

Output based incentive schemes for public hospitals—a queueing theory approach*

Hugh Gravelle[†] and Fred Schroyen[‡]

March 30, 2009.

Preliminary.

Abstract We consider patients that fall in need for an elective treatment according to a Poisson process. Patients have the option between queuing up for a free treatment in a public hospital, or going for an immediate but financially costly treatment in a private hospital. We derive the steady state distribution for treatments in the public hospital and show how the hospital can influence this distribution by choosing an effort level. It turns out that by exerting more effort, the hospital will not only increase the average number of treatments per period, but also its variance. If the hospital is reimbursed according to a linear incentive scheme, a too powerful the incentive scheme is shown to have negative effects on the effort level, the reason being that a risk averse hospital tries to contain its income risk exposure by reducing effort. In a next stage, we derive the optimal structure of the hospital's incentive scheme and show that the welfare maximising bonus rate lies below the one minimising the expected waiting time.

*The paper benefited from presentations at the Health Economics Workshop (Bergen, November 2006), the 8th European Health Economics Workshop (Magdeburg, April 2007) and the Joint Meeting of the UK and Nordic Health Economics Study Groups (Aberdeen, August 2008). The comments by the discussants Pierre-Yves Geoffard, Michael Kuhn and Timo Seppälä, are thankfully acknowledged.

[†]National Primary Care Research and Development Centre, Centre for Health Economics, University of York, Heslington, York YO10 5DD. Email: hg8@york.ac.uk.

[‡]Department of Economics, Norwegian School of Economics and Business Administration, Helleveien 30, N-5045 Bergen, and Health Economics Bergen (HEB). Email: fred.schroyen@nhh.no

1 Introduction

In many countries, public health care services are provided at a price below marginal cost. This means that there is an excess demand which is rationed through waiting lists. Patients in need for an elective treatment are put on a waiting list and are called in when they have reached the top of the list. In countries where these waiting lists are significant, a private supply has developed to cater for the demand of patients with a high willingness to pay for avoiding delayed treatment.

At the same time, public hospitals are given stronger incentives to speed up the treatment of patients in order to reduce lengthy waiting times. Such schemes go under the name activity based finance. It is believed that by making a larger part of the hospital's remuneration conditional on its output, hospitals will provide more effort in treating patients, and in this way contribute to lower waiting times.

The purpose of this paper is to investigate the effect of more powerful incentives on the effort exerted by public hospitals, when these hospitals compete with a private sector. Many papers have been written on incentives for health care providers, but the novelty of our approach lies in the use of queueing theory to model the arrival process of patients looking for public treatment. The use of this theory in explaining waiting lists and waiting times is very modest.¹ The problem with the simple queueing model is that it is void of any incentives and decision making: 'arrivals' and 'service completions' occur according to a mechanistic process.² This may explain why this model has been used so rarely to explain non-price rationing mechanisms.

If this theory is to bear any fruit, it has to be complemented with some decision making on the part of patients, or their representative GP, and hospitals. This is what we will try to do in this paper. On the other hand, several features of the queueing theory model describe well the stochastic process that drives the demand for health care. Except for infectious diseases, the process by which persons are hit by illness or accidents is random and independent. If illness may strike to a large group of persons, but if the probability of a single person falling ill is small, then the 'law of rare events' applies, and the occurrence of illness in the population will follow, approximately, a Poisson distribution which has the distinguishing feature that its mean equals its variance.

¹E.g. Worthington (1987, 1991). See also the survey by Fomundam and Herrmann (2007).

²This view was also expressed by Johansen (1987).

But this feature has important implications for the design of incentives for public hospitals. As we will show, when the public hospital is rewarded according to a linear output based incentive scheme, the bonus may have perverse effects on the effort level of the hospital. The reason is that a higher effort on the part of the hospital not only increases the expected number of treatments, but also its variance. When a risk averse hospital is facing a too powerful incentive structure, it is exposed to an income risk which it will try to contain. This can be done by reducing its effort: average demand for its services will fall, but so will the spread around that average. In sum: a more 'powerful' output based financing scheme will incite the hospital to cut its service rate in an attempt to reduce its exposure towards a risky remuneration. The same logic also points at a more important rôle for lump sum grants. To the extent that risk aversion falls with income, will a larger unconditional grant make the hospital more willing to increase its exposure towards risk and therefore its service rate.

The paper is organised along the following lines. In the next section, we describe the stochastic process that rules the demand for medical care, and how a patient chooses between an immediate treatment in the private sector or queueing up for a treatment in the public hospital. There, we also derive the steady state distribution for the number treatments in the public hospital. Next (section 3), we analyse the response of the public hospital manager when facing a linear financing scheme. Given this response, a welfare analysis of the optimal financing scheme is carried out in section 4. A discussion of the main results in section 5 concludes the paper.

2 Sickness incidence and hospital choice

We consider sickness incidence as a stochastic process with the following features:

- (1) the number of people falling ill in disjoint time intervals are independent random variables;
- (2) the number of people falling ill in a time intervals depends only on the length of that interval;
- (3) the probability of at least one person falling ill in a given time interval is proportional to the length of that interval; and
- (4) the probability of two or more persons falling ill in a 'small' interval is of second order to the length of that interval.

Under these assumptions, one can show that the number of sickness incidences in an interval whose size we normalise to unity follows a Poisson distribution with parameter λ , the constant of proportionality mentioned in (3).³ Thus the probability that within a unit time period x people fall ill is

$$f^{Pois}(x|\lambda) \stackrel{\text{def}}{=} \exp(-\lambda) \frac{\lambda^x}{x!} \quad (x = 0, 1, 2, \dots),$$

with $E(x) = \text{var}(x) = \lambda$. This positive relation between mean and variance of the arrival process should be kept in mind as it drives the main result of this paper.

Suppose for simplicity that the public hospital consists of one 'server'—i.e., that no more than one patient can be treated at a time—and that the service time at this server follows a negative exponential distribution with parameter μ .⁴ The reciprocal of this parameter is the mean time it takes to complete the treatment of a patient. Suppose also that the queue discipline is 'first come, first served' (we will comment on this in the final section). If $n - 1$ people have already joined the public queue and one person is undergoing treatment in the public hospital, then the total time w that a new patient joining the queue will spend in the system is comprised of the sum of $n + 1$ iid exponential random variables; it is Gamma distributed with parameters μ and n (Taylor and Karlin 1998, p 550):

$$g(w|n) = \frac{\mu(\mu w)^n \exp(-\mu w)}{n!}.$$

Let us normalise the monetary benefit of an immediate public treatment (net of any copayment rate) to 1, and suppose that the patient has a time preference rate for treatment that is constant and equal to r . (cf Lindsay and Feigenbaum, 1984). Then her discounted utility over waiting time is given by:

$$u(w) \stackrel{\text{def}}{=} \exp(-rw).$$

Therefore, her expected utility of queueing up in the public hospital when n patients are in the system is

$$E_w[u(w)|n] = \int_0^\infty \exp(-rw)g(w|n)dw = \left(\frac{\mu}{\mu + r}\right)^{n+1}.$$

³Formally, if $N((s, t])$ denotes the number of sickness incidences in the half open time interval $(s, t]$, then (i) the random variables $N((t_0, t_1]), N((t_1, t_2]), \dots, N((t_{n-1}, t_n])$ are independent; (ii) $\Pr\{N((s, s + \Delta]) = k\}$ depends only on Δ , not on s ; (iii) there is a positive constant λ such that $\Pr\{N((s, s + \Delta]) \geq 1\} = \lambda\Delta + o(\Delta)$ as $\Delta \rightarrow 0$; and (iv) $\Pr\{N((s, s + \Delta]) \geq 2\} = o(\Delta)$ as $\Delta \rightarrow 0$. In that case $\Pr\{N((s, s + \Delta]) = k\}$ is the Poisson distribution with parameter $\lambda\Delta$. See Taylor and Karlin (1998, p 281-4).

⁴The assumption of a single server can be easily relaxed.

The most natural decision making process is one where a person who has just fallen ill acquires information about the length of the queue (n), and then compares $E_w[u(w)|n]$ with the utility of an immediate private treatment, τ . A patient observing n other patients in the public system will then queue up when

$$\tau \leq \left(\frac{\mu}{\mu + r} \right)^{n+1},$$

where the *rhs* is strictly less than one. Because they face different transportation costs to the hospital, or because of differences in earnings ability which create differences in access to a private health insurance schemes., patients are likely to be heterogeneous in terms of τ . Denoting the distribution of τ by $H(\cdot)$, the number of patients queuing up will be given by $H\left(\left(\frac{\mu}{\mu+r}\right)^{n+1}\right)$.

A significant feature of this set up is that arrival rates to the public system are dependent of the state of the system (n).⁵ Since this dependency makes it hard to give an analytical solution to the system of steady state probabilities, we are going to assume that patients adopt an *ex ante* decision rule to which they stick when falling ill, irrespective of the actual length of the queue. This decision rule is constructed as follows.

Suppose that the expected number of arrivals in the public system is $\hat{\lambda}$. Public arrivals are therefore expected to follow a Poisson process $f^{Pois}(x|\hat{\lambda})$. In Kendall's notation, the 'expected' public queueing system is an $M/M/1$ system, with the steady state probability that there are n people in the system given by

$$P_n = \frac{\mu - \hat{\lambda}}{\mu} \left(\frac{\hat{\lambda}}{\mu} \right)^n, \quad n = 0, 1, 2, \dots \quad (1)$$

(see, e.g., Taylor and Karlin, 1985, p 549). For this equilibrium distribution to exist, it is necessary that $\mu > \hat{\lambda}$.

⁵Because $H\left(\left(\frac{\mu}{\mu+r}\right)^{n+1}\right)$ patients will queue up in the public system, the arrival rate of patients at the public queue becomes Poisson with a mean rate $\lambda H\left(\left(\frac{\mu}{\mu+r}\right)^{n+1}\right)$, which is obviously conditional on n .

Ex ante, the expected utility for a citizen choosing public hospital treatment is then

$$\begin{aligned} V(\mu, \hat{\lambda}, r) &\stackrel{\text{def}}{=} \sum_{n=0}^{\infty} P_n E_w[u(w)|n] = \sum_{n=0}^{\infty} \frac{\mu - \hat{\lambda}}{\mu} \left(\frac{\hat{\lambda}}{\mu}\right)^n \left(\frac{\mu}{\mu + r}\right)^{n+1} \\ &= \frac{\mu - \hat{\lambda}}{\mu - \hat{\lambda} + r}. \end{aligned} \quad (2)$$

A person with characteristic τ will therefore follow the decision rule "join the public queue when ill" if $\tau \leq \frac{\mu - \hat{\lambda}}{\mu - \hat{\lambda} + r}$. Hence, if the expected utility from queueing up in the public sector is $V(\mu, \hat{\lambda}, r)$, the fraction that will effectively queue up is $H\left(V(\mu, \hat{\lambda}, r)\right)$ and the rate at which people will arrive in the public system is $\lambda H\left(V(\mu, \hat{\lambda}, r)\right)$. If $\mu \rightarrow \infty$, $V(\mu, \hat{\lambda}, r) \rightarrow 1$, and the number of people choosing the public option is $\lambda H(1)$. If everybody considers the net benefit of an immediate treatment in the public hospital higher than that of an immediate private treatment, $H(1) = 1$ and the arrival rate to the public hospital will coincide with the illness incidence rate λ . But one can think of situations where people with a generous private insurance contract or with a strong faith in the services of a particular private physician prefer the treatment in the private sector. Then $\lambda H(1) < \lambda$.

In a rational expectations equilibrium, the mean arrival rate to the public system must coincide with the expected arrival rate:

$$\hat{\lambda} = \lambda H\left(V(\mu, \hat{\lambda}, r)\right). \quad (3)$$

For this equilibrium to be stable under educative learning, we need that $\frac{\lambda h r}{(\mu - \hat{\lambda} + r)^2} < 1$. In this case, individuals can compute the fixed point given by (3):

$$\hat{\lambda} = \hat{\lambda}(\lambda, \mu, r). \quad (4)$$

The function $\hat{\lambda}(\lambda, \mu, r)$ is the equilibrium parameter for the Poisson distribution of arrivals to the public system. Implicit differentiation of (3) gives

$$\begin{aligned} \frac{\partial \hat{\lambda}}{\partial \lambda} &= \frac{(\mu - \hat{\lambda} + r)^2}{(\mu - \hat{\lambda} + r)^2 + r \lambda h(\hat{v})} H(\hat{v}), \text{ and} \\ \frac{\partial \hat{\lambda}}{\partial \mu} &= \frac{\lambda r h(\hat{v})}{(\mu - \hat{\lambda} + r)^2 + r \lambda h(\hat{v})}, \end{aligned}$$

where $h(\cdot) = H'(\cdot)$ and \hat{v} is a shorthand for $V(\mu, \hat{\lambda}(\lambda, \mu, r), r)$. Clearly, $0 < \frac{d\hat{\lambda}}{H(\hat{v})d\lambda}, \frac{d\hat{\lambda}}{d\mu} < 1$. Using subscripts with $\hat{\lambda}$ to denote derivatives, the mean equilibrium arrival rate to the public system has the properties

$$\hat{\lambda}(\lambda, 0, r) = 0, \quad \lim_{\mu \rightarrow \infty} \hat{\lambda}(\lambda, \mu, r) = \lambda H(1) \quad (5)$$

$$\hat{\lambda}_\mu(\lambda, 0, r) = \frac{\lambda h(0)}{r + \lambda h(0)}, \quad \lim_{\mu \rightarrow \infty} \hat{\lambda}_\mu(\lambda, \mu, r) = 0 \quad (6)$$

$$\hat{\lambda}_{\mu\mu}(\lambda, \mu, r) = \frac{1 - \hat{\lambda}_\mu}{(\mu - \hat{\lambda} + r)^2 + r\lambda h(\hat{v})} \left[\frac{r h'(\hat{v})}{h(\hat{v})} - 2(\mu - \hat{\lambda} + r)\hat{\lambda}_\mu \right]. \quad (7)$$

The equilibrium arrival rate is therefore a monotonically increasing function of μ with asymptote $\lambda H(1)$. With an infinite service rate in the public hospital, the waiting time vanishes and everybody with a relative private benefit below 1 decides for a public treatment when falling ill. This upper bound means that the function is globally concave. From (7), it is clear that a non-increasing density function for the relative benefit of a private treatment is sufficient for $\hat{\lambda}(\lambda, \mu, r)$ to be strictly concave in μ , something which we assume from now on.

In the figure below, we have graphically derived $\hat{\lambda}$ as a function of μ . The function in quadrant II maps μ into $\lambda H(V(\mu, 0, r))$, the arrival rate if everybody else would go for a private treatment; it has an asymptote at $\lambda H(1)$. This function determines the intercept of the reaction function $\lambda H(V(\mu, \hat{\lambda}, r))$, which is mapped in quadrant I. The intersection with the 45° line provides us with the solution to (3). Finally, quadrant IV provides us with the relationship between the equilibrium arrival rate $\hat{\lambda}$, and the service rate μ in the public hospital; it has an asymptote at $\lambda H(1)$.

bonus β per accomplished treatment, its income is given by

$$y = f + \beta x_p.$$

By Burke's theorem (1956), the departure process of a queuing system with Poisson arrivals and a negative exponential service time distribution is also Poisson with mean departure rate equal to the mean arrival rate. Hence,

$$E(y) = f + \beta \widehat{\lambda}(\lambda, \mu, r), \quad (10)$$

$$\text{var}(y) = \beta^2 \widehat{\lambda}(\lambda, \mu, r). \quad (11)$$

where the second result stems identity of the mean and variance of a Poisson distributed variable.

Assuming a constant marginal cost c for the service rate, and mean variance preferences of the hospital manager, his decision making can be based on

$$U \stackrel{\text{def}}{=} f + \beta \left(1 - \frac{\rho}{2}\right) \widehat{\lambda}(\lambda, \mu, r) - c\mu, \quad (12)$$

where ρ can be interpreted as the manager's coefficient of absolute risk aversion. Thus by increasing the hospital's effort—its service rate μ —the manager can increase mean revenue, but at the same time he will increase the variance of this revenue. The effect on the certainty equivalent therefore depends on the size and sign of what we call the effective bonus rate $b(\beta, \rho) \stackrel{\text{def}}{=} \beta \left(1 - \frac{\rho}{2}\beta\right)$. The effective bonus rate $b(\cdot)$ is a \cap -shaped function of β , and is maximal when $\beta = \frac{1}{\rho}$, the manager's absolute risk tolerance. A nominal bonus beyond that reduces b ; a nominal bonus above $\frac{2}{\rho}$ destroys any incentive to provide effort.

The first order conditions for an optimal choice of effort, μ^* , are given by

$$\begin{aligned} b\widehat{\lambda}_\mu(\lambda, \mu^*, r) - c &\leq 0, \\ \mu^* \left(b\widehat{\lambda}_\mu(\lambda, \mu^*, r) - c \right) &= 0, \\ \mu^* &\geq 0. \end{aligned}$$

Using (6), a necessary and sufficient condition for $\mu^* > 0$ is that $b\frac{\lambda h(0)}{r+\lambda h(0)} > c$. The important implication is that this reduces the range for the nominal bonus down to $[\beta_{\min}, \beta_{\max}] \subset [0, \frac{2}{\rho}]$, where β_{\min} and β_{\max} are the lower and upper roots of $\beta \left(1 - \frac{\rho}{2}\beta\right) = c\frac{r+\lambda h(0)}{\lambda h(0)}$, respectively.⁶

⁶I.e., $\beta_{\min} \stackrel{\text{def}}{=} \frac{1}{\rho} \left(1 - \sqrt{1 - 2\rho c \frac{r+\lambda h(0)}{\lambda h(0)}}\right)$, and $\beta_{\max} \stackrel{\text{def}}{=} \frac{1}{\rho} \left(1 + \sqrt{1 - 2\rho c \frac{r+\lambda h(0)}{\lambda h(0)}}\right)$.

Result 1: The nominal bonus has to exceed $\beta_{\min} > 0$ in order to induce any effort. Neither a positive bonus below β_{\min} , nor a bonus larger than β_{\max} will trigger any effort. In particular, with a large risk aversion, no bonus exist that induces any effort. The reason that large bonus rates do not work is that they expose the hospital manager to too much risk.

4 Welfare analysis

We assume a utilitarian welfare function for citizens. All citizens who follow the decision rule to queue up in the public system when falling ill have an expected utility given by (2). This fraction of the population has size $H(V(\mu, \hat{\lambda}, r))$. Those citizens that commit to a private treatment have a utility that equals their personal τ . Hence, citizen welfare is given by

$$W = \lambda \left[V(\mu, \hat{\lambda}, r) H(V(\mu, \hat{\lambda}, r)) + \int_{V(\mu, \hat{\lambda}, r)}^{\infty} \tau dH(\tau) \right],$$

with

$$\begin{aligned} \frac{\partial W}{\partial \mu} &= \lambda V_{\mu}(\mu, \hat{\lambda}, r) H(V(\mu, \hat{\lambda}, r)) = \lambda \frac{r(1 - \hat{\lambda}_{\mu})}{(\mu - \hat{\lambda} + r)^2} H(V(\mu, \hat{\lambda}, r)), \\ \frac{\partial^2 W}{\partial \mu^2} &= \lambda V_{\mu\mu}(\mu, \hat{\lambda}, r) H(V(\mu, \hat{\lambda}, r)) + \lambda \left[V_{\mu}(\mu, \hat{\lambda}, r) \right]^2 h(V(\mu, \hat{\lambda}, r)). \end{aligned}$$

Since $\hat{\lambda}_{\mu} < 1$, citizen welfare is increasing in μ . It is therefore maximised by setting β equal to $\frac{1}{\rho}$, since this maximises the effective bonus rate b . Note that as $V(0, \hat{\lambda}(\lambda, 0, r) = 0$, and $H(0) = 0$, we have that

$$\frac{\partial W}{\partial \mu} \Big|_{\mu=0} = 0, \text{ and } \frac{\partial^2 W}{\partial \mu^2} \Big|_{\mu=0} \geq 0.$$

This means that a marginal increase in effort has no first order effect on citizen welfare when effort is zero to begin with, and has a positive second order effect to the extent that $h(0) > 0$.

Social welfare, SW , is defined as citizen welfare, W , plus the expected utility of the public hospital, U , minus the expected social cost of the generating the transfer to the public hospital:

$$SW \stackrel{\text{def}}{=} W + U - (1 + \phi)Ey,$$

where ϕ is the marginal cost of social funds. Using (12) and (3), social welfare can be rewritten as

$$SW = W - (1 + \phi) \left(\frac{\rho}{2} \beta^2 \widehat{\lambda}(\lambda, \mu, r) + c\mu \right) - \phi U,$$

and is maximised by setting the utility of the hospital at its reservation utility level, 0.

Suppose first that β equals β_{\min} or β_{\max} . Then $\mu^* = \widehat{\lambda} = 0$, and social welfare equals $\lambda E(\tau)$, the average citizen utility when everybody chooses a private health care provider.

Result 2. If β equals β_{\min} or β_{\max} , $SW = \lambda E(\tau)$.

Next, we inquire how SW locally behaves around the end points for the nominal bonus rate. Suppose first that $\beta = \beta_{\min}$. Then $\mu^* = 0$. The effect on social welfare of a marginal increase in the nominal bonus rate β , is then given by

$$\begin{aligned} \frac{\partial SW}{\partial \beta} \Big|_{\beta=\beta_{\min}} &= \frac{\partial W}{\partial \mu} \Big|_{\mu=0} \frac{\partial \mu^*}{\partial \beta} \Big|_{\beta=\beta_{\min}} \\ &\quad - (1 + \phi) \left(\rho \beta_{\min} \widehat{\lambda}(\lambda, 0, r) + \left[\frac{\rho}{2} \beta_{\min}^2 \widehat{\lambda}_{\mu}(\lambda, 0, r) + c \right] \frac{\partial \mu^*}{\partial \beta} \Big|_{\beta=\beta_{\min}} \right). \end{aligned}$$

Since $\frac{\partial W}{\partial \mu} \Big|_{\mu=0} = 0$, the first *rhs* term vanishes. We are thus left with the second term which (using (5) and the definition for β_{\min}) reduces to

$$\frac{\partial SW}{\partial \beta} \Big|_{\beta=\beta_{\min}} = -(1 + \phi) \beta_{\min} \widehat{\lambda}_{\mu}(\lambda, 0, r) \frac{\partial \mu^*}{\partial \beta} \Big|_{\beta=\beta_{\min}}.$$

The effect $\frac{\partial \mu^*}{\partial \beta}$ can be decomposed as $\frac{\partial \mu^*}{\partial b} \cdot \frac{\partial b}{\partial \beta}$. By the second order condition, the first component is strictly positive. The second component is $1 - \rho\beta$. Since $\beta_{\min} < \frac{1}{\rho}$, we can conclude that $\frac{\partial SW}{\partial \beta} \Big|_{\beta=\beta_{\min}} < 0$.

Result 3. Suppose that $\beta = \beta_{\min}$. Then marginally increasing the nominal bonus rate will increase the hospital service rate but decrease social welfare.

By the same token,

$$\frac{\partial SW}{\partial \beta} \Big|_{\beta=\beta_{\max}} = -(1 + \phi) \beta_{\max} \widehat{\lambda}_{\mu}(\lambda, 0, r) \frac{\partial \mu^*}{\partial \beta} \Big|_{\beta=\beta_{\max}}.$$

Since $\beta_{\max} > \frac{1}{\rho}$, $\frac{\partial SW}{\partial \beta}|_{\beta=\beta_{\max}} > 0$, so that social welfare will fall when lowering β below β_{\max} .

Result 4. Suppose that $\beta = \beta_{\max}$. Then marginally lowering the nominal bonus rate will lower the hospital service rate and decrease social welfare.

The intuition for these results is simple. If the service rate in the public hospital is zero, every citizen will commit to choosing a private treatment when falling ill. Inducing the hospital to exert effort will then have no welfare effect on citizens, while it increases the need for (distortionary) taxation.

Alternatively, assume that $\beta = \frac{1}{\rho}$. This nominal bonus rate provides the hospital with maximal incentives to exert effort, and will therefore minimise the expected waiting time. Let the corresponding service rate be denoted by μ_{\max} . The effect of a marginal increase in β on social welfare is now

$$\begin{aligned} \frac{\partial SW}{\partial \beta}|_{\beta=\frac{1}{\rho}} &= \frac{\partial W}{\partial \mu}|_{\mu=\mu_{\max}} \frac{\partial \mu^*}{\partial \beta}|_{\beta=\frac{1}{\rho}} \\ &\quad - (1 + \phi) \left(\widehat{\lambda}(\lambda, \mu_{\max}, r) + \left[\frac{\rho}{2} \beta_{\min}^2 \widehat{\lambda}_{\mu}(\lambda, 0, r) + c \right] \frac{\partial \mu^*}{\partial \beta}|_{\beta=\frac{1}{\rho}} \right) \end{aligned}$$

Since $\frac{\partial \mu^*}{\partial \beta}|_{\beta=\frac{1}{\rho}} = 0$, this effect reduces to $\frac{\partial SW}{\partial \beta}|_{\beta=\frac{1}{\rho}} = -(1 + \phi) \widehat{\lambda}(\lambda, \mu_{\max}, r) < 0$.

Result 5. The nominal bonus rate that maximises the hospital service rate and minimises the expected waiting time is not optimal. In particular, social welfare can be increased by choosing a smaller nominal bonus rate.

We have now shown that social welfare is below the level under a pure private system when β approaches β_{\min} from above or β_{\max} from below. We have also shown that the rate providing maximal incentives, $\frac{1}{\rho}$ cannot be a local optimum. To explore the global optimum, we provide a numerical example.

5 A numerical example

In this section, we assume that the relative benefit τ is uniform on the interval $[0, 1]$. In that case, an explicit solution for (4) is available, i.e.,

$$\widehat{\lambda}(\lambda, \mu, r) = \frac{1}{2} \left(\mu + r + \lambda - \sqrt{(\mu + r + \lambda)^2 - 4\lambda\mu} \right).$$

For an interior solution, the optimal service rate for the public hospital is given by

$$\mu^* = \mu(\lambda, r, b, c) = \frac{b - 2c}{\sqrt{c}} \sqrt{\frac{r\lambda}{b - c}} - r + \lambda, \quad (13)$$

and citizen welfare is given by

$$W = \frac{\lambda}{2} \left(1 + \left(\frac{\mu - \hat{\lambda}}{\mu - \hat{\lambda} + r} \right)^2 \right).$$

Furthermore, we assume that on average 10 people fall ill each period ($\lambda = 10$), that the time preference rate is 20% ($r = .2$), that the hospital's coefficient of absolute risk aversion is 2 ($\rho = 2$), and the marginal cost of providing a treatment is $\frac{1}{6}$ of benefit of an immediate public treatment ($c = \frac{1}{6}$). The shadow cost of public funds is set at 10% ($\phi = .1$).

The figure below displays $\hat{\lambda}(\lambda, \mu, r)$. It is strictly increasing and concave in μ with an asymptote at $\lambda H(1) = 10$.

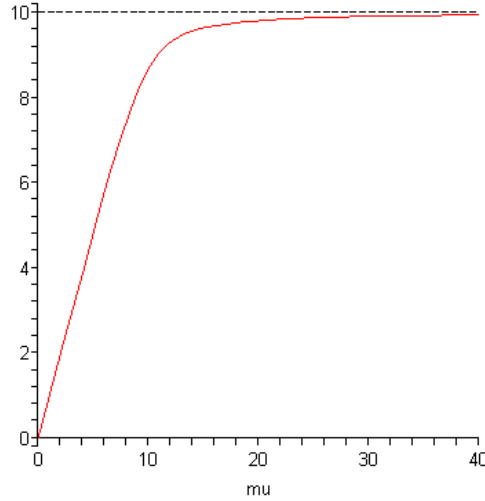


Figure 2. The function $\hat{\lambda}(10, \mu, .2)$.

The relevant range where the nominal bonus rate gives positive incentives is $[\cdot 217, \cdot 783]$. Outside this range, the public hospital provides zero effort and social welfare is equal to $\lambda E(\tau) = 10 \frac{1}{2} = 5$. The figure below depicts the level of social welfare as a function of the nominal bonus rate β when the hospital selects μ^* according to (13).

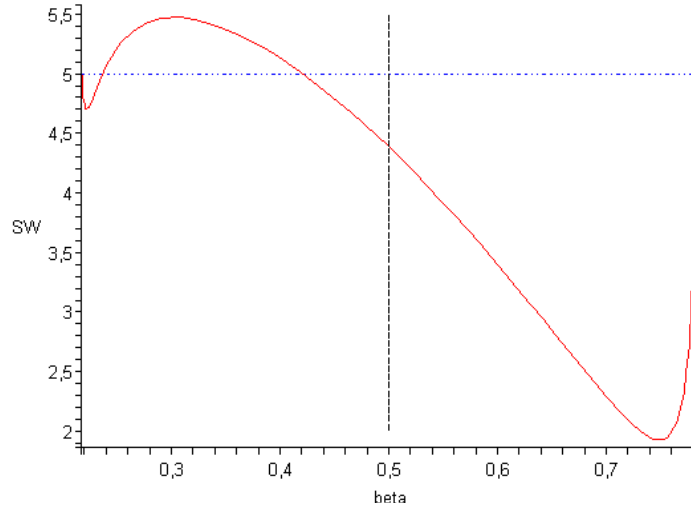


Figure 3. Social welfare as a function of β over the interval $[\beta_{\min}, \beta_{\max}]$.

The figure shows a non-monotonic behaviour for social welfare. The global maximum is at a bonus rate of 0.304 yielding a welfare level of 5.475. The bonus rate that maximises the service rate ($\frac{1}{2}$), and therefore minimises expected waiting time, yields a welfare level that is about 80% of the maximal welfare level. A sensitivity analysis shows that for marginal costs above 18.7% of the benefit of an immediate public treatment, social welfare will fall short of that with only a private sector.

6 Conclusion

In this paper, we studied the arrival process to a public hospital when illness incidence follows a Poisson process and patients can opt out of the public waiting list by choosing a private hospital. Under a commitment assumption, we showed that patients will queue up at the public hospital at a rate that is also Poisson distributed, with mean and variance determined by the service rate in the public hospital. When the public hospital is financed by a reimbursement scheme that depends linearly on completed treatments, we asked how the choice of the service rate is determined by the bonus parameter of this scheme. We showed that a too large bonus will trigger a lower effort level, the intuition being that a higher effort not only boosts the expected number of treatments, but also the variance of that number. For a given effort level will a higher bonus expose the hospital to a more risky income prospect. A risk averse hospital realises that it can offset this increased exposure to risk by choosing a lower effort level.

In the analysis above, it was assumed that citizens make an *ex ante* choice of provider when falling ill. In an earlier version of the paper, we made the more natural assumption that citizens, upon falling ill, observe the length of the queue in the public system, n , and then make a choice of provider. A consequence of this assumption is that the arrival rate to the public system will depend on n and that it is no longer possible to obtain simple formulae for the steady state probabilities. However, numerical simulations showed that with *ex post* decision making, the unconditional arrival process for the public system has a mean and variance that are both increasing in the service rate, μ .⁷ Therefore, the *ex ante* choice approach pursued in this paper preserves the important feature of a more realistic setting, i.e., that the variance of the hospital output depends positively on its effort level. And therefore that effort may be lowered in response to a too high bonus rate.

Marchand and Schroyen (2005) asked under which conditions a mixed health care system can be desirable on equity grounds. They considered a continuum of citizens, differing in labour earning ability, and a government that cares for redistribution. People are hit by illness with a certain probability, and choose whether to go to a public or private hospital. Treatment in the former is free of charge but a waiting time equilibrates demand with treatment capacity. The public hospital receives in a lump sum grant (no issues of moral hazard) which is raised through a linear income tax on citizens. Finally, waiting time reduces the time endowment of those individuals opting for a treatment in the public hospital. The authors find that the introduction of a small public health care sector is never welfare improving and that the welfare gains of a mixed system are relatively small compared to a first best redistribution policy. Our results share some of the qualitative features of Marchand and Schroyen (2005): (i) the result that incentivising the public hospital is welfare deteriorating when the activity level there is close to zero, and (ii) the non-monotone shape of social welfare as a function of the policy parameter, with the possibility of absence of an interior solution when the marginal cost gets too large. The second best nature of our model derives from the inability to control the hospital's effort directly. In Marchand and Schroyen (2005), it derives from the unobservability of earnings ability by the government.

In the present model, public and private hospitals provided the same kind of (elective) treatment. But usually, public hospitals are in addition assigned the responsibility to provide emergency care. Emergency care patients will

⁷Though mean and variance are no longer identical, except at an infinity service rate where everybody chooses the public hospital.

typically not make any choice of provider. When they arrive at the public hospital, they are given (preemptive) priority. This arrival of emergency cases to the public hospital will have implications for the steady state distribution of people in the elective line. Though the steady state distribution of the number of priority patients in the system is similar to (1), solving for the steady state distribution for elective patients is a complex matter. We leave it as a topic for future research what the implications are for the design of the reimbursement system for hospitals that are also responsible for emergency treatments.

References

- [1] Burke P (1956) "The output of a queuing system", *Operations Research* **4**, 699-704.
- [2] Fomundam S and J Herrmann (2007) "A survey of queuing theory applications in healthcare" ISR report 2007-24, University of Maryland.
- [3] Johansen L (1987) "Queues (and 'rent-seeking') as non-cooperative games, emphasizing mixed strategy solution", ch 43 in F Førsund (ed.) *Collective works of Leif Johansen, vol 2* (Amsterdam: North Holland)
- [4] Lindsay C and B Feigenbaum (1984) "Rationing by waiting lists", *American Economic Review* **74**, 404-417.
- [5] Marchand M and F Schroyen (2005) "Can a mixed health care system be desirable on equity grounds?" *Scandinavian Journal of Economics* **107**, 1-23
- [6] Taylor H M and S Karlin (1994) *An introduction to stochastic modelling* (3th edition) (New York: Academic Press).
- [7] Worthington D J (1987) "Queueing models for hospital waiting lists", *Journal of the Operational Research Society* **38**, 413-422.
- [8] Worthington D J (1991) "Hospital waiting list management models", *Journal of the Operational Research Society* **42**, 833-843.