
Article:
Gravelle, Hugh Stanley Emrys orcid.org/0000-0002-7753-4233 and Schroyen, Fred (2020) Optimal hospital payment rules under rationing by waiting. Journal of Health Economics. 102277. ISSN 0167-6296

https://doi.org/10.1016/j.jhealeco.2019.102277

Journal of Health Economics 70 (2020) 102277

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Optimal hospital payment rules under rationing by waiting☆

Hugh Gravelle[a], Fred Schroyen[b,*]

[a] Centre for Health Economics, University of York, United Kingdom
[b] Department of Economics, Norwegian School of Economics, Norway

ARTICLE INFO

Article history:
Received 14 March 2019
Received in revised form 24 October 2019
Accepted 13 December 2019
Available online 10 January 2020

JEL classification:
I11
I13
I18
L51
D81

Keywords:
Rationing
Waiting times
Queues
Prospective payment
Hospitals

ABSTRACT

We derive optimal rules for paying hospitals for non-emergency care when providers choose quality and capacity, and patient demand is rationed by waiting time. Waiting for treatment is costly for patients, so that hospital payment rules should take account of their effect on waiting time as well as on quality. Since deterministic waiting time models imply that profit maximising hospitals will never choose to have both positive quality and positive waiting time, we develop a stochastic model of rationing by waiting in which both quality and expected waiting are positive in equilibrium. We use it to show that, although a prospective output price gives hospitals an incentive to attract patients by raising quality and reducing waiting times, it must be supplemented by a price attached to hospital decisions on quality or capacity or to a performance indicator which depends on those decisions (such as average waiting time, or average length of stay). A prospective output price by itself can support the optimal quality and waiting time distribution only if the welfare function respects patient preferences over quality and waiting time, if patients' marginal rates of substitution between quality and waiting time are independent of income, and if waiting for treatment does not reduce the productivity of patients. If these conditions do not hold, supplementing the output price with a reward linked to the hospital's cost can increase welfare, though it is possible that costs should be taxed rather than subsidised.

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 elective (non-emergency) treatment is rationed by waiting times which are often long and a source of concern to both patients and policy makers (Cullis et al., 2000; Siciliani and Iversen, 2012).[1] 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; Milstein and Schreyoegg, 2016; Sutton et al., 2012).

In this paper we derive optimal rules for paying hospitals in a public health care system in which patient demand

☆ The paper has benefitted from advice from two anonymous referees. We are also grateful for comments and suggestions from the Editor, Pau Olivella, Gjermund Grimsby, Philippe Bontem and participants at the Annual Conference of the Norwegian Economic Association (Bergen) and at the European Health Economics Workshop (Toulouse) Declaration of interest: none.
* Corresponding author.
E-mail addresses: hugh.gravelle@york.ac.uk (H. Gravelle), fred.schroyen@nhh.no (F. Schroyen).

[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.

https://doi.org/10.1016/j.jhealeco.2019.102277

for elective health care is rationed by waiting time and hospitals can choose quality and make capacity decisions that change the distribution of waiting times facing patients.

In the normative literature on hospital payment systems the policy objective is to induce welfare maximising hospital behaviour: treatment of an optimal number of patients with optimal quality at minimum cost. (See Chalkley and Malcomson (2000) for a review.) In this literature it is assumed that payment cannot be linked directly to unverifiable or unobserved quality or to cost reducing effort. Policy makers are restricted to setting a price for output and to reimbursing hospitals for their costs.

The key paper is Chalkley and Malcomson (1998) which shows that, if there is only one dimension of quality and it is optimal to treat all patients who demand care at the optimal quality, so that there is no rationing, then first best quality and output can be achieved at minimum cost with a single instrument: a prospective output price.[2] Because higher quality attracts more patients and increases revenue, hospitals respond to a higher price by increasing quality. It is 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. 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 all patients have the same marginal rates of substitution between different quality dimensions.

The insights from this literature are obtained from models which do not take account of a salient feature of most public health care systems: rationing by waiting time for non-emergency treatment.[3] Longer waiting times impose costs on patients: their health gain from treatment is delayed, longer delays can worsen treatment outcomes (Nikolova et al., 2015; Reichert and Jacobs, 2018), and patients may be unable to work whilst waiting (Aakvik et al., 2015). These costs from 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) or to forgo care entirely (Gravelle et al., 2002; Martin and Smith, 1999; Windmeijer et al., 2005). In addition to their effects on output and quality, design of payment schemes should therefore take account of their effects on waiting time and on the costs waiting times impose on patients and on income insurance schemes.

Recognition that waiting time affects patient utility from treatment and that there may be production losses if patient labour supply is reduced when waiting for treatment means that, even if there is only one dimension of quality of care, there are two dimensions of hospital decisions which affect welfare. Thus the optimal payment mechanism needs at least one instrument in addition to a prospective output price.

But waiting time cannot be analysed as though it is just another type of quality: it is determined by hospital supply decisions *and* by patient decisions about whether to demand treatment. Analysis of the effect of hospital payment regimes must therefore take account of how equilibrium is established in this market. Demand is uncertain because of the uncertain incidence of illness, and length of stay is also uncertain because of unobserved patient characteristics and supply shocks. Thus waiting times are uncertain and their distribution is determined by both patient and hospital decisions.

Following Lindsay and Feigenbaum (1984), almost all economic models of rationing by waiting in health care assume that demand and supply, and hence waiting time, are certain.[4] In these models 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 not for the analysis of hospital responses to payment regimes when demand is affected by quality as well as waiting time.

To see this, consider a hospital that faces a fixed price for output and does not care about quality *per se* and which has a certain and positive waiting time determined by the equality of certain demand and its certain output. If it reduces quality, holding its output constant, its waiting time will fall to keep demand equal to output. Profit will increase: the reduction in quality will reduce cost and revenue is unchanged because output is unchanged. 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 even if such a combination maximises welfare. 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 intrinsic provider concern with quality.

By contrast, models in which demand and supply, and hence waiting times, are stochastic can be used to analyse hospital choices which affect quality and waiting times. In such a setting, the mean number treated is equal to the

[2] Chalkley and Malcomson (1998) examine three types of solution: (i) optimal output is equal to the number demanding care at the optimal quality ("demand constrained"); (ii) optimal output is equal to the exogenous fixed capacity ("capacity constrained"); (iii) optimal output is less than the minimum of capacity and demand ("unconstrained"). In cases (ii) and (iii), where demand exceeds output, a price attached to the number of patients demanding treatment is required in addition to the price on output, i.e., on the number of patients treated.

[3] In their two solution types where demand exceeds supply Chalkley and Malcomson (1998, pp. 1106–7) note that welfare depends on how patients are rationed but only consider the implications of perfect rationing (all patients treated have higher benefits that those who are not selected for treatment) and of random rationing. 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 by the probability of treatment.

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

mean demand and is strictly less than the capacity of the hospital. The equilibrium mean waiting time is always positive since only an infinitely large capacity can result in all patients having a zero realised waiting time. A reduction in quality with unchanged capacity will reduce hospital cost but, as we show, it will also reduce the equilibrium mean demand and hence the mean number treated and the mean revenue. Hence expected profit could decrease or increase when quality is reduced. The equilibrium of the system will always have positive expected waiting time and, if quality is not too costly, also have positive quality.[5]

Two papers in the health economics literature have considered stochastic waiting time models with demand depending on the distribution of waiting times.[6] 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. The resulting complicated expressions for the steady state probabilities of the number of patients in the system and for 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 simpler queueing model to examine competition between GPs via their choice of quality and expected waiting time.

Our first contribution is to develop a model of rationing by random waiting times which has firm welfare foundations, and is analytically tractable. Because we use our model 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, as in Goddard et al. (1995) and Iversen and Lurås (2002). We use 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 more general specification of the queueing model and of individual preferences.[7] 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 with Goddard et al. (1995), our specification yields an equilibrium steady state distribution of waiting times for the public system with reasonably simple properties.[8]

Our second contribution is to extend Chalkley and Malcomson (1998) by using our waiting time model to derive first and second-best payment schemes for a hospital treating publicly funded elective patients who are rationed by waiting. The payment schemes depend on their effects on the hospital's quality and capacity decisions, and the resulting impacts on the equilibrium waiting time distribution. The hospital bears the costs of capacity and of treating patients but it does not take full account of the benefits of treatment for patients. It also ignores any output losses due to patients being less productive whilst ill and waiting for treatment and the fact that the payments to the hospital are financed by distortionary taxation. A welfare maximum can be achieved with a payment scheme which ensures that the marginal revenues for the hospital from capacity and quality decisions are equal to their residual marginal welfare effects: the marginal social benefits and costs which the hospital would otherwise ignore.

If there is a single quality dimension and a single hospital capacity decision affecting the distribution of patient waiting times, achieving the first best quality and distribution of waiting times requires that the prospective output price must be combined with another instrument, for example, a payment related to average length of stay, or to quality, or to the mean waiting time. The optimal price per patient treated is higher the greater are the marginal social benefits of capacity and quality, the weaker is hospital altruism, the smaller is the marginal cost of public funds, and the greater is the effect of waiting lists and waiting times on the costs of insuring patients against lost earnings whilst waiting for care. If the other price is attached to quality, as in a Pay for Performance scheme, it should be less than the residual marginal welfare effect of quality because the price attached to output is already indirectly incentivising quality given that demand increases with quality.

In the second best where the prospective output price is the only instrument it should exceed the first best price because it has to incentivise two decisions (quality and capacity). In the absence of other policy instruments an output price can support first best quality and capacity decisions only if marginal and infra-marginal patients are willing to trade off waiting times and quality at the same rate, and if there is no loss of output whilst patients are waiting for treatment. If these strong conditions do not hold it is likely that supplementing the output price with a payment linked to hospital costs will increase welfare. However, because welfare depends on both quality and capacity and, at the second best output price, the marginal welfare effects of quality and capacity may have opposite signs, it is possible that hospital costs should be taxed rather than subsidised.

In Section 2 we present the waiting time model and patient choices between public and private treatment, examine the effects of hospital choice of quality and capacity decisions on the equilibrium demand and the waiting

[5] We contrast the equilibria of deterministic and stochastic waiting time models diagrammatically in footnote 22 in Section 2.4. For a fuller comparison of stochastic and deterministic rationing by waiting see Gravelle and Schroyen (2016).

[6] There are stochastic waiting time models of hospitals in the operations research literature (see the survey by Fomundam and Herrmann (2007). But none of these allow for balking: patient decisions to join the waiting list being affected by the distribution of waiting times. Some of the queuing literature does consider balking (Hassin and Haviv, 2003). Here 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 Hildebrand (1975), Naor (1969), and Littlechild (1974).

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

[8] The *ex ante* formulation 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.

time distribution, and set out the welfare function. In Section 3 we derive the first best hospital payment scheme when a prospective price per patient treated is combined with a price linked to a performance measure affected by quality or capacity decisions. Second best pricing rules and cost subsidisation are considered in Section 4. Section 5 concludes.

2. Model

2.1. Queueing model

Our model of the queueing process is general and, unlike most stochastic queueing models, we allow demand (the arrival process) to depend on the distribution of waiting times.[9] Our focus is on obtaining a tractable model of the resulting market equilibrium in order to analyse welfare maximising payment schemes for public hospitals.

We assume that the event that an individual becomes ill and requires elective treatment is identically and independently distributed with probability σ. All patients have the same illness severity and the same health gain from treatment. Those who choose to be treated in the public hospital are placed on a waiting list and are treated in order of arrival: the queue discipline is "first come, first served".

The time w from joining the list to discharge after treatment varies with the number of patients already on the list which depends on the random process generating arrivals and on the random length of stay of patients once admitted. The random arrival rate is determined by the illness probability and by patient decisions about whether to be treated in the public hospital. The mean rate of arrivals at the public hospital per unit of time is λ and we assume that λ completely describes the distribution of arrivals per unit of time.[10] In this and the next section we treat λ as exogenous and then in Section 2.3 we explain how it is determined by the decisions of patients about whether they wish to be treated in the public hospital when ill.

The hospital can influence the distribution of length of stay by decisions on the number of beds, operating theatres, staffing levels, and managerial effort to improve coordination between different departments. We denote these decisions by s and will usually assume that s is a scalar and refer to it as capacity.[11]

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

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

where increases in the arrival rate λ and reductions in capacity s produce first degree stochastic dominating changes in the distribution of waiting times. The mean waiting time is increasing in the arrival rate and decreasing in capacity:

$$\overline{w}(\lambda,s) \overset{\text{def}}{=} \int_0^\infty w \mathrm{d}H(w;\lambda,s), \quad \overline{w}_\lambda > 0, \quad \overline{w}_s < 0. \tag{2}$$

Main symbols and definitions are given in Table 1.

2.2. Patients

Compulsory public health insurance covers the costs of treatment in the public hospital and a public earnings insurance scheme reimburses some or all of earnings lost due to illness. Both schemes are funded from general taxation. Individuals have the same preferences but differ in their incomes and those with a sufficiently high income choose to take out supplementary private insurance to cover the cost of treatment in a private hospital.[12]

Income per unit of time y when well is distributed over $[y_{\min}, y_{\max}]$ with distribution function $F(y)$. When ill, earnings are reduced by $\ell(y)$ and reimbursement $r(\ell(y)) \in [0, \ell(y)]$ is received from the public insurance scheme, so that income when ill is

$$\begin{aligned} y^L(\mathrm{y}) &= \mathrm{y} - \ell(y) + \mathrm{r}(\ell(y)), \\ \frac{\mathrm{d}y^L}{\mathrm{d}y} &= 1 - \ell'(y)(1 - r'(\ell(y))) \geq 0, \end{aligned} \tag{3}$$

where we assume that income when ill is non-decreasing in income when well.[13]

Utility for a patient treated in a public hospital who has a wait of w days from illness to discharge after treatment is

$$u(y,q,w) \overset{\text{def}}{=} U^1(y^L(y)) \int_0^w \delta(t)\mathrm{d}t + U^2(y,q) \int_w^\infty \delta(t)\mathrm{d}t, \tag{4}$$

where $u_y > 0$, $u_q > 0$, $u_w < 0$, U^1 and U^2 are flow utility per day whilst waiting and post-treatment, $\delta(t)$ is the discount factor and q is hospital quality.[14] We assume that individuals with more income when well have higher utility if they fall ill ($u_y > 0$). We also assume that treatment makes the

[9] See Taylor and Karlin (1998, Ch. 9) or Gross et al. (2008, Ch. 2) for expositions of queueing theory.

[10] As in queuing theory where the most common assumption is that the arrival rate has a Poisson distribution. And, as in this literature, we assume that the population from which arrivals come is infinite, so that the probability of an arrival in any time interval is independent of the number of previous arrivals (see Gross et al., 2008: 85).

[11] We discuss the implications of multiple hospital decisions affecting waiting time in Section 3.3 and Appendix F.

[12] Allowing income to affect the choice between public and private care is realistic (Besley et al., 1999). It also simplifies the model since with a population of identical individuals an equilibrium with some individuals choosing public and some choosing private care would require the representative individual to play a mixed strategy between public and private care.

[13] For example, the average sickness insurance replacement rate in Canada between 2000 and 2011 (for a single 40 year old worker earning the average production worker wage in manufacturing) was 36% and the scheme covered 79% of the labour force. For other countries the replacement rates and coverage were: 88% and 100% (France), 89% and 85% (Germany), 100% and 100% (Norway), and 22% and 88% (Great Britain) See Scruggs et al. (2017).

[14] The utility functions are cardinal and unique up to the same linear transformation. We discuss the implications of multiple quality dimensions in Appendix C. We assume that patients can observe quality possibly via advice from their primary care doctor or via public websites, such as NHS Choices in England or Helsedirektoratet in Norway, which publish hospital quality indicators.

Table 1
Main symbols and definitions.

σ	Probability of ill health
q, s	Quality, capacity in public hospital
w; $H(w;\lambda,s)$, $H_\lambda<0$, $H_s>0$	Waiting time; waiting time distribution function
y, $F(y)$, $f(y)$	Income per day when well, income cdf fn, density fn
$y^L(y)=y-\ell(y)+r(\ell(y))$	y^L: income when ill; ℓ: lost earnings; r: compensation
$u(y,q,w)$	Realised utility when ill if treated in public hospital
$\overline{u}(y,q,\lambda,s)=\int_0^\infty u(y,q,w)dH$	Expected utility when ill if treated in public hospital
$u^N(y)$	Utility if not ill
$v(y,q,\lambda,s)=\sigma\overline{u}+(1-\sigma)u^N$	Expected utility if will choose public hospital if ill
$\lambda^e(q,s)$, $\lambda_z>0$, $(z=q,s)$	Equilibrium mean daily demand for public hospital
$v^e(y,q,s)=v(y,q,\lambda^e(q,s),s)$	Expected utility if will choose public hospital if ill
$v^o(y)$	Expected utility if will choose private hospital if ill
$\hat{y}^e(q,s)$	Threshold income: public hospital chosen if $y\le\hat{y}^e$
$B^e(q,s)=B(q,\lambda^e(q,s),s)$	Aggregate patient welfare
$c^e(q,s)=c(q,\lambda^e(q,s),s)$	Expected cost of public hospital
$c^{le}(q,s)=c^l(q,\lambda^e(q,s),s)$	Expected total earnings compensation
θ	Marginal deadweight cost of taxation
$B^e-(1+\theta)(c^e+c^{le})$	Welfare function
$\alpha\ge 0$	Public hospital altruism parameter
$R_z^e=\beta B_z^e-c_z^{le}$, $(z=q,s)$	Residual marginal welfare ignored by hospital
$\beta=\frac{1-\alpha(1+\theta)}{1+\theta}$	

patient better off than whilst ill and on the list ($U^2>U^1$), so that $u_w=[U^1-U^2]<0$, and that higher quality of care increases post-discharge utility ($U_q^2>0$), so that $u_q>0$.[15] The marginal rate of substitution between quality and waiting time is

$$mrs(q,w;y)\overset{\text{def}}{=}\frac{\partial w}{\partial q}\Big|_{du=0}=-\frac{u_q(y,q,w)}{u_w(y,q,w)}. \tag{5}$$

Expected utility when ill for a patient treated in the public hospital is the expectation of (4):

$$\overline{u}(y,q,\lambda,s)\overset{\text{def}}{=}\int_0^\infty u(y,q,w)\mathrm{d}H(w;\lambda,s),$$
$$\overline{u}_y>0,\quad \overline{u}_q>0,\quad \overline{u}_\lambda<0,\quad \overline{u}_s>0. \tag{6}$$

The first order stochastic dominance properties of the distribution of w with respect to λ and s and the assumption that $u_w<0$ imply that expected utility is decreasing in the arrival rate λ and increasing in capacity s.[16]

Utility when in good health and not requiring hospital treatment $u^N(y)$ is an increasing function of income and $u^N(y)>u(y,q,0)$, so that even immediate treatment does not make a patient better off than if healthy. Expected utility for individuals who decide not to take out supplementary private health care insurance and to be treated in the public hospital when ill is

$$v(y,q,\lambda,s)\overset{\text{def}}{=}\sigma\overline{u}(y,q,\lambda,s)+(1-\sigma)u^N(y), \tag{7}$$

where $v_y>0$, $v_\lambda<0$, $v_z>0$ $(z=q,s)$.

The private hospital provides a care package which, if it had a zero price, would always be preferred to the public hospital.[17] Individuals who know they will prefer to use the private sector when ill buy full cover supplementary private insurance at an actuarially fair premium γ. Their utility when ill is $u^o(y-\gamma)$ and when in good health is $u^N(y-\gamma)$. Expected utility from the outside option of taking out private insurance and being treated in the private hospital when ill is

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

2.3. Rational expectations equilibrium

We assume that private health care is a normal good in that there is a threshold income level $(y_{\min},y_{\max})$ defined by

$$v(\hat{y},q,\lambda,s)-v^o(\hat{y})=0, \tag{8}$$

with $\hat{y}_q>0$, $\hat{y}_\lambda<0$, $\hat{y}_s>0$.

All individuals with $y\le\hat{y}$ choose not to have private insurance and to be treated in the public hospital when ill. Since individuals fall ill at the rate σ and when ill a proportion $F(\hat{y}(q,\lambda,s))$ demand care in the public sector, the expected demand (arrival rate) λ at the public hospital

[15] Although patients may distinguish between time spent on the waiting list and time spent in the hospital being treated (their length of stay), allowing for this, and for quality to affect utility whilst in hospital as well as post treatment, would make no essential difference to the results.

[16] $\overline{u}(y,q,\lambda,s)$ is a reduced form summary of the factors determining expected utility when ill and treated in the public hospital. Patients care about the distribution of waiting times (or, depending on $u(y,q,w)$, about sufficient statistics of the distribution such as the mean wait). We assume that they observe this distribution, not that they know λ and s.

[17] To keep the analysis tractable we assume that private hospital decisions on the premium, quality, and waiting time are not affected by decisions in the public hospital. As Grassi and Ma (2011) and Laine and Ma (2017) illustrate, even in a context with no rationing by waiting, models of the interactions between public and private providers can be complex and generate multiple types of equilibria.

is $\sigma F(\hat{y}(q, \lambda, s))$. Hence the equilibrium expected demand $\lambda^e(q, s)$[18] is implicitly defined by

$$\lambda^e - \sigma F(\hat{y}(q, \lambda^e, s)) = 0.$$

This embodies the rational expectations assumption: the arrival rate in (8) upon which patients' decision about joining the waiting list for the public hospital are based is the arrival rate to which their decisions give rise. Expected demand is increasing in hospital quality and capacity since both increase the utility of the marginal patient:

$$\lambda^e_z(q, s) = \frac{\sigma f(\hat{y})\hat{y}_z}{1 - \sigma f(\hat{y})\hat{y}_\lambda} > 0, \quad (z = q, s). \tag{9}$$

In equilibrium, the distribution function for waiting time is

$$H^e(w; q, s) \overset{\text{def}}{=} H(w; \lambda^e(q, s), s). \tag{10}$$

From (1) and (9) an increase in q produces a stochastically dominating shift in $H^e(w; q, s) : H^e_q = H_\lambda \lambda^e_q < 0$. We assume that the positive direct effect of a capacity expansion exceeds the negative induced demand effect, thereby producing a first order stochastically dominated shift in the equilibrium waiting time distribution: $H^e_s = H_s + H_\lambda \lambda^e_s > 0$. Hence, substituting $\lambda^e(q, s)$ in (2), the equilibrium expected time on the waiting list is $\overline{w}^e(q, s) = \overline{w}(\lambda^e(q, s), s)$ increases with quality ($\overline{w}^e_q = \overline{w}_\lambda \lambda^e_q > 0$) and falls with capacity ($\overline{w}^e_s = \overline{w}_\lambda \lambda^e_s + \overline{w}_s < 0$).

Using $H^e(w; q, s)$ in (6), and hence in (7), the equilibrium expected utility for individuals who will use the public sector when ill is[19]

$$v^e(y, q, s) \overset{\text{def}}{=} v(y, q, \lambda^e(q, s), s). \tag{11}$$

Our assumption that an increase in s produces a leftward shift in the equilibrium waiting time distribution implies that $v^e_s > 0$. We also make the plausible assumption that the positive direct effect of q on expected utility is bigger than the negative indirect effect via the induced rightward shift in waiting time distribution, so that an increase in quality increases equilibrium expected utility: $v^e_q > 0$.

The effect of hospital decisions about q and s on individuals varies with their income. An increase in capacity s will make *all* public patients better off since it induces a preferred distribution of waiting times and has no direct effect on utility. However, it is possible that an increase in quality q will make *infra-marginal* patients with $y < \hat{y}$ worse off if the utility loss from a worse waiting time distribution caused by the increase in demand is greater than the direct effect of quality on utility. But the *marginal* patient with income $\hat{y}$ must have been made better off by an increase in quality since otherwise demand would not have increased and demand can only increase if the marginal patient is made better off choosing the public hospital.[20]

A patient's marginal rate of substitution of quality for capacity at the equilibrium distribution of w is

$$MRS^e(q, s; y) \overset{\text{def}}{=} -\left.\frac{\partial s}{\partial q}\right|_{dv^e=0} = \frac{v^e_q(y, q, s)}{v^e_s(y, q, s)}, \tag{12}$$

and in general the MRS^e varies with patient income. Since hospital decisions shift the distribution of waiting times facing patients and determine the quality they experience once in hospital, MRS^e plays an important role in determining the optimal payment regime.

Note that $mrs(q, w; y)$ in (5) is the rate at which a patient is willing to trade off realised ex post waiting time for quality whereas $MRS^e(q, s; y)$ in (12) is the rate at which they are willing to trade off the ex ante distribution of waiting times for quality.

The equilibrium critical income level $\hat{y}^e$ that divides individuals into those who are treated in the public hospital and those who take out supplementary health care insurance and are treated in the private sector is defined (see (8)) by $v(\hat{y}, q, \lambda^e(q, s)) = v^o(\hat{y})$ as

$$\hat{y}^e(q, s) = \hat{y}(q, \lambda^e(q, s), s). \tag{13}$$

Any changes in q and s that leave critical income and thus the expected utility of the marginal public patient unchanged will also leave demand unchanged (see Appendix A):

$$\begin{aligned} -\left.\frac{\partial s}{\partial q}\right|_{dv^e(\hat{y}^e, q, s)=0} &= \frac{v^e_q(\hat{y}^e, q, s)}{v^e_s(\hat{y}^e, q, s)} = \frac{\hat{y}^e_q}{\hat{y}^e_s} = \frac{\lambda^e_q}{\lambda^e_s} \\ &= -\left.\frac{\partial s}{\partial q}\right|_{d\lambda^e=0} \end{aligned} \tag{14}$$

Thus the $MRS^e(q, s; \hat{y}^e)$ of the marginal public sector patient, but not necessarily the $MRS^e(q, s; y)$ of infra-marginal patients with $y < \hat{y}^e$, is revealed by the marginal demand responses to quality and capacity.

2.4. A simple special case

We can illustrate the derivation of the rational expectations equilibrium with an instructive special case with simple preferences and queueing mechanism. Assume that the period utility functions in (4) are $U^1 = a(y)b^1$ and $U^2 = a(y)b^2(q)$, individuals have finite lifetime of T, the discount factor δ is constant, and that income is not affected by illness. Then realised utility when ill for an individual who is treated in the public hospital is $u(y, q, w) = a(y)\{b^2(q)T - [b^2(q) - b^1]w\}\delta$ and expected utility (7) for an individual

[18] We use superscript ^{e} to indicate equilibrium values of variables and functions.

[19] We assume that patients observe quality but not that they observe s. $v^e(y, q, s)$ is a reduced form showing the dependence of expected utility at the REE on y, q and s. The reduced form is derived from the more primitive utility $u(y, q, w)$ from public hospital treatment (4), the distribution function of waiting times $H(w; \lambda, s)$, which together give expected utility (6) when treated in the public hospital when ill, the probability of falling ill, and the REE expected demand $\lambda^e(q, s)$. We assume that patients observe the equilibrium waiting time distribution $H^e(w; q, s) = H(w; \lambda^e(q, s), s)$. Or, if their primitive preferences $u(y, q, w)$ imply that they only care about some sufficient statistics of the distribution, as in Section 2.4 where they care only about the mean wait, we assume that they observe these sufficient statistics.

[20] See Appendix C for a more formal argument.

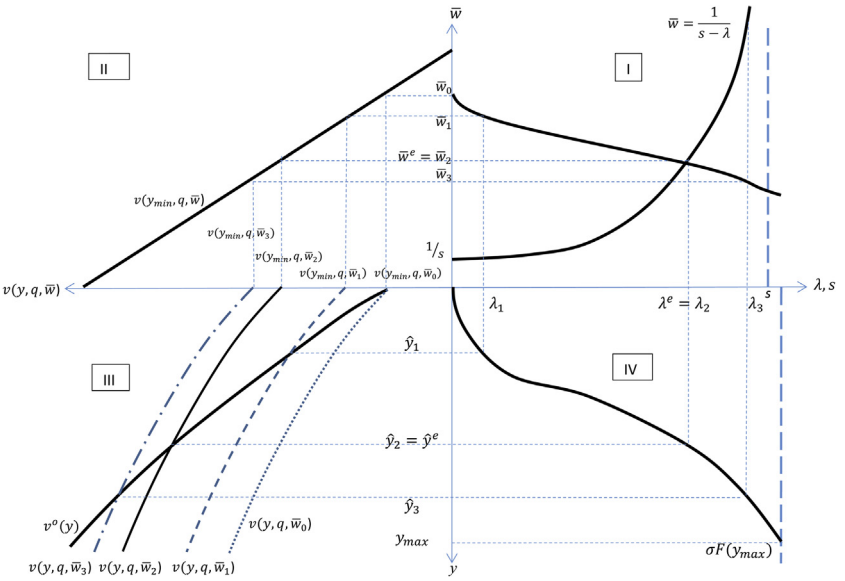

Fig. 1. Derivation of the rational expectation equilibrium. $v(y, q, \overline{w})$: expected utility from public hospital. $v^o(y)$: expected utility from private hospital. $1/s$: expected length of stay in public hospital. λ expected arrival rate at public hospital. $F(y)$: income distribution function (minimum income normalized to zero: $y_{\min} = 0$). $\hat{y}$ critical income at which $v(y, q, \overline{w}) = v^o(y)$.

who decides to be treated in the public hospital if they fall ill is linear in the expected wait $\overline{w}$:

$$v(y, q, \bar{w}) = \sigma a(y)\{b^2(q)T - [b^2(q) - b^1]\overline{w}\}\delta + (1 - \sigma)u^N(y).$$

Assume also that the length of stay when admitted to hospital is negatively exponentially distributed with parameter s, so that the average length of stay in hospital is $1/s$, and that the hospital has a single bed. Patients fall ill and join the waiting list according to a Poisson process with average arrival rate λ. These assumptions define the simplest system in queueing theory which has a steady state in which the waiting time w from falling ill to completion of treatment has a negative exponential distribution with expectation $\overline{w} = \frac{1}{s-\lambda}$ (Taylor and Karlin, 1998, p 551).

In Panel I of Fig. 1, the bold upward sloping line shows, for a given service rate s, the relationship between the expected wait $\overline{w}$ and the average arrival rate for treatment λ. As λ approaches the capacity of the system (s), the expected waiting time tends to infinity.

We next derive the demand curve. Panel II of the figure maps the expected waiting time into an expected utility $v(0, q, \overline{w})$ for individuals with the lowest income level which for convenience we normalise to zero ($y_{\min} = 0$). Panel III then maps the income level to the expected utility $v(y, q, \overline{w})$ if public treatment is chosen and to expected utility $v(y, q^o, \overline{w}^o)$ when the private hospital is chosen. These two curves intersect at the critical income $\hat{y}(q, \overline{w})$ where public and private hospital yield the same expected utility. Finally, Panel IV maps the critical income level into the number of individuals $\sigma F(\hat{y})$ who choose public treatment. This is the expected arrival rate $\left(\lambda = \sigma F(\hat{y})\right)$ for the public hospital, which then maps into an expected waiting time $\overline{w} = \frac{1}{s-\lambda}$ in Panel I.

The demand curve for public treatment is derived as follows. Consider an anticipated expected wait $\overline{w}_1$. This maps into a critical income level $\hat{y}_1$ in Panel III which in Panel IV yields the arrival rate rate λ_1. Thus $(\lambda_1, \overline{w}_1)$ is a point on the demand curve in Panel I. But it is not an equilibrium since the low referral rate λ_1 results in an expected wait $\frac{1}{s-\lambda_1} < \overline{w}_1$. Similarly $(\lambda_3, \overline{w}_3)$ is also on the demand schedule in Panel I and is also not an equilibrium since the high arrival rate λ_3 would yield an expected wait of $\frac{1}{s-\lambda_3} > \overline{w}_3$. Given the public hospital decisions on the parameters (q, s) of the $M/M/1$ queueing model, the unique equilibrium is given by $\left(\lambda^e, \overline{w}^e\right) = (\lambda_2, \overline{w}_2)$. When citizens anticipate an expected wait $\overline{w}_2$ at the public hospital their choices between public and private treatment yield an arrival rate at the public hospital which results in the expected waiting time they anticipated: $\frac{1}{s-\lambda_2} = \overline{w}_2$. The rational expectations equilibrium at given quality and capacity is at the intersection of the downward sloping demand curve and the upward sloping expected waiting time curve $\overline{w}$.

Now consider the effects of changes in the public hospital decisions on s and q. An increase in the hospital service rate s shifts the upward sloping $\overline{w} = 1/(s - \lambda)$ to the right and so results in higher equilibrium demand $\lambda^e(q, s)$ and lower equilibrium expected waiting time $\overline{w}^e(q, s)$.

An reduction in q shifts the straight line in Panel II plotting $v(0, q, \overline{w})$ against $\overline{w}$ to the right and makes it steeper in $(\overline{w}, v)$-space.[21] In Panel III the three functions $v(\cdot, q, \overline{w}_i)$ $(i = 1, 2, 3)$ plotting expected utility from the

[21] Our assumptions imply $\partial v(0, q, \overline{w})/\partial q = \sigma a(0) b_q^2(q)(T - \overline{w}) > 0$, and $\partial^2 v(0, q, \overline{w})/\partial q \partial \overline{w} = -\sigma a(0) b_q^2(q) < 0$. Hence the line $v(0, q, \overline{w})$ becomes flatter in $(v, \overline{w})$-space (steeper in $(\overline{w}, v)$-space).

public hospital against income at different expected waiting times but with the same quality will also shift to the right. Their intersections with the curve for expected utility from private treatment $v(\cdot, q^o, \overline{w}^o)$ will be pushed further up that curve, thereby reducing the critical incomes $\hat{y}_i$ given $\overline{w}$ and thus reducing $\lambda_i = \sigma F(\hat{y}_i)$. Hence the quality reduction shifts the demand curve in Panel I to the left, resulting in a lower equilibrium demand $\lambda^e(q, s)$ and a shorter equilibrium expected wait $w^e(q, s)$.[22]

In this simple specification of preferences the expected wait is a sufficient statistic for the entire waiting time distribution as far as patients are concerned and could be interpreted as a price for public treatment. The assumption that income is unaffected by illness (either because of full insurance or because illness does not affect productivity) implies that income is not affected by hospital decisions and simplifies the identification of the critical income earner. The graphical representation of the rational expectations equilibrium would be much more complicated, or impossible, without these assumptions. However, as we will show in Section 3.4, both assumptions have very strong implications for the optimal financing rules for the public hospital. In what follows we therefore revert to the more general assumptions about preferences and the effect of waiting on income.

2.5. Welfare

2.5.1. Individual welfare

To focus on the implications of rationing by waiting we assume that the welfare function is individualistic and respects patient preferences. Total individual welfare is[23]

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

Since patients with income $\hat{y}^e(q, s)$ are indifferent between public and private hospital care

$$\begin{aligned} B^e_z(q,s) &= \int_{y^{\min}}^{\hat{y}^e(q,s)} v^e_z(y,q,s)d\mathrm{F}(y) + f(\hat{y}^e)[v^e(\hat{y}^e,q,s) - v^o(\hat{y}^e)] \\ &= \int_{y^{\min}}^{\hat{y}^e(q,s)} v^e_z(y,q,s)d\mathrm{F}(y), \quad (z = q, s). \end{aligned}$$

Some of the expressions for optimal payments in Section 3 involve the change in aggregate patient welfare from a change in quality when demand is held constant by a reduction in capacity (i.e., when the change in quality is accompanied by a change in capacity $\mathrm{d}s = -(\lambda^e_q/\lambda^e_s)\mathrm{d}q$):

$$\begin{aligned} \frac{dB^e}{dq}\Big|_{\mathrm{d}\lambda^e=0} &= B^e_q - B^e_s\frac{\lambda^e_q}{\lambda^e_s} = \int_{y^{\min}}^{\hat{y}^e(q,s)} v^e_s(y,q,s)[\frac{v^e_q(y,q,s)}{v^e_s(y,q,s)} \\ &\quad - \frac{\lambda^e_q(q,s)}{\lambda^e_s(q,s)}]\mathrm{d}F(y) \\ &= \int_{y^{\min}}^{\hat{y}^e(q,s)} v^e_s(y,q,s)[\frac{v^e_q(y,q,s)}{v^e_s(y,q,s)} \\ &\quad - \frac{v^e_q(\hat{y}^e,q,s)}{v^e_s(\hat{y}^e,q,s)}]\mathrm{d}F(y), \end{aligned}$$

where the last step uses (14). This expression is positive, negative or zero as $MRS^e(q, s; y) = v^e_q(y, q, s)/v^e_s(y, q, s)$ decreases, increases, or is unaffected by income. In Appendix B we prove

Proposition 1. *(i) The marginal rate of substitution between quality and capacity $MRS^e(q, s; y)$(12) is independent of preillness income y if and only if the marginal rate of substitution between quality and waiting time $mrs(q, w; y)$(5) is independent of income. (ii) $mrs(q, w; y)$ is independent of income if and only if (a) utility per day when waiting and utility per day post treatment can be written as $U^1 = a(y^1)b^1$ and $U^2 = a(y^2)b^2(q)$, and (b) income is not affected by treatment: $y^1 = y^2$.*

We can interpret b^1 and b^2 as functions of health status whilst waiting for treatment and post treatment. Condition (a) is a condition on preferences: multiplicative separability between income and health status.[24] Condition (b) is a condition on income. Since we have assumed that there is no loss of earnings post treatment condition (b) requires either that there is no loss of earnings whilst on the waiting list, $\ell(y) = 0$, or that the insurance scheme fully compensates all lost earnings, $r(\ell(y)) \equiv \ell(y)$. Given (a) and (b), the utility of income can be factored out from u_q and u_w, so that $mrs(q, w; y)$ is independent of income. Conditions (a) and (b) also imply that utility of income can be factored out from v^e_q and from v^e_s so that $MRS^e(q, s; y)$ is also independent of income.

[22] We can use Panel I to contrast the stochastic waiting time model with the equilibrium in a deterministic waiting time model in which demand and supply and hence waiting time w are certain. With the same preferences the demand curve, now plotting demand against the certain wait, is unchanged. In the stochastic model we can interpret $1/s$ as the average length of stay (days per patient). In the deterministic model we can interpret s as the certain supply (patients per day) and the equilibrium waiting time is determined by the equality of certain supply and certain demand: $s = \lambda(q, w)$. In Panel I the certain equilibrium waiting time would be determined by the intersection of the demand curve and a vertical supply curve at s. A reduction in quality would shift the demand curve downward, reducing the certain waiting time. But, since supply is not changed neither is the number of patients treated. Thus in the deterministic waiting time model, if q and w are positive, a reduction in quality will push the equilibrium down the vertical supply curve. This will reduce costs but leave revenue unchanged, thereby increasing profit. Hence there cannot be an equilibrium with positive quality and waiting time in a deterministic waiting time model unless the provider had an altruistic concern for quality.

[23] In Fig. 1, given the strong assumptions about preferences and income, and assuming a uniform income distribution, the total welfare of those using the public hospital is the area in quadrant III between the curve $v(y, q, \overline{w}^e) = v(y, q, \overline{w}_2)$ and the income axis up to $\hat{y}^e = \hat{y}_2$. The welfare of those chosing the private hospital is area between the curve $v^o(y, q^o, \overline{w}^o)$ and the income axis for $y > \hat{y}^e = \hat{y}_2$.

[24] Additive separability would also ensure that the marginal rate of substitution did not vary with income but would imply that all individuals would make the same choice of public or private treatment. Hence we require multiplicative separability in Proposition 1 so that individuals differ in some parameter which affects utility from treatment but not the marginal rate of substitution between q and s.

Patient decisions on whether to take out private health insurance and be treated in the private sector when ill or to avoid paying for private health insurance and to be treated in the public hospital when ill are made ex ante: on the basis of their beliefs about the equilibrium distribution of public hospital waiting times $H^e(w; q, s)$. It is expected utility in equilibrium $v^e(y, q, s)$, not realised utility $v(y, q, w)$ on joining the waiting list, which is part of the welfare function used to evaluate alternative hospital payment schemes. But patient preferences over quality q and the realised waiting time w ($u(y, q, w)$ (4)) determine their preferences over quality and the equilibrium distribution of waiting time ($v^e(y, q, s)$ (11)).

2.5.2. Costs

Treatment cost. In equilibrium the public hospital's expected output is equal to expected demand (λ^e) from patients. The hospital's expected cost is $c(q, \lambda, s)$. Increasing quality is costly ($c_q > 0$) as are increases in capacity ($c_s > 0$). Expected hospital cost also increase with the expected number of patients treated ($c_\lambda > 0$), for example because each patient treated requires drugs and other consumables.[25] In equilibrium, expected cost is

$$c^e(q, s) = c(q, \lambda^e(q, s), s),$$

with $c^e_z = c_z + c_\lambda \lambda_z > 0$, $(z = q, s)$. We ignore, until Section 4.2, the possibility that expected hospital cost also depends on cost reducing effort.

Insurance cost Patients who lose earnings whilst waiting receive compensation (3) from the public insurance fund. The expected public sector payment to an ill individual with income y when well who is treated in the public hospital and has an expected waiting time of $\overline{w}^e(q, s)$ days is $\overline{w}^e(q, s)r(\ell(y))$. The scheme also compensates individuals who choose to be treated privately and who have a total time (exogenous and short) from becoming ill to discharge of $w^o < \overline{w}^e$. The expected total payment from the earnings insurance fund is[26]

$$c^{Ie}(q, s) = \sigma \overline{w}^e(q, s) \int_{y_{\min}}^{\hat{y}^e(q,s)} r(\ell(y))\mathrm{d}F(y) + \sigma w^o \int_{\hat{y}^e(q,s)}^{y_{\max}} r(\ell(y))\mathrm{d}F(y).$$

An increase in q or s alters expected insurance cost by changing the expected time to completion of treatment for public patients and by changing the expected number treated in the public sector[27] :

$$c^{Ie}_z = \sigma \overline{w}^e_z \int_{y_{\min}}^{\hat{y}^e(q,s)} \mathrm{r}(\ell(\mathrm{y}))\mathrm{dF(y)} + [\overline{w}^e(\mathrm{q, s}) - w^o]\mathrm{r}(\ell(\hat{y}^e(\mathrm{q, s})))\lambda^e_z, \quad (z = q, s). \tag{15}$$

The first *rhs* term is the *waiting time effect* and the second is the *waiting list effect*. The waiting list effect is positive on the plausible assumption that expected total time to discharge is greater in the public hospital than in the private hospital. Since an increase in quality increases the waiting time ($\overline{w}^e_q = \overline{w}_\lambda \lambda^e_q > 0$) the waiting time effect is also positive for quality increases and so $c^{Ie}_q > 0$. But notice that because $\overline{w}^e_s < 0$ the sign of c^{Ie}_s is ambiguous: capacity increases have a positive waiting list effect but a negative waiting time effect. However, if an increase in s is accompanied by a reduction in q to keep $\lambda^e(q, s)$ constant, the waiting list effects are zero since there is no change in demand. Then the effect on the mean wait is $\overline{w}_s + \overline{w}_\lambda(\lambda^e_s + \lambda^e_q\left(-\lambda^e_s/\lambda^e_q\right)) = \overline{w}_s < 0$ and we get (see Appendix C)

$$\left.\frac{\mathrm{d}c^{Ie}}{\mathrm{ds}}\right|_{d\lambda^e=0} = \sigma \overline{w}_s \int_{y_{\min}}^{\hat{y}^e(q,s)} r(\ell(y))\mathrm{d}F(y) = \left(\left.\frac{\mathrm{d}c^{Ie}}{\mathrm{dq}}\right|_{\mathrm{d}\lambda=0}\right)\left(-\frac{\lambda^e_q}{\lambda^e_s}\right) < 0. \tag{16}$$

Thus increasing s and simultaneously reducing q to keep demand constant reduces expected insurance cost and conversely increasing q and reducing s increases expected insurance cost.

3. Optimal payment schemes

3.1. First best quality and capacity

The regulator's welfare function is[28]

$$A^e(q, s) \overset{\text{def}}{=} B^e(q, s) - \left(1 + \theta\right)\left[c^e(q, s) + c^{Ie}(q, s)\right]. \tag{17}$$

[25] We assume that the expectation of the cost w.r.t. the number of patients treated can be expressed as the cost of the expected number of patients. Letting n be the random number of patients treated, this requires that the cost function is of the form $c^1(n)c^2(q, s)$, and that either $c^1(n)$ is linear in n or that $c^1(n)$ is a polynomial in n and the arrival rate of patients, and therefore the output rate, follows a Poisson process.

[26] There is some debate in the economic evaluation literature about whether the costs of lost earnings should be measured by the human capital (Weisbrod, 1961) or friction cost methods (Koopmanschap et al., 1995) but there is agreement that they should be taken into account (Drummond et al., 2015). We avoid the double counting problem (Pritchard and Sculpher, 2000) by separating out the cost imposed on the public sector via the insurance of lost earnings and the utility cost of uncompensated lost earnings imposed on the patient. An example of public concerns about the social insurance costs of long waiting list is the motivation behind the Faster Return to Work scheme in Norway (see Aakvik et al. (2015)). Note that if the cost of lost production whilst waiting for treatment fell on private sector firms, rather than on workers, the welfare function should still include c^{Ie}, though not scaled up by the marginal welfare cost of taxation θ, as in (17) below.

[27] Using (9) and (13), it can be shown that $\lambda^e_z = \sigma f(\hat{y}^e)\hat{y}^e_z$. See Appendix A.

[28] 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 the provider just breaks even financially after any incentive payments. Or we can assume that welfare is the sum of patient benefit and the hospital utility and the lump sum tax or subsidy drives hospital utility to zero. We ignore here the implications of the regulator being unable to impose lump sum taxes or subsidies. In analyses available on request we show that a hospital breakeven constraint would then imply that the optimal prices also depend on inverse demand elasticities, as in Boiteux (1956). If hospital managerial effort affected quality or the monetary cost of production and had a non-monetary cost this should be reflected in the hospital objective function and the welfare function and would affect the precise form of the optimal incentive scheme. But it would not affect the basic message of our simpler specification that achieving the first best when there is rationing by waiting requires an additional policy instrument. We briefly consider the implication of cost reducing effort in Section 4.2.

where θ is the marginal cost of public funds. Ignoring corner solutions, first best[29] quality and capacity levels satisfy the first order conditions

$$A_z^e = B_z^e - (1+\theta)c_z^e - (1+\theta)c_z^{Ie} = 0, \quad (z = s, q). \tag{18}$$

The first best can be achieved if the regulator has as many policy instruments with linearly independent effects on hospital decisions as the hospital has decision variables. The hospital takes decisions on q and s which result in an expected number of treatments, $\lambda^e(q, s)$. We assume that it is always possible to observe output and so to set a prospective price p_λ.[30] In Section 3.2 we derive first best payment schemes in which a prospective price per patient treated p_λ is combined with price p_m attached to another observable performance measure $m(q, s)$ affected by hospital decisions. In Section 3.3 we consider multiple quality and capacity dimensions and the implications of patients being concerned only about the average waiting time.

In Section 3.4 we consider the restrictive assumptions under which the first best can be achieved when only output is observable so that the prospective price is the only policy instrument. In Section 4.1 we derive the second best prospective price when the restrictive assumptions set out in Section 3.4 are not satisfied. In Section 4.2 we then allow for the possibility that unobservable hospital effort affects cost and consider a cost reimbursement rule combined with a prospective output price.

3.2. First best prospective price and performance payment

Suppose the risk neutral public hospital receives a payment per patient treated, p_λ and a payment p_m per unit of some other observable measure of activity $m(q, s)$ that depends monotonically on the hospital's decisions on quality and capacity. It also receives a lump sum transfer Υ to ensure that it breaks even.[31]

Examples of measures of performance are (i) $m = s$ (e.g., a price per bed), (ii) $m = q$ (a P4P quality incentive scheme), (iii) $m = \overline{w}^e(q, s)$ (a price, possibly negative, on expected waiting time), and (iv) $m = c^{Ie}$ (the hospital bears a proportion of the sickness leave insurance cost). In general, any measure $m(q, s)$ will do as long as it is observable and its gradient is linearly independent of the gradient of $\lambda^e(q, s)$.

We discuss examples (i) and (ii) later in this section, while examples (iii) and (iv) are examined in Appendix D.

As in Ellis and McGuire (1990), the hospital is assumed to be risk neutral and to maximise a weighted sum of expected profit and patient welfare, with the weight $\alpha \geq 0$ reflecting its concern for patient welfare. Allowing for such concerns is common in models of hospital behaviour and, whilst he degree of altruism affects the magnitudes of payments related to the hospital's decisions, it does not affect their essential structure. The hospital solves

$$\max_{q,s} \; p_\lambda \lambda(q, s) + p_m m(q, s) - c^e(q, s) + \alpha B^e(q, s) + \Upsilon,$$

and first order conditions for an interior solution are

$$p_\lambda \lambda_q + p_m m_q + \alpha B_q^e = c_q^e,$$
$$p_\lambda \lambda_s + p_m m_s + \alpha B_s^e = c_s^e.$$

When choosing q and s the hospital takes into account marginal cost c_z^e and a fraction α of marginal patient benefit B_z^e. But the hospital ignores the remaining fraction $(1-\alpha)$ of B_z^e, all of the marginal cost of insuring lost earnings c_z^{Ie}, and the fact that public funds have a marginal cost of θ. Define the *residual marginal social benefit* (RMSB) of hospital decision z as

$$R_z^e \overset{\text{def}}{=} \beta B_z^e - c_z^{Ie}, \quad \beta \overset{\text{def}}{=} \frac{1-\alpha(1+\theta)}{1+\theta}. \tag{19}$$

R_z^e is that part of the marginal welfare effect of decision z which is not internalised by the hospital.

To achieve the first best the regulator sets prices p_λ^{FB} and p_m^{FB} so that the hospital marginal revenues from decisions on q and s equal their RMSBs:

$$p_\lambda^{FB}\lambda_q^e + p_m^{FB} m_q = R_q^e, \tag{20}$$
$$p_\lambda^{FB}\lambda_s + p_m^{FB} m_s = R_s^e. \tag{21}$$

The hospital will then take full account of the marginal social benefits and costs choosing q and s to satisfy

$$\begin{aligned} p_\lambda^{FB}\lambda_q + p_m^{FB} m_q + \alpha B_q^e - c_q^e &= (\beta B_q^e - c_q^{Ie}) + \alpha B_q^e - c_q^e \\ &= \frac{B_q^e}{(1+\theta)} - c_q^e - c_q^{Ie} = 0, \end{aligned}$$

$$\begin{aligned} p_\lambda^{FB}\lambda_s + p_m^{FB} m_s + \alpha B_s^e - c_s^e &= (\beta B_s^e - c_s^{Ie}) + \alpha B_s^e - c_s^e \\ &= \frac{B_s^e}{(1+\theta)} - c_s^e - c_s^{Ie} = 0, \end{aligned}$$

so that (18) holds. Solving (20) and (21) for p_λ^{FB} and p_m^{FB} gives

Proposition 2. *The first best prices per treated patient and for the performance measure $m(q, s)$ are*

$$p_\lambda^{FB} = \frac{R_q^e m_s - R_s^e m_q}{\lambda_q^e m_s - \lambda_s^e m_q} = \frac{R_q^e - R_s^e \frac{m_q}{m_s}}{\lambda_q^e - \lambda_s^e \frac{m_q}{m_s}}, \tag{22}$$

$$p_m^{FB} = \frac{R_s^e \lambda_q^e - R_s^e \lambda_s^e}{\lambda_q^e m_s - \lambda_s^e m_q} = \frac{R_s^e - R_q^e \frac{\lambda_s^e}{\lambda_q^e}}{m_s - m_q \frac{\lambda_s^e}{\lambda_q^e}}, \tag{23}$$

where all terms are evaluated at the first best q and s.

The output price (22) is set so that marginal revenue $(p_\lambda^{FB}(\lambda_q^e - \lambda_s^e \frac{m_q}{m_s}))$ from increasing quality *when capacity is adjusted to keep the performance measure $m(q, s)$ constant* is equal to the marginal residual social benefit of quality $(R_q^e - R_s^e \frac{m_q}{m_s}))$ when s adjusted to keep $m(q, s)$ constant.[32]

[29] Strictly speaking, the term "first best" is a misnomer since we ignore the externality that arises because each patient does not take account of the effect of her decision to join the waiting list on the waiting times of other patients. See Naor (1969), Littlechild (1974), and Edelson and Hildebrand (1975) on policies to control decisions to join the queue.

[30] With a risk neutral hospital a prospective price per patient treated is equivalent to payment for the expected number of treatments.

[31] With the presence of substantial fixed costs, this transfer will be positive.

[32] If $m_s \to 0$, as in Example (2) below, L'Hospital's rule gives $p_\lambda^{FB} = \frac{R_s^e}{\lambda_s^e}$.

Likewise, p_m^{FB} is set equal to the marginal residual social benefit of capacity per unit of measure m by raising capacity *when quality is adjusted to keep demand* $\lambda^e(q,s)$*constant*. Intuitively, p_λ and p_m incentivise different margins and their optimal levels are contingent on the other margin being kept at its optimal level.

We now consider two specific examples of first best payment regimes. The first combines the prospective output price with a price p_s on capacity s. If hospital capacity only depends on the number of beds, then a price per bed, together with a price per patient treated can support the first-best allocation. Since $m_s = 1$ and $m_q = 0$, Proposition 2 implies

Example 1. The first best prices on treated patients and capacity s are

$$p_\lambda^{FB1} = \frac{R_q^e}{\lambda_q^e} = \frac{\beta B_q^e}{\lambda_q^e} - \frac{c_q^{le}}{\lambda_q^e}, \tag{24}$$

$$p_s^{FB1} = R_s^e - R_q^e \frac{\lambda_s^e}{\lambda_q^e} = \beta\left(B_s^e - B_q^e \frac{\lambda_s^e}{\lambda_q^e}\right) - \left(c_s^{le} - c_q^{le}\frac{\lambda_s^e}{\lambda_q^e}\right) \tag{25}$$

$$= \beta\left(B_s^e - B_q^e \frac{\lambda_s^e}{\lambda_q^e}\right) - \sigma \overline{w}_s \int_{y^{\min}}^{\hat{y}^e(q,s)} r(\ell(y))\mathrm{d}F(y), \tag{26}$$

where all terms are evaluated at the first best q and s.

The prospective output price (24) reflects the fact that rewarding output incentivises quality. The first best output price p_λ^{FB1} is less than the marginal social benefit per patient attracted by higher quality $\left(B_q^e/\lambda_q^e\right)$ to the extent that (i) hospitals are intrinsically motivated and raising public funds is costly (which imply $\beta < 1$), and (ii) a quality increase results in greater compensation for lost earnings because it attracts more patients 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, or no compensation for loss of earnings, so that $c^{le} = 0$. Then (24) can be written as $p_\lambda^{FB}\lambda_q^e = B_q^e/\left(1+\theta\right)$, so that the public hospitals' marginal revenue from increasing quality should be less than the marginal patient welfare from higher quality only because of the marginal cost of public funds.

The optimal reward for capacity (p_s^{FB1}) is less than its *RMSB* (R_s^e) because the prospective payment for patients treated already provides some incentive to increase capacity in order to shift the distribution of waiting times to the left.[33] Since demand is controlled through the effect of p_λ on choice of quality, the optimal marginal reward for capacity (26) ensures the optimal mix between q and s with demand held constant. As we noted in Section 2.5.1, if the marginal rate of substitution between quality and capacity is the same at all income levels, then $B_s^e = B_q^e \frac{\lambda_s^e}{\lambda_q^e}$ and p_s^{FB1} is solely determined by the effect of capacity on the expected cost of sickness insurance when quality is adjusted to hold demand constant. With demand constant c^{le} is reduced by an increase in capacity (see (16)) since it reduces the mean waiting time. Hence capacity should be subsidised.

In many systems pay for performance (P4P) schemes link payments to hospital quality. For example, in the English NHS hospitals are paid higher prices for some treatments if they follow stipulated best practice guidelines and are financially penalised if too many of their patients have an emergency readmission within 30 days of discharge (Meacock et al., 2014; Kristensen et al., 2013). In our setting, a price attached to quality implies $m_s = 0$ and $m_q = 1$ and we obtain

Example 2. The first best prices per patient and for quality are

$$p_\lambda^{FB2} = \frac{R_s^e}{\lambda_s^e} = \beta\frac{B_s^e}{\lambda_s^e} - \frac{c_s^{le}}{\lambda_s^e}, \tag{27}$$

$$p_q^{FB2} = R_q^e - R_s^e\frac{\lambda_q^e}{\lambda_s^e} = \beta\left(B_q^e - B_s^e\frac{\lambda_q^e}{\lambda_s^e}\right) + \frac{\lambda_q^e}{\lambda_s^e}\sigma\overline{w}_s\int_{y^{\min}}^{\hat{y}^e(q,s)} r(\ell(y))\mathrm{d}F(y), \tag{28}$$

where all terms are evaluated at the first best q and s.

Since s is not directly rewarded, it is incentivised by the price per treated patient attracted by the fact that higher s results in a more favourable distribution of waiting times (27). The marginal reward for quality (28) is less than its RMSB R_q^e because quality is also indirectly incentivised through the prospective output price. The quality reward is thus given by the demand constant RMSB of quality.

If all patients have the same marginal rate of substitution of quality for capacity, the responses of demand to quality and capacity would transmit the correct signal to the hospital about patient preferences. The first term on the right hand side of (28) would be zero and the only reason for wishing to change quality is because it also affects the expected waiting time and hence the cost of providing insurance against lost earnings. At constant demand (and thus constant waiting list), the reduction in capacity required to keep demand constant when quality increases results in a longer waiting time ($-\overline{w}_s\frac{\lambda_q^e}{\lambda_s^e} > 0$) and higher insurance costs. Thus a *penalty* on quality is required to correct for the implicit over-rewarding of quality through p_λ. This is a stark illustration of how allowing for the costs imposed by rationing by waiting affects the form of the optimal payment scheme.

These two examples of first best payment regimes combine a prospective price with a reward directly targeted either q or s. As we show in Appendix D it is also possible to achieve the first best by combining the prospective price with a price on measures such as the average waiting time $\overline{w}^e(q,s)$, or social insurance costs $c^{le}(q,s)$, which are functions of both decision variables.

3.3. Multiple quality and supply decisions

As we show in Appendix F.1, our results also hold when there are multiple quality ($n^q > 1$) and capacity ($n^s > 1$) dimensions. Achieving the first best will in general require

[33] Notice that in (25) we can use (24) to get $p_s^{FB1} = R_s^e - p_\lambda^{FB1}\lambda_s^e$.

an output price plus $n^q + n^s - 1$ prices attached to $n^q + n^s - 1$ performance measures which are linearly independent and have at least one of the measures affected by each of $n^q + n^s - 1$ quality and capacity decisions.

The variety of pay for performance schemes which incentivise different aspects of quality (Jha et al., 2012; Milstein and Schreyoegg, 2016; Sutton et al., 2012) suggests that finding sufficient performance measures related to quality decisions may not be a problem. But the distribution of waiting times may depend on hard to observe factors such as efforts to reduce patient non-attendance at clinics, the extent of coordination between different hospital specialities, and hospital liaison with local social service departments to reduce bed-blocking by patients who are medically fit for discharge.

However, the required set of first best prices is drastically reduced if patients care only about quality and the mean waiting time.[34] In Appendix E we prove

Proposition 3. *If patients care only about the mean wait* $\overline{w}^e(q_1, \ldots, q_{n^q}, s_1, \ldots, s_{n^s})$ *then the first best can be achieved (a) with an output price, prices attached to* $n^q - 1$ *quality dimensions, and a price attached to the mean waiting time or (b) with an output price and* n^q *prices on the quality dimensions.*

The intuition is straightforward. Although the distribution of waiting times depends on all the quality and supply dimensions, at given quality the mean waiting time is a sufficient statistic for the waiting time distribution and the n^s decisions affecting it. Thus there are only $n^q + 1$ hospital characteristics which affect patient welfare and so only n^q prices are required in addition to the output price.

3.4. First best with only a prospective output price?

We now return to the world in which quality and capacity each have a single dimension in order to examine the circumstances in which the first best can be achieved using only the prospective output price. Thus it is not possible to observe q or s or any performance measure $m(q, s)$ affected by them. Proposition 2 implies that the first best can then be achieved using just the prospective output price p_λ^{FB} only if the first best price p_m^{FB} on a performance measure $m(q, s)$ is zero. Inspection of the numerator of (23) shows that this requires (see (25) and (26))

$$R_s^e - R_q^e \frac{\lambda_s^e}{\lambda_q^e} = \beta \left(B_s^e - B_q^e \frac{\lambda_s^e}{\lambda_q^e} \right)$$

$$- \sigma \overline{w}_s \int_{y^{\min}}^{\hat{y}^e(q,s)} r(\ell(y)) \mathrm{d}F(y) = 0. \tag{29}$$

From Proposition 1 the first term is zero if and only if the marginal rate of substitution $MRS^e(q, s; y)$ between $q$and s is independent of income. As we showed this requires (a) that income is multiplicatively separable from quality and waiting time and (b) there is no loss of income whilst waiting. Since the second term in (29) is zero if and only if there is no sickness insurance against lost earnings, this combined with (b) implies that treatment must not affect earnings. Hence, we have

Proposition 4. *The first best allocation is achievable using only a prospective output price if and only if (a) quality and waiting time are multiplicatively separable from income in patient preferences and (b) treatment does not affect earnings* ($\ell(y) \equiv 0$).

We can illustrate the proposition diagrammatically in Fig. 2 since the conditions stated in Proposition 3 satisfy the assumptions required to provide the simple illustration of the rational expectations equilibrium in Fig. 1.

Suppose that the hospital initially faces the first best price p_λ^{FB} defined in Proposition 2 and there are no other financial incentives. The hospital chooses q_1 and s_1 which result in the rational expectations equilibrium in Panel I in which the equilibrium expected wait is $\overline{w}_1 = \overline{w}^e(q_1, s_1)$ days and the equilibrium demand is $\lambda^e(q_1, s_1)$. Consider the introduction of a small reward $p_s > 0$ for capacity accompanied by a small reduction in p_λ which lead the hospital to increase capacity ($\mathrm{d}s > 0$) but to reduce quality ($\mathrm{d}q < 0$), so that the equilibrium demand (and hence expected number of patients treated) does not change ($\mathrm{d}q = -\frac{\lambda_s^e}{\lambda_q^e}\mathrm{d}s < 0$).

The effect of $\mathrm{d}q < 0$ is to shift the $v(y_{\min}, q, \overline{w}_1)$ line in Panel II to the right and to make it steeper in $(\overline{w}, v)$ space. The utility function $v(\cdot, q, \overline{w}_1)$ in Panel III shifts to the right from $v(\cdot, q_1, \overline{w}_1)$ to $v(\cdot, q_2, \overline{w}_1)$ and the critical income falls to $\hat{y}(q_2, \overline{w}_1)$. The reduction in quality shifts the demand curve in Panel I down at all expected waiting times. But, by construction, the increase in capacity from s_1 to s_2 shifts the expected waiting time function $\overline{w} = 1/(s - \lambda)$ downward so that the expected waiting time falls and equilibrium demand is unchanged: $\lambda^e(q_1, \overline{w}^e(q_1, s_1)) = \lambda^e(q_2, \overline{w}^e(q_2, s_2))$.

In Panel III the reduction in $\overline{w}$ to $\overline{w}_2$ shifts the $v(\cdot, q_2, \overline{w})$ curve back to the left, offsetting the reduction in quality, so that the critical income (and so demand) is unchanged: $\hat{y}(q_1, \overline{w}_1) = \hat{y}(q_2, \overline{w}_2)$. The expected utility of the marginal public hospital patient with this critical income is unchanged: $v(\hat{y}, q_1, \overline{w}_1) = v(\hat{y}, q_2, \overline{w}_2) = v(\hat{y}, q_1 + \mathrm{d}q, \overline{w}_1 + \mathrm{d}\overline{w})$ which implies that the marginal rate of substitution of the marginal patient is $-v_q(\hat{y}, q, \overline{w}_2)/v_{\overline{w}}(\hat{y}, q, \overline{w}_2)$. But the assumption (a) on preferences and patient income in Proposition 3 means that, from Proposition 1, the marginal rate of substitution between quality and waiting time is the same at all income levels. Hence the changes $\mathrm{d}q$,$\mathrm{d}\overline{w}$ which hold the expected utility of the marginal patient unchanged also do not change the expected utility of infra-marginal patients. Thus reducing the prospective price below its first best level and introducing a capacity incentive does not change the expected utility of any patient.

The other reason for introducing a capacity incentive is to change the expected cost of insuring individuals against loss of income whilst waiting for treatment because this is ignored by the partially altruistic public hospital. The

[34] We show in Appendix F.2 that this requires that $u(y, q, w)$ (4) is linear in w.

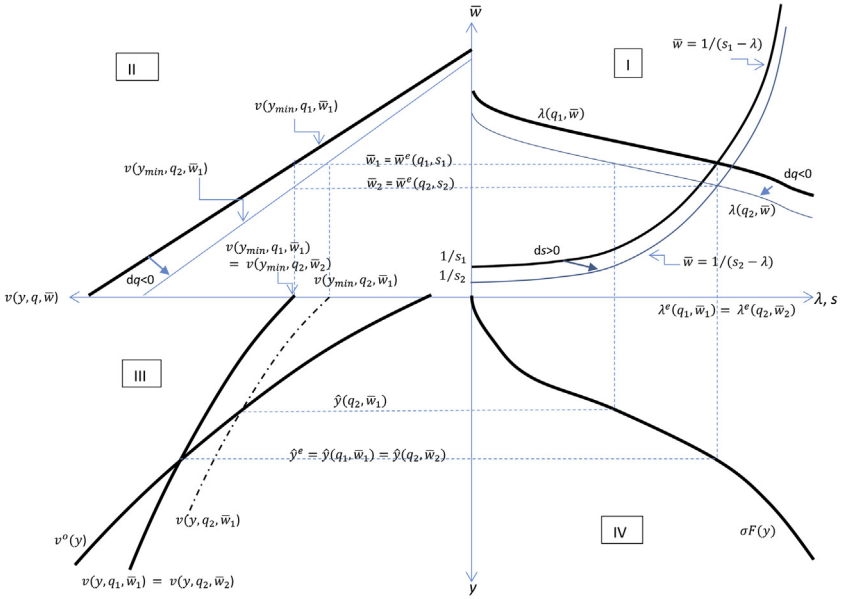

Fig. 2. Effect of ($dq < 0$, $ds > 0$) which keep equilibrium expected demand $\lambda^e(q, s)$ constant. Initial equilibrium at q_1 , s_1 has expected waiting time $\bar{w}_1 = \bar{w}^e(q_1, s_1)$, equilibrium expected demand $\lambda^e(q_1, s_1)$. New equilibrium at $q_2 = q_1 + dq < q_1$, $s_2 = s_1 + ds > s_1$ has expected waiting time $\bar{w}_2 = \bar{w}^e(q_2, s_2) < \bar{w}_1 = \bar{w}^e(q_1, s_1)$, equilibrium expected demand $\lambda^e(q_2, s_2) = \lambda^e(q_1, s_1)$. $v(y, q, \bar{w})$ and $v^o(y)$ are expected utility for individuals who chose the public hospital and the outside option of private treatment when ill, respectively. $\hat{y}(q, \bar{w})$ critical income of marginal public hospital patient.

diagrammatic representation of the rational expectations equilibria in Fig. 2 requires, in addition to the restrictions on preferences given in part (ii)(a) in Proposition 1, that waiting for treatment does not affect patient income (part (ii)(b)). This implies either that patients are fully compensated by the insurance scheme for loss of *earnings* or that there is no loss of earnings. But if there is loss of earnings which is fully compensated from the insurance fund, then the reduction in waiting time from the introduction of a reward for capacity will reduce the expected insurance cost and welfare will increase. Thus the first best can be achieved using only a prospective output price only with strong restrictions on preferences (requirement (a) in Proposition 3) *and* if there is no loss of earnings when waiting for treatment (requirement (b) in Proposition 3).

The strong conditions on preferences and the effect of illness on incomes imply that a welfare loss is likely when the only instrument is the prospective output price. The output price leads the hospital to take account of the effect of its quality and capacity decisions because patient demand is affected by quality and the distribution of waiting times. The effects of quality and capacity decisions on demand depend on the preferences of the marginal patients with income $\hat{y}^e(q, s)$ who are indifferent between the public and private sectors. The responses of these marginal patients will not convey the right information about the marginal value of the public hospital quality and capacity decisions for the infra-marginal patients with $y < \hat{y}^e(q, s)$ unless all patients have the same marginal rates of substitution between quality and waiting times.

As Spence (1975) noted, this type of problem will arise in all markets, competitive or not, where consumers care about attributes of the commodity other than its price. In Spence (1975) consumers pay for the commodity and a monopoly will produce the socially optimal quality only if all consumers have the same marginal rate of substitution between income and quality so that the demand response to higher quality, from the marginal patient, conveys accurate information about its effect on infra-marginal patients. In our case patients "pay" for public hospital care by waiting and an output price will induce the hospital to chose the optimal mix of quality and capacity only if marginal and infra-marginal patients have the same marginal rate of substitution between quality and waiting time.

Consider, for example, minor skin procedures. Patients will be able to work whilst waiting for treatment (so that $c^{le} \equiv 0$). If lower income patients are more willing to sacrifice quality (cosmetic effects) for a shorter wait, then the demand response of the marginal patient ($\lambda_s^e/\lambda_q^e = -\frac{dq}{ds}\Big|_{d\lambda^e=0}$) will provide a misleadingly high signal about the willingness of infra-marginal patients to accept a longer wait in exchange for higher quality: quality will be too high and waiting times too long.

The second reason why a prospective output price may not implement the first best is the effect of waiting on earnings. If patients are less productive whilst waiting for treatment ($\ell(y) > 0$) and there is incomplete earnings insurance ($r(\ell(y)) < \ell(y)$) then, from Proposition 1, the marginal rate of substitution between quality and capacity of marginal and infra-marginal patients will differ and the hospital will receive the wrong demand signals about the mix of quality and capacity. But, even if there is full compensation for loss of earnings ($r(\ell(y)) = \ell(y)$), the hospital will still choose the wrong mix of q and s because it

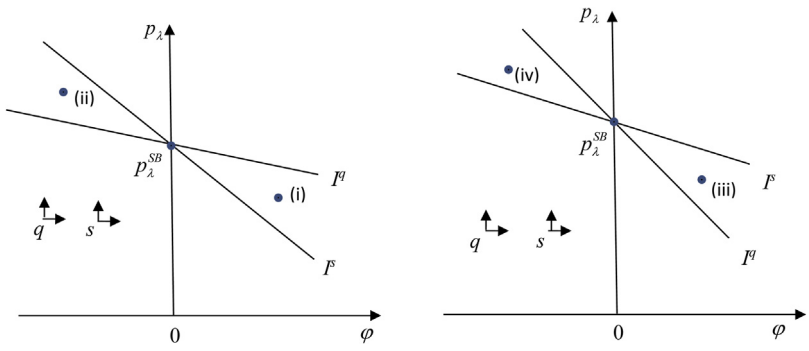

Fig. 3. Welfare increasing cost subsidy ($\varphi > 0$) or penalty ($\varphi < 0$). p_λ^{SB} second best price with no cost subsidy. I^s, I^q contours of $s(p_\lambda, \varphi)$, $q(p_\lambda, \varphi)$ chosen by provider. If $A_q^e < 0$, $A_s^e > 0$ welfare increased by cost subsidy at (i) and by cost penalty at (iv). If $A_q^e > 0$, $A_s^e < 0$ welfare increased by cost penalty at (ii) and by cost subsidy at (iii).

ignores the effect of its decisions on the cost of insurance $c^{Ie}(q, s)$.[35]

4. Second best incentives

We now consider a second best world in which quality and capacity decisions are not observable and the assumptions set out in Proposition 4 do not hold. First, in Section 4.1 we derive the second best optimal output price on the assumption that this is only policy instrument available. Then, in Section 4.2 we allow for the possibility that managerial effort can affect hospital costs and show that linking payment to both output and hospital cost is welfare increasing, though, in contrast to the previous literature, it is possible that hospital cost should be taxed rather than subsidised.

4.1. Second best output price

When the regulator's only instrument is the prospective output price p_λ the hospital's objective function is $p_\lambda\lambda(q, s) + \alpha B^e(q, s) - c^e(q, s)$ and its choices $q(p_\lambda)$ and $s(p_\lambda)$ satisfy the first order conditions

$$p_\lambda \lambda_z^e + \alpha B_z^e = c_z^e, \quad (z = q, s). \tag{30}$$

The regulator chooses her only instrument p_λ to satisfy

$$0 = \frac{\mathrm{d}A^e}{\mathrm{d}p_\lambda} = A_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + A_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda}. \tag{31}$$

Using the definition of the RMSBs (19), the hospital's first order conditions (30), and the definition $\frac{\mathrm{d}\lambda^*}{\mathrm{d}p_\lambda} \overset{\text{def}}{=} \lambda_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + \lambda_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda}$, (31) is equivalent to

$$\begin{aligned} 0 &= (R_q^e + \alpha B_q^e - c_q^e)\frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + (R_s^e + \alpha B_s^e - c_s^e)\frac{\mathrm{d}s}{\mathrm{d}p_\lambda} \\ &= (R_q^e - p_\lambda \lambda_q^e)\frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + (R_s^e - p_\lambda \lambda_s^e)\frac{\mathrm{d}s}{\mathrm{d}p_\lambda} \\ &= R_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + R_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda} - p_\lambda \frac{\mathrm{d}\lambda^*}{\mathrm{d}p_\lambda}. \end{aligned}$$

Replacing R_z^e by $p_\lambda^{FB}\lambda_z + p_m^{FB} m_z$ $(z = q, s)$ (cf (20) and (21)) evaluated at second best quantities and defining $\frac{\mathrm{d}m^*}{\mathrm{d}p_\lambda} \overset{\text{def}}{=} m_q(q, s)\frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + m_s(q, s)\frac{\mathrm{d}s}{\mathrm{d}p_\lambda}$ where $m(q, s)$ is the generic performance measure used in the first best scheme in Proposition 2, yields

Proposition 5. *The second best price for output is*

$$p_\lambda^{SB} = \frac{R_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + R_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda}}{\mathrm{d}\lambda^*/\mathrm{d}p_\lambda} \tag{32}$$

$$= \tilde{p}_\lambda^{FB} + \tilde{p}_m^{FB} \frac{\mathrm{d}m^*/\mathrm{d}p_\lambda}{\mathrm{d}\lambda^*/\mathrm{d}p_\lambda}, \tag{33}$$

where $\tilde{p}_\lambda^{FB}$, $\tilde{p}_m^{FB}$ are the first best prices attached to output and the generic measure m, specified by Proposition 1 but evaluated at the second best decision levels for (q, s).

The first expression (32) shows p_λ^{SB} as a weighted sum of the RMSBs from quality and capacity, emphasising the compromise that the second best output price strikes between incentives for these two hospital decisions.[36] In (33) $\frac{\mathrm{d}m^*/\mathrm{d}p_\lambda}{\mathrm{d}\lambda^*/\mathrm{d}p_\lambda}$ can be interpreted as the induced marginal rate of transformation between the performance measure m and hospital output.[37] In the second best, m can no longer be directly incentivised and so the price on output should take over some of the rôle played by p_m in first best.

4.2. Cost reducing effort

We have so far ignored the possibility that managerial effort by the hospital can reduce its production cost c. When the regulator has sufficient instruments to control quality and capacity she does not need an additional instrument to control managerial effort. Since managerial effort and production costs are both borne by the hospital, effort will be chosen efficiently to minimise the sum of production and effort cost.

If the only policy instrument which can be linked to quality and capacity is the output price p_λ then an incen-

[35] The hospital will also ignore the cost to employers if they insure workers by maintaining their income whilst waiting for treatment.

[36] $\frac{R_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + R_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda}}{\mathrm{d}\lambda^*/\mathrm{d}p_\lambda} = \left(\frac{R_q^e}{\lambda_q^e}\lambda_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + \frac{R_s^e}{\lambda_s^e}\lambda_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda}\right) / \left(\lambda_q^e \frac{\mathrm{d}q}{\mathrm{d}p_\lambda} + \lambda_s^e \frac{\mathrm{d}s}{\mathrm{d}p_\lambda}\right)$.

[37] In the case where $m(s, q) = s$, $\frac{\mathrm{d}m^*/\mathrm{d}p_\lambda}{\mathrm{d}\lambda^*/\mathrm{d}p_\lambda} = \frac{\mathrm{d}s/\mathrm{d}p_\lambda}{\mathrm{d}\lambda^*/\mathrm{d}p_\lambda}$, which is the marginal capacity requirement per patient attracted.

tive linked to production cost could be welfare increasing despite its distorting effect on managerial effort. Subsidising production cost will reduce the marginal costs of quality and capacity as perceived by the hospital and so will provide an additional instrument to supplement the prospective output price. But a cost subsidy will also reduce managerial effort and so increase hospital cost at given quality and capacity. The overall welfare effect of the cost subsidy will depend on the magnitudes of the resulting changes in cost, quality and capacity.

Suppose that the hospital cost function is $c^e(q, s, \tau)$ where τ is cost-reducing effort which imposes a non-verifiable cost $g(\tau)$ on the hospital and that the hospital is refunded a fraction φ of its production cost.[38] Hospital decisions on q, s and τ will depend on φ and p_λ. At $\varphi = 0$, the optimal price p_λ^{SB} (given in Proposition 5) satisfies

$$\left.\frac{dA^e}{dp_\lambda}\right|_{\varphi=0} = A_q^e \frac{\partial q}{\partial p_\lambda} + A_s^e \frac{\partial s}{\partial p_\lambda} + A_\tau^e \frac{\partial t}{\partial p_\lambda}$$
$$= A_q^e \frac{\partial q}{\partial p_\lambda} + A_s^e \frac{\partial s}{\partial p_\lambda} = 0, \qquad (34)$$

since at $\varphi = 0$ the hospital chooses τ so that $c_\tau^e + g_\tau = 0$ which implies $A_\tau^e = -(1+\theta)[c_\tau^e + g_\tau] = 0$. The welfare effect of the introduction of a cost subsidy is

$$\left.\frac{dA^e}{d\varphi}\right|_{\varphi=0} = A_q^e \frac{\partial q}{\partial \varphi} + A_s^e \frac{\partial s}{\partial \varphi} + A_\tau^e \frac{\partial \tau}{\partial \varphi} = A_q^e \frac{\partial q}{\partial \varphi} + A_s^e \frac{\partial s}{\partial \varphi}. \qquad (35)$$

In the second best neither A_q^e or A_s^e are zero and a cost subsidy will change welfare because it induces changes in quality and capacity.

The comparative static responses of q and s to p_λ and φ are in general ambiguous (see Appendix G) but even if we make the intuitively plausible assumptions that q and s both increase with p_λ and φ this is insufficient to determine the sign of (35). Thus it is possible that the optimal second best cost subsidy is negative: the hospital should pay a tax on its costs.

To illustrate, use (34) to substitute for A_q^e in (35) and rearrange to get

$$\left.\frac{dA^e}{d\varphi}\right|_{\varphi=0} = (-A_s^e \frac{\partial s}{\partial p_\lambda} / \frac{\partial q}{\partial p_\lambda}) \frac{\partial q}{\partial \varphi} + A_s^e \frac{\partial s}{\partial \varphi}$$
$$= A_s^e \frac{\partial s}{\partial p_\lambda} [-(\frac{\partial q}{\partial \varphi} / \frac{\partial q}{\partial p_\lambda}) + (\frac{\partial s}{\partial \varphi} / \frac{\partial s}{\partial p_\lambda})]$$
$$= A_s^e \frac{\partial s}{\partial p_\lambda} [\left.\frac{\partial p_\lambda}{\partial \varphi}\right|_{dq=0} - \left.\frac{\partial p_\lambda}{\partial \varphi}\right|_{ds=0}].$$

The terms $\left.(\partial p_\lambda/\partial \varphi)\right|_{dq=0}$ and $\left.(\partial p_\lambda/\partial \varphi)\right|_{ds=0}$ are the slopes of the contours of the hospital's "supply" functions $q(\varphi, p_\lambda)$ and $s(\varphi, p_\lambda)$. In Fig. 3 we make the plausible assumption that increases in p_λ and φ both increase quality and capacity so that the contours I^q and I^s of $q(\varphi, p_\lambda)$ and $s(\varphi, p_\lambda)$ are negatively sloped. With no cost sharing welfare is maximised at $(0, p_\lambda^{SB})$ in both parts of the figure. In part (a) I^s has a more negative slope than I^q and if $A_s^e > 0$ (and hence $A_q^e < 0$) then moving to (i) by introducing a cost subsidy and reducing p_λ will increase welfare. Conversely, if $A_s^e < 0$ (and hence $A_q^e > 0$) welfare is increased by moving to (ii) with a cost penalty and $p_\lambda > p_\lambda^{SB}$. In part (b) I^q has a steeper negative slope than $I^s$and welfare is increased at (iii) by a cost subsidy if $A_s^e < 0, A_q^e > 0$ and by a cost penalty at (iv) if $A_s^e > 0, A_q^e < 0$.[39]

There is one case in which a cost subsidy or penalty is not a useful policy instrument in the second best. If the hospital is a pure profit maximiser ($\alpha = 0$), its choice of q and s will equate their marginal revenues $(p_\lambda \lambda_z^e)$ to their marginal cost $((1-\varphi)c_z^e)$. If the cost function is additively or multiplicatively separable between effort τ and (q, s) so that $c^e(q, s, \tau) = c^1(q, s)c^2(\tau) + c^3(\tau)$ and $c_z^e = c_z^1(q, s)c^2(\tau)$, $(z = q, s)$, we can write the first order conditions on q and s as

$$\frac{p_\lambda}{(1-\varphi)c^2(\tau)} \lambda_z^e(q, s) = c_z^1(q, s), \quad (z = q, s).$$

Thus, as far as choice of (q, s) is concerned, changing the cost subsidy or penalty is equivalent to changing the output price and so does not provide an additional means of controlling (q, s). And since $\varphi \neq 0$ leads to an inefficient choice of effort a cost subsidy or penalty will reduce welfare in this case (Appendix G sets out the detailed analysis). Hence

Proposition 6. *If an output price is insufficient to achieve the first best then welfare can be increased by a cost subsidy or tax provided that the cost function is not separable between managerial effort and* (q, s).

The conclusion that subsidising, or taxing, hospital cost may increase welfare when there are insufficient instruments to control hospital decisions with a direct effect on patients has a straightforward intuition. The welfare loss due to the reduction in effort from a small cost subsidy is small because the effort level is initially optimally chosen by the hospital, whereas the welfare gains from the changes in the hospital decisions directly affecting patients are non-trivial. In previous analyses of cost sharing (surveyed in Chalkley and Malcomson (2000)) it is assumed that these hospital decisions have positive marginal welfare effects and are increased by the cost subsidy so that welfare is increased by subsidising cost. For example, in the seminal paper on prospective pricing (Ellis and McGuire, 1986) the number of patients requiring treatment is exogenously determined, so that with a prospective price the hospital, unless it is perfectly altruistic, will skimp on quality because it 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 since partial cost reimbursement also reduces the incentive for cost reducing effort 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 in the case considered in Ellis and McGuire (1986) because, with a fixed number of patients,

[38] The hospital now maximises $\alpha B^e(q, s) + p_\lambda \lambda^e(q, s) - [(1-\varphi)c^e(q, s, \tau) + g(\tau)]$ and the welfare function is $A^e = B^e(q, s) - (1+\theta)[c^e(q, s, \tau) + g(\tau) + c^{le}(q, s)]$.

[39] In Appendix G we provide an alternative characterisation of $\left.\frac{dA^e}{d\varphi}\right|_{\varphi=0}$ in terms of p_s^{FB}.

the prospective price is equivalent to a lump sum payment with no incentive properties: the only instrument which affects the hospital quality decision is a cost subsidy.

But when there is more than one other hospital decision in addition to cost reducing effort (quality and capacity in our case) and the regulator has insufficient instruments to control them all, the marginal welfare effects of some of these hospital decisions will be positive and some negative, so that it is possible that cost should be taxed rather than subsidised in order to increase welfare. Proposition 6 thus generalises Ellis and McGuire (1986).

5. Conclusion

We have analysed optimal hospital payment schemes for elective procedures, extending previous analyses to take account of a salient and ubiquitous feature of public health care systems previously ignored in the literature: rationing by waiting time. Longer waiting times delay treatment, can reduce the health benefit from treatment, and can increase output losses if patients are less productive whilst waiting.

We developed a new, general, and analytically tractable, queueing model with rational expectations. The hospital chooses quality and capacity (beds, staffing, ...) taking account of their effects on its costs and on demand from patients. Patient decisions to demand care by joining the waiting list depend on quality and the equilibrium distribution of waiting times and their decisions give rise to the equilibrium distribution which thus depends on hospital and patient decisions.

In general, even in the simplest case in which there is a single dimension of quality and a single dimension of capacity, the first best can only be achieved if the prospective output price is supplemented with an additional instrument. Candidate instruments include payments linked to capacity or to quality, or to the average waiting time, or making the hospital bear some of the cost of public earnings insurance (Proposition 2 and Examples 1 and 2). Our results thus have implications for the design of Pay for Performance schemes linking hospital revenue to measures of quality. They also provide a justification for direct regulation of quality or the imposition of waiting times targets (Propper et al., 2010).

Even when quality only has a single dimension a prospective price alone is, in general, insufficient to achieve the first best. There are two difficulties. The first is Spence (1975) problem: if the hospital is rewarded only via the price, its incentives to adjust its quality and capacity decisions depend on their effects on demand and thus on the preferences over quality and the distribution of waiting times of the *marginal* patients. But the welfare consequences of quality and capacity decisions depend on their effects on the *infra-marginal* patients. Marginal and infra-marginal patients will have the same marginal rate of substitution between quality and waiting times only under strong separability assumptions about preferences and only if earnings are not affected by the length of wait for treatment (Proposition 1). The second problem is that the hospital will ignore the effect of its decisions on the costs of insuring patients for income lost whilst waiting for treatment. Together these problems imply that a prospective price alone will support the first best only under the strong assumptions that preferences are separable and that waiting for treatment has no effect on patient productivity so that there is no insurance of patient income (Proposition 4).

If there are n^q dimensions of quality affecting patient demand and n^s hospital supply decisions which directly alter the distribution of waiting times then achieving the first best will require $n^q + n^s - 1$ prices in addition to a prospective output price. But if patients care only about the mean waiting time, so that it is a sufficient statistic for the distribution of waiting times, then the first best can be achieved by supplementing the prospective output price with $n^q - 1$ prices on quality dimensions and a price, probably negative, on the mean waiting time or by using the output price and n^q prices on the quality dimensions (Proposition 3).

In the second best, when the regulator can only link payment to the number of patients treated, the optimal price per patient in general exceeds the first-best price to reflect the extra revenue the hospital would have obtained under a first best incentive payment scheme by increasing capacity to attract extra patients by reducing waiting times (Proposition 5).

When the hospital can exert effort to reduce its costs it is not necessary to directly incentivise such effort in the first best when there are a sufficient number of prices attached to hospital decisions since the hospital bears both production and effort costs. In the second best, welfare can be increased by linking reward to the hospital's cost and so distorting incentives for cost reducing effort. But, because patients are affected by both quality and waiting time, rather than just quality, it is possible that hospital cost should be surcharged rather than partially reimbursed. (Proposition 6).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jhealeco.2019.102277.

References

Aakvik, A., Holmås, T.H., Kjerstad, E., 2015. Prioritization and the elusive effect on welfare – a Norwegian health care reform revisited. Soc. Sci. Med. 128, 290–300.

Besley, T., Hall, J., Preston, I., 1999. The demand for private health insurance: do waiting lists matter? J. Public Econ. 72, 155–181.

Boiteux, M., 1956. Sur la gestion des monopoles publiques astreints à l'équilibre budgétaire. Econometrica 24, 22–40 (translated as: Boiteux M. (1971). On the management of public monopolies subject to budget constraints, *Journal of Economic Theory* 3, 219–240.).

Brekke, K., Siciliani, L., Straume, O., 2008. Competition and waiting times in hospital markets. J. Public Econ. 92, 1607–1628.

Chalkley, M., Malcomson, J., 1998. Contracting for health services with unmonitored quality. Econ. J. 108, 1093–1110.

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

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.

Drummond, M., Sculpher, M., Claxton, K., Stoddart, G., Torrance, G., 2015. Methods for the Economic Evaluation of Health Care Programmes, 4th ed. Oxford University Press.
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. J. Health Econ. 5, 129–151.
Ellis, R., McGuire, T., 1990. Optimal payments systems for health services. J. Health Econ. 8, 375–396.
Fomundam, S., Herrmann, J., 2007. A Survey of Queuing Theory Applications in Healthcare. University of Maryland, ISR Report 2007-24.
Goddard, J., Malek, M., Tavakoli, M., 1995. An economic model for hospital treatment for non-urgent conditions. Health Econ. 4, 41–55.
Grassi, S., Ma, A., 2011. Optimal public rationing and price response. J. Health Econ. 30, 1197–1206.
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. J. Health Econ. 21, 423–449.
Gravelle, H., Schroyen, F., 2016. Optimal Hospital Payment Rules Under Rationing by Random Waits, DP SAM-08-2016 (Norwegian School of Economics) or CHE Research Paper 130. Centre for Health Economics.
Gravelle, H., Siciliani, L., 2008. Optimal quality, waits and charges in health insurance. J. Health Econ. 27, 663–674.
Gross, D., Shortle, J., Thompson, J., Harris, C., 2008. Fundamentals of Queueing Theory, 3rd ed. Wiley, New York.
Hassin, R., Haviv, M., 2003. To Queue or not to Queue: Equilibrium Behavior in Queueing Systems. Springer.
Iversen, T., Lurås, H., 2002. Waiting time as a competitive device: an example from general medical practice. Int. J. Health Care Finance Econ. 2, 189–204.
Iversen, T., Siciliani, L., 2011. Non-price rationing and waiting times. In: Glied, S., Smith, P. (Eds.), The Oxford Handbook of Health Economics. OUP, Oxford.
Jha, A., Joynt, K., Orav, E., Epstein, A., 2012. The long-term effect of Premier pay for performance on patient outcomes. N. Engl. J. Med. 366, 1606–1615.
Kristensen, S., McDonald, R., Sutton, M., 2013. Should pay for performance schemes be locally designed? Evidence from the commissioning for quality and innovation (CQUIN) framework. J. Health Serv. Res. Policy 18 (2 Suppl), 38–49.
Koopmanschap, M., Rutten, F., van Ineveld, B., van Roijen, L., 1995. The friction cost method for measuring indirect costs of disease. J. Health Econ. 14, 171–189.
Laine, L., Ma, A., 2017. Quality and competition between public and private firms. J. Econ. Behav. Organ. 140, 336–353.
Lindsay, C., Feigenbaum, B., 1984. Rationing by waiting lists. Am. Econ. Rev. 74, 404–417.
Littlechild, S., 1974. Optimal arrival rate in a simple queueing system. Int. J. Prod. Res. 12, 391–397.
Marchand, M., Schroyen, F., 2005. Can a mixed health care system be desirable on equity grounds? Scand. J. Econ. 107, 1–23.
Martin, S., Smith, P., 1999. Rationing by waiting lists: an empirical investigation. J. Public Econ. 71, 141–164.
Meacock, R., Kristensen, S., Sutton, M., 2014. Paying for improvements in quality: recent experience in the NHS in England. Nord. J. Health Econ. 2 (1), 239–255.
Milstein, R., Schreyoegg, J., 2016. Pay for performance in the inpatient sector: a review of 34 P4P programs in 14 OECD countries. Health Policy 120, 1125–1140.
Naor, P., 1969. The regulation of queue size by levying tolls. Econometrica 37, 15–24.
Nikolova, S., Harrison, M., Sutton, M., 2015. The impact of waiting time on health gains from surgery: evidence from a national patient-reported outcome dataset. Health Econ. 25, 955–968.
Paris, V., Devaux, M., Wei, L., 2010. Health Systems Institutional Characteristics: A Survey of 29 OECD Countries. OECD Publishing, http://dx.doi.org/10.1787/5kmfxfq9qbnr-en, OECD Health Working Papers, No. 50.
Pritchard, C., Sculpher, M., 2000. Productivity costs: principles and practice in economic evaluation. Off. Health Econ., Monograph Number 000464.
Reichert, A., Jacobs, R., 2018. The impact of waiting time on patient outcomes: evidence from early intervention in psychosis services in England. Health Econ., http://dx.doi.org/10.1002/hec:3800.
Propper, C., Sutton, M., Whitnall, C., Windmeijer, F., 2010. Incentives and targets in hospital care: evidence from a natural experiment. J. Public Econ. 94, 318–335.
Scruggs, L., Jahn, D., Kuitto, K., 2017. Comparative Welfare Entitlements Dataset 2 Version 2017–09. University of Connecticut and University of Greifswald.
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: Jones, A. (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 Econ. 21, 444–456.
Spence, M., 1975. Monopoly, quality, and regulation. Bell J. Econ. 6, 417–429.
Sutton, M., Nikolova, S., Boaden, R., Lester, H., McDonald, R., Roland, M., 2012. Reduced mortality with pay for performance in England. N. Engl. J. Med. 379, 1821–1828.
Taylor, H., Karlin, S., 1998. An Introduction to Stochastic Modelling, 3rd ed. Academic Press, New York.
Weisbrod, B.A., 1961. The valuation of human capital. J. Polit. Econ. 69, 425–436.
Windmeijer, F., Gravelle, H., Hoonhout, P., 2005. Waiting lists, waiting times and admissions: an empirical analysis. Health Econ. 14, 971–985.