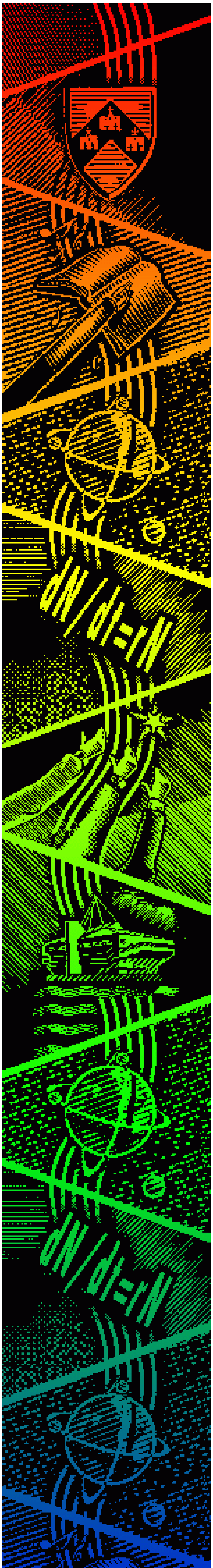

THE UNIVERSITY of York

Discussion Papers in Economics

No. 2007/22

Third Degree Waiting Time Discrimination: Optimal Allocation of a Public Sector Health Care Treatment Under Rationing by Waiting

By

Hugh Gravelle and Luigi Siciliani

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

Third degree waiting time discrimination: optimal allocation of a public sector health care treatment under rationing by waiting

Hugh Gravelle[*]   Luigi Siciliani[†]

Version 1f. 30 June 2007

Abstract

In many public health care systems treatment is rationed by waiting time. We examine the optimal allocation of a fixed supply of a treatment between different groups of patients. Even in the absence of any distributional aims welfare is increased by third degree waiting time discrimination. Because waiting time imposes dead weight losses on patients, lower waiting times should be offered to groups with higher marginal waiting time costs and with less elastic demand for the treatment.

Keywords: Waiting times; Prioritisation; Rationing.

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

Waiting times are used as a rationing mechanism for elective surgery 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 common procedures, such as hip and knee replacement, cataract surgery or varicose veins, of six months are not uncommon (Siciliani and Hurst, 2005).

The question we consider is whether different types of potential patients (young vs old, men vs women, northerners vs southerners) waiting for the same treatment should face different waiting times. We do not investigate whether types with different demands (at a given waiting time) for the treatment should have different amounts allocated to them. Rather we ask whether different types should face different waits, so that the marginal patients of different types have different benefits from the treatment. We show that in general it is welfare increasing to set different waiting times for different groups of patients waiting for the same treatment.

When treatment is rationed by waiting time there are no public sector revenue consequences of changing the allocation of a given amount of treatment between two groups of potential patients. Choosing the allocation of a given supply of treatment between the two groups is equivalent to determining their waiting times.

Individuals differ in their health gains from treatment and the marginal patient in a group places a value on the health gain which is equal to their waiting time cost. Suppose that the two groups have the same preferences but have different distributions of health gains from treatment. With equal waiting times for the two groups, the health gains of the marginal treated patient in each group are equal. Switching a unit of treatment from group A to group B will not alter production costs. Nor will it change the total health gain from treatment since the marginal patient in each group has the same health gain. But the switch will lead to a higher waiting time in group A and a lower waiting time in group B.

Waiting time is a deadweight loss: it imposes costs on patients which are not offset by gains to anyone else. If group A has a more waiting time elastic demand than group B then the total waiting time of group A patients will fall by more than the increase in total waiting time of group B patients and total waiting time over all patients will fall. Hence waiting time discrimination in favour of the group with the less elastic is welfare increasing. If the two groups have different preferences as well as a different distribution of health gains then the direction of discrimination will also depend on the marginal cost of waiting per treated patient in the two groups. But the conclusion is the same: in general it is welfare increasing to offer different waiting times to different groups of patients. Notice that the discrimination does not arise from any wish to favour one group over another on distributional grounds nor does it arise from a need to raise revenue as might be the case if care was rationed by user charges.

The results provide support for the prioritisation schemes such as those in Canada, New Zealand, Spain and Sweden (Siciliani and Hurst, 2005) which set different waiting times for different types of patient waiting for the same treatment. They also have implications for the allocation of resources across regions and debates over post code rationing. If patients in different regions have different preferences or distributions of benefits from treatment then optimal geographical resource allocation will yield different waiting times for the same treatment in different regions.

In a previous paper (Gravelle and Siciliani, 2007a) we examined linear prioritisation rules where the waiting time for patients varied linearly according to some characteristic such as age. We showed that linear prioritisation was welfare increasing unless the age and health benefit had a particular type of joint distribution. The categorical prioritisation scheme considered in this paper is more general in that it does not require that waiting times are linear, or even monotonic, with respect to observable patient characteristics. With categorical prioritisation waits could be high for the old and young and low for the middle aged. Since categorical prioritisation is more general, the conditions under which it is welfare increasing are weaker than with linear prioritisation. All that is necessary is that is possible to define categories of patients with either different preferences or different distributions of health gain so that they have either different marginal costs of waiting or different demand elasticities.

There are two related literatures. The first examines the optimal allocation of a health care budget across different treatments (Garber, 2000, Gravelle and Siciliani, 2007b; Hoel, 2007; Smith, 2005). In this paper we examine allocation within a treatment to different groups of patients. The allocation rule used within a treatment will affect the marginal value of resources devoted to a treatment and hence the optimal allocation of a budget across treatments. The second literature examines the optimal mix of money and time prices to ration treatment (Gravelle and Siciliani, 2007c; Hoel and Saether 2003; Marchand and Schroyan, 2004). Our conclusion that waiting time discrimination within treatments increases welfare from rationing by waiting has implications for the optimal mix of money and waiting time prices.

Section 2 describes patient preferences and demand and sets out the welfare function. Section 3 shows that the optimal allocation requires third degree waiting time discrimination. Section 4 shows that the results hold in more general specifications when demand for public sector care depends on income and patients also have the option of buying care in the private sector where there is no wait. Section 5 concludes.

2 Rationing by waiting time

2.1 Preferences and demand

All health care is produced in a public health care system. We consider the implications of a private sector alternative in Section 4. Demand for treatment in the public sector is rationed by waiting. To reduce notational clutter we initially assume there are no user charges and relax this assumption in Section 4. $h \in [h^{\min}, h^{\max}]$ is the health gain or benefit from treatment. There are two types or groups of individual $j = A, B$ with different density and distribution functions $f_j(h)$ and $F_j(h)$. The total population is normalised to 1. The proportion of the population in group j is π_j and $\sum_j \pi_j = 1$.

We assume that all individuals of a given type have the same preferences but different types may have different preferences. Allowing for differences in preferences within groups merely complicates the analysis and does not alter the results. If ill and not treated an individual of type j has utility v_j^{NT}. Treated patients in group j have a wait of w_j before receiving one unit of treatment which produces health gain h. Utility if treated $v_j^T(h, w_j)$ is increasing in health gain and decreasing in the wait: $\partial v_j^T(h, w_j)/\partial h = v_{jh}^T > 0$, $\partial v_j^T(h, w_j)/\partial w_j = v_{jw}^T < 0$.[1] We do not need to consider utility or welfare for well individuals since allocating a fixed amount of a treatment between groups has no effect on the utility or welfare of individuals who are not ill. 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

[1] We are assuming that health is separable from other variables affecting utility, for example consumption of other private goods, and that these other variables are unaffected by whether the individual is treated or not. We relax this assumption in Section 4.

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 patients 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).[2]

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).

An individual of type j demands treatment if and only if

$$v_j^T(h, w_j) - v_j^{NT} \geq 0 \Longleftrightarrow h \geq \hat{h}_j = \hat{h}_j(w_j) \tag{1}$$

where $\hat{h}_j(w_j)$ is the threshold health gain such that all those with a smaller gain do not seek treatment. The threshold is increasing in the waiting time since

$$\partial\hat{h}_j/\partial w_j = \hat{h}_{jw}(w_j) = -v_{jw}^T(\hat{h}_j, w_j)/v_{jh}^T(\hat{h}_j, w_j) > 0 \tag{2}$$

Note that if the two groups have the same preferences and differ only in the distribution of health gain, all patients would have the same threshold at a given waiting time, irrespective of their group. Conversely, if they differ in the utility from treatment or non-treatment and have the same distributions, their thresholds will differ. For example if group A has more utility from treatment, then its threshold is lower: a member of group A would be more likely to seek treatment. However, the marginal individuals in both groups would have the same utility from treatment if their non-treatment utility is the same, irrespective of their treatment utility functions and their distribution functions.

Expected demand for treatment from individuals of type j is

$$\pi_j D_j(w_j) = \pi_j \int_{\hat{h}_j} dF_j(h) = \pi_j \left[1 - F_j(\hat{h}_j)\right] \tag{3}$$

[2] Lindsay and Feigenbaum (1984) assume a utility function from treatment equal to $he^{-w} - c$, where c is the fixed cost of getting on the list. Our specification is more general since we do not impose a negative exponential discount rate.

where $D_j(w_j)$ is demand per person of type j. From (2) demand is decreasing in the waiting time

$$D_{jw}(w_j) = -\hat{h}_{jw}(w_j) f_j(\hat{h}_j(w_j)) < 0 \tag{4}$$

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).

2.2 Determination of waiting time

The supply of treatment allocated to group j is z_j and total supply is $z = \sum_j z_j$. The waiting time $w_j(z_j)$ for group j is determined by

$$\pi_j D_j(w_j) - z_j \leq 0, \quad w_j \geq 0, \quad w_j \left[\pi_j D_j(w_j) - z_j\right] = 0 \tag{5}$$

and is decreasing in supply

$$\frac{\partial w_j}{\partial z_j} = w_{jz} = \frac{1}{\pi_j D_{jw}} < 0 \tag{6}$$

We assume $\sum_j \pi_j \ D_j(0) > z$.

When there is no prioritisation patients in the two groups face the same waiting time $\bar{w}$ which is determined by

$$\sum_j \pi_j D_j(\bar{w}) = z \tag{7}$$

where $D_j(\bar{w}) > 0$. At $\bar{w}$ the different types have the same treatment threshold if and only if they have the same preferences. With a common waiting list and equal waiting times the amount of capacity allocated to type j is $\bar{z}_j$ $= \pi_j D_j(\bar{w})$. Note that although waiting time is the same, supply differs in equilibrium for the different types.

2.3 Welfare

The planner knows the distribution functions but cannot prioritise individual patients on the basis of their health gain. Expected utility for potential patients of type j is

$$V_j(w_j) = \int^{\hat{h}_j(w_j)} v_j^{NT} dF_j + \int_{\hat{h}_j(w_j)} v_j^T(h, w_j) dF_j \tag{8}$$

If one group's expected utility had a greater social value than the other it would be unsurprising that the groups would be offered different waiting times. To focus on other factors determining their waiting times we assume that welfare is the sum of the expected utilities of the two types

$$S(w_A, w_B) = \sum_j \pi_j V_j(w_j) \tag{9}$$

Increasing w_j reduces welfare from treatment:

$$\begin{aligned} \frac{\partial S}{\partial w_j} &= \pi_j \left\{ \left[(v_j^{NT} - v_j^T(\hat{h}_j(w_j), w_j) \right] \hat{h}_{jw}(w_j) + \int_{\hat{h}_j} v_{jw}^T(h, w_j) dF_j \right\} \\ &= \pi_j \int_{\hat{h}_j} v_{jw}^T(h, w_j) dF_j < 0 \qquad (10) \end{aligned}$$

where the term in the square brackets on the first line is zero because individuals make privately optimal decisions about demanding care and the welfare function respects these decisions. The marginal welfare cost of waiting per treated patient in group j is

$$\kappa_j = -\int_{\hat{h}_j} v_{jw}^T(h, w_j) dF_j / D_j > 0 \tag{11}$$

3 Optimal third degree waiting time discrimination

The cost of treatment does not depend on the type of patient. Hence the problem of allocating a given amount of treatment requires that welfare S is maximised by choice of z_j subject to $\sum_j z_j \leq z$.

Using the constraint to substitute $z - z_A$ for z_B, recalling (2), (4) and (6), the marginal welfare effect of prioritising type A by increasing z_A (and thereby reducing w_A) is

$$\begin{aligned} \frac{dS}{dz_A} &= \sum\nolimits_j \pi_j w_{jz} \frac{dz_j}{dz_A} \int_{\hat{h}_j(w_j)} v_{jw}^T(h, w_j) dF_j(h) \\ &= \frac{v_{Ah}^T(\hat{h}_A(w_A), w_A)}{v_{Aw}^T(\hat{h}_A(w_A), w_A) f_A(\hat{h}_A(w_A))} \int_{\hat{h}_A} v_{Aw}^T(h_A, w_A) dF_A \\ &\quad - \frac{v_{Bh}^T(\hat{h}_B(w_B), w_B)}{v_{Bw}^T(\hat{h}_B(w_B), w_B) f_B(\hat{h}_B(w_B))} \int_{\hat{h}_B} v_{Bw}^T(h_B, w_B) dF_B \qquad (12) \end{aligned}$$

The signs of the first and second derivatives (see Appendix) of the welfare function with respect to z_A depend on the specifics of the preferences and distributions of the two groups. But by inspecting (12) at $z_A = \bar{z}_A$ where $w_A = w_B = \bar{w}$ we see that

Proposition 1 *Offering different waiting times to groups waiting for a treatment (third degree waiting time discrimination) is always welfare increasing unless the groups have the same preferences and the same distribution of health gains from treatment.*

Although the direction of prioritisation depends on the details of preferences and distributions of health gain we can provide an example where there is a definite conclusion. An alternative way to write (12) when the initial allocation has equal waiting times $w_A = w_B = \bar{w}$ is:

$$\left.\frac{dS}{dz_A}\right|_{z_A=\bar{z}_A} = \frac{1}{D_{Aw}} \int_{\hat{h}_A(\bar{w})} v^T_{Aw}(h,\bar{w})dF_A(h) - \frac{1}{D_{Bw}} \int_{\hat{h}_B(\bar{w})} v^T_{Bw}(h,\bar{w})dF_B(h) \tag{13}$$

If both groups have the same preferences $v^T_{Aw}(h,w) = v^T_{Bw}(h,w) = v^T_w(h,w)$ then $\hat{h}_A(\bar{w}) = \hat{h}_B(\bar{w}) = \hat{h}(\bar{w})$, but $D_{Aw} = -\hat{h}_w f_A(\hat{h}) \neq D_{Bw} = -\hat{h}_w f_B(\hat{h})$. We have

$$\begin{aligned}\left.\frac{dS}{dz_A}\right|_{z_A=\bar{z}_A} &= \frac{1}{D_{Aw}} \int_{\hat{h}(\bar{w})} v^T_w(h,\bar{w})dF_A(h) - \frac{1}{D_{Aw}} \int_{\hat{h}(\bar{w})} v^T_w(h,\bar{w})dF_B(h) \\ &\quad + \left(\frac{1}{D_{Aw}} - \frac{1}{D_{Bw}}\right) \int_{\hat{h}(\bar{w})} v^T_w(h,\bar{w})dF_B(h) \end{aligned} \tag{14}$$

Integrating by parts, we obtain

$$\begin{aligned}\left.\frac{dS}{dz_A}\right|_{z_A=\bar{z}_A} &= \frac{1}{-D_{Aw}}\left[v^T_w(\hat{h}(\bar{w}),\bar{w}) \left[F_A(\hat{h}(\bar{w})) - F_B(\hat{h}(\bar{w}))\right] \right. \\ &\quad \left. + \int_{\hat{h}(\bar{w})} v^T_{wh}(h,\bar{w}) \left[F_A(h) - F_B(h)\right] dh \right] \\ &\quad + \frac{D_{Bw} - D_{Aw}}{D_{Aw}D_{Bw}} \int_{\hat{h}(\bar{w})} v^T_w(h,\bar{w})dF_B(h) \end{aligned} \tag{15}$$

If the distribution of benefit of group A strictly first order stochastically dominates that of group B, $F_A(h) < F_B(h)$, $h \in (h^{\min}, h^{\max})$, and if the marginal disutility of waiting is higher for patients with higher benefit, $v^T_{wh} < 0$, then the first term is positive. If the demand of group B is more responsive than group A, i.e. $-D_{Bw} > -D_{Aw}$, then the second term is also positive, and prioritising group A is always welfare improving.

Proposition 2 *If (a) the distribution of benefit of group A strictly first order stochastically dominates that one of group B $(F_A(h) < F_B(h), h \in (h^{\min}, h^{\max}))$; (b) the marginal disutility of waiting is higher for patients with higher expected benefit $\left(v_{wh}^T < 0\right)$; (c) demand of group B is more responsive than group A $(-D_{Bw} > -D_{Aw})$; then welfare is higher for some degree of prioritisation of group A $(w_A < \bar{w} < w_B)$ than with no prioritisation $(w_A = \bar{w} = w_B)$.*

Using the definitions of the marginal welfare cost of waiting (11), we can write (12) in a more intuitively appealing form:

$$\frac{dS}{dz_A} = w_B \frac{\kappa_B}{\varepsilon_B} - w_A \frac{\kappa_A}{\varepsilon_A} \tag{16}$$

where $\varepsilon_j = D_{jw} w_j / D_j$ is the elasticity of demand by group j with respect to waiting time. Hence

Proposition 3 *The optimal prioritisation scheme for patient groups A and B when both receive treatment at a positive waiting time, satisfies*

$$\frac{w_A}{w_B} = \frac{\varepsilon_{Aw}}{\varepsilon_{Bw}} \frac{\kappa_B}{\kappa_A} \tag{17}$$

Thus, if the marginal cost of waiting is identical across the two groups $(\kappa_A = \kappa_B)$, the group with the less elastic demand will be prioritised by getting lower waiting times. If the elasticity of demand is identical across the two groups $(\varepsilon_{Aw} = \varepsilon_{Aw})$, the group with the lower marginal cost of waiting will be prioritised by getting higher waiting times.

Figure 1 provides an intuition for the welfare effects of discrimination in a simple case where the two groups have the same size $(\pi_A = \pi_B = 1/2)$, and the same preferences. The utility of treated patients is $v_j^T = h - kw_j$. Thus the per patient marginal welfare cost of an increase in waiting time (32) is the same for both groups: $\kappa_A = \kappa_B = k$.

Total supply is the horizontal distance between the vertical axes. Supply to group A is measured rightwards from the left origin. All patients for whom $h \geq kw_j$ demand treatment and the demand curve for group j is $\pi_j D_j(w_j) = \pi_j [1 - F_j(kw_j)]$. The waiting time for each group is determined by the intersection of their demand curve and the vertical supply curve at z_j.

The expected health gain from treatment for group j, net of waiting time costs, is $\pi_j \int_{\hat{h}_j} (h - kw_j) dF_j$ which is the area below their demand curve and above the waiting time price line. Since waiting time costs are deadweight losses, the allocation of the given total supply of treatment between the two groups should maximise their combined expected net health gain $\sum_j \pi_j \int_{\hat{h}_j} (h - kw_j) dF_j$.

Consider moving from an initial allocation $z_A = \bar{z}_A$ with equal waiting times ($\bar{w}$), to an allocation $z_A^o < \bar{z}_A$ where group A has reduced supply and hence a higher waiting time than group B: $w_A^o > \bar{w} > w_B^o$. The loss to group A which has the more elastic demand is $a + b$ which is less than the gain to group B of $d + e + f$. Since total production cost is unchanged, welfare is increased by the discrimination in favour of group B which has less elastic demand at the initial no discrimination allocation.

The gross potential benefit from treatment for the groups, before taking account of waiting time costs, is the area under their demand curves. Thus it is not necessarily welfare increasing to favour the group with higher potential benefits $\left(\pi_j \int_{h^{\min}} hdF_j\right)$ with a lower waiting time. Nor should priority necessarily be given to the group with greater benefits to treated patients $\left(\pi_j \int_{\hat{h}_j(\bar{w})} hdF_j\right)$. Account must also be taken of how the two groups will respond to the changes in waiting times when prioritisation is introduced and capacity is shifted from one to the other.

Waiting time is a deadweight cost: it rations demand by imposing costs on users which are not offset by a gain to anyone else (Barzel, 1974). Money prices also ration by imposing costs on users but these costs are offset by the revenue received by the producer. If the payments by patients have the same welfare weight as the resulting revenue and all patients have the same social weight, third degree price discrimination would be welfare reducing. The prices charged to the two groups should aim to maximise the sum of consumer surplus and producer profit. Given that the total supply is fixed this is equivalent to maximising willingness to pay or the sum of the areas under the demand curves in (money price, quantity) space. Thus, if we temporarily interpret w as a money price, so that Figure 1 plots demand curves for groups A and B in (money price, quantity) space, any move from the initial allocation $\bar{z}_A$ with equal prices reduces the sum of the willingness to pay of the two groups. For example, the reallocation from $\bar{z}_A$ to z_A^o gives a lower money price to group B and would reduce welfare by the area $b + c$.[3] Intuitively, with a fixed supply to allocate and with no distributional

[3] Consumers of group B gain $d + e + f$, and the producer gains from group B is $g + h + i - e - f$. Net welfare gain from group B is $g + h + i + d$. Consumers of group A loose $a + b$, and the producer loss from group A is $c + d + g + h + i - a$. Net welfare loss from group A is $c + d + g + h + i + b$. The difference between the net welfare loss from group B and the net welfare gain from group A is $(c + d + g + h + i + b) - (g + h + i + d) = b + c$.

considerations, optimal allocation when there is rationing by money price maximises willingness to pay and when there is rationing by waiting time, it maximises consumer surplus. The difference arises because waiting time imposes a deadweight loss.

Figure 1 is also useful in illustrating the possibility of corner solutions and non-concavity of the welfare function even with plausible simple assumptions about preferences. A small shift from equal waiting times to higher waits for group A increases welfare. But this is only a local welfare improvement. Recall that the utility of treated patients are $v_j^T = h - kw_j$. Then, the marginal welfare effect of increasing the allocation for group A is, using (6),

$$\begin{aligned} \frac{dS}{dz_A} &= -\pi_A k w_{Az} \int_{\hat{b}_A} dF_A + \pi_B k w_{Bz} \int_{\hat{b}_B} dF_B \\ &= \frac{1 - F_A(kw_A)}{f_A(kw_A)} - \frac{1 - F_B(kw_B)}{f_B(kw_B)} \qquad (18) \end{aligned}$$

If the distributions have positive density over their supports then

$$\begin{aligned} \left.\frac{dS}{dz_A}\right|_{z_A=0} &= -\frac{1 - F_B(kw_B(z))}{f_B(kw_B(z))} < 0 \qquad (19) \\ \left.\frac{dS}{dz_A}\right|_{z_A=z} &= \frac{1 - F_A(kw_A(z))}{f_A(kw_A(z))} > 0 \qquad (20) \end{aligned}$$

so that there are local optima at both extreme prioritisation allocations where all the treatment is given to one group. There will also be an allocation where the first order condition is satisfied but, if there is only one such allocation, it is a global minimum. In the example in Figure 1 welfare is maximised by giving all the supply to group B (setting them a wait of w_B^*) rather than giving it all to group A with a wait of w_A^*.

We can use (18) to state another set of assumptions about preferences which yields a definite conclusion about the direction of welfare increasing prioritisation. Distributions can be uniquely characterised by their hazard functions $f_j/(1 - F_j)$. A distribution A is "strictly more favourable" than distribution B if

$$\frac{f_A(h)}{1 - F_A(h)} < \frac{f_B(h)}{1 - F_B(h)}, \qquad h \in (h^{\min}, h^{\max}) \qquad (21)$$

If (21) holds then the distribution F_A first order stochastically dominates distribution F_B (De Fraja, 2005, p.1014) which in turn implies that individuals in group A have a greater expected health gain those in group B.

Write (18) as

$$\begin{aligned}\frac{dS}{dz_A} &= \left(\frac{1-F_A(kw_A)}{f_A(kw_A)} - \frac{1-F_B(kw_A)}{f_B(kw_A)}\right) \\ &\quad + \left(\frac{1-F_B(kw_A)}{f_B(kw_A)} - \frac{1-F_B(kw_B)}{f_B(kw_B)}\right) \qquad (22)\end{aligned}$$

At $w_A = w_B = \bar{w}$ the term in the second line is zero and so (22) is positive if F_A is more favourable than F_B. Now evaluate (22) at $w_A < \bar{w} < w_B$. The first term is positive. The second term is non-negative if the group B hazard rates is monotonically non-decreasing. Many common distributions such as the uniform, logistic, chi-squared, exponential and Laplace have non-decreasing hazard rates (Laffont and Tirole, 1993, p.66).

This establishes

Proposition 4 *(a) If both groups have the same separable preferences with linear waiting time costs: $v^T(h, w_j) = h - kw_j$ and group A has a more favourable distribution of benefits in the sense of (21), then some degree of prioritisation in favour of group A is welfare increasing compared with prioritisation. If (b) in addition, the hazard rates for group B is monotonically non-decreasing $(d[f_B/(1-F_B)]/dh \geq 0)$ then group A should be given complete priority ie it should get the entire supply of treatment.*

4 Extension

We have so far made a number of simplifying assumptions: no effect of treatment on other factors affecting utility, zero charges for public sector care, and no alternative private sector supply of health care. We now show that relaxing these assumptions leaves the basic result about the welfare gain from third degree waiting time discrimination unchanged.

Suppose that utility for a member of group j when treated in the public sector is $v_j^T(y - p, h, w_j)$ where $y \in [y^{\min}, y^{\max}]$ and p is the fixed user charge for public sector health care. There is a private sector where health care treatment with no wait is available at price m. Utility if treated in the private health care sector is $v_j^T(y - m, h, 0)$. We assume that $m > p$ otherwise no patient ever demands care in the public sector. However we do not need to assume that $p \geq 0$: individuals could be paid to consume public health care. Utility if not treated is $v_j^{NT}(y)$.[4] All utility functions are concave in income and v_j^T is increasing in health gain and decreasing in the

[4] The formulation allows for the possibility that receiving treatment can increase in-

waiting time. We also assume that treatment increases the marginal utility of income $v^{T}_{jy}(y,h,w) > v^{NT}_{jy}(y)$ and that marginal utility of income is not increased by a positive wait: $v^{T}_{jy}(y,h,w) \leq v^{T}_{jy}(y,h,0)$.

Patients choose public treatment to no treatment if $v^{T}_{j}(y-p,h,w_j) \geq v^{NT}_{j}(y)$. The health gain threshold above which patients prefer public (government) treatment to no treatment is $\hat{h}^{GN}_{j}(y,p,w_j)$. The threshold is increasing in the waiting time and the user charge:

$$\hat{h}^{GN}_{jw} = -\frac{v^{T}_{jw}}{v^{T}_{jh}} > 0, \qquad \hat{h}^{GN}_{jp} = \frac{v^{T}_{jy}}{v^{T}_{jh}} > 0 \tag{23}$$

and decreasing in income:

$$\hat{h}^{GN}_{jy} = -\frac{v^{T}_{jy}(y-p,h,w_j) - v^{NT}_{jy}(y)}{v^{T}_{jh}} < 0 \tag{24}$$

Patients prefer public to private treatment if $v^{T}_{j}(y-p,b,w_j) \geq v^{T}_{j}(y-m,b,0)$. $\hat{h}^{GP}_{j}(y,p,m,w_j)$ is the threshold below which patients prefer public treatment to private treatment. It is decreasing in waiting time and in income

$$\hat{h}^{GP}_{jy} = -\frac{v^{T}_{jy}(y-p,h,w_j) - v^{T}_{jy}(y-m,h,0)}{v^{T}_{jh}} < 0 \tag{25}$$

Patients prefer private treatment to no treatment if $v^{T}_{j}(y-m,h,0) \geq v^{NT}_{j}(y)$. The health gain threshold above which patients prefer private treatment to no treatment is $\hat{h}^{PN}_{j}(y,m)$. The threshold is decreasing in income

$$\hat{h}^{PN}_{jy} = -\frac{v^{T}_{jy}(y-m,h,0) - v^{NT}_{jy}(y)}{v^{T}_{jh}} < 0 \tag{26}$$

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.

come. Let y be income if treated, let income if not treated be $y-\ L(y)$, and denote utility if not treated as $\hat{v}^{NT}(y-L(y))$. Then we can write $v^{NT}(y) = \hat{v}^{NT}(y-L(y))$. It also allows for a fixed user charge p for the treatment. Let utility if treated be $\tilde{v}^{T}(y-p,h,w_j)$ and write $v^{T}(y,h,w_j) = \tilde{v}^{T}(y-p,h,w_j)$ without loss of generality since p does not vary across groups or with income or with waiting time.

The joint density and distribution functions over health gain and income for group j are $f_j(h, y)$ and $F_j(h, y)$. The expected demand for public treatment from group j is

$$\pi_j D(w_j) = \pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{\hat{h}_j^{GN}(w_j)}^{\hat{h}_j^{G}(y,w_j)} f_j(h, y) dh dy \tag{27}$$

where $\hat{h}_j^G(y, w_j) = \max[\min[\hat{h}_j^{GP}, h^{\max}], \hat{h}_j^{GN}]$.

Increases in waiting time reduce demand for public treatment:

$$\pi_j D_w(w_j) = -\pi_j \int\limits_{y^{\min}}^{y^{\max}} \frac{\partial \hat{h}^{GN}(w_j)}{\partial w_j} f_j(\hat{h}_j^{GN}, y) dy + \pi_j \int\limits_{y^{\min}}^{y^{\max}} \frac{\partial \hat{h}_j^{G}(w_j)}{\partial w_j} f_j(\hat{h}_j^{G}, y) dy < 0$$

When waiting time increases some public sector patients decide not to be treated (first term) and some patients opt for the private sector (second term).

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

$$S^{GT} = \sum\nolimits_j \pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{\hat{h}_j^{GN}(w_j)}^{\hat{h}_j^{G}(y,w_j)} v_j^T(y - p, h, w_j) f_j(h, y) dh dy \tag{28}$$

$$S^{PT} = \sum\nolimits_j \pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{\hat{h}_j^{P}}^{b^{\max}} v_j^T(y - m, h, 0) f_j(h, y) dh dy \tag{29}$$

$$S^{NT} = \sum\nolimits_j \pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{h^{\min}}^{\hat{h}_j^{N}} v_j^{NT}(y) f_j(h, y) dh dy \tag{30}$$

where $\hat{h}_j^P(y, w_j) = \max[\min[\hat{h}_j^{GP}, h^{\max}], \hat{h}_j^{PN}]$ and $\hat{h}_j^N(y, w_j) = \min[\hat{h}_j^{GN}, \hat{h}_j^{PN}]$.

The marginal social value of an increase in waiting time for group j is

$$\begin{aligned}
\frac{\partial S}{\partial w_j} &= \pi_j \int\limits_{y^{\min}}^{y^{\max}} \left[v_j^{NT}(y) - v_j^T(y-p, \hat{h}_j^{GN}, w_j) \right] \hat{h}_{jw}^{GN} f_j(\hat{h}_j^{GN}, y) dy \\
&\quad + \pi_j \int\limits_{y^{\min}}^{y^{\max}} \left[v_j^T(y-p, \hat{h}_j^{GP}, w_j) - v_j^T(y-m, \hat{h}_j^{GP}, 0) \right] \hat{h}_{jw}^{GP} f_j(\hat{h}_j^{GP}, y) dy \\
&\quad + \pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{\hat{h}_j^{GN}}^{\hat{h}_j^{G}} v_{jw}^T(y-p, h, w_j) f_j(h, y) dh dy \\
&= \pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{\hat{h}_j^{GN}}^{\hat{h}_j^{G}} v_{jw}^T(y-p, h, w_j) f_j(h, y) dh dy < 0 \qquad (31)
\end{aligned}$$

The first term is the effect on public sector patients who decide not to be treated when the waiting time for the public sector increases. The second term is the effect on public sector patients who decide to switch to the private sector. But since the welfare function respects the choices of patients both of these terms are zero. Thus the welfare effect of an increase in waiting time is via its effect on the utility of patients treated in the public sector.

Dividing (31) through by the number of treated patients ($\pi_j D_j$) we can write the marginal social cost of an increase in waiting time for group j in the same way as (11) in Section 2.3:

$$\kappa_j = -\pi_j \int\limits_{y^{\min}}^{y^{\max}} \int\limits_{\hat{h}_j^{GN}}^{\hat{h}_j^{G}} v_{jw}^T(y-p, h, w_j) f_j(h, y) dh dy / \pi_j D_j \qquad (32)$$

The constraint on the welfare problem of allocating a fixed total supply of public sector care between the two groups is the same as in Section 3 and so Propositions 1 and 3 hold when there are user charges in the public sector, a private sector alternative, utility is not separable, and treatment affects income.

5 Conclusions

In many countries public sector health care treatments are rationed by waiting. We have investigated whether different groups of patients, defined by their preferences or distribution of benefits from treatment, waiting for the same treatment should be given different waiting times. We have shown that, even with a simple utilitarian social welfare function which respects individual choices and places the same weight on the utility of members of different groups of patients waiting for a treatment, it is in general welfare increasing to give different waiting times to the different groups of patients. The rationale is that rationing by waiting imposes deadweight losses: the cost of waiting imposed on patients is not offset by gains elsewhere. It is not necessarily the case that the groups with higher expected benefit from the treatment should have shorter waits. Shorter waiting times should be offered to groups with higher marginal waiting time costs and with less elastic demand for the treatment.

6 References

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

De Fraja, G., 2005, "Reverse Discrimination and Efficiency in Education", *International Economic Review*, 46, 1009-1031.

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

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

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

Gravelle, H. and L. Siciliani, 2007a, “Is waiting time prioritisation welfare improving?”, *Health Economics*, forthcoming.

Gravelle, H., and L. Siciliani, 2007b, "Ramsey waits: allocating public health service resources when there is rationing by waiting", June, Department of Economics Discussion Paper 07/15.

Gravelle, H., and L. Siciliani, 2007c, “Optimal waits and charges in health insurance”, February, Department of Economics Discussion Paper

07/02.

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

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.

Laffont, J.J., and J. Tirole, 1993, *A Theory of Incentives in Procurement and Regulation*, Cambridge, MA: MIT Press.

Lindsay, C.M., and B. Feigenbaum, 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.

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

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

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

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

Appendix: concavity of welfare function

The second derivative of the welfare function with respect to z_A is

$$\begin{aligned}\frac{d^2S(w_A(z_A),w_B(z_B))}{dz_A} &= \frac{d^2S(w_A(z_A),w_B(z-z_A))}{dz_A} \\ &= S_{w_Aw_A}(w_{Az})^2 - S_{w_Aw_B}w_{Az}w_{Bz} + S_{w_A}w_{Azz} \\ &\quad +S_{w_Bw_B}(w_{Bz})^2 + S_{w_Bw_A}w_{Bz}w_{Az} + S_{w_B}w_{Bzz}\end{aligned}$$

Since the cross partials of S with respect to w_A and w_B are zero, the welfare function is concave in z_A only if $S_{w_jw_j}(w_j')^2 + S_{w_j}w_j'' < 0$ for $j = A$ or B. Now

$$
\begin{aligned}
S_{w_j w_j}(w_j')^2 + S_{w_j} w_j'' \quad = \quad & \frac{1}{(D_{jw})^2}\left[\int_{\hat{h}_j} v_{jww}^T(h, w_j)\,dF_j(h)\right. \\
& \qquad\qquad \left. - \, v_{jw}^T(\hat{h}_j, w_j) f_j(\hat{h}_j)\hat{h}_{jw}\right] \\
& -\frac{D_{jww}}{(D_{jw})^2}\int_{\hat{h}_j(\bar{w})} v_{jw}^T(h, w_j)\,dF_j(h)
\end{aligned}
$$

The first term in the square brackets is negative if $v_{jww}^T(h, \bar{w}) < 0$ but the second term is positive. Since

$$D_{jww} = -\hat{h}_{jw}(w_j) f_j(\hat{h}_j(w_j)) - \hat{h}_j(w_j) f_j'(\hat{h}_j\,(w_j))$$

and

$$\hat{h}_{jw}(w_j) = -v_{jw}^T(\hat{h}_j, w_j)/v_{jh}^T(\hat{h}_j, w_j) > 0$$

the sign of the term on the second line is also in general indeterminate without further restrictions on the distribution function and preferences. Note that in the case of the plausible seeming preferences used in Lindsay and Feigenbaum (1986) where $v^T = he^{-rw} - a$, the utility function is convex in waiting time: $v_{ww}^T = r^2 e^{-rw} > 0$.

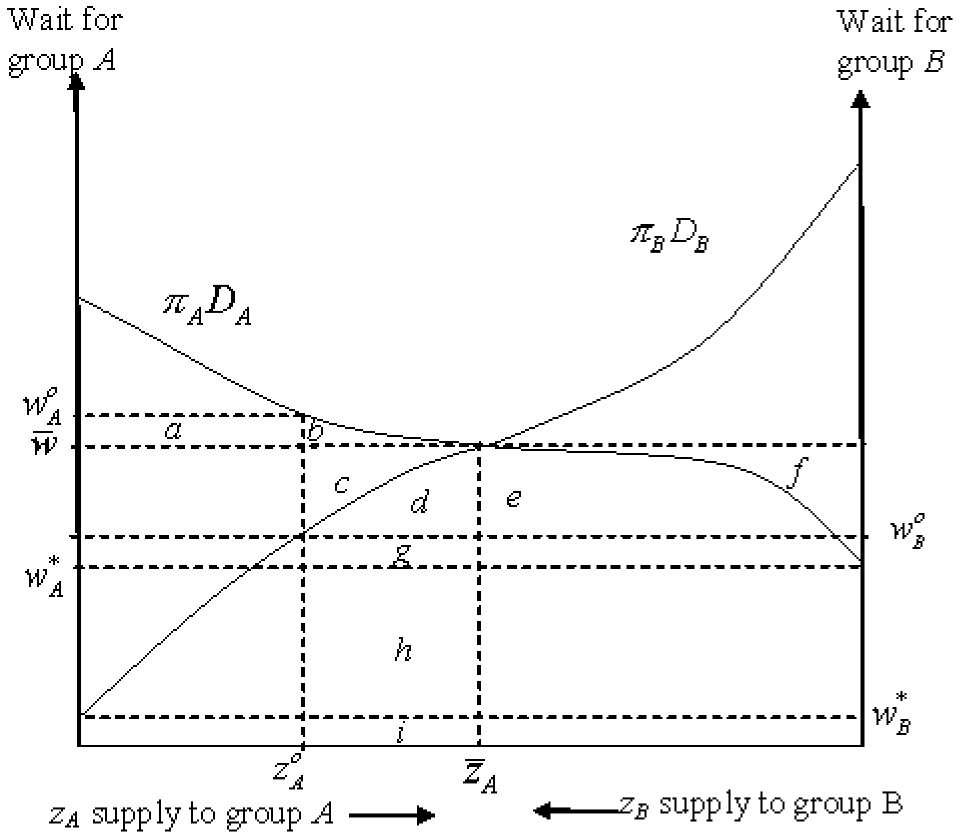

Figure 1. Allocation of treatment to two groups: third degree waiting time discrimination increase welfare since the loss to group A of $(a + b)$ is more than offset by the gain to group B of $(d + e + f)$.