THE UNIVERSITY of York

Discussion Papers in Economics

No. 2007/15

Ramsey Waits: Allocating Public Health Service Resources when there is Rationing by Waiting

By

Hugh Gravelle and Luigi Siciliani

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

Ramsey waits: allocating public health service resources when there is rationing by waiting

Hugh Gravelle* Luigi Siciliani†

5 June 2007

Abstract

The optimal allocation of a public health care budget across treatments must take account of the way in which care is rationed within treatments since this will affect their marginal value. We investigate the optimal allocation rules for health care systems where user charges are fixed and care is rationed by waiting. The optimal waiting time is higher for treatments with demands more elastic to waiting time, higher costs, lower charges, smaller marginal welfare loss from waiting by treated patients, and smaller marginal welfare losses from underconsumption of care. The results hold for a wide range of welfarist and non-welfarist objective functions and for systems in which there is also a private health care sector. They imply that allocation rules based purely on cost effectiveness ratios are suboptimal because they assume that there is no rationing within treatments.

Keywords: Waiting times; Prioritisation; Rationing; Cost effectiveness ratios.

JEL numbers: H21, H42, I11, I18

*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, and Centre for Health Economics, University of York, Heslington, York YO10 5DD, UK; C.E.P.R., 90-98 Goswell Street, London EC1V 7DB, UK. Email: ls24@york.ac.uk.

1 Introduction

We investigate how a fixed health care budget should be allocated across treatments when patient charges are fixed and care is rationed by waiting.[1] Since an increase in the supply of a treatment reduces the time patients wait for it, optimal allocation of the budget across treatments is equivalent to determining the optimal waiting times for the different treatments.

Waiting times for elective surgery are used as a rationing mechanism in many countries with tax or public health insurance finance. Examples include Australia, Canada, Denmark, Finland, the Netherlands, Spain and the United Kingdom. Average waiting times for procedures, such as hip and knee replacement, cataract surgery, and varicose veins, of six months are not uncommon (Siciliani and Hurst, 2004). There is considerable variation in waiting times across treatment. Some examples from England in 2003/4 are 125 days for cataracts; 65 days for mouth or throat procedures category 2; 77 days for percutaneous transluminal coronary angioplasty; and 223 days for primary hip replacement.

The literature on the optimal allocation of the health care budget across treatments considers the mix of treatments which will maximise total health gain (Garber, 2000). It suggests that resources should be allocated to different treatments according to their cost effectiveness ratios (CERs): the ratio of cost to health benefit. The literature assumes that there is an exogenously determined number of patients for each treatment, all of whom have the same health gain from the treatment. Treatments are ranked by their CERs and treatments with the lowest CER are funded until the budget is exhausted. The CER allocation rule implies that there is no wait for treatment for funded treatments and an infinite wait for the unfunded.

The key assumption underlying the CER allocation rule is that marginal health benefit from a treatment is constant. But it is difficult to reconcile this assumption with two types of evidence. First, sub-group analyses in trials show that benefit typically varies across sub-groups receiving a treatment (Douglas, Buxton and O'Brien, 2003). Second, if all patients had the same gains from treatment, demand curves for health care would be horizontal up to the number of patients. But utilisation studies show that demand curves are negatively sloped with respect to time and money prices (Martin and Smith, 1999; Martin et al., 2007; Newhouse et al, 1993; Ringel et al, 2002).

Given the differences in health benefit for patients receiving a treatment, the assumption of constant marginal benefit is equivalent to implicitly as-

[1] Gravelle and Siciliani (2007) consider the optimal choice of patient charges and waiting times. They show that under plausible assumptions about patient preferences the optimal waiting time is zero if the welfare function is utilitarian.

suming that patients are drawn at random from those who could benefit. But there are almost no examples of random rationing in public health care systems. Publicly funded health care is generally rationed by money or waiting time prices. Such rationing implies that treated patients have a higher perceived benefit than those not treated and that the health gain of the marginal patient falls as supply is increased and the money or waiting time price is reduced. Thus CER based rules for allocating budgets across treatments rest on an incorrect specification of the way the supply of a given treatment is rationed. Even if one accepts the value judgements underlying them, CER based allocation rules require modification because the rationing mechanism within treatments affects the marginal value of treatments.

Only one paper on priority setting across treatments has taken account of the way in which patients are rationed within treatments.[2] Smith (2005) considers a health care system which is financed by a mixture of an exogenous budget and user charge revenue. The volumes of different treatments are determined by the patient demands at the user charges set by the health planner. It is assumed that all patients have same health benefit from a given treatment but, because patients have different incomes and utility of income is concave, increases in the user charge reduce demand by discouraging poorer potential patients. The optimal prices which maximise total health gain are determined by the exogenous CERs and by the money price elasticity of demands. Treatments with a high CER are not provided, those with a low CER are provided at no charge, and those with intermediate CERs are provided at a below cost price. In contrast to the usual Ramsey public sector pricing rules (Ramsey, 1927), charges are higher (and the volume of treatment smaller) the more elastic is demand with respect to the charge.

The result is similar to that derived by Besley (1988) for optimal health insurance contracts. In the usual Ramsey case prices are set to raise a given sum in revenue and so must exceed marginal cost. In the case of health care, either because of risk aversion (Besley, 1988) or because of an extra welfarist

[2]Hoel (2007) shows that the simple CER based allocation rules require modification when some patients will buy the treatment in the private sector if it is not provided in the public sector. Consequently not all treatments provided in the public sector have lower CERs than those not provided. For example, a treatment for which there is a private sector alternative may not be provided in the public sector even though it has a lower CER than a treatment which is provided but for which there is no private sector alternative. The intuition is that when a private sector alternative is available the gain from public provision is the cost saving to the patients who would have bought it in the private sector. This cost saving is less than the utility gain from treatment compared with no treatment. Thus the average welfare gain from a treatment is lower if it has a private sector alternative. Hoel (2007) does however assume that if a treatment is provided it is provided to all patients. Hence he does not need to consider how care is actually rationed within treatments.

desire to increase consumption of health care (Smith, 2005), the aim is to spend a fixed sum and so prices are less than marginal cost. But, whether the constraint is to raise a given revenue or to spend a given budget, there is less welfare loss in departing from marginal cost prices the less elastic is demand. In the standard Ramsey revenue raising case, prices are raised further above marginal cost for goods with less elastic demands. In the budget spending case of health care, prices are lowered further below marginal cost the less elastic is demand. Hence prices (or coinsurance rates) are higher for goods with more elastic demands.

In this paper we examine allocation across treatments for a heath care system where, in contrast to Smith (2005), waiting time rather than money price rations demand. We also allow for the possibility that patient health benefit from a treatment varies across patients and we consider a more general welfare function which includes health gain maximisation as a special case.

Unlike a money price, a waiting time price imposes a pure deadweight loss: it is a cost to patients which is not offset by any gain to the providers. Longer waiting times reduce the value of treatment because of lost expected benefit, temporary discomfort and pain, and, for some pathologies, the higher risk of permanent reductions in health (see Hoel and Saether (2003) for some detailed examples of the cost of waiting).

The change in waiting time costs incurred by patients when supply of a treatment is reduced is determined by the increase in the wait and the reduction in the number of patients who wait. The reduction in the number waiting is equal to the reduction in the supply of treatment. A given reduction in the supply of treatment will generate a smaller increase in the waiting time the more elastic the demand with respect to waiting time. Thus the more elastic the demand with respect to waiting time, the smaller the increase in waiting time costs when supply is reduced and the waiting time increased. Hence, when there is rationing by waiting within treatments, the optimal allocation of resources across treatments should result in longer waiting times for treatments where demand is more elastic with respect to waiting time. Waiting times should also be longer for treatments where treated patients suffer less from increased waiting times.

The above argument holds whether or not the welfare function is paternalistic. If the welfare function does not respect individual choices between treatment and no treatment there is a further factor influencing the waiting time: the welfare loss associated with indifferent patients displaced as waiting time increases. With a paternalistic welfare function patients may place too low a value on treatment relative to no treatment, so that there is a welfare loss associated with the marginal patient's decision not to be treated. We show that waiting times should be lower, ceteris paribus, for treatments

where there is a larger welfare loss for the marginal patient.

Section 2 sets out the model. Section 3 derives and discusses the main results for allocation rules between treatments. Section 4 extends the analysis by allowing patients to opt for a private sector and shows that the existance of a private sector makes no essential difference to the results. Section 5 concludes.

2 Model specification

2.1 Preferences, demand, and waiting times

All health care is produced in the public health care system. Demand for treatment is rationed by waiting. To reduce notational clutter we assume there are no user charges. We show in section 3.3 that allowing for positive fixed user charges makes no essential difference to the results.

Individuals have the same preferences but differ ex ante in having different incomes $y \in [y^{\min}, y^{\max}]$ and ex post in having different health. Allowing for difference in preferences merely complicates analysis and does not alter the results. With probability π_i individuals are ill with condition i and will benefit from treatment i. No individual has more than one condition. If ill with condition i and not treated an individual has utility $v_i^{NT}(y)$, $v_{iy}^{NT} > 0$.[3] Treated patients have a wait of w_i before receiving one unit of treatment which produces a benefit (health gain) $b \in [b_i^{\min}, b_i^{\max}]$. Utility if treated is $v_i^T(y, b, w_i)$, which is increasing in income and health gain and decreasing in the wait: $v_{iy}^T > 0$, $v_{ib}^T > 0$, $v_{iw}^T < 0$.[4]

Health benefit and income for patients with condition i is distributed with density $f_i(b, y)$ and distribution function $F_i(b, y)$.The marginal distribution function for income is $F_y(y)$. The total population is normalised to 1. The planner knows the distribution functions but cannot prioritise individual patients on the basis of their health gain or their income.

[3]The formulation allows for the possibility that receiving treatment can increase income. Let y be income if treated, assume condition i reduces income if not treated by $L_i(y)$, and denote utility if not treated as $\hat{v}_i^{NT}(y - L_i(y))$. Then we can write $v_i^{NT}(y) = \hat{v}_i^{NT}(y - L_i(y))$.

[4]We assume that waiting times vary across treatments but are constant within treatments. Gravelle and Siciliani (2006) analyse the effect of waiting-time prioritisation within a treatment when benefit is partially observable through a continous variable (like age). They find that prioritisation within treatments can increase welfare but has ambiguous effects on the marginal value of treatment. Thus its effect on allocations across treatments is also ambiguous.

The key assumption is that increases in waiting time reduce the utility from treatment compared with the no treatment alternative. The most salient form of rationing by waiting time is rationing by waiting list for elective care. 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; Martin and Smith, 1999; Farnworth, 2003). In some systems there is rationing by waiting in line (queues). Waiting for treatment has an opportunity cost of forgone work or leisure time, as well as possible effects on the health gain. Rationing by waiting line can be used for minor ailments in hospital accident and emergency rooms and for general practitioner consultations. Our specification encompasses both rationing by waiting list and by waiting line and does not restrict the way in which longer waits reduce the utility of treatment relative to no treatment (Hoel and Saether, 2003).[5] The empirical evidence shows that increases in waiting time reduce demand for health care (Gravelle, Smith and Xavier, 2003; Martin and Smith, 1999; Martin et al, 2007).

An individual with illness i demands treatment i if and only if

$$v_i^T(y, b, w_i) - v_i^{NT}(y) \geq 0 \Longleftrightarrow b \geq \hat{b}_i(w_i, y) \tag{1}$$

where $\hat{b}_i(w_i, y)$ is the threshold benefit level such that all those with a smaller benefit do not seek treatment i. The threshold is increasing in the waiting time since $\hat{b}_{iw}(w_i, y) = -v_{iw}^T / v_{ib}^T > 0$.

The effect of income on the benefit threshold is

$$\hat{b}_{iy}(w_i, y) = -\frac{v_{iy}^T - v_{iy}^{NT}}{v_{ib}^T} \tag{2}$$

which may negative or positive depending on the effect of illness and treatment on the marginal utility of income.[6] If income is additively separable

[5] Our specification includes the original Lindsay and Feigenbaum (1984) formulation of the model of rationing by waiting if we write $v_i^T = v_i(y) + be^{-rw} - a_i$ and $v_i^{NT} = v_i(y)$, where a_i is the cost of getting on the list.

[6] Empirical studies suggest that, controlling for a wide range of morbidity and other socio-economic characteristics, income has little impact on the number of GP visits but is positively associated with specialist visits and hospital stays (van Doorslaer et al, 2004; Morris et al, 2005).

from health and waiting time, the income sub-utility function is unaffected by treatment, ill health has the same effect on the income of the treated and untreated, then (2) is zero and those on higher incomes demand care as much as individuals with low income.

Demand for treatment i is

$$\pi_i D_i(w_i) = \pi_i \int \int_{\hat{b}_i(w_i,y)} f_i(b,y) db dy \tag{3}$$

and

$$D_{iw}(w_i) = -\int \hat{b}_{iw}(w_i,y) f_i(\hat{b}_i(w_i,y),y) dy < 0 \tag{4}$$

where D_i is the per capita demand for treatment i from individuals with illness i. The supply of treatment i is z_i. The waiting time for treatment i is determined by the market clearing condition

$$\pi_i D_i(w_i) - z_i \leq 0, \qquad w_i \geq 0, \qquad w_i[\pi_i D_i(w_i) - z_i] = 0 \tag{5}$$

The equilibrium waiting time, when positive, is determined by (5) as

$$w_i = w_i(z_i, \pi_i), \qquad w_{iz} = 1/(\pi_i D_{iw}) < 0 \tag{6}$$

Resource allocation decisions result in three categories of treatment. Some treatments are not provided by the health service, which is equivalent to setting a sufficiently high waiting time ($w_i \geq w_i^o$, $\pi_i D_i(w_i^o) = 0$) to drive demand to zero. Treatments which are provided are either emergencies, where the level of supply is such that there is no waiting time ($\pi_i D_i(0) = z_i > 0$), or electives, where there is a positive wait ($w_i > 0$, $\pi_i D_i(w_i) = z_i > 0$).

2.2 Welfare

Welfare for an individual with income y is

$$s(y, \mathbf{w}) = \sum\nolimits_i \pi_i \left[\int^{\hat{b}_i(w_i,y)} s_i^{NT}(y) f_i(b,y) db + \int_{\hat{b}_i(w_i,y)} s_i^T(y,b,w_i) f_i(b,y) db \right] \tag{7}$$

where s_i^{NT}, s_i^T are welfare if the individual has condition i and is not treated or treated and $\mathbf{w}$ is the vector of waiting times. The social welfare function is

$$S(\mathbf{w}) = \int s(y; \mathbf{w}) dF_y(y) \tag{8}$$

The welfare formulation is quite general and is compatible with utilitarianism, more complicated Bergsonian individualistic welfare functions, and with extra welfarist (Culyer, 1991) value judgements, including simple health gain maximisation as in the allocation literature. The only restriction we impose is that welfare for a treated individual is increasing in health gain and decreasing in the waiting time: $s^T_{ib} > 0$, $s^T_{iw} < 0$.

The welfare maximising benefit threshold b^S_i for consumption of health care by individuals with income y is defined by

$$s^T_i(y, b, w_i) - s^{NT}_i(y) \geq 0 \Longleftrightarrow b \geq b^S_i(w_i, y) \tag{9}$$

Individuals may choose to consume too much ($b^S_i > \hat{b}_i$) or too little ($b^S_i < \hat{b}_i$) care. It is not possible to control use directly, so that it is determined by $\hat{b}_i(w_i, y)$, not by $b^S_i(w_i, y)$. But $\hat{b}_i(w_i, y)$ and $b^S_i(w_i, y)$ vary with waiting time, so that the welfare maximising allocation may be influenced by the effect of waiting time on the discrepancy between welfare maximising and actual benefit thresholds.

In the utilitarian welfare function $s^T_i = v^T_i$, $s^{NT}_i = v^{NT}_i$, so that individuals' decisions on health care consumption are respected in the sense that $\hat{b}_i(w_i, y) = b^S_i(w_i, y)$. Even if it was possible to directly alter the benefit thresholds (and hence demand) at given waiting times, there would be no welfare gain from doing so.

Policy in health care markets often reflects paternalistic value judgements which imply that individuals consume too little or too much health care (Musgrave, 1959; Sandmo, 1983). A common example is the belief that the use of health care should depend only on "need", defined as capacity to benefit from health care b, and not on characteristics such as income. We can capture this notion, which is closely related to horizontal equity, with a welfare function in which

$$s^{NT}_i(y) = v^{NT}_i(y^o), \quad s^T_i(y, b, w_i) = v^T_i(y^o, b, w_i) \tag{10}$$

The welfare maximising treatment threshold defined by (9) is $b^S_i(w_i, y) = \hat{b}_i(w_i, y^o)$ which is the same for individuals irrespective of their income level (and any other characteristics deemed irrelevant in assessing need). y^o is the reference income required to generate the welfare maximising treatment threshold given the waiting time. Individual decisions on consuming health care will not be welfare maximising except for those with $y = y^o$. If $\hat{b}_{iy}$ is negative, expected use will be too high for those with $y > y^o$ and too low for those with $y < y^o$.

The literature on CER allocation rules embodies the extra welfarist value judgement that the aim of the health care system is to maximise the total

health gain of the population (Garber 2000; Smith 2005). We could capture this value judgement by writing welfare as discounted health gain

$$s_i^T(y, b, w_i) = \delta_i(w_i)b, \qquad s_i^{NT}(y) = 0 \tag{11}$$

where b is the health gain from health care with no waiting, and $\delta_i(w_i)$ is a discount factor satisfying

$$\delta_i(w_i) \in (0, 1], \qquad \delta_i(0) = 1, \qquad r_i(w_i) \equiv -\delta_{iw}(w_i)/\delta_i(w_i) > 0 \tag{12}$$

with r_i being the proportionate rate of decrease of the discount factor.

3 Priority setting between treatments

3.1 Optimal waiting times

Decision makers in the public health care service allocate resources amongst different treatments subject to an exogenously determined health service budget constraint:

$$M - \sum_i \pi_i c_i D_i(w_i) \geq 0 \tag{13}$$

where c_i is the constant average and marginal cost of treatment i, and M the fixed health service budget. Health service decision makers allocate treatment (z_i), or equivalently choose waiting times, to maximise the social welfare function. The cost of meeting all demand generated by zero waiting times exceeds the budget: $\sum_i \pi_i c_i D_i(0) > M$, so that the budget constraint binds.

The Lagrangean is

$$L = \int s(y; \mathbf{w}) dF_y(y) + \lambda \left[M - \sum_i \pi_i c_i D_i(w_i) \right] \tag{14}$$

and the optimal allocation satisfies

$$\begin{aligned} \partial L/\partial w_i = \pi_i \int \left(s_i^{NT} - s_i^T\right) \hat{b}_{iw}(w_i, y) f_i(\hat{b}(w_i, y), y) dy \\ + \pi_i \int \int_{\hat{b}_i} s_{iw}^T(y, b, w_i) f_i(b, y) db dy - \lambda \pi_i c_i D_{iw} \leq 0, w_i \geq 0 \end{aligned} \tag{15}$$

with complementary slackness, for all i.[7]

[7] Although there always exists a vector of waiting times such that the constraint is satisfied as a strict inequality, sufficiency of the first order conditions requires restrictions on the second derivatives of the social welfare function, utility functions and the distribution functions.

Increasing w_i has two effects on welfare. First, increasing the waiting time reduces welfare from treatment. The marginal welfare cost of waiting per treated patient is

$$\kappa_i = -\int\int_{\hat{b}_i} s_{iw}^T(y, b_i, w_i) f_i db dy / D_i > 0 \tag{16}$$

Second, there is a welfare loss arising from the patients who decide not to seek treatment when the waiting time increases. The marginal threshold cost per treated patient is

$$\psi_i = \int (s_i^T - s_i^{NT}) \hat{b}_{iw}(w_i, y) f_i(\hat{b}_i(w_i, y), y) dy / D_i \tag{17}$$

$s_i^T(y\ , \hat{b}_i(w_i, y), w_i)\ - s_i^{NT}(y)$ is the social welfare reduction when the marginal patient with income y fails to seek treatment. When the waiting time increases, the benefit threshold at which patients seek treatment is increased ($\hat{b}_{iw} > 0$) and $s_i^T - s_i^{NT}$ is forgone.

The marginal patient has zero personal benefit from treatment ($v_i^T(y, \hat{b}_i(w_i, y),\ w_i) = v_i^{NT}(y)$) so that, if the welfare function respects individual relative valuations of treatment and no treatment, $\psi_i = 0$. If all individuals place too low a value on treatment relative to no treatment then $s_i^T(y, \hat{b}_i(w_i, y), w_i) > s_i^{NT}(y)$ at all income levels and $\psi_i > 0$. For example, suppose that we care about need and the social welfare function is given by (10) and that the reference income is $y_i^o = y^{\max}$. Then all individuals ought to consume the same expected amount of care as the richest individual. If $v_{iy}^T > v_{iy}^{NT}$, so that the probability of consumption care when ill increases with income, then all individuals consume too little care and $\psi_i > 0$. With a smaller reference income some individuals have too high an expected consumption and some too low, so that it is possible that $\psi_i < 0$.

We can use (16) and (17) to write (15) as

$$-(\psi_i + \kappa_i) - \lambda c_i \varepsilon_{iw} w_i^{-1} \leq 0, \quad w_i \geq 0 \tag{18}$$

where $\varepsilon_{iw} = D_{iw} w_i / D_i < 0$ is the elasticity of demand for treatment i with respect to the waiting time.

Proposition 1 *The optimal allocation has three types of treatment: (i) treatment i is not provided: $w_i \geq w_i^o$, where $D_i(w_i^o) = 0$, and*

$$\lim_{w_i \to w_i^o} -(\psi_i + \kappa_i) - \lambda c_i \varepsilon_{iw} w_i^{-1} \geq 0 \tag{19}$$

(ii) treatment i is an emergency treatment: $w_i = 0$, and

$$\lim_{w_i \to 0} -(\psi_i + \kappa_i) - \lambda c_i \varepsilon_{iw} w_i^{-1} \leq 0 \tag{20}$$

(iii) treatment i is an elective treatment: $w_i \in (0, w_i^o)$ and

$$w_i = \frac{-\lambda c_i}{\kappa_i + \psi_i} \varepsilon_{iw} \tag{21}$$

The optimal waiting time for elective care is higher the greater the cost of treatment and lower the greater the marginal cost of waiting (κ_i) and the marginal threshold welfare cost (ψ_i) from patients displaced by higher waiting times.

Waiting time should be higher if demand is more elastic with respect to waiting time. A given reduction in provision will generate a smaller increase in waiting time the more elastic the demand. Waiting time is a deadweight loss: none of additional cost to consumers of having to wait longer accrues as a gain to anyone else (by contrast user charges paid by patients accrue to funders). Hence we should be more willing to reduce supply for treatments with more elastic demands.

3.2 Allocation in terms of cost effectiveness ratios

Suppose that welfare from treatment is additively separable in income and multiplicatively separable in the benefit and waiting time as:

$$s_i^T(y, b, w_i) = \tilde{s}_i^T(y) + \delta_i(w) b \tag{22}$$

The marginal cost of waiting per treated patient is

$$\begin{aligned} \kappa_i &= -\int \int_{\hat{b}_i} \delta_{iw}(w_i) b f_i db dy / D_i = r_i(w) \delta_i(w_i) \int \int_{\hat{b}_i} b f_i db dy / D_i \\ &= r_i(w) \delta_i(w_i) B_i(w_i) \end{aligned} \tag{23}$$

where $r_i(w_i) = -\delta_{iw}(w_i)/\delta_i(w_i)$ is the proportionate rate of decline in the discount factor δ_i and $B_i(w_i)$ is the undiscounted health gain per treated patient. Note that B_i depends on the wait because w_i affects the number of patients being treated through $\hat{b}_i$.

The marginal threshold welfare cost of displaced patients per treated patient is

$$\psi_i = \int \left[\tilde{s}_i^T(y) + \delta_i(w) \hat{b}_i - s_i^{NT}(y) \right] \hat{b}_{iw} f_i(\hat{b}_i(y, w_i)) dy / D_i \tag{24}$$

When $\tilde{s}_i^T(y) = s_i^{NT}(y)$, the marginal threshold cost is positive because the welfare function takes account only of discounted health gains from treatment and ignores the factors which influence patient decisions on whether to seek treatment.

An increase in expenditure on treatment of £1 increases supply of the treatment by $1/c_i$. A unit increase in supply reduces the waiting time by $1/\pi_i D_{iw}$. A unit reduction in waiting time reduces waiting time costs at the rate $\kappa_i \pi_i D_i = r_i \delta_i B_i \pi_i D_i$ and welfare arising from additional patients treated increases at the rate $\psi_i \pi_i D_i$. Thus, when the welfare function has (22), the marginal social value in terms of health gain from an additional £1 spent on treatment i is

$$\frac{1}{c_i} \frac{-1}{\pi_i D_{iw}} \left[\psi_i \pi_i D_i + \kappa_i \pi_i D_i\right] = \frac{-D_i}{D_{iw}} \left(\frac{\psi_i}{c_i} + r_i \frac{\delta_i B_i}{c_i}\right) \tag{25}$$

Note that the marginal value of expenditure depends on the benefit-cost ratio B_i/c_i which is the inverse of the CER.

We can restate the conditions describing an optimal allocation of the budget across treatments in terms of the marginal value of expenditure on treatment i (25) and the shadow value of the budget constraint (λ):

Proposition 2 *If welfare from treatment is additive separable in income and multiplicatively separable in waiting time and benefit then at the optimal allocation*
(i) if treatment i is not provided then

$$\lim_{w_i \to w_i^o} \left[\frac{-D_i}{D_{iw}} \left(\frac{\psi_i}{c_i} + r_i \frac{\delta_i B_i}{c_i}\right)\right] \leq \lambda \tag{26}$$

where w_i^o is the waiting time at which demand is zero ($D_i(0) = 0$)
(ii) if treatment i is provided with zero wait then

$$\lim_{w_i \to 0} \left[\frac{-D_i}{D_{iw}} \left(\frac{\psi_i}{c_i} + r_i \frac{\delta_i B_i}{c_i}\right)\right] \geq \lambda \tag{27}$$

(iii) if treatment i is provided with a positive wait then

$$\frac{-D_i}{D_{iw}} \left(\frac{\psi_i}{c_i} + r_i \frac{\delta_i B_i}{c_i}\right) = \lambda \tag{28}$$

A pure CER rule would provide treatment i if and only if $B_i(0)/c_i$ exceeded the shadow marginal value of the health sector budget λ. Although the optimal allocation depends on the cost-effectiveness ratios for different

treatments, the CER based rule of allocating resources in reverse order of cost effectiveness is not optimal. Since patients are not drawn at random from the pool of potential patients the undiscounted health benefit per treated patient ($B_i(w_i)$) declines with the number of patients treated. Thus the CER for a treatment is determined by the allocation decision rather than determining it.

Suppose that all patients would have the same benefit b_i from a treatment, so that undiscounted health gain per treated patient $B_i = b_i$ is constant and does not depend on the amount of treatment i. Suppose that $v_i^T = u_i^T(y) + \delta_i(w_i)b_i - a_i$ where a_i is an access cost of getting treatment. If there exists an income $\hat{y}_i \in (y^{\min}\ y^{\max})$ such that $v_i^T(\hat{y}_i, b_i, w_i) = v_i^{NT}(\hat{y}_i)$ and if $v_{iy}^T - v_{iy}^{NT}$ is positive, then demand for treatment i is decreasing in the waiting time. Then even when the constant b_i replaces $B_i(w_i)$ in Proposition 2, the decision to provide treatments is not based solely on their benefit cost ratios. Although a higher benefit cost ratio makes it more likely that a treatment will be provided it is also necessary to take account of the responsiveness of demand to waiting time, the marginal threshold welfare loss and the rate at which discounted health benefits decline with the waiting time. A higher discount rate r_i makes it more likely that a treatment will be supplied since the cost of smaller supply (implying a longer wait) are greater. A higher threshold cost for patients displaced by a smaller supply will also increase supply. Finally, the more responsive demand is to waiting time the smaller will be the supply and the higher the waiting time.

3.3 User charges

The assumption that user charges were zero was made to simplify the exposition and notation. Allowing for positive fixed user charges makes very little difference. Thus suppose that treatment i carries a charge of $p_i < c_i$. Then the threshold benefit level is $\hat{b}_i(y, p_i, w_i)$ determined by $v_i^T(y - p_i, b, w_i) = v_i^{NT}(y)$ and the demand functions also depend on the charge $D_i(w_i, p_i)$. After inserting $\hat{b}_i(y, p_i, w_i)$ and $D_i(w_i, p_i)$ in the welfare function (8) and the budget constraint (13), the resulting first order conditions differ only in having $c_i - p_i$ replace c_i. The only change to Proposition 1 required is that production cost c_i is replaced with the net financial cost $c_i - p_i$ to the public sector the net financial cost to the public sector.

With the social welfare function satisfying (22), the value of an additional £1 spent on treatment i (25) becomes

$$\frac{-D_i}{D_{iw}}\left(\frac{\psi_i}{c_i} + r_i\frac{\delta_i B_i}{c_i}\right)\left(\frac{1}{1-\theta_i}\right) \tag{29}$$

where $\theta_i = p_i/c_i$ is the copayment rate: the proportion of the unit cost of treatment recovered by a charge to the patient. Proposition 2 now holds with (29) replacing (25). A treatment is more likely to be provided, or have a larger supply, if it has a higher exogenous copayment rate.

4 Public sector allocation in the presence of a private sector

In many countries with public health care systems patients also have the option of buying treatment from private sector providers. We now show that the existance of the private sector affects the form of the demand functions for public care but does not alter our main conclusions about the factors determining the optimal allocation of resources across treatments in the public sector.

Individuals with condition i can obtain private sector treatment with no wait at price of m_i or public sector treatment at no charge but after a wait of w_i. We assume that utility is separable in income and health, and concave in income, to simplify the derivation of the results, though they do not depend on the assumption.

Public treatment yields utility $v_i^{GT} = v_i^T(y, b, w_i) = \upsilon(y) + u_i(b, w_i) - a_i$. Utility from private treatment is $v_i^{PT} = v_i^T(y - m_i, b, 0) = \upsilon(y - m_i) + u_i(b, 0) - a_i$. We assume that marginal utility from health gain is higher if the wait is smaller. Utility from no treatment is $v_i^{NT}(y) = \upsilon(y)$.

Patients prefer public to no treatment if $v_i^{GT} > v_i^{NT}$. The benefit threshold above which patients prefer public treatment to no treatment is $\hat{b}_i^{GN}(w_i)$, defined by $u_i(b, w_i) - a_i = 0$.

Patients prefer public to private treatment if $v_i^{GT}(y, b, w_i) > v_i^{PT}(y - m_i, b, 0)$. $b_i^{GP}(y; w_i, m_i)$ is the benefit threshold below which patients prefer public treatment to private treatment. It is decreasing in waiting time and income.

Patients prefer private to no treatment if $v_i^T(y - m_i, b, 0) > v_i^{NT}(y)$. The benefit threshold above which patients prefer private treatment to no treatment is $b_i^{PN}(y, m_i)$. The threshold is decreasing in income.

Figure 1 shows the choices of different types of patients. The locus where patients are indifferent between private and public treatment (b_i^{GP}) and the locus where they are indifferent between private and no treatment (b_i^{PN}) are downward sloping in (b, y) space. The locus where they are indifferent between public treatment and no treatment ($\hat{b}_i^{GN}$) is horizontal since it depends only on the public sector waiting time. Patients who are indifferent between

public and private treatment and indifferent between public and no treatment, must also be indifferent between private and no treatment. Hence b_i^{GP} and b_i^{PN} intersect on $\hat{b}_i^{GN}$. It can also be shown that the locus b_i^{GP} cuts the locus b_i^{PN} from above.

Patients with high benefit and low income demand public treatment. Patients with high benefit and high income demand private treatment. Patients with low benefit demand no treatment.

The demand for public treatment is

$$D(w_i) = \pi_i \int_{y^{\min}}^{y^{\max}} \int_{\hat{b}_i^{GN}(w_i)}^{\hat{b}_i^G(y,w_i)} f_i(b,y) dbdy \tag{30}$$

where $\hat{b}_i^G(y,w_i) = \max[\min[b_i^{GP}, b^{\max}], \hat{b}_i^{GN}]$.

Increases in waiting time reduce demand for public treatment:

$$D_{w_i}(w_i) = -\pi_i \int_{y^{\min}}^{y^{\max}} f_i(\hat{b}_i^{GN}, y) \frac{\partial \hat{b}^{GN}(w_i)}{\partial w_i} dy + \pi_i \int_{y^{\min}}^{y^{\max}} f_i(\hat{b}_i^{G}, y) \frac{\partial \hat{b}^{G}(w_i)}{\partial w_i} dy < 0 \tag{31}$$

When waiting time increases some public sector patients decide not to be treated (first term in (31)) and some patients opt for the private sector (second term in (31)). In terms of Figure 1, an increase in w_i shifts the locus $\hat{b}_i^{GN}$ upward and so patients along the locus who have incomes $y \in [y^{\min}, y_i^{GN}]$ switch out of the public sector and are not treated. The increase in w_i also shifts the locus b_i^{GP} downward. Hence patients along the locus with $y_i \in [y_i^{GP}, y_i^{GN}]$ decide to switch from the public into the private sector.

Total welfare is the sum of the utility of public patients, private patients and patients with no treatment: $S = S^{GT} + S^{PT} + S^{NT}$ where

$$S^{GT} = \sum\nolimits_i \pi_i \int_{y^{\min}}^{y^{\max}} \int_{\hat{b}_i^{GN}(w_i)}^{\hat{b}_i^G(y,w_i)} s_i^{GT}(y,b,w_i) f_i(b,y) dbdy \tag{32}$$

$$S^{PT} = \sum\nolimits_i \pi_i \int_{y^{\min}}^{y^{\max}} \int_{\hat{b}_i^{P}}^{b^{\max}} s_i^{PT}(y - m_i, b, 0) f_i(b,y) dbdy \tag{33}$$

$$S^{NT} = \sum\nolimits_i \pi_i \int_{y^{\min}}^{y^{\max}} \int_{b^{\min}}^{\hat{b}_i^{N}} s_i^{NT}(y) f_i(b,y) dbdy \tag{34}$$

where $\hat{b}_i^P(y, w_i) = \max[\min[b_i^{GP}, b^{\max}], b_i^{PN}]$ and $\hat{b}_i^N(y, w_i) = \min[\hat{b}_i^{GN}, b_i^{PN}]$.

An increase in waiting time has no effect on the inframarginal untreated patients or private patients. It reduces the utility of the public patients directly and also changes the thresholds at which people choose public treatment rather than private treatment or no treatment. Although the marginal patients are indifferent, there will be a welfare gain or loss associated with the change in the number of patients treated in the public sector unless the social welfare function respects the patient decisions.

The marginal social value of an increase in waiting time for treatment i is, via the Lagrangean $S - \lambda \sum_i c_i \pi_i D_i$,

$$\begin{aligned}
&\pi_i \int_{y^{\min}}^{y^{GN}} \left[s_i^{NT}(y) - s_i^{GT}(y, \hat{b}_i^{GN}, w_i) \right] \hat{b}_{iw}^{GN} f_i(\hat{b}_i^{GN}, y) dy \\
&\qquad + \pi_i \int_{y^{GP}}^{y^{GN}} \left[s_i^{GT}(y, b_i^{GP}, w_i) - s_i^{PT}(y - m_i, \hat{b}_i^{GP}, 0) \right] \hat{b}_{iw}^{GP} f_i(\hat{b}_i^{GP}, y) dy \\
&\qquad + \pi_i \int_{y^{\min}}^{y^{\max}} \int_{\hat{b}_i^{GN}}^{\hat{b}_i^{G}} s_{iw}^{GT}(y, b, w_i) f_i(b, y) db dy - \lambda c_i \pi_i D_{iw} \qquad (35)
\end{aligned}$$

The first term is the effect on welfare via the displacement of patients who are forced out of the public sector and who do not get treated. The second is the effect via the displacement of patients from the public into to private sector. Dividing (35) through by the number of treated patients ($\pi_i D_i$) we can write the first order conditions on waiting times in an analogous manner to (18):

$$-\left(\psi_i^{GN} + \psi_i^{GP} + \kappa_i\right) - \lambda c_i \varepsilon_{iw} w_i^{-1} \leq 0, \quad w_i \geq 0 \qquad (36)$$

where ψ_i^{GN}, ψ_i^{GP} are the marginal threshold costs for patients pushed out of the public sector into no treatment or into the private sector.

ψ_i^{GP} likely to be zero even with an extra welfarist welfare function since the marginal individual is merely shifting between consumption of private and public health care. Thus the form of the rules for allocating a given public health care sector across treatments are unaffected by the existance of a private sector alternative to public care.

5 Conclusions

We have investigated the optimal allocation of a fixed health care budget across treatments when there is rationing by waiting and user charges are exogenous. Our main finding is that the optimal waiting time is higher for treatments with demands which are more elastic to waiting time, higher costs, lower charges, and smaller marginal disutility from waiting. In addition, if the welfare function does not respect patient choices between treatment and no treatment, the waiting time should be lower for treatments where under-consumption of health care has a greater welfare cost.

The general message is that optimal allocation across sectors must take account of way care is rationed within sectors. Hence, allocation rules based purely on cost effectiveness ratios are suboptimal because they assume that there is no rationing within treatments.

References

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

van Doorslaer, E., Koolman, X. and Jones, A.M. 2004. "Explaining income-related inequalities in doctor utilisation in Europe". *Health Economics*, 13, 629-647.

Culyer A.J. 1991. "The normative economics of health care finance and provision", in A. McGuire, P. Fenn & O. Mayhew (eds.), *Providing Health Care: The Economics of Alternative Systems of Finance and Delivery*, Oxford Economic Press.

Douglas, C., Buxton, M.J., O'Brien, B.J, 2003, "Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria", Health Economics, 12, 5, 421-427.

Garber, A.M. 2000. "Advances in cost-effectiveness analysis", in A. J. Culyer and J. P. Newhouse (eds), *Handbook on Health Economics*, Amsterdam: Elsevier

Farnworth, M.G., 2003, "A game theoretic model of the relationship between prices and waiting times", *Journal of Health Economics*, 22(1), 47-60.

Gravelle, H., P.C. Smith and Xavier, A., 2003, "Performance signals in the public sector: the case of health care", *Oxford Economic Papers*, 55, 81-103.

Gravelle, H. and Siciliani, L. 2006. "Is waiting time prioritisation welfare improving?", Department of Economics Discussion Paper 06/13. to appear in *Health Economics*.

Gravelle, H. and Siciliani, L. 2007. "Optimal waits and charges in health insurance", February. Department of Economics Discussion Paper 07/02.

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.

Hoel, M., 2007,"What should (public) health insurance cover?", *Journal of Health Economics*, 26(2), 251-262.

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

Martin, S., and Smith, P.C., 1999, "Rationing by waiting lists: an empirical investigation", *Journal of Public Economics*, 71, 141-64.

Martin, S., Rice, N., Jacobs, R., Smith, P.C. 2007. "The market for elective surgery: Joint estimation of supply and demand", *Journal of Health Economics,* 26, 263-285

Morris, S., Sutton, M., Gravelle, H. 2005. "Inequity and inequality in the use of health care in England: an empirical investigation". *Social Science and Medicine*, 60, 1251-1266.

Musgrave, R. 1959. *The theory of public finance*, McGraw-Hill, New York.

Newhouse, J. P., and the Insurance Experiment Group. 1993. *Free For All? Lessons from the Health Insurance Experiment*, Harvard University Press, Cambridge.

Pauly, M., and Blavin, F., 2007, "Value based cost sharing meets the theory of moral hazard: medical effectiveness in insurance benefit design", NBER Working Paper 13044.

Ramsey, F., 1927, "A Contribution to the Theory of Taxation", *Economic Journal*, 37,145, 47-61.

Ringel, J.S., Hosek, S.D., Vollard, B.A., and Mahnovski, S. 2002. *The Elasticity of Demand for Health Care: A Review of the Literature and Its Application to the Military Health System.* Rand Organisation.

Sandmo, A., 1983, "Ex post welfare economics and the theory of merit goods", *Economica*, 50, 19– 33.

Siciliani, L., and Hurst, J., 2004, "Explaining waiting times variations for elective surgery across OECD countries", *OECD Economic Studies*, 38(1), 1-23.

Smith, P.C. 2005, "User charges and priority setting in health care: balancing equity and efficiency", *Journal of Health Economics*, 24,1018-1029.

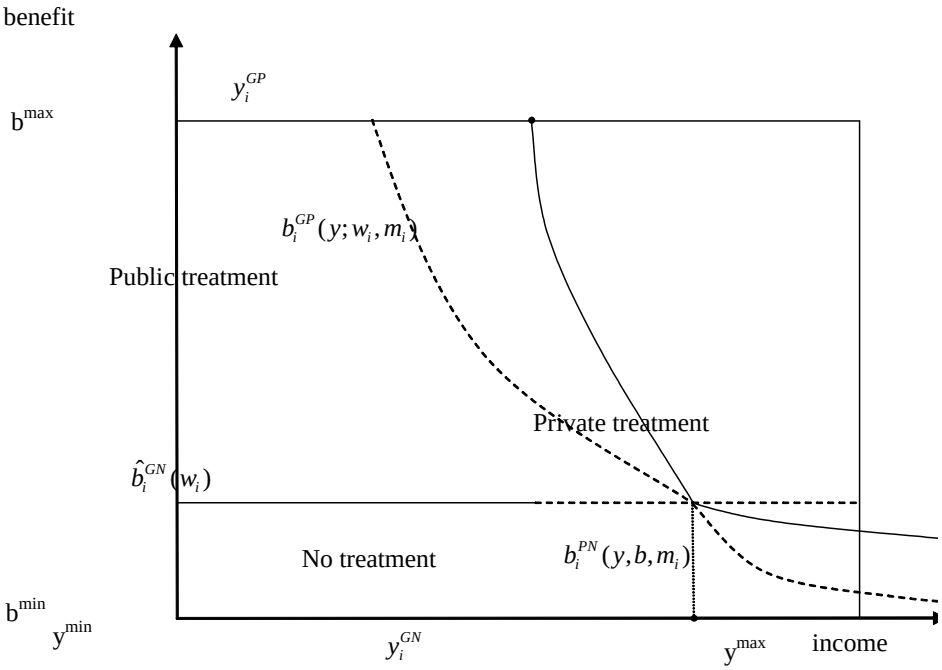

Figure 1. Patient characteristics and choice amongst public treatment, private treatment, and no treatment for condition *i*.