THE UNIVERSITY of York

Discussion Papers in Economics

No. 2007/02

Optimal Waits and Charges in Health Insurance

by

Hugh Gravelle and Luigi Siciliani

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

Optimal waits and charges in health insurance

Hugh Gravelle* Luigi Siciliani†

13 October 2006

Abstract

Waiting times are commonly used in the health sector to ration demand. We show that when money charges (coinsurance rates) are optimally set and there are no redistributional considerations, it is never optimal to have a positive waiting time if the marginal cost of waiting is higher for patients with greater benefits from health care. Although waiting time provides an additional instrument to control demand it does not mitigate the conflict between efficient risk bearing and efficient consumption of health care.

Keywords: Waiting times; Rationing; Optimal pricing; Insurance
JEL classification codes: H21, H42, I11, I18

1 Introduction

Rationing of health care by waiting is common in many OECD countries. It is an apparently inefficient method of allocating resources: it imposes a cost on the consumer in the form of delayed health improvement or reduced productive capacity. But unlike rationing by price the cost to the consumer is not offset by any gain for the producer, or indeed anyone else. The costs imposed on patients are thus a deadweight loss (Barzel, 1974).

The usual welfare justification for rationing by waiting is that policy makers have distributional motives but face constraints on their tax and

*National Primary Care Research and Development Centre, Centre for Health Economics, University of York. Email: hg8@york.ac.uk. NPCRDC receives core funding from the Department of Health. The views expressed are those of the authors and not necessarily those of the DH.

†Department of Economics and Related Studies, University of York. Email: ls24@york.ac.uk

price instruments. Bucovetsky (1984) showed that providing rationed access to a good priced below cost pushes rich individuals into the private sector, thereby reducing the cost of financing the subsidised good consumed by the poor. Besley and Coate (1991) make a similar argument in which a good is produced at lower quality in the public sector than in the private sector. Hoel and Saether (2003) show that if distributional objectives are sufficiently strong, rationing by waiting can increase welfare even if governments can use non-linear tax schedules.

The welfare gains from rationing by waiting seem to be rather small (Marchand and Schroyan, 2004). In this paper we consider whether there is an additional welfare argument for rationing by waiting. The welfare justifications so far do not take full account of a salient feature of health care: the uncertainty of ill health and the role of insurance. Ideal insurance would have efficient risk bearing (equal marginal utility of income across health states) and efficient consumption of care (marginal value of care equal to its marginal cost). But insurers, whether public or private sector, can observe only expenditure, not the health state. Consequently health insurance, apart from a limited amount of "dread disease" cover, is insurance against expenditure, not health state, leading to ex-post moral hazard which insurers can attempt to mitigate by providing partial reimbursement of health care expenditures. This suggests that it may be possible to increase welfare by using an additional instrument - waiting time - to control utilisation, thereby improving the trade-off between efficient risk bearing and efficient use.

To focus on the potential role of waiting time as a means of mitigating ex post moral hazard we use a welfare model whose specification rules out previous rationales for rationing by waiting: individuals have identical preferences, income, and health risks and there is no private sector provision of health care. We show that in these circumstances a positive waiting time may increase welfare and we discuss the factors influencing the optimal combination of price and waiting. However, we also show that a necessary condition for an optimal positive waiting time is that the marginal cost of waiting is smaller for patients with higher benefit from treatment. This seems implausible. It implies that a patient whose treatment will produce a 1 QALY gain in health would be willing to pay more for a reduction in the waiting time from six months to one month than an otherwise identical patient with a QALY gain of 10. We conclude that a positive waiting time is unlikely to improve the trade-off between efficient risk bearing and efficient utilisation of health care.

Although made in the context of public-sector health-care system, the arguments carry over to the design of optimal insurance contracts by private insurers and HMOs. Insurers cannot offer a contract preferred by insureds

under which insureds are required to use particular providers with positive waiting times.

2 Model specification

The most salient form of rationing by waiting time is rationing by waiting list for elective care.[1] Individuals bear a cost in getting on the waiting list for treatment. In systems with gatekeeping general practitioners, patients first have to consult their general practitioner to get a referral and then incur further costs in attending hospital outpatient department to be seen by a specialist who will then place them on a waiting list. The longer the time potential patients know they will have to wait on the list, the less the discounted value of the treatment and the less likely are they to be willing to incur the initial costs of joining the list (Lindsay and Feigenbaum, 1984).

We assume individuals are ex ante identical. Their number is normalised to 1. There is a single treatment provided only in the public sector at a constant unit cost of c. Health care is financed by a lump-sum tax T and a charge p for treatment. Equivalently there is compulsory public health insurance with a premium of T and a coinsurance rate of p/c.

Ex post individuals are either ill (probability π) or well. If ill they either join the waiting list and are treated or decide not to join the list and are not treated. Individual illness probabilities are independent.

If ill and treated utility is $v_1(y_1(w,h) - T - p, h, w)$ where y_1 is income, w is the waiting time and h is the health gain from immediate treatment. $h \in (h_{\min}, h_{\max})$ has distribution function $F(h)$. We assume that $y_{1w} \leq 0$ and $y_{1h} \geq 0$ to allow for the possibilities that waiting for treatment may reduce income and that income increases with health. Utility is increasing in income and health benefit and non-increasing in waiting time: $v_{1w} \leq 0$. The specification allows waiting time to impose a cost on the patient because it reduces income, or because it reduces health gain when actually treated or because it has a direct utility cost.

Utility if well is $v_0(y_0 - T)$ and if ill and untreated is $v_2(y_2 - T)$ where y_0, y_2 are exogenous. We assume ill health does not increase income: $y_0 \geq y_1(0, h_{\max}) \geq y_2$.[2]

[1] In some systems there is rationing by waiting in line (queues) where the opportunity cost to individuals arises from the more valuable work or leisure uses of the time spent waiting. Such rationing can be used for minor ailments in hospital accident and emergency rooms and for general practitioner consultations.

[2] We could allow for endogenous labour supply to affect income y_s $(s = 0, 1, 2)$ but this merely complicates the exposition and makes no difference to the results since the

Individuals are averse to consumption or income risk ($v_{syy} < 0$, $s = 0, 1, 2$). The specification permits marginal utility of income to depend on whether the individual is ill and on the health gain and the waiting time.

When ill the utility gain from joining the waiting list, paying p for treatment and waiting w is[3]

$$\Delta(T, p, w, h) = v_1(y_1(w, h) - T - p, h, w) - v_2(y_2 - T) \tag{1}$$

Since $dv_1/dh = v_{1y}y_{1h} + v_{1h} > 0$ there exists a threshold health gain $\hat{h} = \hat{h}(T, p, w)$ such that individuals join the list and are treated if and only if $h \geq \hat{h}(T, p, w)$. We rule out uninteresting cases where either everyone is treated ($\hat{h} = h_{\min}$) or no one is ($\hat{h} = h_{\max}$) and so $\hat{h}(T, p, w, h)$ is defined by $\Delta(T, p, w, h) = 0$. Demand per ill individual is

$$D(T, p, w) = 1 - F(\hat{h}(T, p, w)) \tag{2}$$

with

$$D_T = -f(\hat{h})\hat{h}_T = -f(\hat{h})\left[\hat{v}_{1y} - v_{2y}\right]\left[d\hat{v}_1/dh\right]^{-1} \tag{3}$$

$$D_p = -f(\hat{h})\hat{h}_p = -f(\hat{h})\hat{v}_{1y}\left[d\hat{v}_1/dh\right]^{-1} < 0 \tag{4}$$

$$D_w = -f(\hat{h})\hat{h}_w = -f(\hat{h})\left[\hat{v}_{1y}y_{1w} + \hat{v}_{1w}\right]\left[d\hat{v}_1/dh\right]^{-1} < 0 \tag{5}$$

where the hat on $\hat{v}_{1y}, \hat{v}_{1w}, d\hat{v}_1/dh$ indicates that the derivatives are evaluated at $\hat{h}$.

The policy problem is to choose the charge p, the wait w, and the lump-sum tax or premium T to maximise

$$(1 - \pi)v_0(y_0 - T) + \pi\left\{\int^{\hat{h}} v_2\left(y_2 - T\right)dF + \int_{\hat{h}} v_1(y_1(w, b) - T - p, h, w)dF\right\} \tag{6}$$

subject to the budget constraint

$$T + \pi(p - c)D(w, p, T) \geq 0 \tag{7}$$

marginal effects of the insurance policy on labour supply are irrelevant by virtue of the envelope theorem. We could interpret v_s as an indirect utility functions in which we have suppressed the wage rate.

[3]We ignore the complications arising from the possibility that patients and their general practitioners are uncertain about the diagnosis and hence whether a specialist will recommend joining the waiting list and the fact there may be a wait to see the specialist. These make no essential difference to the properties of the rationing by waiting mechanism (Gravelle, Dusheiko and Sutton, 2002).

We do not restrict the sign of p and T: in addition to the conventional insurance contract ($p \in (0, c)$, $T > 0$) it is feasible to pay sick individuals to be treated ($p < 0, T > 0$) or to make a profit on the sale of health care ($p > c$) which can be paid to all individuals as a lump sum dividend ($T < 0$).

Crucially, it is not possible to make payments contingent on the states of the world (being well, being ill) of any individual. If it was the optimal policy would be to set the charge equal to marginal cost and to compensate sick individuals with a payment, whether or not they consume care, thereby decoupling insurance (the equalisation of marginal utilities across states of the world) from incentives (the efficient consumption of health care). By varying the charge and the premium it is only feasible to make transfers of income between the event "not ill, or ill and not treated" and the event "ill and treated".

We assume that preferences, income and medical technologies, and the distribution of health gain are such that the welfare function is concave in the instruments, and the feasible set convex.

3 Optimal insurance

Define L and λ respectively as the Lagrangian function and multiplier. The necessary and sufficient first order conditions are

$$L_T = -(1-\pi)v_{0y} - \pi\left[(1-D)v_{2y} + D\bar{v}_{1y}\right] + \lambda[1 + (p-c)\pi D_T] = 0 \quad (8)$$

$$L_p = -\pi D\bar{v}_{1y} + \lambda\left[\pi D + (p-c)\pi D_p\right] = 0 \quad (9)$$

$$L_w = \pi\int_{\hat{h}}\left(v_{1y}y_{1w} + v_{1w}\right)dF + \lambda\left[(p-c)\pi D_w\right] \leq 0, \quad w \geq 0, \quad L_w w = 0 \quad (10)$$

$$L_\lambda = T + \pi(p-c)D(w,p,T) \geq 0, \quad \lambda \geq 0, \quad L_\lambda\lambda = 0 \quad (11)$$

where $\bar{v}_{1y} = \int_{\hat{h}} v_{1y}dF/D = \bar{v}_{1y}(T,p,w)$ is expected utility if ill and treated. The budget constraint will bind and $\lambda > 0$ since the marginal utility of income is positive ($v_{sy} > 0, s = 0, 1, 2$).

We did not constrain the optimal policy to have $T > 0, p < c$. Suppose that there is no insurance ($p = c$, $T = 0$). Then, using (9),

$$\begin{aligned} L_T &= \lambda - (1-\pi)v_{0y} - \pi\left[Fv_{2y} + (1-F)\bar{v}_{1y}\right] \\ &= \bar{v}_{1y}[(1-\pi) + \pi F] - [(1-\pi)v_{0y} + \pi F v_{2y}] \end{aligned} \quad (12)$$

Hence a sufficient condition for the optimal policy to be a conventional insurance policy ($T > 0, p < c$) is that

$$\bar{v}_{1y}(0,c,w) > \frac{(1-\pi)v_{0y}(y_0) + \pi F\left(\hat{h}(0,c,w)\right)v_{2y}(y_2)}{(1-\pi) + \pi F\left(\hat{h}(0,c,w)\right)} \quad (13)$$

so that when there is no insurance the expected marginal utility of income in the event "ill and treated" is greater than average expected marginal utility of income from the event "not ill, or ill and not treated". We will assume that the condition (13) is satisfied.

One set of assumptions that ensures (13) is $v_0 = v(y_1)$, $v_1 = v(y_1 - T - p) + u(h, w)$, and $v_2 = v(y_2)$, with y_1 not affected by h or w and $y_1 \leq \min\{y_0, y_2\}$.

Since the optimal contract has $T > 0$, $p < c$ the marginal value of relaxing the budget constraint is

$$\lambda = \frac{Ev_{sy}}{1 + \pi(p - c)D_T} < \bar{v}_{1y} \tag{14}$$

i.e. the expected marginal utility of income (Ev_{sy}) adjusted for the feedback effect of additional income on the budget via demand. From (9) this is less than the expected marginal utility of income for treated patients.

We can use the first order conditions to describe the optimal contract

Proposition 1 *(a)The optimal charge satisfies*

$$\frac{c - p}{p} = \frac{(\lambda - \bar{v}_{1y})}{\lambda} \frac{1}{\varepsilon_p} \tag{15}$$

(b) If a positive wait for treatment is optimal then it satisfies

$$w = \lambda(c - p)\varepsilon_w \frac{1}{\kappa} \tag{16}$$

and (c) the optimal charge and waiting time satisfy

$$\frac{w}{p} = \frac{\varepsilon_w}{\varepsilon_p} \frac{(\lambda - \bar{v}_{1y})}{\kappa} \tag{17}$$

where $\varepsilon_p, \varepsilon_w$ are the demand elasticities with respect to money price and waiting time and $\kappa = \int_{\hat{h}} (v_{1y} y_{1w} + u_w)\, dF/D$ is the per patient expected marginal disutility from waiting.

The condition for the optimal charge (15) appears similar to the Boiteux-Ramsey pricing rule (Atkinson and Stiglitz, 1980) but differs in crucial respects (Besley, 1988). The Boiteux-Ramsey rule arises when it is not possible to use non-distorting lump sum taxes to finance public expenditure. As a consequence the shadow value of the public sector budget constraint (λ) exceeds the marginal utility of income of consumers. Hence $p > c$ and price is higher the less elastic is demand. Here there are lump-sum taxes and all individuals have the same income, but the shadow value of the budget

constraint is less than the marginal utility of income of treated individuals because moral hazard rules out full insurance which equalises the marginal utilities of individuals. The optimal price is less than marginal cost ($p < c$) and is *lower* the less elastic is demand.

If individuals were risk neutral with respect to income, the assumptions of feasible lump-sum taxes and identical individuals, would imply that the optimal price was equal to marginal cost and hence that the optimal wait was zero (see (10)). Thus risk aversion leads to an insurance scheme which subsidises care and may also ration care by waiting.

Raising the price and reducing the waiting time so as to leave supply unchanged transfers income from ill treated patients to the ill and not treated and to the well. If income has the same marginal value for all individuals the transfer would have no welfare consequence, since the reduction in the waiting time reduces a deadweight loss and welfare is increased. But this argument does not work because imperfect information prevents insurance equalising ex post marginal utilities of income. The marginal utility of income of patients paying a charge is greater than the marginal utility of the well and the ill but untreated. Hence the transfer of charge revenue from ill and treated patients to the well and to the ill but untreated has a net social cost. If this is greater than the deadweight loss imposed by waiting then it is optimal to have positive waiting times for care.

As (17) indicates, the wait will be higher relative to the charge the greater the difference between the marginal utility of income of patients ($\bar{v}_{1y}$) and the marginal utility of income of taxpayers as a whole (λ) because the larger is $\bar{v}_{1y} - \lambda$ the greater the utility cost of controlling demand by raising price. Waiting time will be higher relative to price the more elastic is demand with respect to waiting time and the less the elastic is demand with respect to price.

4 When is the optimal waiting time zero?

Substituting $D(\bar{v}_{1y} - \lambda)\pi/D_p = (p - c)\pi\lambda$ from (9) gives

$$L_w = \pi \int_{\hat{h}} (v_{1y} y_{1w} + v_{1w})\, dF + \pi D(\bar{v}_{1y} - \lambda)\frac{D_w}{D_p} \tag{18}$$

The first term is negative: increases in w make the treated patient worse off directly (v_{1w}) and via reduced income ($v_{1y}y_{1w}$). The second term is positive: increasing the waiting time permits a reduction in the price at the rate D_w/D_p with demand, and hence the cost of health care which must be covered by the premium, is unchanged. The net financial effect is then just the reduction in

the price. A £1 reduction in the price increases expected utility by $(\pi D\bar{v}_{1y})$ but tightens the budget constraint since charge revenue has fallen by πD. The reduction in charge revenue reduces expected utility at the rate $\pi D\lambda$ so that the transfer of income across individuals increases expected utility since $\bar{v}_{1y} > \lambda$.

This explains how a positive wait might increase welfare but it does not establish whether it does. Substituting in (18) for D_p and D_w from (4) and (5) and rearranging gives

$$\begin{aligned} L_w &= \pi \int_{\hat{h}} (v_{1y}y_{1w} + v_{1w})\, dF + \pi D(\lambda - \bar{v}_{1y}) \frac{(\hat{v}_{1y}\hat{y}_{1w} + \hat{v}_{1w})}{\hat{v}_{1y}} \\ &= \lambda\pi D \frac{(\hat{v}_{1y}\hat{y}_{1w} + \hat{v}_{1w})}{\hat{v}_{1y}} \\ &\quad + \pi \int_{\hat{h}} v_{1y} \left[\left(y_{1w} + \frac{v_{1w}}{v_{1y}} \right) - \left(\hat{y}_{1w} + \frac{\hat{v}_{1w}}{\hat{v}_{1y}} \right) \right] dF \end{aligned} \tag{19}$$

where, recall, the hats indicate that the derivatives are evaluated at the health gain threshold $\hat{h}$. Now

$$- \left(y_{1w}(w,h) + \frac{v_{1w}(y_1(w,h) - p - T, h, w)}{v_{1y}(y_1(w,h) - p - T, h, w)} \right) \tag{20}$$

is the marginal cost of waiting for a patient with a health gain of h. The first part of the marginal cost is the loss of income from a longer wait. The second part is the monetary value of the utility loss arising either because the longer wait reduces the final achieved health or because waiting is disliked per se. Thus

Proposition 2 *The optimal wait is zero if the marginal cost of waiting is non-decreasing in the health gain from immediate treatment*

$$\frac{d}{dh} \left[- \left(y_{1w} + \frac{v_{1w}}{v_{1y}} \right) \right] \geq 0 \tag{21}$$

We suggest that (21) is likely to be satisfied in most cases. Consider what would properties of the income function or preferences for it not to hold: either $y_{1wh} > 0$ or $d(v_{1w}/v_{1y})/dh > 0$. $y_{1wh} = y_{1hw} > 0$ means that the income gain from a treatment with a wait of one month is less than the income gain from a treatment with the same health effect with a wait of six months. There may be cases in which treatment at very short notice would disrupt income generating activity more than if the patient had longer to arrange her affairs but in most instances longer waits will not increase the

income gain from treatment. The implausibility of $d(v_{1w}/v_{1y})/dh > 0$ can be seen if we consider a simple case in which utility is additive separable: $v_1 = v(y_1 - T - p) + u(h, w)$. $d(v_{1w}/v_{1y})/dh > 0$ is then equivalent to $u_{hw} = u_{wh} > 0$: the utility gain from treatment increases the longer one waits or the marginal cost of waiting is higher for patients with greater benefits from health care.

5 Conclusion

Waiting times are used in many OECD countries to ration demand in health care (Siciliani and Hurst, 2005). Besley and Coate (1991), Bucovetsky (1984), Hoel and Saether (2003), and Marchand and Schroyan (2005) have provided a justification for a positive waiting time based on the desire to redistribute in the presence of distorting taxes. These arguments do not take account of the facts that individual expenditure on health care is uncertain and that because of information asymmetries neither public nor private insurers can provide insurance with both efficient risk bearing and efficient consumption. We have investigated whether positive waits can be justified as an additional means of controlling demand and thereby improving the trade-off between efficient risk bearing and efficient consumption of care.

We find that the plausible assumption that patient willingness to pay for a reduction in waiting time is non decreasing in the health gain from treatment, implies that in the absence of distributional considerations and if the price of care (or coinsurance rate) is set optimally, there is no welfare gain from having a positive waiting time. Marchand and Schroyan (2005) have shown that even when policy makers do wish to distribute the welfare gains from rationing by waiting, these are rather small. Coupled with our finding, this suggests that other normative rationales for the long waits observed in many countries are required, perhaps based on extra-welfarist objectives (Culyer, 1989).

References

Atkinson, A., Stiglitz, J. 1980. *Lectures on Public Economics*, McGraw Hill.

Barzel, Y. 1974. A theory of rationing. *Journal of Law and Economics*, 17, 73-95.

Besley, T. 1988. Optimal reimbursement health insurance and theory of Ramsey taxation. *Journal of Health Economics*, 7, 321-336.

Besley, T., Coate, S. 1991. Public provision of private goods and the redistribution of income. *American Economic Review*, 18(4), 979–984.

Bucovetsky, S. 1984. On the use of distributional waits. *Canadian Journal of Economics*, 17(4), 699-717.

Culyer, A. 1989. The normative economics of health finance and provision. *Oxford Review of Economic Policy*, 5(1), 34-58.

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

Hoel, M., Saether, E.M. 2003. Public health care with waiting time: the role of supplementary private health care. *Journal of Health Economics,* 22, 599–616.

Lindsay, C.M., Feigenbaum, B. 1984. Rationing by waiting lists. *American Economic Review,* 74(3), 404-417.

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

Siciliani, L., Hurst, J. 2005. Tackling excessive waiting times for elective surgery: a comparison of policies in twelve OECD countries. *Health Policy,*72, 201-215.