda_\mu}{\lambda_q}) \frac{w_k}{w_\mu} \end{pmatrix},$$

which gives (29)-(31). ■

Proof of Proposition 6 When k cannot be priced, the regulator's FOCs (A.9) on p_λ and p_w are

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} \end{pmatrix} \left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda \\ p_w \\ 0 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Use (A.9) to replace $\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix}$ by $\begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \\ p_k^{FB++} \end{pmatrix}$, and let $^{++}$ indicate that prices are evaluated at the SB values for q, μ, k. Then

$$\begin{aligned} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = & \begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} \end{pmatrix} \left[\begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \\ p_k^{FB++} \end{pmatrix} \right. \\ & \left. - \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{SB} \\ p_w^{SB} \\ 0 \end{pmatrix} \right] \end{aligned}$$

This system of equations will hold in two distinct cases.

(i) the FB price for k is zero. This is the case discussed in section 4.2**.** In that case the square bracket term is the zero vector when also $p_\lambda^{SB} = p_\lambda^{FB}$ and $p_w^{SB} = p_w^{FB}$ (the (3×3) matrix being invertible).

(ii) when the FB price for k is different from zero. Then the (3×1) square bracket vector should be orthogonal to the two row vectors of responses w.r.t. the prices p_λ and p_w. This is equivalent to

$$\begin{aligned} & \begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} \end{pmatrix} \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \\ p_k^{FB++} \end{pmatrix} \\ = & \begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} \end{pmatrix} \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{SB} \\ p_w^{SB} \\ 0 \end{pmatrix}. \end{aligned}$$

Using the definitions $\frac{\partial \lambda^*}{\partial p_i} \stackrel{\text{def}}{=} \sum_{z=q,k,\mu} \lambda_z \frac{\partial z}{\partial p_i}$ $(i = \lambda, w)$ and $\frac{\partial w^{e*}}{\partial p_i} \stackrel{\text{def}}{=} \sum_{z=q,k,\mu} w_z^e \frac{\partial z}{\partial p_i}$ $(i = \lambda, w)$ (i.e., the optimal demand and expected wait responses that follows from the optimal

responses in q, k, μ) gives:

$$\begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial k}{\partial p_w} \end{pmatrix} \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \\ p_k^{FB++} \end{pmatrix} = \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} \end{pmatrix} \begin{pmatrix} p_\lambda^{SB} \\ p_w^{SB} \end{pmatrix}$$

implying

$$\begin{aligned} \begin{pmatrix} p_\lambda^{SB} \\ p_w^{SB} \end{pmatrix} &= \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial k}{\partial p_w} \end{pmatrix} \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \\ p_k^{FB++} \end{pmatrix} \\ &= \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \end{pmatrix} + \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial k}{\partial p_w} \end{pmatrix} p_k^{FB++} \\ = \begin{pmatrix} p_\lambda^{FB++} \\ p_w^{FB++} \end{pmatrix} &+ \frac{1}{\frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial w^{e*}}{\partial p_w} - \frac{\partial \lambda^*}{\partial p_w}\frac{\partial w^{e*}}{\partial p_\lambda}} \begin{pmatrix} \frac{\partial w^{e*}}{\partial p_w} & -\frac{\partial w^{e*}}{\partial p_\lambda} \\ -\frac{\partial \lambda^*}{\partial p_w} & \frac{\partial \lambda^*}{\partial p_\lambda} \end{pmatrix} \begin{pmatrix} \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial k}{\partial p_w} \end{pmatrix} p_k^{FB++}. \end{aligned}$$

Then we obtain the following SB pricing rules:

$$\begin{aligned} p_\lambda^{SB} &= p_\lambda^{FB++} + \frac{\frac{\partial w^{e*}}{\partial p_w}\frac{\partial k}{\partial p_\lambda} - \frac{\partial k}{\partial p_w}\frac{\partial w^{e*}}{\partial p_\lambda}}{\frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial w^{e*}}{\partial p_w} - \frac{\partial \lambda^*}{\partial p_w}\frac{\partial w^{e*}}{\partial p_\lambda}} p_k^{FB++}, \\ p_w^{SB} &= p_w^{FB++} + \frac{-\frac{\partial \lambda^*}{\partial p_w}\frac{\partial k}{\partial p_\lambda} + \frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial k}{\partial p_w}}{\frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial w^{e*}}{\partial p_w} - \frac{\partial \lambda^*}{\partial p_w}\frac{\partial w^{e*}}{\partial p_\lambda}} p_k^{FB++}, \end{aligned}$$

which upon using the definitions (33)-(34) can be written as (35)-(36). ■

Proof of Proposition 7 Using (29)-(31), the requirement for $p_k^{FB+} = 0$ amounts to

$$\begin{aligned} p_k^{FB+} &= S_k^e - p_\lambda^{FB+}\lambda_k - p_w^{FB+}w_k \\ &= S_k^e - S_q^e\frac{\lambda_k}{\lambda_q} - \left(S_\mu^e - S_q^e\frac{\lambda_\mu}{\lambda_q}\right)\frac{w_k}{w_\mu} = 0. \end{aligned}$$

Expanding by using the definition $S = \beta B - c^I$ gives

$$\begin{aligned} &\beta\left(B_k - B_q\frac{\lambda_k}{\lambda_q}\right) - \beta\left(B_\mu - B_q\frac{\lambda_\mu}{\lambda_q}\right)\frac{w_k}{w_\mu} \\ = \; & c_k^I - c_q^I\frac{\lambda_k}{\lambda_q} - \left(c_\mu^I - c_q^I\frac{\lambda_\mu}{\lambda_q}\right)\frac{w_k}{w_\mu}. \end{aligned}$$

The RHS can be written as

$$\begin{aligned}
& \sigma\left(w_k - w_q\frac{\lambda_k}{\lambda_q}\right)\int_{y_{\min}}^{\widehat{y}} y\mathrm{d}F(y) - \left(\sigma\left(w_\mu - w_q\frac{\lambda_\mu}{\lambda_q}\right)\int_{y_{\min}}^{\widehat{y}} y\mathrm{d}F(y)\right)\frac{w_k}{w_\mu} \\
= & \left(\left(w_k - w_q\frac{\lambda_k}{\lambda_q}\right) - \left(\left(w_\mu - w_q\frac{\lambda_\mu}{\lambda_q}\right)\right)\frac{w_k}{w_\mu}\right)\sigma\int_{y_{\min}}^{\widehat{y}} y\mathrm{d}F(y) \\
= & \left(w_k - w_q\frac{\lambda_k}{\lambda_q} - w_k + w_q\frac{\lambda_\mu}{\lambda_q}\frac{w_k}{w_\mu}\right)\sigma\int_{y_{\min}}^{\widehat{y}} y\mathrm{d}F(y) \\
= & -\frac{w_q}{\lambda_q}\left(\lambda_k - \lambda_\mu\frac{w_k}{w_\mu}\right)\sigma\int_{y_{\min}}^{\widehat{y}} y\mathrm{d}F(y) = 0,
\end{aligned}$$

where the last equality follows from the fact that q has no direct effect on w ($w_q \equiv 0$). So then we are left with the condition

$$\left(B_k - B_q\frac{\lambda_k}{\lambda_q}\right) = \left(B_\mu - B_q\frac{\lambda_\mu}{\lambda_q}\right)\frac{w_k}{w_\mu}. \tag{A.10}$$

The round bracket terms can be expanded as

$$B_s - B_q\frac{\lambda_s}{\lambda_q} = \int_{y^{\min}}^{\hat{y}^e(q,k,\mu)}\left(b_s - b_q\frac{\lambda_s}{\lambda_q}\right)\mathrm{d}F(y), \qquad s = \mu, k,$$

and under individual and social preferences (3) and (16), this becomes

$$B_s - B_q\frac{\lambda_s}{\lambda_q} = \sigma\int_{y^{\min}}^{\hat{y}^e(q,k,\mu)} m^1(y)\mathrm{d}F(y)m^2(R)\left(R_s - R_q\frac{\lambda_s}{\lambda_q}\right) = 0, \qquad s = \mu, k,$$

because $\frac{\lambda_s}{\lambda_q} = \frac{\widehat{v}_s}{\widehat{v}_q} = \frac{\sigma a^2(y)R_s}{\sigma a^2(y)R_q}$ $(s = \mu, k)$.

Moreover, if the expected wait is a sufficient statistic for the waiting time distribution w.r.t. preferences, (10)-(11) imply $\frac{w_k}{w_\mu} = \frac{\lambda_\mu}{\lambda_q}$ and condition (A.10) reduces to $B_k - B_\mu\frac{\lambda_k}{\lambda_\mu} = 0$, which will be satisfied if also social preferences depend only on the expected wait. ■

Proof of Proposition 8 If w cannot be priced either, the regulator's FOCs reduce to the first FOC:

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \end{pmatrix}\left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix}\begin{pmatrix} p_\lambda^{TB} \\ 0 \\ 0 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Use (A.9) to replace $\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix}$ by $\begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB+++} \\ p_w^{FB+++} \\ p_k^{FB+++} \end{pmatrix}$, $^{+++}$ indicating that prices are evaluated at the TB values for q, μ, k. Then

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} \end{pmatrix} \left[\begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB+++} \\ p_w^{FB+++} \\ p_k^{FB+++} \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & 0 \\ \lambda_k & w_k^e & 1 \\ \lambda_\mu & w_\mu^e & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{TB} \\ 0 \\ 0 \end{pmatrix} \right] = 0.$$

Assuming that $(p_w^{FB+++}, p_k^{FB+++}) \neq (0,0)$, and using the definitions of $\frac{\partial \lambda^*}{\partial p_\lambda}$ and $\frac{\partial w^{e*}}{\partial p_\lambda}$ gives

$$\begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \end{pmatrix} \begin{pmatrix} p_\lambda^{FB+++} \\ p_w^{FB+++} \\ p_k^{FB+++} \end{pmatrix} = \frac{\partial \lambda^*}{\partial p_\lambda} p_\lambda^{TB}$$

$$\Downarrow$$

$$p_\lambda^{TB} = p_\lambda^{FB+++} + \frac{\frac{\partial w^{e*}}{\partial p_\lambda}}{\frac{\partial \lambda^*}{\partial p_\lambda}} p_w^{FB+++} + \frac{\frac{\partial k}{\partial p_\lambda}}{\frac{\partial \lambda^*}{\partial p_\lambda}} p_k^{FB+++}.$$

Using the definitions in (37), this can be rewritten as (38). ■

Proof of Proposition 9 From (38) the third best output price will coincide with its first best value when both $p_w^{FB+} = 0$ and $p_k^{FB+} = 0$. Using (29), the condition that $p_w^{FB+} = 0$ is equivalent to

$$\begin{aligned} \beta \left(B_\mu - B_q \frac{\lambda_\mu}{\lambda_q} \right) &= \left(c_\mu^I - c_q^I \frac{\lambda_\mu}{\lambda_q} \right), \\ &= \sigma w_\mu \int_{y_{\min}}^{\widehat{y}} y \mathrm{d}F, \end{aligned}$$

which will only hold when preferences are quasi-separable and insurance costs are absent such that both sides of the expression equal zero. The condition that $p_k^{FB+} = 0$ then reduces to (see (31))

$$\begin{aligned} \beta \left(B_k - B_q \frac{\lambda_k}{\lambda_q} \right) &= \left(c_k^I - c_q^I \frac{\lambda_\mu}{\lambda_q} \right) \\ &= \sigma w_k \int_{y_{\min}}^{\widehat{y}} y \mathrm{d}F, \end{aligned}$$

which also holds under the same conditions. ■

Proof of Proposition 10 The proof consists of three parts. In the first part, we derive the first best pricing and cost-sharing rules. The second part then derives the second-best rules (giving (39)). The last part derives the third-best rules (giving (40)).

(a) We first solve for the first best pricing/cost-sharing scheme. When the hospital

faces prices p_λ, p_k and p_w and is refunded a fraction φ of its costs it solves

$$\max_{q,k,\mu} p_\lambda \lambda(q,k,\mu) + p_k k + p_w w^e(q,k,\mu) - (1-\varphi) c^{He}(q,k,\mu,t) + \alpha B^e(q,k,\mu) - g(t) + T, \quad \text{(A.11)}$$

where T is the lump sum amount that the regulator pays the hospital up front in order that it breaks even after accounting for the cost of cost-reducing effort, g:[31]

$$T = g - p_\lambda \lambda - p_k k - p_\mu \mu + (1-\varphi) c^{He}(q,k,\mu).$$

The first order conditions for an interior solution are

$$p_\lambda \lambda_z + p_k \mathbf{1}_{(z=k)} + p_w w_z^e + \alpha B_z^e = (1-\varphi) c_z^{He} \quad (z = q,k,\mu), \text{ and} \quad \text{(A.12)}$$

$$-(1-\varphi) c_t^{He} = g'(t). \quad \text{(A.13)}$$

and yield optimal hospital decisions $q(p_\lambda, p_k, p_w, \varphi)$, $k(p_\lambda, p_k, p_w, \varphi)$, $\mu(p_\lambda, p_k, p_w, \varphi)$, and $t(p_\lambda, p_k, p_w, \varphi)$.

The regulation problem is then

$$\max_{p_\lambda, p_k, p_w} B^e(q,k,\mu) - (1+\theta)[p_\lambda \lambda(q,k,\mu) + p_k k + p_w w(q,k,\mu) + \varphi c^{He}(q,k,\mu,t) + T + c^{Ie}(q,k,\mu)]$$

which since $T = g(t) - p_\lambda \lambda(q,k,\mu) - p_k k - p_w w(q,k,\mu) + (1-\varphi) c^{He}(q,k,\mu,t)$ is

$$\begin{gathered}
\max_{p_\lambda, p_k, p_w} B^e(q(p_\lambda, p_k, p_w, \varphi), k(p_\lambda, p_k, p_w, \varphi), \mu(p_\lambda, p_k, p_w, \varphi)) \\
-(1+\theta) c^{He}(q(p_\lambda, p_k, p_w, \varphi), k(p_\lambda, p_k, p_w, \varphi), \mu(p_\lambda, p_k, p_w, \varphi), t(p_\lambda, p_k, p_w, \varphi)) \\
-(1+\theta) c^{Ie}(q(p_\lambda, p_k, p_w, \varphi), k(p_\lambda, p_k, p_w, \varphi), \mu(p_\lambda, p_k, p_w, \varphi)) \\
-(1+\theta) g(t(p_\lambda, p_k, p_w, \varphi)).
\end{gathered}$$

The regulator's FOCs are

$$\begin{pmatrix}
\frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\
\frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} & \frac{\partial t}{\partial p_w} \\
\frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \\
\frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k} & \frac{\partial t}{\partial p_k}
\end{pmatrix}
\begin{pmatrix}
B_q^e - (1+\theta)[c_q^{He} + c_q^{Ie}] \\
B_k^e - (1+\theta)[c_k^{He} + c_k^{Ie}] \\
B_\mu^e - (1+\theta)[c_\mu^{He} + c_\mu^{Ie}] \\
0 - (1+\theta)[c_t^{He} + g']
\end{pmatrix}
=
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The hospital FOC w.r.t. z $(z = q,k,\mu)$ is

$$p_\lambda \lambda_z + p_k \mathbf{1}_{(z=k)} + p_w w_z^e + \alpha B_z^e + \varphi c_Z^{He} = c_z^{He} \quad (z = q,k,\mu),$$

so that

$$B_z^e - (1+\theta)[c_z^{He} + c_z^{Ie}] = [1 - \alpha(1+\theta)] B_z^e - (1+\theta)\left(p_\lambda \lambda_z + p_k \mathbf{1}_{(z=k)} + p_w w_z^e + \varphi c_z^{He}\right)$$

[31] The break-even condition applies to the monetary part of the revenue, i.e., excluding αB^e.

or equivalently

$$\frac{B_z^e}{1+\theta} - [c_z^{He} + c_z^{Ie}] = \beta B_z^e - \left(p_\lambda \lambda_z + p_k \mathbf{1}_{(z=k)} + p_w w_z^e + \varphi c_z^{He}\right) - c_z^{Ie}.$$

Then the regulator's FOCs can be written as (since $S_\varphi^e = 0$)

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} & \frac{\partial t}{\partial p_w} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \\ \frac{\partial q}{\partial p_k} & \frac{\partial k}{\partial p_k} & \frac{\partial \mu}{\partial p_k} & \frac{\partial t}{\partial p_k} \end{pmatrix} \left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \\ 0 \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs} \\ p_w^{FBcs} \\ \varphi^{FBcs} \\ p_k^{FBcs} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Since the (4×4) matrix of responses is invertible, these FOCs are equivalent to

$$\begin{aligned} \begin{pmatrix} p_\lambda^{FBcs} \\ p_w^{FBcs} \\ \varphi^{FBcs} \\ p_k^{FBcs} \end{pmatrix} &= \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix}^{-1} \begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{w_\mu^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} & 0 & -\frac{w_q^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} \\ -\frac{\lambda_\mu}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} & 0 & \frac{\lambda_q}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} \\ 0 & 0 & 0 \\ -\frac{\lambda_k w_\mu^e - \lambda_\mu w_k^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} & 1 & \frac{\lambda_k w_q^e - \lambda_q w_k^e}{\lambda_q w_\mu^e - \lambda_\mu w_q^e} \end{pmatrix} \begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \end{pmatrix}, \end{aligned} \tag{A.14}$$

which gives $\varphi^{FBcs} = 0$ (and therefore $(p_\lambda^{FBcs}, p_w^{FBcs}, p_k^{FBcs})$ as derived earlier in Proposition 5).

(b) When beds can no longer be rewarded, the regulator has three instruments: p_λ, p_w and φ. The hospital's objective function reduces to

$$p_\lambda \lambda(q,k,\mu) + p_w w^e(q,k,\mu) + \alpha B^e(q,k,\mu) - (1-\varphi) c^{He}(q,k,\mu)$$

and in (A.12) the term $p_k \mathbf{1}_{(z=k)}$ is absent.

When k cannot be priced, the regulator's FOCs reduce to the first three rows:

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} & \frac{\partial t}{\partial p_w} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \\ 0 \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{SBcs} \\ p_w^{SBcs} \\ \varphi^{SBcs} \\ 0 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

We can then use (A.14) to replace $\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \\ 0 \end{pmatrix}$ by $\begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs++} \\ p_w^{FBcs++} \\ 0 \\ p_k^{FBcs++} \end{pmatrix}$,

with $^{++}$ indicating that prices are evaluated at the SB values for q, μ, k. Then

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} & \frac{\partial t}{\partial p_w} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \times$$
$$\left[\begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs++} \\ p_w^{FBcs++} \\ 0 \\ p_k^{FBcs++} \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{SBcs} \\ p_w^{SBcs} \\ \varphi^{SBcs} \\ 0 \end{pmatrix} \right].$$

This system of equations will hold in two distinct cases.

(i) the FB price for k is zero. This is the case discussed in section 4.2**.** In that case the square bracket term is the zero vector when also $p_\lambda^{SBcs} = p_\lambda^{FBcs}$ and $p_w^{SBcs} = p_w^{FBcs}$ and $\varphi^{SBcs} = 0$ (the (4×4) matrix being invertible).

(ii) when the FB price for k is different from zero. Then the (4×1) square bracket vector should be orthogonal to the three row vectors of responses w.r.t. the prices p_λ, p_w and φ. This is equivalent to

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} & \frac{\partial t}{\partial p_w} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs++} \\ p_w^{FBcs++} \\ 0 \\ p_k^{FBcs++} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial p_w} & \frac{\partial k}{\partial p_w} & \frac{\partial \mu}{\partial p_w} & \frac{\partial t}{\partial p_w} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{SBcs} \\ p_w^{SBcs} \\ \varphi^{SBcs} \\ 0 \end{pmatrix}.$$

Defining $\frac{\partial \lambda^*}{\partial m} \stackrel{\text{def}}{=} \sum_{z=q,k,\mu} \lambda_z \frac{\partial z}{\partial m}$, $\frac{\partial w^{e*}}{\partial m} \stackrel{\text{def}}{=} \sum_{z=q,k,\mu} w_z^e \frac{\partial z}{\partial m}$ and $\frac{\partial c^{He*}}{\partial m} \stackrel{\text{def}}{=} \sum_{z=q,k,\mu,t} c_z^{He} \frac{\partial z}{\partial m}$ $(m = p_\lambda, p_w, \varphi)$ (i.e., the demand, expected wait and cost responses that follow from the hospital's optimal responses in q, k, μ), we get:

$$\begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial c^{He*}}{\partial p_w} & \frac{\partial k}{\partial p_w} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} & \frac{\partial k}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs++} \\ p_w^{FBcs++} \\ 0 \\ p_k^{FBcs++} \end{pmatrix} = \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial c^{He*}}{\partial p_w} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{SBcs} \\ p_w^{SBcs} \\ \varphi^{SBcs} \end{pmatrix}.$$

Below in Lemma A.3 we show that the (3×3) matrix on the RHS is positive definite and

thus invertible. Then

$$\begin{pmatrix} p_\lambda^{SBcs} \\ p_w^{SBcs} \\ \varphi^{SBcs} \end{pmatrix} = \begin{pmatrix} \frac{\partial\lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} \\ \frac{\partial\lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial c^{He*}}{\partial p_w} \\ \frac{\partial\lambda^*}{\partial\varphi} & \frac{\partial w^{e*}}{\partial\varphi} & \frac{\partial c^{He*}}{\partial\varphi} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial\lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial\lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial c^{He*}}{\partial p_w} & \frac{\partial k}{\partial p_w} \\ \frac{\partial\lambda^*}{\partial\varphi} & \frac{\partial w^{e*}}{\partial\varphi} & \frac{\partial c^{He*}}{\partial\varphi} & \frac{\partial k}{\partial\varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs++} \\ p_w^{FBcs++} \\ 0 \\ p_k^{FBcs++} \end{pmatrix}$$

$$\begin{pmatrix} p_\lambda^{SB} \\ p_w^{SB} \\ \varphi^{SB} \end{pmatrix} = \begin{pmatrix} p_\lambda^{FBcs++} \\ p_w^{FBcs++} \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{\partial\lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} \\ \frac{\partial\lambda^*}{\partial p_w} & \frac{\partial w^{e*}}{\partial p_w} & \frac{\partial c^{He*}}{\partial p_w} \\ \frac{\partial\lambda^*}{\partial\varphi} & \frac{\partial w^{e*}}{\partial\varphi} & \frac{\partial c^{He*}}{\partial\varphi} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial k}{\partial p_w} \\ \frac{\partial k}{\partial\varphi} \end{pmatrix} p_k^{FB++}.$$

We now combine the two outcome variables (λ, w) in the vector Λ' and define the corresponding price vector $p'_\Lambda = (p_\lambda, p_w)$. Then

$$\begin{pmatrix} p_\Lambda^{SBcs} \\ \varphi^{SBcs} \end{pmatrix} = \begin{pmatrix} p_\Lambda^{FBcs++} \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda} & \frac{\partial c^{He*}}{\partial p_\Lambda} \\ \frac{\partial\Lambda^{*\prime}}{\partial\varphi} & \frac{\partial c^{He*}}{\partial\varphi} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial k}{\partial p_\Lambda} \\ \frac{\partial k}{\partial\varphi} \end{pmatrix} p_k^{FBcs++}.$$

Using the formulae for partitioned matrix inversion we can write the solution as

$$\begin{pmatrix} p_\Lambda^{SBcs} \\ \varphi^{SBcs} \end{pmatrix} = \begin{pmatrix} p_\Lambda^{FBcs++} \\ 0 \end{pmatrix} + \begin{pmatrix} D^{-1}\left[\frac{\partial k}{\partial p_\Lambda} - \frac{\partial c^{He*}}{\partial p_\Lambda}\left(\frac{\partial c^{He*}}{\partial\varphi}\right)^{-1}\frac{\partial k}{\partial\varphi}\right] \\ C^{-1}\left[\frac{\partial k}{\partial\varphi} - \frac{\partial\Lambda^{*\prime}}{\partial\varphi}\left(\frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda}\right)^{-1}\frac{\partial k}{\partial p_\Lambda}\right] \end{pmatrix} p_k^{FBcs++},$$

where $C = \frac{\partial c^{He*}}{\partial\varphi} - \frac{\partial\Lambda^{*\prime}}{\partial\varphi}\left(\frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\Lambda}$ and $D = \frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda} - \frac{\partial c^{He*}}{\partial p_\Lambda}\left(\frac{\partial c^{He*}}{\partial\varphi}\right)^{-1}\frac{\partial\Lambda^{*\prime}}{\partial\varphi}$.

Therefore

$$\begin{aligned} p_\Lambda^{SBcs} &= p_\Lambda^{FBcs++} + \left[\frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda} - \frac{\partial\Lambda^{*\prime}}{\partial\varphi}\left(\frac{\partial c^{He*}}{\partial\varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\Lambda}\right]^{-1} \\ &\quad \times \left[\frac{\partial k}{\partial p_\Lambda} - \frac{\partial k}{\partial\varphi}\left(\frac{\partial c^{He*}}{\partial\varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\Lambda}\right] p_k^{FB++} \\ \varphi^{SBcs} &= \frac{\frac{\partial k}{\partial\varphi} - \frac{\partial k}{\partial p'_\Lambda}\left(\frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda}\right)^{-1}\frac{\partial\Lambda^*}{\partial\varphi}}{\frac{\partial c^{He*}}{\partial\varphi} - \frac{\partial c^{He*}}{\partial p'_\Lambda}\left(\frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda}\right)^{-1}\frac{\partial\Lambda^*}{\partial\varphi}} p_k^{FBcs++}. \end{aligned}$$

Define $\frac{\partial k}{\partial\varphi}|_{\mathrm{d}\Lambda=0} = \frac{\partial k}{\partial\varphi} - \frac{\partial k}{\partial p'_\Lambda}\left(\frac{\partial\Lambda^*}{\partial p_\Lambda}\right)^{-1}\frac{\partial\Lambda^*}{\partial\varphi}$ and $\frac{\partial c^{He*}}{\partial\varphi}|_{\mathrm{d}\Lambda=0} = \frac{\partial c^{He*}}{\partial\varphi} - \frac{\partial c^{He*}}{\partial p'_\Lambda}\left(\frac{\partial\Lambda^{*\prime}}{\partial p_\Lambda}\right)^{-1}\frac{\partial\Lambda^*}{\partial\varphi}$. Then

$$\varphi^{SBcs} = \frac{\frac{\partial k}{\partial\varphi}|_{\mathrm{d}\Lambda=0}}{\frac{\partial c^{He*}}{\partial\varphi}|_{\mathrm{d}\Lambda=0}} p_k^{FB++}.$$

Below in Lemma A.3 we show that $\frac{\partial c^{He*}}{\partial\varphi}|_{\mathrm{d}\Lambda=0} > 0$ (although smaller than $\frac{\partial c^{He*}}{\partial\varphi} > 0$ by the LeChatelier principle). Suppose that a higher cost sharing rate increases the number of installed beds. Then $\varphi^{SBcs} > 0$. Notice that $\frac{\frac{\partial k}{\partial\varphi}|_{\mathrm{d}\Lambda=0}}{\frac{\partial c^{He*}}{\partial\varphi}|_{\mathrm{d}\Lambda=0}}$ can be interpreted as the

inverse of a 'induced response' marginal cost of beds, i.e., $MC_k^{\rm ir} \stackrel{\rm def}{=} \frac{\frac{\partial c^{He*}}{\partial \varphi}|_{{\rm d}\Lambda=0}}{\frac{\partial k}{\partial \varphi}|_{{\rm d}\Lambda=0}}$. Then $\varphi^{SBcs} MC_k^{\rm ir} = p_k^{FBcs++}$. This proves (39).

The SB prices for λ and w can also be written as

$$\begin{aligned}\left[\frac{\partial \Lambda^{*\prime}}{\partial p_\Lambda} - \frac{\partial \Lambda^{*\prime}}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\Lambda}\right] p_\Lambda^{SBcs} &= \left[\frac{\partial \Lambda^{*\prime}}{\partial p_\Lambda} - \frac{\partial \Lambda^{*\prime}}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\Lambda}\right] p_\Lambda^{FBcs++} \\ &\quad + \left[\frac{\partial k}{\partial p_\Lambda} - \frac{\partial k}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\Lambda}\right] p_k^{FBcs++} \\ &\Downarrow \\ \frac{\partial \Lambda^{*\prime}}{\partial p_\Lambda}|_{{\rm d}c^H=0} p_\Lambda^{SBcs} &= \frac{\partial \Lambda^{*\prime}}{\partial p_\Lambda}|_{{\rm d}c^H=0} p_\Lambda^{FBcs++} + \frac{\partial k}{\partial p_\Lambda}|_{{\rm d}c^H=0} p_k^{FBcs++}.\end{aligned}$$

i.e., the SB marginal increase in hospital revenue from patient and waiting time reward should equal the FB revenue increase from these rewards plus the FB revenue increase from the bed reward. The triggered responses are conditioned on hospital costs being kept constant (since that outcome is rewarded separately).

(c) When w cannot be priced either, the regulator's FOCs reduce to the first and last FOCs (because p_w is no longer a decision variable):

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \left[\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \\ 0 \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{TBcs} \\ 0 \\ \varphi^{TBcs} \\ 0 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Use (A.9) to replace $\begin{pmatrix} S_q^e \\ S_k^e \\ S_\mu^e \\ 0 \end{pmatrix}$ by $\begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs+++} \\ p_w^{FBcs+++} \\ 0 \\ p_k^{FBcs+++} \end{pmatrix}$, $^{+++}$ indicating that prices are evaluated at the TB values for q, μ, k. Then

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \times$$

$$\left[\begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs+++} \\ p_w^{FBcs+++} \\ 0 \\ p_k^{FBcs+++} \end{pmatrix} - \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{TBcs} \\ 0 \\ \varphi^{TBcs} \\ 0 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This system of equations will hold in two distinct cases.

(i) the FB price for k and w are zero. In that case the square bracket term is the zero vector when also $p_\lambda^{TB} = p_\lambda^{FB}$ and $\varphi^{TB} = 0$(the (4×4) matrix being invertible).

(ii) when the FB prices for k and w are different from zero. Then the (4×1) square bracket vector should be orthogonal to the two row vectors of responses w.r.t. the prices

p_λ and φ. This is equivalent to

$$\begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{FB+++} \\ p_w^{FB+++} \\ 0 \\ p_k^{FB+++} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial q}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} & \frac{\partial \mu}{\partial p_\lambda} & \frac{\partial t}{\partial p_\lambda} \\ \frac{\partial q}{\partial \varphi} & \frac{\partial k}{\partial \varphi} & \frac{\partial \mu}{\partial \varphi} & \frac{\partial t}{\partial \varphi} \end{pmatrix} \begin{pmatrix} \lambda_q & w_q^e & c_q^{He} & 0 \\ \lambda_k & w_k^e & c_k^{He} & 1 \\ \lambda_\mu & w_\mu^e & c_\mu^{He} & 0 \\ 0 & 0 & c_t^{He} & 0 \end{pmatrix} \begin{pmatrix} p_\lambda^{TB} \\ 0 \\ \varphi^{TB} \\ 0 \end{pmatrix}.$$

Using the definitions for $\frac{\partial \lambda^*}{\partial m}$, $\frac{\partial w^{e*}}{\partial m}$ and $\frac{\partial c^{He*}}{\partial m}$ yields:

$$\begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} & \frac{\partial k}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs+++} \\ p_w^{FBcs+++} \\ 0 \\ p_k^{FBcs+++} \end{pmatrix} = \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{TBcs} \\ \varphi^{TBcs} \end{pmatrix}$$

Rearranging rows and columns we have

$$\begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} & \frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial k}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs+++} \\ 0 \\ p_w^{FBcs+++} \\ p_k^{FBcs+++} \end{pmatrix} = \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{TBcs} \\ \varphi^{TBcs} \end{pmatrix}$$

$$\begin{pmatrix} p_\lambda^{TBcs} \\ \varphi^{TBcs} \end{pmatrix} = \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial c^{He*}}{\partial p_\lambda} & \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi} & \frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial k}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_\lambda^{FBcs+++} \\ 0 \\ p_w^{FBcs+++} \\ p_k^{FBcs+++} \end{pmatrix}$$

$$= \begin{pmatrix} p_\lambda^{FBcs+++} \\ 0 \end{pmatrix} + \frac{1}{\frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial c^{He*}}{\partial \varphi} - \frac{\partial c^{He*}}{\partial p_\lambda}\frac{\partial \lambda^*}{\partial \varphi}} \begin{pmatrix} \frac{\partial c^{He*}}{\partial \varphi} & -\frac{\partial c^{He*}}{\partial p_\lambda} \\ -\frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial \lambda^*}{\partial p_\lambda} \end{pmatrix} \times \begin{pmatrix} \frac{\partial w^{e*}}{\partial p_\lambda} & \frac{\partial k}{\partial p_\lambda} \\ \frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial k}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_w^{FBcs+++} \\ p_k^{FBcs+++} \end{pmatrix}$$

$$= \begin{pmatrix} p_\lambda^{FBcs+++} \\ 0 \end{pmatrix} + \frac{1}{\frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial c^{He*}}{\partial \varphi} - \frac{\partial c^{He*}}{\partial p_\lambda}\frac{\partial \lambda^*}{\partial \varphi}} \times \begin{pmatrix} \frac{\partial c^{He*}}{\partial \varphi}\frac{\partial w^{e*}}{\partial p_\lambda} - \frac{\partial c^{He*}}{\partial p_\lambda}\frac{\partial w^{e*}}{\partial \varphi} & \frac{\partial c^{He*}}{\partial \varphi}\frac{\partial k}{\partial p_\lambda} - \frac{\partial c^{He*}}{\partial p_\lambda}\frac{\partial k}{\partial \varphi} \\ -\frac{\partial \lambda^*}{\partial \varphi}\frac{\partial w^{e*}}{\partial p_\lambda} + \frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial w^{e*}}{\partial \varphi} & -\frac{\partial \lambda^*}{\partial \varphi}\frac{\partial k}{\partial p_\lambda} + \frac{\partial \lambda^*}{\partial p_\lambda}\frac{\partial k}{\partial \varphi} \end{pmatrix} \begin{pmatrix} p_w^{FBcs+++} \\ p_k^{FBcs+++} \end{pmatrix}.$$

Therefore

$$p_\lambda^{TB} = p_\lambda^{FB++} + \frac{\frac{\partial w^{e*}}{\partial p_\lambda} - \frac{\partial w^{e*}}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\lambda}}{\frac{\partial \lambda^*}{\partial p_\lambda} - \frac{\partial \lambda^*}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\lambda}} p_w^{FB+++} + \frac{\frac{\partial k}{\partial p_\lambda} - \frac{\partial k}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\lambda}}{\frac{\partial \lambda^*}{\partial p_\lambda} - \frac{\partial \lambda^*}{\partial \varphi}\left(\frac{\partial c^{He*}}{\partial \varphi}\right)^{-1}\frac{\partial c^{He*}}{\partial p_\lambda}} p_k^{FB+++},$$

$$\varphi^{TB} = \varphi^{FB+++} + \frac{\frac{\partial w^{e*}}{\partial \varphi} - \frac{\partial w^{e*}}{\partial p_\lambda}\left(\frac{\partial \lambda^*}{\partial p_\lambda}\right)^{-1}\frac{\partial \lambda^*}{\partial \varphi}}{\frac{\partial c^{He*}}{\partial \varphi} - \frac{\partial c^{He*}}{\partial p_\lambda}\left(\frac{\partial \lambda^*}{\partial p_\lambda}\right)^{-1}\frac{\partial \lambda^*}{\partial \varphi}} p_w^{FB+++} + \frac{\frac{\partial k}{\partial \varphi} - \frac{\partial k}{\partial p_\lambda}\left(\frac{\partial \lambda^*}{\partial p_\lambda}\right)^{-1}\frac{\partial \lambda^*}{\partial \varphi}}{\frac{\partial c^{He*}}{\partial \varphi} - \frac{\partial c^{He*}}{\partial p_\lambda}\left(\frac{\partial \lambda^*}{\partial p_\lambda}\right)^{-1}\frac{\partial \lambda^*}{\partial \varphi}} p_k^{FB+++},$$

or

$$p_\lambda^{TBcs} = p_\lambda^{FB+++} + \frac{\frac{\partial w^{e*}}{\partial p_\lambda}|_{\mathrm{d}c^H=0}}{\frac{\partial \lambda^*}{\partial p_\lambda}|_{\mathrm{d}c^H=0}} p_w^{FB+++} + \frac{\frac{\partial k}{\partial p_\lambda}|_{\mathrm{d}c^H=0}}{\frac{\partial \lambda^*}{\partial p_\lambda}|_{\mathrm{d}c^H=0}} p_k^{FB+++} + \frac{\frac{\partial t}{\partial p_\lambda}|_{\mathrm{d}c^H=0}}{\frac{\partial \lambda^*}{\partial p_\lambda}|_{\mathrm{d}c^H=0}} \varphi^{FB+++},$$

$$\varphi^{TBcs} = \frac{\frac{\partial w^{e*}}{\partial \varphi}|_{\mathrm{d}\lambda=0}}{\frac{\partial c^{He*}}{\partial \varphi}|_{\mathrm{d}\lambda=0}} p_w^{FB+++} + \frac{\frac{\partial k}{\partial \varphi}|_{\mathrm{d}\lambda=0}}{\frac{\partial c^{He*}}{\partial \varphi}|_{\mathrm{d}\lambda=0}} p_k^{FB+++}.$$

Multiplying the last equation through by $MC_k^{\mathrm{ir}}|_{\mathrm{d}\lambda=0} \stackrel{\mathrm{def}}{=} \frac{\frac{\partial c^{He*}}{\partial \varphi}|_{\mathrm{d}\lambda=0}}{\frac{\partial k}{\partial \varphi}|_{\mathrm{d}\lambda=0}}$, and using the definition $\widehat{w}_k^{\mathrm{ir}}|_{\mathrm{d}\lambda=0} \stackrel{\mathrm{def}}{=} \frac{\frac{\partial w^{e*}}{\partial \varphi}|_{\mathrm{d}\lambda=0}}{\frac{\partial k}{\partial \varphi}|_{\mathrm{d}\lambda=0}}$ results in (40). ■

Lemma A.3 *In the second best cost sharing case, the matrix* $\begin{pmatrix} \frac{\partial \lambda^*}{\partial p_\lambda} & \frac{\partial w^*}{\partial p_\lambda} & \frac{\partial c^*}{\partial p_\lambda} \\ \frac{\partial \lambda^*}{\partial p_w} & \frac{\partial w^*}{\partial p_w} & \frac{\partial c^*}{\partial p_w} \\ \frac{\partial \lambda^*}{\partial \varphi} & \frac{\partial w^*}{\partial \varphi} & \frac{\partial c^*}{\partial \varphi} \end{pmatrix}$ *is symmetric and positive definite and* $\frac{\partial c^*}{\partial \varphi}|_{d\lambda^*=dw^*=0}$ *is positive.*

Proof of Lemma A.3 Totally differentiate the FOCs of the hospital. Let H be the negative definite Hessian of the objective function. Let $z = (q, k, \mu)'$ and drop the H-superscript with the hospital cost function:

$$\begin{pmatrix} H_{zz} & H_{zt} \\ H_{tz} & H_{tt} \end{pmatrix} \begin{pmatrix} \mathrm{d}z \\ \mathrm{d}t \end{pmatrix} = -\begin{pmatrix} \lambda_z & w_z & c_z^e \\ 0 & 0 & c_t^e \end{pmatrix} \begin{pmatrix} \mathrm{d}p_\lambda \\ \mathrm{d}p_w \\ \mathrm{d}\varphi \end{pmatrix}$$

$$\begin{pmatrix} \mathrm{d}z \\ \mathrm{d}t \end{pmatrix} = -\begin{pmatrix} H_{zz} & H_{zt} \\ H_{tz} & H_{tt} \end{pmatrix}^{-1} \begin{pmatrix} \lambda_z & w_z & c_z^e \\ 0 & 0 & c_t^e \end{pmatrix} \begin{pmatrix} \mathrm{d}p_\lambda \\ \mathrm{d}p_w \\ \mathrm{d}\varphi \end{pmatrix}$$

Let $K \stackrel{\mathrm{def}}{=} H^{-1}$ so that

$$\begin{pmatrix} \mathrm{d}z \\ \mathrm{d}t \end{pmatrix} = -\begin{pmatrix} K_{zz} & K_{zt} \\ K_{tz} & K_{tt} \end{pmatrix} \begin{pmatrix} \lambda_z & w_z & c_z^e \\ 0 & 0 & c_t^e \end{pmatrix} \begin{pmatrix} \mathrm{d}p_\lambda \\ \mathrm{d}p_w \\ \mathrm{d}\varphi \end{pmatrix}$$

$$\begin{pmatrix} \mathrm{d}z \\ \mathrm{d}t \end{pmatrix} = -\begin{pmatrix} K_{zz}\lambda_z & K_{zz}w_z & K_{zz}c_z^e + K_{zt}c_t^e \\ K_{tz}\lambda_z & K_{tz}w_z & K_{tz}c_z^e + K_{tt}c_t^e \end{pmatrix} \begin{pmatrix} \mathrm{d}p_\lambda \\ \mathrm{d}p_w \\ \mathrm{d}\varphi \end{pmatrix}$$

Define $\Lambda = \binom{\lambda}{w}$ and $\Lambda_z = \binom{\lambda_z'}{w_z'}$ where $\lambda_z = (\lambda_q, \lambda_k, \lambda_\mu)'$ and $w_z = (w_q, w_k, w_\mu)'$. Also define $c_z^{e\prime} = (c_q^e, c_k^e, c_\mu^e)$ Then $\mathrm{d}\Lambda = \Lambda_z \mathrm{d}z$, and $\mathrm{d}c = c_z^{e\prime}\mathrm{d}z + c_t^e \mathrm{d}t$, and we can write

$$\begin{aligned}
\begin{pmatrix} \mathrm{d}\Lambda \\ \mathrm{d}c \end{pmatrix} &= -\begin{pmatrix} \Lambda_z & \underline{0} \\ c_z^{e\prime} & c_t^e \end{pmatrix}\begin{pmatrix} K_{zz} & K_{zt} \\ K_{tz} & K_{tt} \end{pmatrix}\begin{pmatrix} \Lambda_z' & c_z^e \\ \underline{0}' & c_t^e \end{pmatrix}\begin{pmatrix} \mathrm{d}p_\Lambda \\ \mathrm{d}\varphi \end{pmatrix} \\
&= -\begin{pmatrix} \Lambda_z K_{zz}\Lambda_z' & \Lambda_z\left(K_{zz}c_z^e + K_{zt}c_t^e\right) \\ \left(c_z^{e\prime}K_{zz} + c_t^e K_{tz}\right)\Lambda_z' & c_z^{e\prime}K_{zz}c_z^e + 2c_t^e K_{tz}c_z^e + c_t^e K_{tt}c_t^e \end{pmatrix}\begin{pmatrix} \mathrm{d}p_\Lambda \\ \mathrm{d}\varphi \end{pmatrix}
\end{aligned}$$

Since the K matrix is negative definite by the second order condition, the matrix $\begin{pmatrix} \frac{\mathrm{d}\Lambda}{\mathrm{d}p_\Lambda'} & \frac{\mathrm{d}\Lambda}{\mathrm{d}\varphi} \\ \frac{\mathrm{d}c}{\mathrm{d}p_\Lambda'} & \frac{\mathrm{d}c}{\mathrm{d}\varphi} \end{pmatrix}$ is symmetric and positive definite, implying that $\frac{\mathrm{d}\lambda^*}{\mathrm{d}p_w} = \frac{\mathrm{d}w^*}{\mathrm{d}p_\lambda}$, $\frac{\mathrm{d}\lambda^*}{\mathrm{d}\varphi} = \frac{\mathrm{d}c^*}{\mathrm{d}p_\lambda}$, $\frac{\mathrm{d}w^*}{\mathrm{d}\varphi} = \frac{\mathrm{d}c^*}{\mathrm{d}p_w}$ and $\frac{\mathrm{d}\lambda^*}{\mathrm{d}p_\lambda} > 0$, $\frac{\mathrm{d}w^*}{\mathrm{d}p_w} > 0$ and $\frac{\mathrm{d}c^*}{\mathrm{d}\varphi} > 0$.

Now suppose that the vector Λ cannot change: $\mathrm{d}\Lambda = \underline{0}$. Then the prices for output and waiting time need to adjust to keep λ and w constant:

$$\mathrm{d}p_\Lambda = -\left(\Lambda_z K_{zz}\Lambda_z'\right)^{-1}\Lambda_z\left(K_{zz}c_z^e + K_{zt}c_t^e\right)\mathrm{d}\varphi$$

Therefore, the induced response change in hospital cost following $\mathrm{d}\varphi$ when output and waiting time are held fixed is given by

$$\begin{aligned}
\frac{\mathrm{d}c^e}{\mathrm{d}\varphi}|_{\mathrm{d}\Lambda=\underline{0}} &= \left(c_z^{e\prime}K_{zz} + c_t^e K_{tz}\right)\Lambda_z'\left(\Lambda_z K_{zz}\Lambda_z'\right)^{-1}\Lambda_z\left(K_{zz}c_z^e + K_{zt}c_t^e\right) \\
&\quad - \left[c_z^{e\prime}K_{zz}c_z^e + 2c_t^e K_{tz}c_z^e + c_t^e K_{tt}c_t^e\right]
\end{aligned}$$

which is positive but smaller than $\frac{\mathrm{d}c^e}{\mathrm{d}\varphi}$ (LeChatelier principle). ■