THE UNIVERSITY of York

Discussion Papers in Economics

No. 2006/13

Is Waiting-time Prioritisation Welfare Improving?
by

Hugh Gravelle and Luigi Siciliani

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

Is waiting-time prioritisation welfare improving?

Hugh Gravelle* Luigi Siciliani†

Revised version: 9 May 2007 (first version: 3 July 2006)

Abstract

Rationing by waiting time is commonly used in health care systems with zero or low money prices. Some systems prioritise particular types of patient and offer them lower waiting times. We investigate whether prioritisation is welfare improving when the benefit from treatment is the sum of two components, one of which is not observed by providers. We show that positive prioritisation (shorter waits for patients with higher observable benefit) is welfare improving if the mean observable benefit of the patients who are indifferent about receiving the treatment is smaller than the mean observable benefit of the patients who receive the treatment. This is true (a) if the distribution of the unobservable benefit is uniform for any distribution of the observable benefit; or (b) if the distribution of the observable benefit is uniform and the distribution of the unobservable benefit is log-concave. We also show that prioritisation is never welfare increasing if and only if the distribution of unobservable benefit is negative exponential.

Keywords: waiting times, prioritisation, rationing.

JEL: I11

*National Primary Care Research and Development Centre, Centre for Health Economics, University of York. Email: hg8@york.ac.uk. NPCRDC receives long term funding from the Department of Health. The views expressed are not necessarily those of the DH. We are grateful for comments and suggestions by two referees, Fred Schroyan, Rossella Levaggi, and participants in seminars in Constanz, Leuvan and York.

†Department of Economics and Related Studies; Centre for Health Economics, University of York; and C.E.P.R.. Email ls24@york.ac.uk

1 Introduction

In many countries with tax or public health insurance finance elective surgery is rationed by waiting. In 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, vary from three to eight months (Siciliani and Hurst, 2004).

Some countries have explicit waiting-time prioritisation for certain types of treatment. Schemes can have a limited number of priority categories, as in Spain and Sweden (high-priority and low-priority) and in Australia and Italy (recommended admission within 30 days, 90 days and 12 months). More elaborate priority scoring systems, as in New Zealand and Canada, assign points to patients and patients with higher scores have shorter waits (Siciliani and Hurst; 2005, section 3.2.1).

In the New Zealand points scheme for cataract, patients with "lens induced glaucoma" are assigned 71-90 points, whilst those categorised as "cataract extraction required in order to treat posterior segment disease", are assigned 51-70 points. All other patients receive up to 50 points according to the following criteria: "visual acuity score" (max 5 points), "clinical modifiers" (max 5 points), "severity of visual impairment" (max 10 points), "ability to work, give care, live independently" (max 5 points), "other disability" (max 5 points). Patients with the maximum number of points wait four weeks, while patients with only 20 points wait six months. Patients scoring less than 20 are "deferrable" (i.e. they are not offered treatment and are sent back to the GP for "active care and review"). Similar criteria have been developed for hip and knee replacement, and other common elective procedures.[1]

We examine the welfare effects of two possible prioritisation schemes: 1) linear prioritisation, under which a patient's waiting time is linearly related to their observable benefit; 2) threshold rationing, under which all patients whose observable benefit exceeds a threshold are offered immediate treatment and all those with lower observable benefit are subject to linear prioritisation.

[1] For more detailed information see www.electiveservices.govt.nz/guidelines.html. Similarly, the Western Canada Waiting List Project has developed priority criteria for general surgery, cataract, hip and knee replacement. For hip and knee replacement, criteria include: "pain and motion" (0-13 points), "pain at rest" (0-11 points), "ability to walk without significant pain" (0-7points), "other functional limitations" (0-19 points), "potential for progression of disease" (0-20 points). For more details see www.wcwl.org. There are similar schemes in many countries for allocating public sector housing at below market clearing rents.

If patient benefit from a treatment is perfectly observable by providers, rationing by waiting is inefficient since it imposes costs on patients which are not offset by gains to producers. Longer waiting times reduce the value of a treatment because of lost expected benefit, temporary discomfort and pain, and, for some pathologies, the higher risk of permanent reductions in health. With perfect information providers can treat high-benefit patients with no costly wait and refuse treatment to those with low benefit.

But patient benefit from treatment is not fully observable by providers because of unobservable patient preferences or characteristics. Patients know their characteristics and, after consultation with their medical advisors, are better informed about their benefit than providers. They use their information in deciding whether to join the waiting list. When there is no prioritisation all patients face the same waiting time and so the patients who join the list and get treatment have higher benefits than those who do not join the waiting list.

Providers can usually observe some characteristics of the patient, such as age, which convey information about the benefit. Suppose that benefit is negatively correlated with age. The provider can prioritise patients by offering lower waiting times for younger patients. Prioritisation increases the number of young patients treated and reduces the number of old patients treated. The welfare of the young is increased and the welfare of the old is reduced. But since some young patients may have low benefit, and some old patients may have high benefit, and prioritisation reallocates treatment to the young, it is unclear a priori whether prioritisation based on the imperfect signal increases overall welfare compared to no prioritisation.

We show that linear prioritisation improves welfare if the average age of patients who are indifferent between obtaining the treatment after some wait or not obtaining the treatment at all, is higher than the average age of the patients receiving treatment. The reason is that prioritisation changes the total time waited and leads to the treatment of more young people and fewer old people. Treating more young and fewer old has no welfare effect since the marginal old and marginal young patients are indifferent about treatment taking into account their waiting time. Hence prioritisation is welfare increasing if it reduces the total waiting time of the infra-marginal patients. Since waiting time increases with age in our example, total waiting time is reduced if the mean age of the treated is less than the mean age of the indifferent.

The condition that the average age of patients who are indifferent between getting treatment or not, is higher than the average age of the patients receiving treatment is satisfied by several distributions of age and unobserv-

able benefits. Examples include (i) a uniform distribution of the unobservable benefit and (ii) a log concave distribution of unobservable benefit and a uniform distribution of the observable benefit.[2]

We show that linear prioritisation is not welfare increasing if and only if the unobservable component of health gain has a negative exponential distribution. With a negative exponential distribution of the unobservable component, the mean age of patients who are indifferent about joining the waiting list is the same as the mean age of those who do join.

Under *threshold* rationing patients whose observable benefit exceeds a threshold are offered immediate treatment and all those with lower observable benefit are subject to waiting-time prioritisation. We show that, if for a given threshold the minimum wait is zero, an increase in the threshold is welfare improving.

The main results on the effects of prioritisation hold when there is also a private sector alternative providing treatment at a money price but with a zero wait. The rationale is that marginal changes in the prioritisation regime in the public health sector alter individuals' choices between treatment in the public sector, treatment in the private sector, and no treatment. But the welfare loss for these marginal individuals who shift is zero since they were indifferent between their choice before and after the change in the public sector prioritisation rule. Hence the welfare effect of the change is the change in the total time waited by patients in the public sector and so the results for the simpler case where the only alternative to public treatment is no treatment continue to apply.

Most of the theoretical literature on rationing by waiting assumes that all patients have the same waiting time (Lindsay and Feigenbaum, 1984; Bucovetsky, 1984; Hoel and Saether, 2003; Farnworth, 2003; Gravelle, Dusheiko and Sutton, 2002; Iversen, 1997; Martin and Smith, 1999; Siciliani, 2005).[3] An exception is Barros and Olivella (2005) in which the public sector does not treat patients with benefit below a threshold level and patients with benefit above the threshold will wait for treatment in the public sector. However, the threshold is treated as exogenous as the focus is on the incentives for doctors who work in both the private and public sectors to creamskim and the question of whether prioritisation is welfare enhancing is not considered.

Our focus is on the optimal way to use waiting times to allocate a fixed supply of a particular treatment amongst potential patients. It is thus com-

[2] A distribution is log-concave if the hazard rate is monotonically increasing, which is satisfied by many common distributions like the Normal, the Chi-square and Gamma distribution (Bagnoli and Bergstrom, 2005).

[3] For a review of the literature see Cullis, Jones and Propper (2000).

plementary to two related literatures on the optimal allocation of a given health care budget across different treatments (Garber, 2000; Gravelle and Siciliani, 2007b; Smith 2005) and on the use of waiting time and money prices to ration access to treatment (Bucovetsky, 1984; Gravelle and Siciliani, 2007a; Hoel and Saether, 2003; Marchand and Schroyen, 2005; and Olivella, 2003).

Section 2 presents the main features of the model and establishes the effects of prioritisation on welfare; section 3 extends the analysis to threshold rationing; section 4 contains an illustration with uniform distributions. Section 5 shows that the main results hold when there is a private sector alternative to treatment in the public sector, when the social welfare function weights the utility of individuals by their observable benefit, and when the cost of treatment varies with the observable part of patient benefit. Section 6 concludes.

2 Linear prioritisation

2.1 Model specification

We initially assume that the alternative to treatment in the public sector is no treatment. Section 5 considers the case where there is also private sector treatment.

A public sector provider wishes to allocate a fixed supply z of treatment to a population of potential patients to maximise welfare. The health gain from treatment is $b - a$, where a is observable by the provider but b is not. To fix ideas we interpret a as age: old patients have lower health gain, given their other characteristics which determine b, than young patients. b and a are distributed according to the joint density $f(a, b)$ over the support $a \in [0, \overline{a}]; b \in \left[0, \overline{b}\right]$.

Patient utility with treatment is $b - a - w$ where w is waiting time.[4] Patients, possibly after consulting their general practitioner, know both their b and a and join the waiting list if $b - a \geq w$.

Our specification differs from the original formulation of the model of rationing by waiting by Lindsay and Feigenbaum (1984) in two respects. In their model (and in others such as Farnworth (2003)) patients have a cost $c > 0$ of getting on a waiting list for treatment. For example, a patient may incur a cost of attending an outpatient clinic to see a specialist who will

[4] A more general specification $b - a - kw$, where k can be interpreted as the marginal disutility of waiting does not alter the results. A lower k implies a higher waiting time in equilibrium, so that total disutility from waiting is unaffected.

agree that they need treatment and place them on the waiting list. Lindsay and Feigenbaum assume that the benefit from an operation received after a wait of w has a present value of be^{-rw} and that the patient decides to join the list if and only if $be^{-rw} \geq c$. With $c > 0$, increases in the waiting time reduce the demand because the present value of the benefit from treatment is reduced relative to cost of joining the list. If c is zero all patients with $b > 0$ join the waiting list and demand is unaffected by the waiting time. Assuming that there is a positive cost of joining a list which must be incurred before the health benefits are realised seems plausible for rationing by list for elective hospital treatment where patients must first be seen by a specialist in order to join the list.

The evidence suggests that many individuals do not use exponential discounting of health (Cairns and van der Pol, 2000) but instead use a variety of discounting functions, including hyperbolic discounting (Loewenstein and Prelec, 1992) where the discounted health benefit is $b(1 + \alpha w)^{-\beta/\alpha}$.

We assume utility is linear in the waiting time since exponential or hyperbolic discounting do not yield tractable models. Moreover, in the hyperbolic case, the discounted benefit from treatment is quite well approximated by a linear function for positive waiting time as α becomes large. But when utility from treatment is linear in the waiting time, c has no role in the model and we drop it to reduce notational clutter. Thus we assume that b is the unobserved benefit from treatment net of the cost of joining the list $b \equiv b' - c$. b' is distributed between c and some upper limit $\bar{b}'$, so that b is distributed between 0 and $\bar{b} = \bar{b}' - c$.

The linear specification means that the model also applies to the case of rationing by waiting in line (queuing), as for example in accident and emergency departments and GP surgeries. Potential patients will baulk at the queue if it is so long that their benefit from treatment is less than the opportunity cost of the time they would have to spend in the queue (which we assume is proportional to the waiting time). In this case we also do not need to assume there is a cost to joining the queue and can again set $c = 0$.

Since a is observable it is possible to make waiting time of a patient depend on a. We consider a linear points rationing scheme in which patients of age a receive $-p_2 a$ points when placed on the waiting list and then accumulate points at the rate $p_1 w$ the longer they wait.[5] We assume that $p_1 > 0$; otherwise waiting time would have no interesting role in the scheme. There

[5] A more general (nonlinear) prioritisation rule could be in principle be derived by chosing a general function $w(a)$ to maximise the welfare function. However, the resulting optimal control problem is ill-behaved.

is no prioritisation according to age when $p_2 = 0$. If $p_2 > 0$ there is positive prioritisation, so that those with a higher observed benefit (lower age) have a shorter waiting time. It is possible in principle to have negative prioritisation ($p_2 < 0$) so those with a lower observed benefit (higher age) have a shorter waiting time. We focus on the more intuitive case of positive prioritisation, mainly using the logical possibility of negative prioritisation to show that, except in one very special case, some form of prioritisation (positive or negative) is welfare increasing. We do however provide in section 4 an example of negative prioritisation being welfare increasing.

In each period patients with the highest number of points are treated. Patients who get treatment will have accumulated the same number of points P, so that $p_1 w - p_2 a = P$. Hence the waiting time for a patient of age a is $[P + p_2 a]/p_1 = w_0 + pa$ where $w_0 = P/p_1$ is the minimum waiting time and $p = p_2/p_1$. Only patients whose health gain, $b - a$, is at least $w_0 + pa$ will join the waiting list.

Define $\widehat{b}(a; w_0, p) = a(1 + p) + w_0$ as the critical level of private benefit, which makes a patient of given age indifferent between receiving the treatment or not. Define $\widehat{a} = \min\{\frac{\bar{b} - w_0}{1+p}, \bar{a}\}$ as the maximum age at which a patient will want to join the waiting list. The demand for treatment is (see Gravelle and Siciliani, 2006; Appendix)

$$D(w_0, p) = \int_0^{\widehat{a}} \int_{\widehat{b}(a; w_0, p)}^{\bar{b}} f(a, b) db da \tag{1}$$

Since $\widehat{b}(a; w_0, p) = a(1 + p) + w_0$ an increase in p or in w_0 will increase the minimum b at which the patient wants to join the waiting list. Hence increases in p or in w_0 will reduce demand:

$$D_p = -\int_0^{\widehat{a}} a f(a, \widehat{b}\,(a; w_0, p)) da < 0, \quad D_{w_0} = -\int_0^{\widehat{a}} f(a, \widehat{b}(a; w_0, p) da < 0 \tag{2}$$

The system is in equilibrium when the number of patients joining the list in each period equals the supply of treatment:

$$D(w_0, p) - z = 0 \tag{3}$$

which yields the minimum equilibrium wait (for the youngest patients) as $w_0 = w_0(p)$, where we suppress the dependence of w_0 on supply z to reduce notational clutter. Increases in p reduce the minimum wait: $\partial w_0 / \partial p =$

$-D_p/D_{w_0} < 0$. The wait for a patient of age a is $w(a;p) = w_0(p) + pa$. When $p = 0$ there is no prioritisation and all patients who join the list have the same wait $w_0(0)$ irrespective of age.

Figure 1 illustrates. If there was no waiting time all patients with $b \geq a$ would join the list and demand would be equal to the mass of patients above the 45^o line. We assume that demand at zero waiting time always exceeds the supply z. With no prioritisation there is a positive waiting time of $w_0(0)$ = OA for patients of all ages and only those with $b - w_0(0) \geq a$ join the list. Thus demand would be equal to the mass of patients in ACH. With positive prioritisation only patients with $b - w_0(p) \geq a(1+p) > a$ join the list. The minimum wait (enjoyed by patients with $a = 0$) is $w_0(p)$ = OD, and demand is the mass of patients in DEH.

The effect of an increase in p on the waiting time of a patient of age a is

$$\frac{dw(a;p)}{dp} = a + \frac{\partial w_0}{\partial p} = a - \frac{\int_0^{\widehat{a}} af(a, w_0(p) + a(1+p))da}{\int_0^{\widehat{a}} f(a, w_0(p) + a(1+p))da} = a - E(a \mid I) \quad (4)$$

where $E(a \mid I)$ is the mean age of the patients who are indifferent about joining the list. An increase in p increases the wait of older patients ($a > E(a \mid I)$) and reduces the wait of younger patients ($a < E(a \mid I)$).

The effect of increased prioritisation on the total time waited by patients on the list is

$$\begin{aligned}
&\frac{d}{dp}\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} [w_0(p) + pa](a;p)f(a,b)dbda \\
&= -\int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} E(a \mid I)f(a,b)dbda + \int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} af(a,b)dbda - \int_0^{\widehat{a}} w(a,p)\frac{d\hat{b}}{dp}f(a,\hat{b})da \\
&= -D(w_0,p)E(a \mid I) + \int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} af(a,b)dbda - \int_0^{\widehat{a}} w(a,p)\frac{d\hat{b}}{dp}f(a,\hat{b})da \\
&= D(w_0,p)[E(a \mid T) - E(a \mid I)] - \int_0^{\widehat{a}} w(a,p)\frac{d\hat{b}}{dp}f(a,\hat{b})da \qquad (5)
\end{aligned}$$

where

$$E(a \mid T) = \int_0^{\hat{a}} \int_{\hat{b}}^{\bar{b}} af(a,b)dbda/D \tag{6}$$

is the mean age of patients who demand treatment.

The first term in the square bracket in the last line of (5) is the effect of prioritisation on the waiting time of infra marginal patients, and the second the effect on the waiting times of patients who were indifferent. Thus increased prioritisation (an increase in p) may increase or reduce the total time waited by patients on the list.

2.2 Welfare and prioritisation

With fixed supply the utilitarian welfare function is

$$S(w_0(p), p) = \int_0^{\widehat{a}} \int_{\widehat{b}}^{\bar{b}} [b - a - w(a,p)] f(a,b)dbda \tag{7}$$

The policy problem is to choose p to maximise S subject to the constraint that the minimum wait cannot be negative: $w_0 \geq 0$. Since $\partial w_0/\partial p < 0$, the constraint $w_0 \geq 0$ can be rewritten as $\widetilde{p} - p \geq 0$ (where $w_0(\widetilde{p}) = 0$). Without further restrictions on the distribution of private and observable benefit the welfare function may not be concave in p so that first order conditions are necessary rather than sufficient (see Gravelle and Siciliani, 2006, Appendix). There are four logically possible solution types:

a) *Maximal prioritisation ($p^* = \widetilde{p} > 0$ and $w_0(p^*) = 0$).* p is set at the highest possible level and the minimum wait is equal to zero.

b) *Partial prioritisation ($p^* > 0$ and $w_0(p^*) > 0$).* Whilst the young have shorter waits, even the youngest have a positive wait.

c) *No prioritisation ($p^* = 0$ and $w_0(0) > 0$).* All patients have the same wait.

d) *Negative prioritisation* ($p^* < 0$). Younger patients have longer waits.

Negative prioritisation is counter-intuitive and we concentrate on the other cases. We do however provide in section 4.3 an example where negative prioritisation is welfare increasing.

Figure 1 illustrates the effect of positive prioritisation on the demand for treatment and on total waiting time. We assume in Figure 1 that the distributions of a and b are independent and uniform. When there is no prioritisation all potential patients face the same waiting time of OA = $w_0(0)$. The line AC with slope of 1 plots $b = w_0(p) + a$ and patients in

the area ACH where with $b \geq w_0(p) + a$ demand treatment. These are the patients with the highest health benefit $b - a$. If private benefit b were observable it would be optimal to provide treatment, with a zero wait, only to the patients in area ACH and to refuse treatment to all other patients. Thus rationing by waiting with no prioritisation ensures an optimal selection of patients for treatment since the same patients would be treated under full information and with no prioritisation. But with no prioritisation welfare is lower than under full information because of the waiting time costs imposed on the patients in ACH.

With prioritisation all patients whose unobservable benefit b exceeds $w_0(p) + a(1 + p)$ join the waiting list for treatment. The line DE with slope $(1 + p)$ plots $\hat{b} = w_0(p) + a(1 + p)$ and OD $= w_0(p)$. All patients in area DEH demand treatment. Prioritisation increases demand from young patients (with age $a < a_B$) since those in area ABD now join the waiting list. Conversely old patients ($a > a_B$) in area BCE who used to demand treatment now do not join the list. The mass of patients in ABD equals that in BCE since total demand is unchanged. In Figure 1 ABD and BCE have the same area because b and a are independently and uniformly distributed.

With prioritisation the waiting time $(w_0 + pa)$ for a patient aged a is the vertical distance between the 45^o line and the line DE plotting $\hat{b} = a + w_0 + pa$. Thus total time waited with prioritisation is the area ODEF. With no prioritisation total time waited is the area OACG. Area OACG equals ODECG since ABD equals BCE. Since ODECG exceeds ODEF by CEFG, total time waited is reduced by prioritisation. However, the new young patients in ABD have smaller health gains $b - a$ than the displaced old patients in BCE, so that the welfare effect of prioritisation is a priori ambiguous.

A more formal analysis is thus required to investigate the effect of prioritisation. The marginal welfare effect of introducing prioritisation is, using (5) and remembering that $\hat{b} = a + w(a, p)$,

$$\frac{dS(w_0(p), p)}{dp} = \frac{d}{dp}\int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} [b - a] f(a, b) dbda - \frac{d}{dp}\int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} w(a, p) f(a, b) dbda$$

$$
\begin{aligned}
&= -\int_0^{\widehat{a}} [\hat{b} - a] \frac{d\hat{b}}{dp} f(a, \hat{b}) da - D(w_0, p)[E(a \mid T) - E(a \mid I)] \\
&\qquad + \int_0^{\widehat{a}} w(a, p) \frac{d\hat{b}}{dp} f(a, \hat{b}) da \\
&= D\left[E(a \mid I) - E(a \mid T)\right] \qquad (8)
\end{aligned}
$$

Hence we have

Proposition 1 *Starting from a regime with no prioritisation, positive (negative) prioritisation is welfare improving if the average age of the patients who are indifferent between receiving treatment or not, $E(a \mid I)$, is larger (smaller) than the average age of the patients receiving treatment, $E(a \mid T)$. Similarly, if prioritisation is already positive, intensifying (dampening) it is welfare improving if $E(a \mid I) > (<) E(a \mid T)$.*

Varying p has two effects: it changes the total waiting time and it changes the age mix of those treated via the changes in the critical benefit $\hat{b}$ at which individuals of given age seek treatment. The change in the age mix of the treated has no welfare effect since the marginal old and marginal young patients are indifferent about treatment taking into account their waiting time: the first and last terms in the second line of (8) cancel.

Rationing by waiting imposes deadweight losses on patients in order to allocate scarce capacity. Prioritisation increases welfare because it makes the marginal indifferent patients more sensitive to waiting times than the inframarginal. Thus a smaller deadweight loss is required to limit demand. When indifferent patients are older than inframarginal treated patients making the old wait longer will increase the effect of waiting time on demand and so reduce deadweight loss from waiting.

In Figure 1, where a and b are independently uniformly distributed, the average age of treated patients in ACH is less than the average age of the indifferent on AC because the young are more likely to demand treatment. Hence introducing positive prioritisation ($p > 0$: shorter waits for the young) will reduce total waiting time and increase welfare.

It is only under very special assumptions about the distribution of observable and unobservable components of health gain that (positive or negative) prioritisation is never welfare improving:

Proposition 2 *Prioritisation is welfare increasing at some supply level unless $f(a, b) = h(a)\gamma e^{-\gamma b}$.*

When there is no prioritisation the market clearing waiting time faced by all patients of all ages is $w_0(0,z)$ (where we temporarily show the dependence of w_0 on supply). Since, from (8),

$$dS(w_0(p,z),p)/dp = D(w_0(p,z),p)\left[E(a \mid I) - E(a \mid T)\right] \tag{9}$$

welfare can be increased by either a small positive or negative p unless $E(a \mid I) = E(a \mid T)$. Indifferent potential patients lie along the line AC in Figure 1 where $\hat{b} = a + w_0(p,z)$. If the mean age of patients along this line is independent of w_0 (and hence of z) then the mean age of the patients with $b > \hat{b}$ who demand treatment is also independent of w_0 and is equal to the mean age of the indifferent patients. The negative exponential distribution ensures that this special type of invariance of the distributions of age conditional on w_0 holds:

$$E(a \mid I) = \frac{\int_0^{\bar{a}} af(a, a+w_o)da}{\int_0^{\bar{a}} f(a, a+w_o)da} = \frac{\int_0^{\bar{a}} ah(a)\gamma e^{-\gamma a}e^{-w_o}da}{\int_0^{\bar{a}} h(a)\gamma e^{-\gamma a}e^{-w_o}da} = \frac{\int_0^{\bar{a}} ah(a)\gamma e^{-\gamma a}da}{\int_0^{\bar{a}} h(a)\gamma e^{-\gamma a}da}$$

and it is the only distribution with this property.

We can find assumptions about the distribution of a and b which ensure that $E(a \mid I) > E(a \mid T)$ holds for all p, so $dS/dp > 0$ for all p, and therefore maximal prioritisation is optimal. Hence, in this case the more intuitively plausible positive form - shorter waits for the young - maximises welfare.

Proposition 3 *Suppose that a and b are independently distributed over the rectangular support $a \in [0,\overline{a}]; b \in \left[0,\overline{b}\right]$ so that $f(a,b) = h(a)g(b)$. Then maximal prioritisation is optimal ($p^* = \widetilde{p}$ and $w_0^* = 0$) if (i) $h(a)$ is uniform and $g(b)$ is log-concave; or if (ii) $g(b)$ is uniform (for any $h(a)$).*

The requirement of independence of a and b may appear to be restrictive but note that it relates to the two components of health gain, not to the joint distribution of health gain and its observable component. The covariance of a and health gain $b-a$ is $Cov(b-a,a) = Cov(b,a) - Var(a)$ so that even if b and a are uncorrelated a is informative about health gain.

If $h(a)$ is uniform, then the average ages of treated patients and indifferent patients are

$$E(a \mid T) = \int_0^{\widehat{a}} a \frac{G(\overline{b}) - G\left(\widehat{b}(a)\right)}{D} da, \qquad E(a \mid I) = \int_0^{\widehat{a}} a \frac{g(\widehat{b}(a))}{G(\overline{b}) - G\left(w_0\right)} da \tag{10}$$

Intuitively, since $G(\bar{b}) - G\left(\widehat{b}\right)$ is decreasing in a while $g(\widehat{b})$ can take any shape, it follows that the first distribution tends to give a higher weight to patients with low age and it tends to be less favourable compared to $g(\widehat{b})$. The log-concavity of $g(b)$ ensures that this is always the case. Log-concavity is satisfied by many common distributions like the Normal, Chi-squared and Gamma distribution.

If $g(b)$ is uniform then the average age of the *population* in $a \in [0, \hat{a}]$, $b \in [0, \bar{b}]$ is equal to the average age of the *indifferent* patients: $E(a \mid a \in [0, \hat{a}]) = E(a \mid I)$.[6] The average age of the patients *treated* at given b cannot exceed the average age of the population in $a \in [0, \hat{a}]$ at the same b

$$E(\,a|\,a \in [0, \min\{(b - w_0)/(1+p), \hat{a}\}], b) \leq E(\,a|\,a \in [0, \hat{a}], b) = E(a \mid a \in [0, \hat{a}])$$

where the inequality is strict for $b \in [w_0, (1+p)\hat{a}+w_0]$. Hence the average age of the treated over $b \in [w_0, \bar{b}]$ is less than the average age of the population in $a \in [0, \hat{a}]$, and so is less than the average age of the indifferent: $E(a \mid T) < E(a \mid a \in [0, \hat{a}]) = E(a \mid I)$ (see Gravelle and Siciliani, 2006, Appendix, for full details).

3 Threshold rationing

We now consider a more general threshold rationing system. Patients whose age is below a given threshold a_0 (i.e. young patients) are given immediate treatment with zero waiting time. Patients with $a \geq a_0$ are prioritised (as in the previous section) and wait according to $w(a, p, a_0) = w_0 + p\,(a - a_0)$. The previous analysis had $a_0 = 0$ so that threshold rationing is more general in that it allows younger patients to receive treatment without delay. It includes the case in which the threshold is set so high that all the supply is used to treat young patients with no delay and old patients are denied treatment (face an infinite waiting time).

For patients with $a \geq a_0$, the lowest unobserved benefit at which they will demand treatment is $\widehat{b}(a; w_0, p, a_0) = a + p(a - a_0) + w_0$, and $\widehat{a} = \min\{(\bar{b} - w_0 + p a_0)(1+p)^{-1}, \bar{a}\}$ is the highest age. The demand for treatment

[6] If a and b are independently distributed, then: $E(a \mid a \in [0, \hat{a}], b \in [0, \bar{b}]) = \int_0^{\hat{a}} \int_0^{\bar{b}} a h(a) g(b) db da \Big/ \int_0^{\hat{a}} \int_0^{\bar{b}} h(a) g(b) db da = \int_0^{\hat{a}} a h(a) da \Big/ \int_0^{\hat{a}} h(a) da$. If $g(b)$ is uniform, $E(\,a|\,I)$ $= \int_0^{\hat{a}} a h(a) g(\widehat{b}(a)) da \Big/ \int_0^{\hat{a}} h(a) g(\widehat{b}(a)) da = \int_0^{\hat{a}} a h(a) da \Big/ \int_0^{\hat{a}} h(a) da = E(a \mid a \in [0, \hat{a}], b \in [0, \bar{b}])$.

is the sum of demands from those who do not have to wait ($a < a_0$) and those who do:

$$D(w_0, p, a_0) = \int_0^{a_0} \int_a^{\overline{b}} f(a,b)dbda + \int_{a_0}^{\widehat{a}} \int_{\widehat{b}}^{\overline{b}} f(a,b)dbda \tag{11}$$

where $D_{a_0} > 0$, $D_p < 0$ and $D_{w_0} < 0$. The market-clearing condition is $D(a_0, p, w_0) = z$, from which we obtain $w_0 = w_0(p, a_0)$.

In Figure 2, under maximal prioritisation with a zero threshold ($p = p^*$, $w_0 = 0$, $a_0 = 0$) only patients in area OAB demand treatment (and wait $w(a, p, a_0) = p^* a$). If the threshold a_0 is set so that the patients with zero wait get all the supply, then patients in area OCDB are treated.

Welfare is

$$S(a_0, p, w_0) = \int_0^{a_0} \int_a^{\overline{b}} (b-a) f(a,b)dbda + \int_{a_0}^{\widehat{a}} \int_{\widehat{b}}^{\overline{b}} (b-\widehat{b}) f(a,b)dbda \tag{12}$$

with $S_{a_0} > 0$, $S_{w_0} < 0$ and $S_p < 0$ (see Gravelle and Siciliani, 2006, Appendix). For fixed a_0 the optimal p^* maximises the second term in (12) subject to $w_0(p, a_0) \geq 0$ or equivalently $\tilde{p}(a_0) \geq p$, with $\partial \tilde{p}/\partial a_0 = -D_{a_0}(w_0, p, a_0)/D_p(w_0, p, a_0)$. The Lagrangean is

$$L = S(w_0(p, a_0), p, a_0) + \lambda[\tilde{p}(a_0) - p]$$

All the previous results apply to this first stage problem. Once we have optimised with respect to p, we obtain

$$S^* = S(w_0(p^*(a_0), a_0), p^*(a_0), a_0)$$

Using the envelope theorem,

$$\frac{dS^*}{da_0} = \frac{dL^*}{da_0} = \frac{\partial S}{\partial w_0}\frac{\partial w_0}{\partial a_0} + \frac{\partial S}{\partial a_0} + \lambda^* \frac{\partial \tilde{p}}{\partial a_0} \tag{13}$$

If prioritisation is maximal then the minimum wait is zero ($w_0^* = 0$) and

the first term in (13) is zero, and $p^* = \tilde{p}(a_0)$. Hence,

$$\begin{aligned}\frac{dS^*}{da_0} &= \frac{\partial S}{\partial a_0} + \lambda^* \frac{\partial \tilde{p}}{\partial a_0} = \frac{\partial S}{\partial a_0} - \frac{\partial S}{\partial p}\frac{D_{a_0}}{D_{\tilde{p}}} \\ &= p\int_{a_0}^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} f(a,b)dbda - \int_{a_0}^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} af(a,b)dbda \frac{p\int_{a_0}^{\widehat{a}} f(a,\widehat{b}\,(a))da}{\int_{a_0}^{\widehat{a}} af(a,\widehat{b}\,(a))da} \qquad (14)\end{aligned}$$

which can be rearranged as:

$$\frac{dS^*}{da_0} = p\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} af(a,b)dbda \frac{E(a \mid I, a \geq a_0) - E(a \mid T, a \geq a_0)}{E(a \mid I, a \geq a_0) \times E(a \mid T, a \geq a_0)} \qquad (15)$$

$E(a \mid I, a \geq a_0)$ and $E(a \mid T, a \geq a_0)$ are the average age of patients who are indifferent (I) or treated (T) and have an age at least as high as the threshold a_0. Suppose that, for a given a_0, optimal prioritisation is maximal ($p^* = \tilde{p}(a_0)$ *and* $w_0^* = 0$). This implies that $E(a \mid I, a \geq a_0) \geq E(a \mid T, a \geq a_0)$ so that $dS^*/da_0 \geq 0$ and increasing the threshold will increase welfare.

If prioritisation is not maximal for a given a_0, the constraint is not binding, $p^* < \tilde{p}$, $w_0^* > 0$ and $\lambda^* = 0$. Hence, (see Gravelle and Siciliani, 2006, Appendix)

$$\begin{aligned}\frac{dS^*}{da_0} &= \frac{\partial S}{\partial a_0} + \frac{\partial S}{\partial w_0}\frac{\partial w_0}{\partial a_0} = \frac{\partial S}{\partial a_0} - \frac{D_{a_0}}{D_{w_0}}\frac{\partial S}{\partial w_0} \\ &= \int_{a_0}^{a_0+w_0^*} (b - a_0)f(a_0,b)db + w_0^* \int_{a_0+w_0^*}^{\overline{b}} f(a_0,b)db \\ &\quad - \int_{a_0}^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}} f(a,b)dbda \frac{\int_{a_0}^{a_0+w_0^*} f(a_0,b)db}{\int_{a_0}^{\widehat{a}} f(a,\widehat{b})da} \qquad (16)\end{aligned}$$

which is indeterminate as the first two terms are positive while the third term is negative.

Summarising:

Proposition 4 *If at threshold a_0 the optimal prioritisation is maximal ($p^* = \tilde{p}(a_0)$ and $w_0^*(p^*) = 0$) and $E(a \mid I, a \geq a_0) > E(a \mid T, a \geq a_0)$ then welfare can be increased by increasing the threshold.*

The proposition is only local but it has the corollary that a sufficient condition for a positive threshold to be optimal is that when there is no threshold ($a_0 = 0$), the optimal degree of prioritisation is maximal ($p^* = \tilde{p}(0)$ *and* $w_0^*(p^*) = 0$) and the average age of the indifferent is greater than the average age of the treated: $E(a \mid I, a \geq a_0 = 0) > E(a \mid T, a \geq a_0 = 0)$.

4 An illustration: uniform distributions

We now use the simple cases with uniform distributions to illustrate the gains from prioritisation and threshold rationing, and to provide an example of negative prioritisation.

4.1 Prioritisation

Assume that a, b are independently and uniformly distributed with support $a \in [0, 1]$ and $b \in [0, 1]$. Refer to Figure 3. Prioritisation benefits all the young patients in area DBFH: those in ABFH were treated before but now have a shorter wait. The gain from prioritisation for a patient of age a in ABFH who was previously treated is the vertical distance between AB and DB at a. Young patients in ABD were not previously treated but now join the waiting list because they face a shorter waiting time. The average gain from prioritisation for them is given by the vertical distance between the lines PB and DB at that age (where the distance AP = DP because of the uniform distribution of b).

Old patients in BCF lose. Those with $a \in (a_B, a_E]$ in area BFE still join the list but now face a longer wait. Those with $a \in (a_B, a_E]$ in area BLE do not join the list. Their average loss at given a is the vertical distance between BI and BL at that a. Old patients with $a \in (a_E, a_C]$ previously joined the list and were treated but now no longer join the list because of the increased wait they would face. The average loss from prioritisation for displaced old patients of age $a \in (a_E, a_C)$ is the vertical distance between IC and LC at that age.

First, consider the young patients aged $a = a_B - \delta$ in BFGJ and old patients aged $a_B + \delta$ in BEF, where $\delta \in (0, a_E - a_B]$. By similar triangles the loss by an old patient aged $a_B + \delta$ equals the gain of a young patient at $a_B - \delta$ but there are more young patients at any given δ in BFGJ than old patients in BEF so the total gain to young patients in BFGJ exceeds the total loss of old patients in BEF. Second, compare the displaced old patients in BLE aged $a_B + \delta$ and new young patients in BJK aged $a_B - \delta$ where GF equals FE. The average gain to the young patient at $a_B - \delta$ equals the average loss by the old patient at $a_B + \delta$. By similar triangles there are the same number of new young patients at $a_B - \delta$ and displaced old patients at $a_B + \delta$. Hence the gains and losses for these two groups cancel. Third, consider the displaced old patients in CEL and the new young patients in DKJA. Since the total demand is unchanged and area BKJ equals area BLE, we must have the same number of new young patients in DKJA as displaced old patients in CEL. The maximum average loss at any given age for old patients in CEL is the distance IL which is equal to the minimum average gain for new young patients in DKJA. Hence the total gain to new young patients in DKJA exceeds the total loss to displaced old patients in CEL. Finally, prioritisation makes old patients in AJGH better off and there are no old patient losses unaccounted for. Thus, prioritisation increases welfare in the case of independent uniform distributions of a and b.

More formally, if there is no threshold rationing ($a_0 = 0$), the demand function is (1) and

$$D(w_0, p) = (1 - w_0)^2 / 2(1 + p) \tag{17}$$

In equilibrium demand equals supply z, so that

$$w_0(p) = 1 - [2z(1 + p)]^{1/2} \tag{18}$$

Welfare is

$$S(w_0, p) = (1 - w_0(p))^3 / 6(1 + p) \tag{19}$$

Prioritisation is welfare improving since

$$\frac{dS}{dp} = \frac{\partial S}{\partial w_0}\frac{dw_0}{dp} + \frac{\partial S}{\partial p} = \frac{(1 - w_0)^3}{12(1 + p)^2} > 0 \tag{20}$$

Proposition 5 *If a, b are independently and uniformly distributed and there is no threshold ($a_0 = 0$), optimal prioritisation is maximal with $p^* = \frac{1}{2z} - 1$, $w_0^* = 0$, $w^*(a, p^*, z) = \left(\frac{1}{2z} - 1\right) a$.*

Prioritisation reduces the total waiting time of infra-marginal patients but also reduces the average health gain $(b-a)$ of treated patients. If the densities of observed and observed benefit are uniform, the first effect dominates: the higher welfare from a reduction in waiting times outweighs the reduction in welfare from suboptimal selection of patients.

4.2 Threshold rationing

Suppose now that the threshold is positive ($a_0 > 0$) so that patients with $a < a_0$ are given immediate treatment with zero waiting time and patients with $a \geq a_0$ wait according to $w(a) = w_0 + p\,(a - a_0)$. The demand function (11) is

$$D = \frac{(1 - w_0 - a_0)^2}{2\,(1+p)} + a_0 - \frac{a_0^2}{2} \tag{21}$$

and the minimum wait is

$$w_0 = 1 - a_0 - \left\{[2(1+p)][z + a_0^2 2^{-1} - a_0]\right\}^{1/2} \tag{22}$$

and welfare (12) is

$$S = \frac{(1 - w_0 - a_0)^3}{6(1+p)} + \frac{1}{2}a_0 - \frac{1}{2}a_0^2 + \frac{1}{6}a_0^3 \tag{23}$$

Substituting (22) in S, the derivative of S with respect to p is positive and so, for a given a_0, prioritisation is again maximal with

$$p^* = \tilde{p}(a_0) = \frac{1 - 2z}{2z - 2a_0 + a_0^2} > 0 \tag{24}$$

Note that the pure prioritisation solution is a special case of this with $a_0 = 0$. Substituting $\tilde{p}(a_0)$ and $w_0(a_0, \tilde{p}(a_0))$ in the welfare function and differentiating with respect to the threshold we can show (Gravelle and Siciliani, 2006, Appendix) that increases in a_0 increase welfare. The optimal policy is to increase a_0 until all of the available capacity is allocated to young patients with $a < a_0^*$, and old patients with $a \geq a_0^*$ demand none of the capacity because they face an infinite waiting time. Using the market clearing condition to solve for a_0^* we have

Proposition 6 *If a, b are independently and uniformly distributed, it is always optimal to provide the treatment only to patients for whom a is below the threshold level a_0^*, where $a_0^* = 1 - (1-2z)^{1/2}$.*

Figure 4 compares welfare at various capacity levels for the first best, simple rationing by waiting where all patients have the same wait, waiting with optimal prioritisation, and threshold rationing. We express the capacity as a proportion of the maximum potential demand with no waiting time or prioritisation, which is (see the demand functions) 1/2.

The waiting time and welfare under simple rationing by waiting are found by setting $p = 0$ in (18) and (19). In the first best providers can observe both b and a and allocate treatment to those with the highest health gain (b) until capacity is exhausted. The patients treated would be the same as those under simple rationing by waiting, who join the list by comparing their health gain with the waiting time. Hence the first best welfare is the welfare under simple rationing plus the avoided total waiting time of $w_0(0)D(w_0(0), 0)$.

Figure 4 shows that the form of rationing makes more difference the smaller the available capacity relative to the maximum potential demand. For example, with capacity equal to 25% of potential demand, simple rationing by waiting produces about one half of the welfare from prioritisation and about one third of the welfare from threshold rationing. But with capacity of 75%, simple rationing by waiting produces over 4/5th of the welfare from prioritisation and 2/3rd of the welfare from threshold rationing. Notice also that the form of rationing affects the marginal value of additional capacity (the slope of the welfare functions with respect to capacity). The marginal value of capacity is increasing under simple rationing by waiting, constant under prioritisation, and decreasing under pure threshold rationing and the first best.

4.3 Negative prioritisation

We now give a simple example of negative prioritisation. The density $f(a, b)$ is uniform over the non-rectangular support of the $a \in [0, \bar{a}], b \in [0, \bar{b}_0 + da]$. The average age of the indifferent is

$$E(a|\, I) = \int_0^{\bar{a}} af da / \int_0^{\bar{a}} f da = \frac{\bar{a}}{2}$$

The average age of the treated is

$$\begin{aligned} E(a|\,T) &= \int_0^{\overline{a}} \int_{\widehat{b}(a)}^{\overline{b}_0+da} af\,dbda \Big/ \int_0^{\overline{a}} \int_{\widehat{b}(a)}^{\overline{b}_0+da} f\,dbda \\ &= \frac{\left(\overline{b}_0 - w_0\right)\overline{a}^2/2 + \left(d-(1+p)\right)\overline{a}^3/3}{\left(\overline{b}_0 - w_0\right)\overline{a} + \left(d-(1+p)\right)\overline{a}^2/2} \end{aligned}$$

so that $E(a|\,I) > E(a|\,T)$ if $(1+p) > d$. Hence introducing positive prioritisation is welfare improving if $d < 1$ (as we assumed in the previous examples in this section where $d = 0$). If $d > 1$ then introducing negative prioritisation is welfare improving. This is the case illustrated in Figure 5 where the average age of the treated in ABDC exceeds the average age of the indifferent along AB.

This example also provides a tractable case where the observable and unobservable components of health gain are not independently distributed since the expected unobserved benefit conditional on age $E(b|\,a) = \left(\overline{b}_0 + da\right)/2$ is increasing in age. The expected health gain from treatment conditional on age is $E(b-a|\,a) = \left(\overline{b}_0 + (d-2)a\right)/2$ which is decreasing in age if $d < 2$. Thus when $d \in (1,2)$ giving priority to the old is welfare increasing even though expected health benefits conditional on age are decreasing. The example is useful in emphasising that the crucial question to be answered in deciding whether to give shorter waits to the young is whether doing so reduces the total time waited, not whether the expected health gain conditional on age is decreasing or increasing in age.

5 Extensions

5.1 Prioritisation with a private sector

It is possible to show that the main result obtained above still holds when patients also have the option of private treatment at a money price k and no wait. Suppose that patient utility if not treated is $u(y)$, where u is increasing and strictly concave in income y. Utility if treated in the public sector after a prioritised wait of $w_0 + pa$ is $u(y) + b - w_0 - (1+p)a$. Utility if treated in the private sector is $u(y-k) + b - a$. Prioritisation is based on the observable component of the health gain from treatment (a), not income. Age, unobservable benefit, and income are distributed on the support $a \in [0, \overline{a}]; b \in \left[0, \overline{b}\right]; y \in \left[\underline{y}, \overline{y}\right]$ with joint density $f(a,b,y)$.

To demand public treatment patients must, as before, prefer public treatment to no treatment. Hence public patients must have a minimum unobservable benefit of $\widehat{b} = w_0 + (1+p)a$ and a maximum age of $\hat{a}$. Public patients must also prefer public to private treatment:

$$G = u(y) - w_0 - pa - u(y-k) \geq 0 \tag{25}$$

Since $G_y = u_y(y) - u_y(y-k) < 0$, there is a unique income $y^G(w_0 + pa)$ such that all individuals with $y \leq y^G$ prefer public to private treatment and all those with $y > y^G$ prefer private to public treatment. The demand for public treatment is

$$D^G(w_0, p, k) = \int_0^{\widehat{a}} \int_{\widehat{b}}^{\overline{b}} \int_{\underline{y}}^{y^G} f(a, b, y) dy db da \tag{26}$$

which is increasing in the charge for private treatment and decreasing in w_0 and p.

Welfare is the sum of the utilities of public patients, private patients and those who are not treated. The contribution from the utility of patients who choose to be treated in the public sector is

$$S^G(w_0(p), p) = \int_0^{\widehat{a}} \int_{\widehat{b}}^{\overline{b}} \int_{\underline{y}}^{y^G} [b - a(1+p) - w_0] f(a, b, y) dy db da \tag{27}$$

An increase in the degree of prioritisation (an increase in p) has no effect on the utility of private patients and those who decide not to be treated. It has a direct effect on the utility of public patients and it also alters the number of private patients and the untreated. Since individuals make privately optimal decisions about whether to be treated in the public or private sector or not to be treated, the patients who shift between choices as a result of the change in p are no worse or better off: they are indifferent between the public and private treatment or between public treatment and no treatment. Hence the welfare effect of a marginal change in p arises only via the change in (27) holding constant the number indifferent between the public sector and no treatment, and the number indifferent between the public and private sectors. The marginal welfare effect of prioritisation is (see Gravelle and Siciliani, 2006, Appendix)

$$\frac{dS^G}{dp} = D^G \left[E\left(a \middle| I \right) - E\left(a \middle| T \right) \right]$$

Hence we have a generalisation of Proposition 1.

Proposition 7 *When public health care is rationed by waiting and private health care is rationed by price, positive prioritisation in the public sector is welfare increasing if the average age of those indifferent about public treatment is bigger than the average age of those treated in the public sector.*

5.2 Non-utilitarian welfare function

The welfare function (7) is utilitarian: the utility of all individuals have the same welfare weight. It is possible to allow for different value judgements by attaching a weight $m(a,b)$ to the utility functions $b-a-w$. The welfare function is then

$$\begin{aligned} S^m(w_0(p),p) &= \int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}}[b-a-w(a,p)]m(a,b)f(a,b)dbda \\ &= \int_0^{\widehat{a}}\int_{\widehat{b}}^{\overline{b}}[b-a-w(a,p)]f^m(a,b)dbda \end{aligned} \tag{28}$$

where, without loss of generality we scale the weighting function so that f^m integrates to 1. For example, we might want to examine the implications of the judgement that the health gains of old patients should receive a greater weight than those of the young. Proceeding as in section 2.2, we get

$$\frac{dS^m(w_0(p),p)}{dp} = D^m\left[E(a|\,I) - E^m(a|\,T)\right] \tag{29}$$

where $D^m$and $E^m(a|\,T)$ are "demand" and "average age" of treated patients evaluated using the value weighted density function f^m rather than the actual distribution function f. Thus if the weighting function $m(a,b)$ has $m_a > 0$, $m_b = 0$ so that it gives more weight to older patients with any given private benefit b then $E^m(a|\,T) > E(a|\,T)$ and the optimal degree of prioritisation is reduced (see Gravelle and Siciliani, 2006, Appendix).

5.3 Service time differences

We have followed the waiting time literature by assuming that patients with different observed benefits and age have the same service time. More realistically, let the service time required for a patient of type (a,b) be $n(a,b)$ so

that effective demand on the fixed total capacity z is

$$D^n(w_0, p) = \int_0^{\widehat{a}} \int_{\widehat{b}}^{\overline{b}} n(a,b) f(a,b) db da \tag{30}$$

and $\partial w_0 / \partial p = -D_p^n / D_w^n$. The marginal effect of prioritisation on the welfare function (7) is now

$$\frac{dS(w_0(p), p)}{dp} = D\left[E^n(a|\, I) - E(a|\, T)\right] \tag{31}$$

where $E^n(a|\, I)$ is the "average age" of indifferent patients calculated using the conditional age density weighted by service time: $n(a, \hat{b}) f(a, \hat{b}) \,/ \int_0^{\hat{a}} n(a, \hat{b}) f(a, \hat{b}) da$. Thus, if service time increases with age and is unaffected by private benefit (i.e. $n_a > 0$ and $n_b = 0$), then $E^n(a|\, I)$ is greater than it would be with identical service times and the marginal welfare gain from prioritisation is increased (see Gravelle and Siciliani, 2006, Appendix). A greater degree of prioritisation is optimal because old patients have, on average, lower health gains and impose higher opportunity costs via their longer service times.

6 Concluding remarks

We have examined whether waiting-time prioritisation is welfare improving when health benefits from treatment vary among patients and are the difference between an unobservable private benefit and an observable factor reducing the health benefit (for example age).

Starting from a regime with no prioritisation, increasing prioritisation locally (shorter waits for the young) increases a utilitarian welfare function if the average age of the patients who are indifferent about receiving the treatment after some wait and not getting the treatment at all is higher than the average age of the patients who receive the treatment. Moreover, at any level of prioritisation, if the condition on average ages is satisfied then further prioritisation increases welfare.

The average age condition is satisfied for all prioritisation intensities if the distribution of the private benefit is uniform (for any distribution of age) or if the distribution of age is uniform and the distribution of the private benefit is log-concave. In this case, we have proved that the optimal prioritisation is maximal prioritisation. The reason is that in these circumstances

a smaller proportion of patients with low observable benefit demand treatment for a given wait, since the utility of the patient decreases with the observable dimension of benefit (for example age). Prioritisation will then reduce waiting for a large group of patients with high observable benefit (low age) at the cost of an increase in waiting of a small group of patients with low observable benefit (high age). Intuitively, rationing by waiting imposes dead weight costs on patients with no offsetting gain to producers, so that it should be imposed most heavily on the patients whose demand is most responsive to the waiting time. These will usually be those who demand less because they have a lower observable benefit (higher age).

We have also shown that a more general rationing rule (threshold rationing plus prioritisation), where patients whose observable benefit is below a threshold are not eligible for treatment, can bring further increases in welfare. Although the observable benefit is an imperfect signal of total health gain, welfare can be higher if providers set waiting time to zero and use only the imperfect (observable) signal to ration patients.

The analysis supports explicit priority schemes, like those in Canada and New Zealand, that give shorter waits to prioritised patients and the setting of treatment thresholds so that treatment is provided with no wait to those meeting the criteria and treatment is denied to all others. When there is imperfect information on the health benefits of individual patients, prioritisation leads to a welfare loss because some patients with low benefits get treatment whilst some with higher benefit are not treated. But this welfare loss due to a worse allocation of treatment may be more than offset by the welfare gain from reducing the deadweight loss imposed by waiting times on all patients receiving treatment.

Some of our results on the direction of prioritisation rely on the assumption that age and unobservable benefit are independent. Suppose that age and unobservable benefit are negatively correlated. This implies that age becomes a better signal of private benefit. Therefore, we conjecture that prioritisation would be even more desirable as by favouring patients with low age, prioritisation favours patients with high private benefit. Similarly, a positive correlation between age and unobservable benefit would make prioritisation less desirable.

References

Bagnoli, M. and T.C. Bergstrom, 2005, "Log-concave probability and its applications", *Economic Theory*, 26(2), 445 - 469.

Barros, P. and P. Olivella, 2005, "Waiting lists and patient selection", *Journal of Economics and Management Strategy*, 14(3), 623-646.

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

Cairns, J.A and M.M. van der Pol, 2000, "The estimation of marginal time preferences in a UK-wide sample (TEMPUS) project", *Health Technology Assessment*, 4(1), www.ncctha.org.

Cullis, P., J.G. Jones and C. Propper, 2000, "Waiting and medical treatment: analyses and policies", Chapter 28 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.

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., M. Dusheiko and M. Sutton, 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.

Gravelle, H. and L. Siciliani, 2006, "Is waiting-time prioritisation welfare improving?", University of York, *DERS Discussion Paper*, 06/13 revised, at www.york.ac.uk/depts/econ/research/dp/2006.htm.

Gravelle, H. and L. Siciliani, 2007a, "Optimal waits and charges in health insurance", University of York, *DERS Discussion Paper DP,* 07/02 at
www.york.ac.uk/depts/econ/research/dp/2007.htm.

Gravelle, H. and L. Siciliani, 2007b. "Ramsey waits: allocating the public health care budget across treatments when there is rationing by waiting", University of York, *DERS Discussion Paper DP,* 07/xx at
www.york.ac.uk/depts/econ/research/dp/2007.htm.

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

Iversen, T., 1997, "The effect of private sector on the waiting time in a National Health Service", *Journal of Health Economics,* 16, 381-396.

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

Loewenstein, G. and D. Prelec, 1992, "Anomalies in intertemporal choice: evidence and interpretation", *Quarterly Journal of Economics*, 107, 573-597.

Marchand, M. and F. Schroyen, 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.

Olivella, P., 2003, "Shifting public-health-sector waiting lists to the private sector", *European Journal of Political Economy*, 19(1), 103-132.

Siciliani, L., 2005, "Does more choice reduce waiting times?", *Health Economics*, 14(1), 17-23.

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

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.

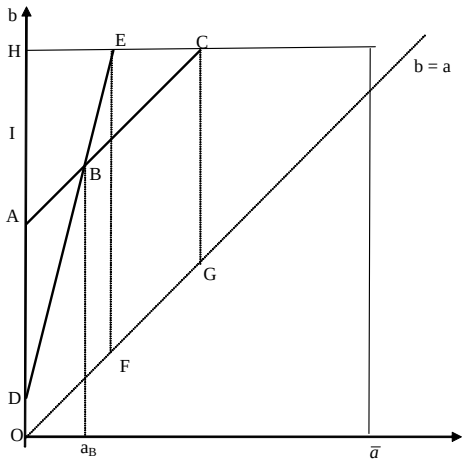

Figure 1. Effect of prioritisation on patients. If there is no prioritisation (all patients have the same wait) patients in ACH join the waiting list. With positive prioritisation (shorter waits for younger patients) patients in DEH join the waiting list. Introducing positive prioritisation is welfare increasing because the average age of the treated in ACH is less than the average age of indifferent on AC.

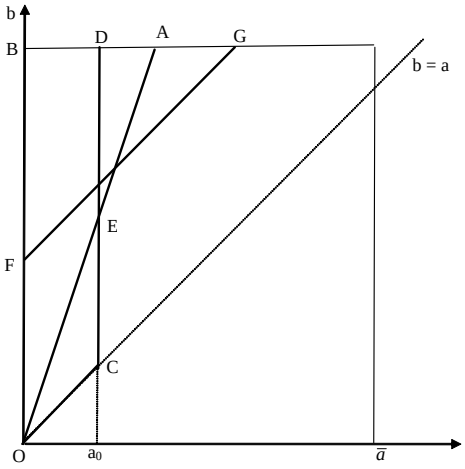

Figure 2. Pure threshold rationing, maximal prioritisation and no prioritisation. Under maximal prioritisation the wait for the youngest patients is zero and all patients in OAB join the list and are treated. Under threshold rationing patients with $a \leq a_0$ have a zero wait and all other potential patients are refused treatment; patients in OCDB are treated. With no prioritisation patients in FGB are treated.

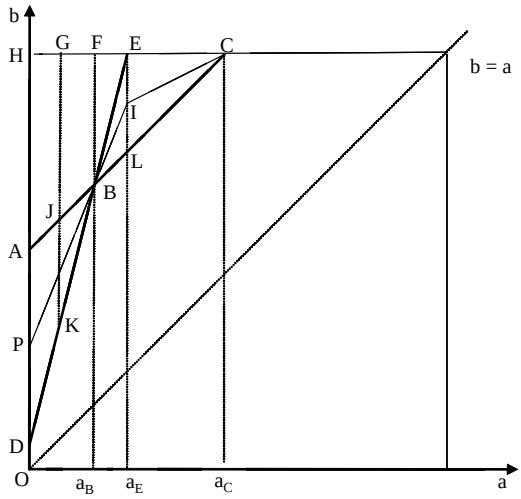

Figure 3. Effect of prioritisation on patients
Prioritisation displaces old patients in BCE and increases the wait for old patients in BEF. The displaced old patients are replaced by young patients in ABD and the waiting time of young patients in ABFH is reduced. The gain to young patients outweighs the loss to old patients.

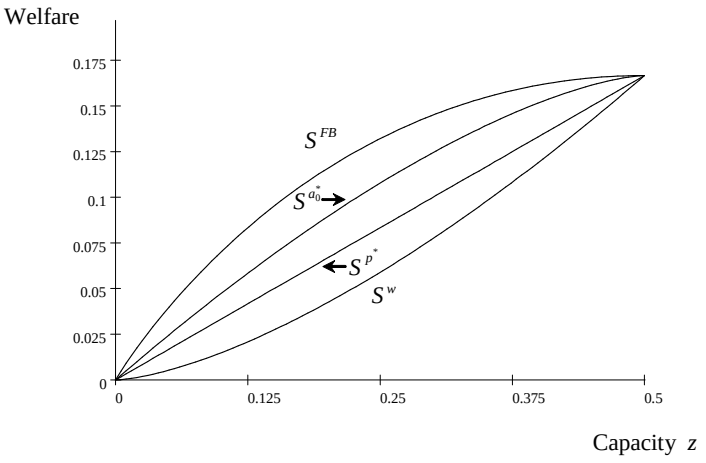

Figure 4. Welfare under no prioritisation (S^w), maximal prioritisation (S^{p^*}), pure threshold rationing ($S^{a_0^*}$) and the first best (S^{FB})

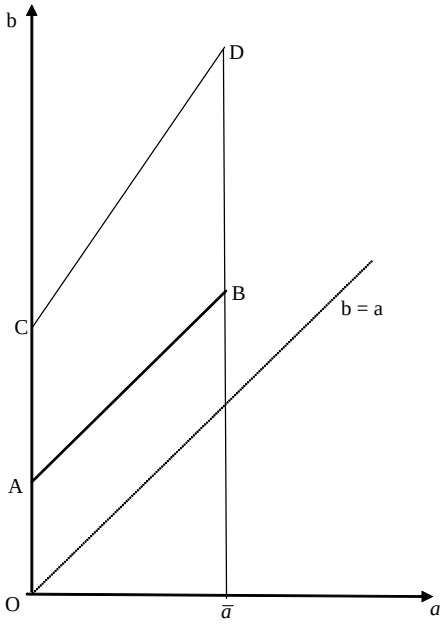

Figure 5. Negative prioritisation. Patients in area ABDC demand treatment. Negative prioritisation (shorter waits for the old) is welfare increasing because the average age of the treated in ABDC exceeds the average age of the indifferent on AB.

Appendices

Notation

b	unobserved component of health gain
a	observed component of health gain
w	waiting time
z	supply of care
$f(a,b)$	joint density of age and benefit
D	demand
S	welfare
$w_0 + pa$	waiting time for patient of age a
k	money price for treatment in the private sector
y	income

Demand and welfare function

From (1), noting that $\widehat{b}\,(\widehat{a}) = \overline{b}$,

$$D_{w_0} = \int_{\widehat{b}(\widehat{a})}^{\overline{b}} f(a,b)db\tfrac{\partial \widehat{a}}{\partial w_0} - \int_0^{\widehat{a}} f(a,\widehat{b}\,(a))da = -\int_0^{\widehat{a}} f(a,\widehat{b}\,(a))da < 0$$

$$D_p = \int_{\widehat{b}(\widehat{a})}^{\overline{b}} f(a,b)db\tfrac{\partial \widehat{a}}{\partial p} - \int_0^{\widehat{a}} af(a,\widehat{b}\,(a))da = -\int_0^{\widehat{a}} af(a,\widehat{b}\,(a))da < 0.$$

and from (7)

$$\begin{aligned}\tfrac{\partial S}{\partial w_0} = &-\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} f(a,b)dbda + \int_{\widehat{b}(\widehat{a})}^{\overline{b}} \left(b - a(1+p) - w_0(p)\right) f(a,b)db\tfrac{\partial \widehat{a}}{\partial w_0} \\ &-\int_0^{\widehat{a}} \left(\widehat{b}\,(a) - a(1+p) - w_0(p)\right) f(a,\widehat{b}\,(a))\tfrac{\partial \widehat{b}(a)}{\partial w_0}da \\ = &-\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} f(a,b)dbda = -D < 0\end{aligned}$$

$$\begin{aligned}\tfrac{\partial S}{\partial p} = &-\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} af(a,b)dbda + \int_{\widehat{b}(\widehat{a})}^{\overline{b}} \left(b - a(1+p) - w_0(p)\right) f(a,b)db\tfrac{\partial \widehat{a}}{\partial p} \\ &-\int_0^{\widehat{a}} \left(\widehat{b}\,(a) - a(1+p) - w_0(p)\right) f(a,\widehat{b}\,(a))\tfrac{\partial \widehat{b}(a)}{\partial p}da\end{aligned}$$

$$= -\int_0^{\widehat{a}} \int_{\widehat{b}(a)}^{\bar{b}} af(a,b)dbda < 0.$$

Second derivative of welfare function

It is possible to show that (extended proof available from the authors)

$$\frac{d^2S}{dp^2} = V(a \mid I) + E(a \mid I)^2 \left(1 - \int_0^{\widehat{a}} f(a, \widehat{b}(a))da\right)$$

$$-\frac{D}{1+p}\frac{f(\widehat{a},\bar{b})[\widehat{a}-E(a|I)]^2}{\int_0^{\widehat{a}} f(a,\widehat{b}(a))da} + D\frac{\int_0^{\widehat{a}} (a-E(a|I))^2 f_b da}{\int_0^{\widehat{a}} f(a,\widehat{b}(a))da}$$

where $V(a \mid I) = \int_0^{\widehat{a}} (a - E(a \mid I))^2 f(a, \widehat{b}\,(a))da$ is the variance of the age of the patients who are indifferent. The first two terms are positive while the third term is negative. The fourth term is unsigned and depends on the sign of $f_b(a, \widehat{b}\,(a))$. Overall the second derivative is unsigned.

Proof of proposition 2

We prove that $dS(p,z)/dp = 0$ for all p and all z if and only if $f(a,b) = h(a)\gamma e^{-\gamma b}$ by showing (i) $dS(0,z)/dp = 0$ for all z if $f(a,b) = h(a)\gamma e^{-\gamma b}$(ii) $(a,b) = h(a)\gamma e^{-\gamma b}$ implies $dS(p,z)/dp = 0$ all p, z.

(i) The marginal value of introducing linear prioritisation is zero if

$$\left.\frac{dS}{dp}\right|_{p=0} = \int_0^{\widehat{a}} \int_{\widehat{b}}^{\bar{b}} f(a,b)dbda \frac{\int_0^{\widehat{a}} af(a,\hat{b})dbda}{\int_0^{\widehat{a}} f(a,\widehat{b})da} - \int_0^{\widehat{a}} a \int_{\widehat{b}}^{\bar{b}} f(a,b)dbda = 0 \tag{32}$$

where $\widehat{b} = a(1+p) + w_0(p,z)$, $\widehat{a} = \min\{\frac{\bar{b}-w_0}{1+p}, \bar{a}\}$, and $w_0(p,z)$ is the waiting time for patients with $a = 0$. We seek conditions on the joint distribution for which (32) holds whatever the equilibrium waiting time is in the absence of prioritisation ($w^o(z) = w_0(0,z)$). Setting $p = 0$, we have $\widehat{b} = a + w^o$ and $\widehat{a} = \min\{\bar{b} - w^o, \bar{a}\}$. Equation (32) must hold for all z and since $w^o(z)$ decreases monotonically with supply z, this implies

$$\left.\frac{d^2S}{dpdz}\right|_{p=0} = \left.\frac{d^2S}{dpdw^o}\right|_{p=0} \frac{dw^o}{dz} = 0$$

and so if (32) is true for all z, then

$$\left.\frac{d^2S}{dpdw^o}\right|_{p=0} = \int_0^{\widehat{a}} af(a,a+w^o)da - \frac{\int_0^{\widehat{a}} af(a,a+w^o)da}{\int_0^{\widehat{a}} f(a,a+w^o)da}\int_0^{\widehat{a}} f(a,a+w^o)da + \int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\bar{b}} f(a,b)dbda\frac{d}{dw^o}\left(\frac{\int_0^{\widehat{a}} af(a,a+w^o)da}{\int_0^{\widehat{a}} f(a,a+w^o)da}\right) = 0$$

at all positive w_0. The first two terms cancel so that

$$\begin{aligned}\left.\frac{d^2S}{dpdw^o}\right|_{p=0} &= \int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\bar{b}} f(a,b)dbda\frac{d}{dw^o}\left(\frac{\int_0^{\widehat{a}} af(a,a+w^o)da}{\int_0^{\widehat{a}} f(a,a+w^o)da}\right)\\ &= D\frac{dE(\,a|\,I)}{dw^o} = 0 \end{aligned} \tag{33}$$

and since $D > 0$ we require that

$$E(\,a|\,I) = \frac{\int_0^{\widehat{a}} af(a,a+w^o)da}{\int_0^{\widehat{a}} f(a,a+w^o)da} \tag{34}$$

does not vary with w_0.

If $\bar{b} < \infty$ then $E(\,a|\,I)$ is decreasing with w^o in some neighbourhood of $\bar{b}$. Thus $\hat{a} = \bar{a}$ and we require

$$E(\,a|\,I) = \frac{\int_0^{\bar{a}} af(a,a+w^o)da}{\int_0^{\bar{a}} f(a,a+w^o)da} = \frac{\int_0^{\bar{a}} am(a,w^o)da}{\int_0^{\bar{a}} m(a,w^o)da} \tag{35}$$

does not vary with w^o. Since $\bar{b}$ must be unbounded from above, f cannot be constant for all b and so $m_w = f_b(a,a+w^o) \neq 0$ for some a. The only functional form compatible with (35) for all w^o is $m(a,w^o) = m_1(a)m_2(w^o)$. But since w^o and a are additive in the second argument in $f(a,a+w^o)$ and $m(a,w^o)$ is multiplicatively separable, we must have $m_1(a) = m_3(a)m_2(a)$ and $f(a,a+w^o) = m(a,w^o) = m_3(a)m_2(a)m_2(w^o) = m_3(a)m_2(a+w^o)$. But $m_2(a)m_2(w^o) = m_2(a+w^o)$ implies $\ln m_2(a) + \ln m_2(w^o) = \ln m_2(a+w^o)$, which implies that $\ln m_2(z) = \theta z$, or $m_2(z) = \exp(\theta z)$ and $m(a,w^o) = m_3(a)\exp(\theta(a+w^o) = f(a,a+w^o) = h(a)g(a+w^o)$. Since h and g are probability densities, $g(a+w^o) = \gamma\exp(-\gamma(a+w^o))$

(ii) We want to prove that if $f(a,b) = h(a)g(b) = h(a)\gamma e^{-\gamma b}$, with $\gamma > 0$, then $\frac{dS}{dp} = 0$. Since $\hat{b} = w_0(p,z) + (1+p)a$

$$\frac{dS}{dp} = D\left[\frac{\int_0^{\widehat{a}} ah(a)g(\widehat{b}(a))da}{\int_0^{\widehat{a}} h(a)g(\widehat{b}(a))da} - \frac{\int_0^{\widehat{a}} ah(a)\left[G(\bar{b}) - G\left(\widehat{b}(a)\right)\right]da}{\int_0^{\widehat{a}} h(a)\left[G(\bar{b}) - G\left(\widehat{b}(a)\right)\right]da}\right] \tag{36}$$

$$= D\left[\frac{\int_0^{\widehat{a}} ah(a)\gamma e^{-\gamma\widehat{b}(a)}da}{\int_0^{\widehat{a}} h(a)\gamma e^{-\gamma\widehat{b}(a)}da} - \frac{\int_0^{\widehat{a}} ah(a)e^{-\gamma\widehat{b}(a)}da}{\int_0^{\widehat{a}} h(a)e^{-\gamma\widehat{b}(a)}da}\right] = 0$$

is true for all p and z.

Proof of proposition 3

(a). We want to prove that if $h(a)$ is uniform and $g(b)$ is strictly log-concave, then

$$\frac{dS}{dp} = D\left[\int_0^{\widehat{a}} a\frac{g(a(1+p)+w_0)}{G(\bar{b}) - G(w_0)}da - \int_0^{\widehat{a}} a\frac{\left[G(\bar{b}) - G(a(1+p)+w_0)\right]}{D}da\right] > 0 \tag{37}$$

First-order stochastic dominance of the distribution of a conditional on indifference over the distribution conditional on joining the list requires:

$$\begin{aligned}\int_a^{\widehat{a}} \frac{g(w_0 + \widetilde{a}(1+p))}{G(\bar{b}) - G(w_0)}d\widetilde{a} &\equiv \frac{G(\bar{b}) - G(w_0 + a(1+p))}{G(\bar{b}) - G(w_0)} \\ &\geq \int_a^{\widehat{a}} \frac{G(\bar{b}) - G(w_0 + \widetilde{a}(1+p))}{D}d\widetilde{a}\end{aligned} \tag{38}$$

for all $a \in [0, \widehat{a}]$ with strict inequality for some $a \in [0, \widehat{a}]$. We can rewrite (38) as

$$\frac{\int_a^{\widehat{a}} \left[G(\bar{b}) - G(w_0 + \widetilde{a}(1+p))\right]d\widetilde{a}}{G(\bar{b}) - G(w_0 + a(1+p))} \leq \frac{D}{G(\bar{b}) - G(w_0)} \tag{39}$$

The left-hand side of (39) is the "mean residual lifetime function", which is strictly monotone decreasing if $g(b)$ is strictly log-concave (see Bagnoli and Bergstrom, 2005, Theorem 6, p.451). At $a = 0$,

$$\frac{\int_0^{\widehat{a}} \left[G(\bar{b}) - G(w_0 + \widetilde{a}(1+p))\right] d\widetilde{a}}{G(\bar{b}) - G(w_0)} = \frac{D}{G(\bar{b}) - G(w_0)}$$

and so (39) holds strictly for $a \in (0, \widehat{a}]$.

(b)**.** If $g(b)$ is uniform, then for any $h(a)$:

$$\begin{aligned}
\frac{dS}{dp} &= \frac{\left[\int_0^{\widehat{a}} h(a)\left[1 - w_0 - (1+p)\,a\right] da\right] \int_0^{\widehat{a}} ah(a)da}{\int_0^{\widehat{a}} h(a)da} \\
&\quad - \int_0^{\widehat{a}} ah(a)\left[1 - w_0 - (1+p)\,a\right] da \\
&= \frac{\left[(1-w_0)\int_0^{\widehat{a}} h(a)da - (1+p)\int_0^{\widehat{a}} ah(a)da\right] \int_0^{\widehat{a}} ah(a)da}{\int_0^{\widehat{a}} h(a)da} \\
&\quad - (1-w_0)\int_0^{\widehat{a}} ah(a)da + (1+p)\int_0^{\widehat{a}} a^2 h(a)da \\
&= (1+p)\int_0^{\widehat{a}} a^2 h(a)da - (1+p)\left(\int_0^{\widehat{a}} ah(a)da\right)^2 \Big/ \int_0^{\widehat{a}} h(a)da \\
&= (1+p)\left[\int_0^{\widehat{a}} a^2 h(a)da - \left[E(a\,|I)\right]^2 \int_0^{\widehat{a}} h(a)da\right] \\
&= (1+p)\left[V(a) + \left[E(a\,|I)\right]^2 \left[1 - H(\widehat{a})\right]\right] > 0 \qquad (40)
\end{aligned}$$

Threshold rationing: demand and welfare function

Recall $\widehat{b}(a) = a + pa - pa_0 + w_0$; $\widehat{a} = \min\left[\frac{\bar{b}-w_0+pa_0}{1+p}, \bar{a}\right]$ and $\widehat{b}(\widehat{a}) = \bar{b}$

The derivatives of (11) are

$$D_{a_0} = \int_{a_0}^{\bar{b}} f(a_0, b)db - \int_{a_0+w_0}^{\bar{b}} f(a, b)db + \int_{\widehat{b}(\widehat{a})}^{\bar{b}} f(\widehat{a}, b)db\frac{\partial\widehat{a}}{\partial a_0} - \int_{a_0}^{\widehat{a}} f(a, \widehat{b}(a))\frac{\partial\widehat{b}(a)}{\partial a_0}da$$

$$= \int_{a_0}^{a_0+w_0} f(a_0, b)db + p\int_{a_0}^{\widehat{a}} f(a, \widehat{b}(a))da > 0$$

$$D_{w_0} = -\int_{a_0}^{\widehat{a}} f(a, \widehat{b}(a))da,\ D_p = -\int_{a_0}^{\widehat{a}} af(a, \widehat{b}(a))da.$$

and of (12) are

$$S_{a_0} = \int_{a_0}^{\bar{b}} (b - a_0)f(a_0, b)db - \int_{a_0+w_0}^{\bar{b}} (b - a_0 - w_0)\, f(a_0, b)db$$

$$+p\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\bar{b}} f(a, b)dbda + \int_{\widehat{b}(\widehat{a})}^{\bar{b}} \left(b - \widehat{b}(\widehat{a})\right) f(\widehat{a}, b)db$$

$$= \int_{a_0}^{a_0+w_0} (b - a_0)f(a_0, b)db + w_0 \int_{a_0+w_0}^{\bar{b}} f(a_0, b)db$$

$$+p\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\bar{b}} f(a, b)dbda > 0,$$

$$S_{w_0} = -\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\bar{b}} f(a, b)dbda < 0\ ,\ S_p = -\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\bar{b}} af(a, b)dbda < 0.$$

If prioritisation is not maximal,

$\frac{dS^*}{da_0} = \frac{\partial S}{\partial a_0} - \frac{D_{a_0}}{D_{w_0}}\frac{\partial S}{\partial w_0}$

$$= \int_{a_0}^{a_0+w_0} (b - a_0)f(a_0, b)db + w_0 \int_{a_0+w_0}^{\bar{b}} f(a_0, b)db$$

$$+p\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} f(a,b)dbda - \frac{\int_{a_0}^{a_0+w_0} f(a_0,b)db + p\int_{a_0}^{\widehat{a}} f(a,\widehat{b}(a))da}{\int_{a_0}^{\widehat{a}} f(a,\widehat{b}(a))da}\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} f(a,b)dbda$$

$$= \int_{a_0}^{a_0+w_0}(b-a_0)f(a_0,b)db + w_0\int_{a_0+w_0}^{\overline{b}} f(a_0,b)db$$

$$-\int_{a_0}^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} f(a,b)dbda\frac{\int_{a_0}^{a_0+w_0} f(a_0,b)db}{\int_{a_0}^{\widehat{a}} f(a,\widehat{b}(a))da}$$

Prioritisation: uniform distributions

Patients demand treatment if $b-a-w\geq 0$, where $w=w_0+pa$. The patient is indifferent if $b=a+w_0+pa$. The demand function is

$$\begin{aligned} D(w_0,p) &= \int_0^{\frac{1-w_0}{1+p}}\int_{a+w_0+pa}^{1} dbda \\ &= \int_0^{\frac{1-w_0}{1+p}} (1-w_0-a-pa)\,da = \frac{(1-w_0)^2}{2\,(1+p)} \end{aligned} \tag{41}$$

and welfare is

$$S(w_0,p) = \int_0^{\frac{1-w_0}{1+p}}\int_{a+w_0+pa}^{1} (b-a-w_0-pa)\,dbda = \frac{(1-w_0)^3}{6(1+p)} \tag{42}$$

The equilibrium condition $D(w_0,p)=z$ implies that $w_0=1-[2z\,(1+p)]^{1/2}$. Hence

$$S=\frac{[2z\,(1+p)]^{3/2}}{6(1+p)} \tag{43}$$

and $\frac{dS}{dp}>0$, then $w_0^*=0$. At the optimum, substituting $w_0^*=0$, we obtain $p^*=\frac{1}{2z}-1$.

Threshold rationing: uniform distributions

The demand function is

$$\begin{aligned} D(w_0,p,a_0) &= \int_0^{a_0}\int_a^1 1dbda + \int_{a_0}^{\frac{1-w_0+pa_0}{1+p}} \int_{a+pa-pa_0+w_0}^{1} 1dbda \\ &= \int_0^{a_0} (1-a)\,da + \int_{a_0}^{\frac{1-w_0+pa_0}{1+p}} (1-a-pa+pa_0-w_0)\,da \\ &= \frac{(1-w_0-a_0)^2}{2(1+p)} + a_0 - \frac{a_0^2}{2} \end{aligned} \tag{44}$$

with $D_{w_0} = -\frac{1-a_0-w_0}{1+p} < 0$, $D_p = -\frac{(1-a_0-w_0)^2}{2(1+p)^2} < 0$ and $D_{a_0} = \frac{w_0+p(1-a_0)}{1+p} > 0$, where $(1-a_0-w_0) > 0$ (notice that if $b=1$ and $a=a_0$, then $u = 1-w_0-a_0-pa_0+pa_0 = 1-w_0-a_0 > 0$). In equilibrium, demand equals supply $D(w_0,p,a_0) = z$ and

$$w_0 = 1 - a_0 - \left\{[2(1+p)][z + a_0^2 2^{-1} - a_0]\right\}^{1/2} \tag{45}$$

The welfare function is

$$\begin{aligned} S(a_0,w_0,p) &= \int_0^{a_0}\int_a^1 (b-a)dbda \\ &\quad + \int_{a_0}^{\frac{1-w_0+pa_0}{1+p}} \int_{a+pa-pa_0+w_0}^{1} (b-a-w_0-pa+pa_0)\,dbda \\ &= \frac{(1-w_0-a_0)^3}{6(1+p)} + \frac{1}{2}a_0 - \frac{1}{2}a_0^2 + \frac{1}{6}a_0^3 \end{aligned} \tag{46}$$

where $\frac{dS}{da_0} = \frac{(1-a_0)^2}{2} - \frac{(1-w_0-a_0)^2}{2(1+p)} > 0$, $\frac{dS}{dw_0} = -\frac{(1-w_0-a_0)^2}{2(1+p)} < 0$, $\frac{dS}{dp} = -\frac{(1-w_0-a_0)^3}{6(1+p)^2} < 0$. Substituting (22)

$$\begin{aligned} S(a_0,w_0(a_0,p),p) &= \frac{\left\{[2(1+p)z][z + a_0^2 2^{-1} - a_0]\right\}^{3/2}}{6(1+p)} \\ &\quad + \frac{1}{2}a_0 - \frac{1}{2}a_0^2 + \frac{1}{6}a_0^3 \end{aligned}$$

which is increasing in p for given a_0. Hence at given a_0, p should be increased until $w_0 = 0$ and so $w_0(a_0, p^*(a_0)) = 0$ and $p^*(a_0) = \frac{1-2z}{2z-2a_0+a_0^2} > 0$.

With $w_0(a_0, p^*(a_0)) = 0$, the market-clearing condition reduces to $\frac{(1-a_0)^2}{2(1+p)}$ $+a_0 - \frac{a_0^2}{2} = z$, so that $\frac{dp^*}{da_0} = \frac{2(1+p^*)}{1-a_0} > 0$, and the welfare function is

$$S(a_0, w_0(a_0, p^*(a_0)), p^*) = \frac{(1-a_0)^3}{6(1+p^*)} + \frac{1}{2}a_0 - \frac{1}{2}a_0^2 + \frac{1}{6}a_0^3 \tag{47}$$

where $\frac{\partial S}{\partial a_0} = \frac{(1-a_0)^2}{2(1+p^*)} > 0$ and $\frac{\partial S}{\partial p^*} = -\frac{(1-a_0)^3}{6(1+p^*)^2} < 0$. The effect of an increase in a_0 on welfare is

$$\frac{dS}{da_0} = \underset{+}{\frac{\partial S}{\partial a_0}} + \underset{-}{\frac{\partial S}{\partial p}} \underset{+}{\frac{\partial p}{\partial a_0}} = \frac{(1-a_0)^2}{6\,(1+p)} > 0 \tag{48}$$

Hence a_0 should be increased until the entire supply is given to individuals with $a \leq a^*(z) = 1 - (1-2z)^{1/2}$.

Prioritisation with a private sector

a, b and y are distributed according to the joint density $f(a, b, y)$ over the support $a \in [0, \overline{a}]; b \in \left[0, \overline{b}\right]$; $y \in \left[\underline{y}, \overline{y}\right]$ where y denotes income. 1) Patients prefer public to no treatment if $b > w_0 + (1+p)a$. Define $\widehat{b}(a)$ as the level of b such that $\widehat{b}(a) = w_0 + (1+p)a$. 2) Patients prefer public to private if: $u(y) - u(y-k) > w_0 + pa$. Define $y^G(a)$ as the level of income such that the patient is indifferent between public and private: $u(y^G) - u(y^G - k) = w_0 + pa$ with $\frac{\partial y^G}{\partial a} = \frac{p}{u_y(y^G) - u_y(y^G - k)} < 0$, $\frac{\partial y^G}{\partial w_0} = \frac{1}{u_y(y^G) - u_y(y^G - k)} < 0$ and $\frac{\partial y^G}{\partial p} = \frac{a}{u_y(y^G) - u_y(y^G - k)} < 0$.
3) Patients prefer private to no treatment if: $b > a + u(y) - u(y-k)$. Define $\widetilde{b}(a, y)$ as the level of private benefit such that $\widetilde{b}(a, y) = a + u(y) - u(y-k)$.

There are three groups of patients:
a) *Public treatment*: $b \geq \widehat{b}(a)$ (better public than no treatment) and $y \leq y^G(a)$ (better public than private treatment). b) *Private treatment*: $b > \widetilde{b}(a, y)$ (better private than no treatment) and $y > y^G(a)$ (better private than public treatment).
c) *No treatment*: $b < \widehat{b}(a)$ (better no treatment than public) and $b < \widetilde{b}(a, y)$ (better no treatment than private).

Notice that $\widehat{b}(a) < \widetilde{b}(a)$ if $w_0 + pa < u(y) - u(y-k)$, so that again y^G is such that $w_0 + pa = u(y^G) - u(y^G - k)$. As before, $\widehat{a} = \frac{\overline{b} - w_0}{1+p}$ (so that $\widehat{b}(\widehat{a}) = \overline{b}$). Also, define $y^*(a)$ such that $\widetilde{b}(a, y^*) = \overline{b}$ so that $u(y^*) - u(y^* - k) = \overline{b} - a$. $\widehat{a}$ can also be computed such that $y^*(\widehat{a}) = y^G(\widehat{a})$ or $w_0 + p\widehat{a} = \overline{b} - \widehat{a}$ so that $\widehat{a} = \frac{\overline{b} - w_0}{1+p}$. Define $\widehat{\widehat{a}} = \overline{b} - (u(\overline{y}) - u(\overline{y} - k))$ as the level of a such that $y^*(\widehat{\widehat{a}}) = \overline{y}$.

The demand for public treatment is:

$$D^G(w_0, p) = \int_0^{\widehat{a}} \int_{\widehat{b}(a)}^{\overline{b}} \int_{\underline{y}}^{y^G(a)} f(a, b, y) dy db da \tag{49}$$

where

$$\begin{aligned} D^G_{w_0} &= \int_{\widehat{b}(\widehat{a}, w_0)}^{\overline{b}} \int_{\underline{y}}^{y^G(\widehat{a}, w_0)} f(\widehat{a}, b, y) dy db \frac{\partial \widehat{a}}{\partial w_0} - \int_0^{\widehat{a}(w_0) y^G(a, w_0)} \int_{\underline{y}} f(a, \widehat{b}(a, w_0), y) dy da \\ &\quad + \int_0^{\widehat{a}(w_0)} \int_{\widehat{b}(a, w_0)}^{\overline{b}} f(a, b, y^G(a, w_0)) \frac{\partial y^G}{\partial w_0} db da \end{aligned}$$

$$= - \int_0^{\widehat{a}(w_0) y^G(a, w_0)} \int_{\underline{y}} f(a, \widehat{b}(a, w_0), y) dy da + \int_0^{\widehat{a}(w_0)} \int_{\widehat{b}(a, w_0)}^{\overline{b}} f(a, b, y^G(a, w_0)) \frac{\partial y^G}{\partial w_0} db da$$

$$\begin{aligned} D^G_p &= \int_{\widehat{b}(\widehat{a}(p), p)}^{\overline{b}} \int_{\underline{y}}^{y^G(\widehat{a}(p), p)} f(\widehat{a}(p), b, y) dy db \frac{\partial \widehat{a}}{\partial p} - \int_0^{\widehat{a}(p)} a \int_{\underline{y}}^{y^G(a, p)} f(a, \widehat{b}(a, p), y) dy da \\ &\quad + \int_0^{\widehat{a}(p)} \int_{\widehat{b}(a, p)}^{\overline{b}} f(a, b, y^G(a, p)) \frac{\partial y^G}{\partial p} db da \\ &= - \int_0^{\widehat{a}(p)} a \int_{\underline{y}}^{y^G(a, p)} f(a, \widehat{b}(a, p), y) dy da + \int_0^{\widehat{a}(p)} \int_{\widehat{b}(a, p)}^{\overline{b}} f(a, b, y^G(a, p)) \frac{\partial y^G}{\partial p} db da \end{aligned}$$

The demand for private treatment is:

$$D^{private} = \int_0^{\widehat{a}} \int_{y^G(a)}^{\overline{y}} \int_{\widetilde{b}(a,y)}^{\overline{b}} f(a,b,y)dbdyda + \int_{\widehat{a}}^{\widehat{\widehat{a}}} \int_{y^*(a)}^{\overline{y}} \int_{\widetilde{b}(a,y)}^{\overline{b}} f(a,b,y)dbdyda \quad (50)$$

The welfare function is:

$$\begin{aligned}
&S(w_0(p),p) \quad (51)\\
&= \int_0^{\widehat{a}} \int_{\widehat{b}(a)}^{\overline{b}} \int_{\underline{y}}^{y^G(a)} [b - a(1+p) - w_0 + u(y)]f(a,b,y)dydbda\\
&\qquad \text{public patients}\\
&+ \int_0^{\widehat{a}} \int_{y^G(a)}^{\overline{y}} \int_{\widetilde{b}(a,y)}^{\overline{b}} [b - a + u(y-k)]f(a,b,y)dbdyda\\
&\qquad \text{private patients}\\
&+ \int_{\widehat{a}}^{\widehat{\widehat{a}}} \int_{y^*(a)}^{\overline{y}} \int_{\widetilde{b}(a,y)}^{\overline{b}} [b - a + u(y-k)]f(a,b,y)dbdyda\\
&\qquad \text{private patients}\\
&+ \int_0^{\widehat{a}} \int_a^{\widehat{b}(a)} \int_{\underline{y}}^{y^G(a)} u(y)f(a,b,y)dydbda + \int_{\widehat{a}}^{\overline{a}} \int_a^{\overline{b}} \int_{\underline{y}}^{y^G(a)} u(y)f(a,b,y)dydbda\\
&\qquad \text{no treatment and poor}\\
&+ \int_0^{\widehat{a}} \int_{y^G(a)}^{\overline{y}} \int_a^{\widetilde{b}(a,y)} u(y)f(a,b,y)dbdyda + \int_{\widehat{a}}^{\widehat{\widehat{a}}} \int_{y^*(a)}^{\overline{y}} \int_a^{\widetilde{b}(a,y)} u(y)f(a,b,y)dbdyda\\
&\qquad \text{no treatment and rich}\\
&+ \int_{\widehat{a}}^{\widehat{\widehat{a}}} \int_{y^G(a)}^{y^*(a)} \int_a^{\overline{b}} u(y)f(a,b,y)dbdyda + \int_{\widehat{\widehat{a}}}^{\overline{a}} \int_{y^G(a)}^{\overline{y}} \int_a^{\overline{b}} u(y)f(a,b,y)dbdyda\\
&\qquad \text{no treatment and rich}
\end{aligned}$$

Differentiating with respect to w_0, we obtain:

$$
\begin{aligned}
S_{w_0} \quad = \quad & -\int_0^{\widehat{a}} \int_{\widehat{b}(a)}^{\overline{b}} \int_{\underline{y}}^{y^G(a)} f(a,b,y)dydbda \text{ [E1]} \\
& -\int_0^{\widehat{a}} \int_{\underline{y}}^{y^G(a)} u(y)f(a,\widehat{b},y)dyda \quad \text{[E2]} \\
& +\int_0^{\widehat{a}} \int_{\widehat{b}(a)}^{\overline{b}} [b-a(1+p)-w_0+u(y^G(a))]f(a,b,y^G(a))\frac{\partial y^G(a)}{\partial w_0}dbda \text{ [E3]} \\
& -\int_0^{\widehat{a}} \int_{\widetilde{b}(a,y^G)}^{\overline{b}} [b-a+u(y^G(a)-k)]f(a,b,y^G(a))\frac{\partial y^G(a)}{\partial w_0}dbda \text{ [E4]} \\
& +\int_{\widehat{a}}^{\overline{b}} \int_{\underline{y}}^{y^G(\widehat{a})} u(y)f(\widehat{a},b,y)dydb \quad \text{[E5]} \\
& -\int_{\widehat{a}}^{\overline{b}} \int_{\underline{y}}^{y^G(\widehat{a})} u(y)f(a,b,y)dydbda \quad \text{[E6]} \\
& +\int_0^{\widehat{a}} \int_{\underline{y}}^{y^G(a)} u(y)f(a,\widehat{b}\,(a)\,,y)dyda \quad \text{[E7]} \\
& +\int_0^{\widehat{a}} \int_a^{\widehat{b}(a)} u(y^G)f(a,b,y^G)\frac{\partial y^G(a)}{\partial w_0}dbda \quad \text{[E8]} \\
& +\int_{\widehat{a}}^{\overline{a}} \int_a^{\overline{b}} u(y^G)f(a,b,y^G)\frac{\partial y^G(a)}{\partial w_0}dbda \quad \text{[E9]}
\end{aligned}
$$

$$-\int_0^{\widehat{a}}\int_a^{\widetilde{b}(a,y^G)} u(y^G)f(a,b,y^G)\frac{\partial y^G(a)}{\partial w_0}dbda \quad [E10]$$

$$-\int_{\widehat{\widehat{a}}}^{\widehat{a}}\int_a^{\overline{b}} u(y^G)f(a,b,y^G)\frac{\partial y^G(a)}{\partial w_0}dbda \quad [E11]$$

$$-\int_{y^*(\widehat{a})}^{\overline{y}}\int_{\widehat{a}}^{\widetilde{b}(\widehat{a},y)} u(y)f(\widehat{a},b,y)dbdy\frac{\partial\widehat{a}}{\partial w_0} \quad [E12]$$

$$-\int_{y^G(\widehat{a})}^{y^*(\widehat{a})}\int_{\widehat{a}}^{\overline{b}} u(y)f(a,b,y)dbdy \quad [E13]$$

$$-\int_{\widehat{\widehat{a}}}^{\overline{a}}\int_a^{\overline{b}} u(y^G)f(a,b,y^G)db\frac{\partial y^G(a)}{\partial w_0}da \quad [E14]$$

$$+\int_{y^G(\widehat{a})}^{\overline{y}}\int_{\widehat{a}}^{\widetilde{b}(\widehat{a},y)} u(y)f(\widehat{a},b,y)dbdy\frac{\partial\widehat{a}}{\partial w_0} \quad [E15]$$

Recall that $y^G(a) : u(y^G) - u(y^G - k) = w_0 + pa$ so that $\widehat{b}\,(a) = \widetilde{b}(a, y^G)$ and [E3] and [E4] cancel out; similarly [E8] and [E10] cancel out; [E5] and [E6] cancel out; [E2] and [E7] cancel out; [E9], [E11] and [E14] cancel out. Notice that $y^*(\widehat{a}) = y^G(\widehat{a})$ as $y^*(\widehat{a})$: $u(y^*) - u(y^* - k) = \overline{b} - \widehat{a}$ and $y^G(\widehat{a})$: $u(y^G) - u(y^G - k) = w_0 + p\widehat{a}$, where $\overline{b} - \widehat{a} = \frac{p\overline{b}+w_0}{1+p}$ and $w_0 + p\widehat{a} = \frac{w_0+p\overline{b}}{1+p}$ so that [E13] is zero. For the same reason [E12] and [E15] cancel out. Therefore

$$S_{w_0} = -\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}}\int_{\underline{y}}^{y^G(a)} f(a,b,y)dydbda < 0$$

Similarly, we can obtain:

$$S_p = -\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}}\int_{\underline{y}}^{y^G(a)} af(a,b,y)dydbda < 0$$

Substituting into: $\frac{dS}{dp} = S_{w_0}\frac{\partial w_0}{\partial p} + S_p = D\left(\frac{D_p}{D_{w_0}} + \frac{S_p}{D}\right)$, the main result is obtained.

Non-utilitarian welfare function

Recall that $E^*(\,a|\,T) = \dfrac{\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} am(a)f(a,b)dbda}{\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} m(a)f(a,b)dbda}$ and $E(\,a|\,T) = \dfrac{\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} af(a,b)dbda}{\int_0^{\widehat{a}}\int_{\widehat{b}(a)}^{\overline{b}} f(a,b)dbda}$.

Define $q(a) = \int_{\widehat{b}(a)}^{\overline{b}} f(a,b)db$, so that $E^*(\,a|\,T) = \int_0^{\widehat{a}} am(a)q(a)da / \int_0^{\widehat{a}} m(a)q(a)da$ and $E(\,a|\,T) = \int_0^{\widehat{a}} aq(a)da / \int_0^{\widehat{a}} q(a)da$.

The density function $m(a)q(a)$ is more *favourable* than $q(a)$ if $\frac{q(a)m(a)}{\int_0^a m(a)q(a)da} \geq \frac{q(a)}{\int_0^a q(a)da}$ (hazard rate dominance, see Laffont and Tirole, 1993, *A theory of incentives in procurement and regulation*, p.77). Integrating by parts, we obtain: $\frac{q(a)m(a)}{Q(a)m(a)-\int_0^a m_a(a)Q(a)da} \geq \frac{q(a)}{Q(a)}$ or $q(a)m(a)Q(a) \geq q(a)m(a)Q(a) - q(a)\int_0^a m_a(a)Q(a)da$, which is always satisfied when $m_a > 0$. Hazard rate dominance implies first-order stochastic dominance (Une M., T. Sajo, 1995, *Economics Letters*, 47(1), p.109-110), which implies that the average age is higher when $\frac{q(a)m(a)}{\int_0^a m(a)q(a)da} \geq \frac{q(a)}{\int_0^a q(a)da}$, so that $E^m(\,a|\,T) > E(\,a|\,T)$.

Service time differences

Recall that $E^n(\,a|\,I) = \int_0^{\widehat{a}} an(a)f(a,\widehat{b}\,(a))da / \int_0^{\widehat{a}} n(a)f(a,\widehat{b}\,(a))da$ and $E(\,a|\,I) = \int_0^{\widehat{a}} af(a,\widehat{b}\,(a))da / \int_0^{\widehat{a}} f(a,\widehat{b}\,(a))da$. Following the same line of proof as in the previous section, the density function $f(a,\widehat{b}\,(a))$ is more *favourable* than $n(a)f(a,\widehat{b}\,(a))$ if $\frac{n(a)f(a,\widehat{b}(a))}{\int_0^a n(a)f(a,\widehat{b}(a))da} \geq \frac{f(a,\widehat{b}(a))}{\int_0^a f(a,\widehat{b}(a))da}$.

Integrating by parts, this is equivalent to $\frac{n(a)f(a,\widehat{b}(a))}{n(a)F(a,\widehat{b}(a))-\int_0^a n_a(a)F(a,\widehat{b}(a))da} \geq$

$\frac{f(a,\widehat{b}(a))}{F(a,\widehat{b}(a))}$ or $0 \geq -f(a,\widehat{b}\,(a))\int_0^a n_a(a)F(a,\widehat{b}\,(a))da$, which is always satisfied. Therefore, $E^n(\,a|\,I) > E(\,a|\,I)$.