Explicit vs implicit rationing in health care provision: a welfare approach

Laura Levaggi [*] Rosella Levaggi [†]

January 31, 2016

We study the welfare properties of direct restrictions based on cost-effectiveness measures against indirect methods represented by waiting lists, as a policy instrument used to improve equity in the access and finance of a public health care system. Health care is supplied for free, but with some restrictions by the public health sector. Patients can choose to address their demand elsewhere by stipulating a private health care insurance policy. Our model shows that if the individual response to treatment is independent of income and cannot be observed by the patient, the choice of opting out simply depends on income and in this respect both instruments are quite similar. The study of the welfare properties shows that in general there is not a superior instruments. Restrictions may improve welfare of relatively rich individuals. In general, for an equal number of individuals opting out of the public system, explicit restrictions produce a lower welfare loss than implicit instruments.

Keywords: Waiting lists Explicit restrictions Welfare analysis.

[*] Faculty of Science and Technology, Free University of Bolzano-Bozen, Piazza Università 1, 39100 Bolzano-Bozen, Italy, laura.levaggi@unibz.it

[†] (Corresponding author) – Department of Economics and Management, University of Brescia, Via S. Faustino, 74b, 25122 Brescia (Italy). E-mail: rosella.levaggi@unibs.it. Tel: 0039-0302988825. Fax: 0039-0302988837.

1 Introduction

Public health care systems rely on the assumption that health care, being a paternalistic good (Schnellenbach (2012); Kirchgassner (2015)) should be made available independently of ability to pay. However, free health care provision may not be advisable because it may lead to an inappropriate use. In most countries, national authorities implement controls (e.g. budget impact limitations) and incentives (e.g. prescription limitations to be followed by physicians) to restrict the use of public health care and improve average effectiveness. Access to care may be explicitly restricted on the basis of the health gain the patient is expected to obtain (Appleby et al., 2009; Levaggi and Levaggi, 2011a) or care may be supplied with a delay or a copayment (or both),with the aim of discouraging patients with a low benefit to demand care (Gravelle and Siciliani, 2008a,b; Cullis et al., 2000; Lindsay and Feigenbaum, 1984; Martin and Smith, 1999). In a pure public system patients that are not eligible would not receive care; if a parallel private care sector exists patients may choose their preferred provider. This consideration gave rise to a new strand of literature that studies the coexistence of private and public health care (Gouveia (1997); Besley et al. (1999); Cullis et al. (2000); Cuff et al. (2012)). However, since only sufficiently rich people can afford this option, the redistributive properties of waiting lists and other fiscal instruments should be investigated (Besley and Coate (1991); Marchand and Schroyen (2005); Gravelle and Siciliani (2008a,b)), but in this respect the literature has not found a unique answer (Hoel and Saether (2003); Fossati and Levaggi (2008); Zweifel (2015)).

In this paper we compare the welfare properties of direct restrictions based on cost-effectiveness measures against indirect methods represented by waiting lists. The two instruments will be evaluated against the "no restriction" benchmark. In both cases these instruments produce an increase in the expected utility of individuals at the top end of the income distribution, but they do not improve total welfare, a result that contradicts Hoel and Saether (2003) conclusions. The comparison between the two instruments shows that only individuals at the lower end of the income distribution may prefer waiting list to explicit restrictions. In a second best environment explicit restrictions should be preferred to the use of waiting list to optimise total welfare, but they may cause equity problems.

2 The model

The community is made of N individuals, normalised to one, with a fixed exogenous income y distributed with density $g(y)$ in the support (b, a) (the average income will be denoted by y_m) and an endowment H of health which produces a marginal money equivalent utility φ. Illness reduces the health stock to δH with a fixed probability π and the individual experiences a reduction in utility equal to $\varphi\delta H$. Health care allows to gain back θH units of health and has a cost[1] equal to p. The term θH is the effectiveness of the treatment; we assume that it depends on θ, a parameter that captures patient's ability to take advantage of health care.[2] It follows a uniform distribution, independent of income, on the support $[0, 1-\delta]$ with $\theta_m = \int_0^{1-\delta} \theta \frac{1}{1-\delta} d\theta = \frac{1-\delta}{2}$ being its expected value. Health care is provided for free, but with some restrictions (either in terms of access to care or in terms of waiting time before treatment) and its cost is financed using a linear income tax. The tax rate t represents the fraction of income that is necessary to finance health care. In a public health care system without restrictions, the tax rate would be written as:

$$t = \frac{\pi p}{y_m} \tag{1}$$

Public health care expenditure can however decrease if some people are not eligible for treatment, or if they opt out of the public health care system, or both. We also assume that there is a time limit S beyond which the treatment is no longer effective. As in Cuff et al. (2012), delayed treatment produces an opportunity cost that can be translated in terms of a money equivalent level of utility. We assume that the loss can be measured by the following linear function: $\beta T y \theta H$ where T is the time elapsing before being treated and βy captures the greater concern that people with higher income have for their health.[3]

Individuals may opt to pay for a supplementary private insurance, fair from an actuarial point of view (i.e. the price is πp), which allows them to receive care without restrictions or delay. As in Hoel and Saether (2003), both sectors are equally efficient, i.e. they have the same production cost equal to p. Ex-post, patients with an health insurance policy always choose to be treated by the private health care sector, even when they would be eligible to receive public health care. Under these assumptions, the state-contingent utility function for the representative consumer can be written as:

$$U = y(1-t) + \begin{cases} \varphi H & \text{if healthy without a private insurance policy,} \\ -\pi p + \varphi H & \text{if healthy with a private insurance policy,} \\ -\beta y \theta H T + \varphi H\left(\delta + \theta\right) & \text{if ill and treated in the public sector} \\ -\pi p + \varphi H\left(\delta + \theta\right) & \text{if ill with a private insurance policy} \\ -\beta y \theta H S + \varphi H \delta & \text{if ill and not treated.} \end{cases} \tag{2}$$

In our model patients do not suffer any direct health loss from waiting: care is equally effective independently on whether it is offered immediately or with

[1] To simplify the model, we assume that one unit of treatment is sufficient to treat the patient.

[2] For example, an active principle that reduces blood pressure might be less effective if used with other drugs, i.e. for class of patients with multiple diseases.

[3] It derives from several characteristics: a greater opportunity cost of leisure and work time spent in bad health, a greater concern for their physical strength.

some delay. The delay affects the money equivalent utility only through the opportunity cost of waiting. Our assumption is in line with Hoel and Saether (2003), but differs from other models that assume the existence of a fixed cost for receiving health care (Gravelle and Siciliani, 2009, 2008b,a; Martin and Smith, 1999). We believe that our assumption captures the empirical evidence showing that rich people are prepared to wait less before being treated, i.e. delay is predominantly an opportunity rather than a health cost. θ is not known at the time the patients have to choose whether to buy the private insurance, but it can be observed at a later stage. We also assume that individuals at the low end of the income distribution are not able to afford to pay for a supplementary health insurance ($b < \pi p$), a condition that justifies public supply in our context. Finally, to simplify the notation, we also assume that S, the limit for the treatment to be effective, is equal to 1.

2.1 Waiting lists

We consider delay in treatment as an implicit form of rationing. Health care is free at the point of use, but the treatment is delayed by an amount T, equal for all the individuals. This delay reduces the utility of the provision of health care, so that some individuals address their demand to the private sector. The delay T, defined as the time the patient has to wait before receiving treatment in the public sector, is an "optional public bad" since everybody wanting a treatment financed by the public sector waits the same amount of time. In this case there is no uncertainty in being treated and the welfare loss is equal to the individual money equivalent cost $\beta y \theta H T$. The expected loss is then $\pi \beta y \theta_m H T$ and patients buy an actuarially fair private health insurance if

$$\pi \beta y \theta_m H T > \pi p.$$

Thus, if income is at least equal to

$$y_T = \frac{p}{\beta \theta_m H T} \tag{3}$$

they will find it convenient to opt out of the public heath care system. As before, this is a choice only if $T \leq \frac{1}{\pi \beta \theta_m H}$, i.e. T is lower than the delay for which the consumer with income πp leaves the market. Any increase beyond this point will certainly decrease welfare because the loss caused by the increased delay is not matched by a gain in terms of reduction in public expenditure.

2.2 Explicit rationing

In a pure public health care system, treatments should be offered to all those needing it. Universal access without any form of restriction would imply that the average effectiveness would be equal to θ_m and the marginal patients would receive no benefit from being treated. In order to improve average and marginal effectiveness, regulators may set a simple criterion $\theta > c$ for eligibility. The level

c can be chosen with reference to cost effectiveness thresholds (Appleby et al. (2009)) or by taking into account the utility derived from health care (Levaggi and Levaggi, 2011b); we assume that its level is set outside the model. This policy restriction implies that the marginal effectiveness will be equal to c, while the average effectiveness, conditional on being treated, is:

$$\theta^c = \frac{\int_c^{1-\delta} \theta \frac{1}{1-\delta} d\theta}{\int_c^{1-\delta} \frac{1}{1-\delta} d\theta} = \frac{1}{2}\left[c + (1-\delta)\right] = \theta_m + \frac{c}{2},$$

From the point of view of the patient explicit rationing implies that:

- if ill and $\theta \geq c$ the patient is treated without delay ($T = 0$).
- if ill and $\theta < c$ the patient is not entitled to receive care and experiences a welfare loss equal to $\theta H (\beta y + \varphi)$.

In other words, an explicit rationing system privileges some individuals (those having a higher ability to recover) and offers nothing to others (those with a low θ). Given that patients do not observe θ, none of them is certain of being treated. From equation (2) we can derive the welfare loss each patient expects from this occurrence:

$$\pi \left(\varphi H (\delta + \theta) + \beta y \theta H - \varphi H \delta\right) = \pi \left(\theta H (\beta y + \varphi)\right)$$

The expected loss depends on the probability of this event, i.e. on θ being below the threshold c:

$$\pi \left(\int_0^c \theta H (\beta y + \varphi) \frac{1}{1-\delta} d\theta \right).$$

Patients can avoid this uncertainty by stipulating an insurance whose premium is equal to πp. Insurance is then bought if $y > \pi p$ and their expected welfare loss is higher than the insurance premium, i.e.:

$$\pi \left(\int_0^c \theta H (\beta y + \varphi) \frac{1}{1-\delta} d\theta \right) \geq \pi p \tag{4}$$

Having assumed that θ and income are not correlated, the decision to buy the insurance policy simply depends on income. From (4) we can find the income level for which buying an insurance policy is convenient:

$$y \geq y_c = \frac{2p}{\beta H} \frac{1-\delta}{c^2} - \frac{\varphi}{\beta}. \tag{5}$$

The threshold y_c is decreasing in c: the higher the expected loss, the lower the threshold for which it is convenient to buy a private health insurance. In what follows we will assume that

$$c \geq c_{\min} = \sqrt{\frac{2p(1-\delta)}{H(\beta a + \varphi)}} \tag{6}$$

i.e. that c should not be set under the level for which $y_c = a$. Consider also that the purchase of an insurance policy is conditional on two events: individuals should find it convenient (willingness to pay) and should also be able to afford the price (ability to pay). As c increases and y_c decreases, there will exist individuals who would like to opt out, but cannot afford the price because their income is below πp. From (5) there is a restriction to the opting out option if $c > \sqrt{\frac{2p(1-\delta)}{H(\beta\pi p+\varphi)}}$, while for

$$c \leq c_{\max} = \sqrt{\frac{2p\,(1-\delta)}{H\,(\beta\pi p + \varphi)}} \tag{7}$$

being treated in the public sector is a free choice.

3 Comparison of the two instruments

The first comparison can be made in terms of the ex post average effectiveness in the public health care system. In this case, for explicit rationing we can write:

$$\theta^c = \theta_m + \frac{c}{2} \tag{8}$$

In the presence of an implicit restriction the effectiveness in the public health care system is equal to

$$\theta^T = \theta_m \tag{9}$$

hence an explicit rationing is always more effective, as one might expect. Explicit rationing increases average effectiveness because individuals for whom the treatment is not sufficiently effective are not treated. When delay is used as a rationing instrument, health care provision is offered independently of its effectiveness.

The second interesting insight that our model shows is that in both cases the decision to be treated in the private health care market simply depends on income. This allows us to compare the welfare properties of the two instruments.

4 Welfare analysis

The literature has long debated on the use of restrictions as tools to improve income redistribution. In what follows, we show that the two instruments we are analysing are pro-rich, i.e. the gainers are rich rather than poor individuals. The existence of gainers and their number depends on several specific characteristics of the community such as the parameters of the income distribution, the shape of the utility function, the characteristics of the health care services (its price, its effectiveness and its cost-utility). The analysis is quite technical and it is presented in the appendix. In this section we present the main results and an intuitive explanation that allow to grasp the main policy implications behind

our findings.
Throughout this section, unless otherwise stated, the analysis will be conducted for values of c satisfying the constraints in (6) and (7).

4.1 Comparison of individual utilities

The welfare of a community is the aggregation of individual preferences; for this reason we start our analysis from the assessment of gainers and losers from health care restrictions.

Let us now move to the use of delay as a policy variable. The effects in terms of individual utility are summarised in Proposition 1.

Proposition 1. *An implicit restriction to health care ($c = 0, T > 0$) may increase the expected utility of individuals at the top end of the income distribution that opt out of the public health care system. As T increases, the number of gainers increases, but it never coincides with all the individuals opting out.*

Proof. See Appendix A.3. □

To understand the intuition behind this proposition, let us divide the community into two groups of individuals: those opting for a private insurance and those that choose public health care. The former suffer no loss in terms of health or delay, because they are always treated immediately, but they may experience a loss in terms of net income, because they have to pay both their tax and insurance bills. However, as T increases, their tax bill decreases because more individuals opt out of the public health care system and for the richest individuals the reduction in taxation might be higher than the prize for the insurance policy. In fact the derivative of the utility difference for these individuals is increasing in both y and T.
Those that do not opt out suffer an expected loss due to the delay in treatment which is never compensated by the reduction in the tax bill. Moreover, increasing T increases the disutility caused by the delay and it can be shown that the derivative with respect to y of the utility difference for this class of individuals is negative.

This result is quite interesting because it contradicts the conclusions of the previous literature. Hoel and Saether (2003) shows that delays may improve welfare of individuals at the bottom end of the income distribution, while we obtain exactly the opposite result.

Let us now move to the comparison between explicit restriction and no restriction. The main results are summarised in Proposition 2.

Proposition 2. *An explicit restriction to health care ($c > 0, T = 0$) may only increase the utility of individuals at the top end of the income distribution. Depending on the parameters there may exist small values of c such that all the people opting out of the health care system and a fraction of the remaining population gain from the introduction of the restriction. However, as c increases and more individuals opt out, only for a fraction of them the difference in expected* *utility will be positive and all the individuals remaining in the public system will be net losers.*

Proof. See Appendix A.3. □

Let us again analyse the situation separately for individuals opting for a private insurance and those that instead choose to be treated by the public health care system. The former suffer no health care related loss since they are treated with no delay irrespective of their level of θ, but have to pay their tax and insurance bills. However, as c increases, their tax bill decreases: some patients are either not eligible ($\theta < c$) or are treated by the private health system.

The individuals that opt for public health care suffer a loss from the benchmark (no restriction) in terms of expected health and income related utility; on the other hand they gain in terms of net income since their tax bill decreases. In Appendix A.3 we show that only in some cases a part of them may benefit from the introduction of a restriction, but there always exist values of c for which all those that do not opt out are net losers, because the gain in terms of reduced tax bills is lower than the loss in terms of health-related utility.

The relationship between c and the number and distribution of gainers is not clearcut. In general, it is possible to prove that for c sufficiently high only a fraction of the opting out individuals has a gain, while all others are net losers. For smaller values of c different scenarios are possible: either all the individuals opting out of the public market and a fraction of the others have a gain, or only some of the people buying an insurance are gainers, while all the others will lose, or it may happen that implementing an explicit restriction with c too low results in a decreased utility for all the population. A key element in understanding the causes of this behaviour is the term $\frac{\theta_m H(\beta a+\varphi)}{p}$, which can be interpreted as the maximum incremental utility-cost level for the considered treatment. For a fixed income distribution on $[b, a]$, if this ratio is low, (i.e the treatment has a relatively high cost with respect to the average utility gain), an explicit restriction with a low threshold for eligibility can increase the expected welfare of all those opting out. As c increases, the probability of not being treated in the public sector increases and individuals with increasingly lower income will buy an insurance, but only part of them gains from the presence of a restriction. As the incremental utility-cost ratio gets higher, the uncertainty in being treated causes a growing expected utility loss. This means that even if c is relatively small the fraction of people opting out that has a gain decreases and in some extreme cases there will even be no gainers (see Appendix A.3 for more analytical details).

The final question to be answered relates to which of the two instruments is superior. The comparison will be made in terms of the same number of individuals opting out of the public health care system, i.e. $y_c = y_T = L$. Using equations (5) and (3) it is possible to define the relationship between T and c for which this happens:

$$\frac{2p(1-\delta)}{\beta c^2 H} - \frac{\varphi}{\beta} = \frac{p}{\beta \theta_m H T}$$

which implies

$$T = T(c) = p \frac{2c^2}{(1-\delta)\left(2p(1-\delta) - c^2 \varphi H\right)}. \tag{10}$$

The results are summarised in Proposition 3.

Proposition 3. *If $T = T(c)$ is chosen as in* (10) *so that the income level for people opting out from a system with an explicit or an implicit restriction is the same, an implicit restriction may only give higher expected utility to people with a low income.*

Proof. See Appendix A.3. □

In this case in fact, since with an explicit restriction only part of the people that do not buy an insurance will be treated in the public sector, public expenditure is lower and thus the tax bill. For individuals that opt out the difference in expected utility is equal to the gain in taxation and thus have a preference for an explicit restriction. For people with lower but relatively high incomes, this gain is still sufficient to counterbalance the possible loss from not being cured, which implies that some individuals treated in the public health system may prefer an explicit to implicit restriction only because of the reduction in taxes it brings about. Ex ante, if we were to consider only health care gains and the opportunity cost of waiting, patients treated by the public health care system would prefer a delay. Depending on the parameters, for small values of c and b sufficiently high, all the population can be better off with an explicit rationing. However, the limit for b tending to zero is negative for all values of c, thus there exist values of the minimum income b for which an implicit rationing will be preferred by people at the low end of the income distribution.

4.2 Welfare comparison

The aggregation of individual preferences allows us to define the welfare of the community. The policymaker might assign a different weight to the utility of different types of individuals. To model this aspect we use a weight function $w(y)$ such that $\int_a^b g(y)w(y)\,\mathrm{d}y = 1$. For $w'(y) > 0$ more weight is assigned to individuals at the top end of the income distribution, but this objective is not compatible with public provision of health care. For $w(y) = 1$ the utility of each individual has the same weight.

Usually, $w'(y) < 0$, i.e. higher weight is assigned to individuals at the lower end of the income distribution. This assumption allows to introduce in the model the idea of non-linear utility (Fossati and Levaggi, 2008) and is in line with the assumption of the literature (Hoel and Saether, 2003).

As shown in Section 4.1, the introduction of a restriction to the access of health care produces gainers and losers and the sign of the welfare difference clearly depends on the weight function. From these results and the analysis in Appendix B the following considerations can be derived:

- for $w(b) = 1$ and $w(y) = 0$ for all $y \neq b$, in general the introduction of a form of restriction produces a welfare loss. Only for very low values of c and sufficiently high b an explicit restriction may be preferred. In this case, the reduction in the tax bill is marginal for this consumer, while the prospect of not being able to receive care produces a relevant loss in terms of health related utility;

- if we assign equal weight only to those opting for public health provision, any restriction produces a welfare loss. Which of the two should be preferred depends on the income distribution and the extent of the restriction;

- if $w(y) = 1$ for all y, i.e. the welfare is the unweighted sum of individual utility, it is possible to show that an implicit restriction always produces a welfare loss. For an explicit restriction the shape of the difference depends on the income distribution, but in general a loss is more likely, since for c high enough only a fraction of the people opting out has a gain. As per the difference between the two restrictions, the true shape depends on the income distribution. In general, it is possible to prove that if c is not too high, an explicit restriction gives a higher welfare.

Explicit rationing allows to reduce public expenditure, but offers timely treatment only to a group of patients and as income decreases, the benefit from a reduction in the tax level decreases (the individual pays a very limited fraction of health care expenditure) and also the disutility from waiting decreases. On the other hand, the expected disutility from not being treated, which is independent of income, is constant. For this reason, individuals with a low income may be better off by waiting a set amount of time and have a relatively higher tax rate (whose cost is relatively low for them) rather than not being treated.

For $c \geqslant \sqrt{\frac{2p(1-\delta)}{H(\pi p\beta+\varphi)}}$ the threshold income is πp and no further patients can opt out. The same is true for $T = \frac{1}{\pi\beta\theta_m H}$. In this case, cost effectiveness and expenditure can be controlled only using an explicit restriction. In fact for the combinations of $c, T \subset \left(c > \sqrt{\frac{2p(1-\delta)}{H(\pi p\beta+\varphi)}}; T > \frac{1}{\pi\beta\theta_m H}\right)$ the individuals that address their demand to the public sector are the same and correspond to everyone having an income below πp. In the presence of delay, they will all be treated, but with an increasing welfare loss because of the increased time delay. On the other hand, when an explicit rationing is used, the number of those treated decreases as c is increased. However, it should be noted that in this case an increasing number of individuals accept public supply because they cannot afford anything better. In this case a welfare comparison cannot be performed because there is no longer a one to one correspondence between c and T in terms of the same number of individuals that buy a health care insurance.

5 Discussion and future research

The analysis presented in this article shows some important features of rationing in the market for health care. The first important feature is that, from a welfare point of view, delay in treatment and explicit rationing have the same redistributive properties if patients cannot observe their level of effectiveness. In this case the two instruments act as pure income redistribution tools since the choice to leave the public health care system simply depends on income. However, our model shows that explicit restrictions are more effective in this context. In fact, explicit restrictions improve the ex-post appropriateness of care (measured through the ex-post level of effectiveness of care) and allow to obtain a higher level of welfare if they are used as a redistributive tool in most settings. This result is robust and does not depend on the income distribution, but it may depend on the weights used to aggregate preferences. In fact, we show that individuals at the bottom end of the income distribution would certainly prefer to wait rather than to be restricted in their access to health care. The second interesting consideration is that the welfare gains mainly arise from the tax cut that explicit rationing allows to do. This point has important policy implications: in a theoretical model like the one we have presented here, the reduction in expenditure always determines a tax cut; in real world health care system this may not be so automatic.

From a policy point of view, two interesting conclusions can be drawn: first of all, the policy of increasing delay to improve the effectiveness of health care treatments should be closely monitored; as delay increases the number of patients asking for treatments decreases because those having a higher ability to pay can address their demand to the private sector. However, there is a limit beyond which patients would be willing to leave the public sector, but their income is not sufficient. Delay should certainly be set below this threshold. The second question relates to the sustainability of public health care system. When restrictions (either in terms of delay or in terms of access to health care) are too high, only those that cannot afford private health care will use it, and in the long run the consensus to public provision may reduce.

Our results are mainly determined by the assumption of orthogonality between income and effectiveness, an assumption that is common in this literature and may be justified on the argument that ex-ante the individual may not have enough information to determine its response to a treatment. In this case, we also show that explicit restrictions should be preferred from a welfare point of view, a result that is not in line with the rest of literature, which often proposes to use delay and waiting times to improve income distribution.

We think that the analysis presented here may be used as a first step in investigating the effects of the combinations of several instruments, such as effectiveness-specific delay Gravelle and Siciliani (2008a,b, 2009), explicit rationing and, possibly co-payment.

References

Appleby, J., Devlin, N., Parkin, D., Buxton, M., Chalkidou, K., 2009. Searching for cost effectiveness thresholds in the {NHS}. Health Policy 91 (3), 239 – 245.
URL http://www.sciencedirect.com/science/article/pii/S0168851008002911

Besley, T., Coate, S., 1991. Public provision of private goods and the redistribution of income. The American Economic Review 81 (4), 979–984.
URL http://www.jstor.org/stable/2006658

Besley, T., Hall, J., Preston, I., 1999. The demand for private health insurance: do waiting lists matter? Journal of Public Economics 72 (2), 155 – 181.
URL http://www.sciencedirect.com/science/article/pii/S004727279800108X

Cuff, K., Hurley, J., Mestelman, S., Muller, A., Nuscheler, R., 02 2012. Public and private health-care financing with alternate public rationing rules. Health Economics 21 (2), 83–100.

Cullis, J. G., Jones, P. R., Propper, C., 00 2000. Waiting lists and medical treatment: Analysis and policies. In: Culyer, A. J., Newhouse, J. P. (Eds.), Handbook of Health Economics. Vol. 1 of Handbook of Health Economics. Elsevier, Ch. 23, pp. 1201–1249.

Fossati, A., Levaggi, R., 2008. Delay is not the answer: waiting time in health care & income redistribution. Working Papers 0801, University of Brescia, Department of Economics.
URL http://ideas.repec.org/p/ubs/wpaper/0801.html

Gouveia, M., 1997. Majority rule and the public provision of a private good. Public Choice 93 (3/4), 221–244.
URL http://www.jstor.org/stable/30024300

Gravelle, H., Siciliani, L., May 2008a. Optimal quality, waits and charges in health insurance. Journal of Health Economics 27 (3), 663–674.

Gravelle, H., Siciliani, L., September 2008b. Ramsey waits: Allocating public health service resources when there is rationing by waiting. Journal of Health Economics 27 (5), 1143–1154.

Gravelle, H., Siciliani, L., 2009. Third degree waiting time discrimination: optimal allocation of a public sector healthcare treatment under rationing by waiting. Health Economics 18 (8), 977–986.
URL http://ideas.repec.org/a/wly/hlthec/v18y2009i8p977-986.html

Hoel, M., Saether, E. M., 2003. Public health care with waiting time: the role of supplementary private health care. Journal of Health Economics 22 (4), 599 – 616.

Kirchgassner, G., 2015. Soft paternalism, merit goods, and normative individualism. European Journal of Law and Economics, 1–28.

Levaggi, L., Levaggi, R., January 2011a. Welfare properties of restrictions to health care based on cost effectiveness. Health Economics 20 (1), 101–110.

Levaggi, L., Levaggi, R., 2011b. Welfare properties of restrictions to health care based on cost effectiveness. Health Economics 20 (1), 101–110.

Lindsay, C. M., Feigenbaum, B., 1984. Rationing by waiting lists. The American Economic Review 74 (3), pp. 404–417.
URL http://www.jstor.org/stable/1804016

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., Smith, P. C., January 1999. Rationing by waiting lists: an empirical investigation. Journal of Public Economics 71 (1), 141–164.
URL http://ideas.repec.org/a/eee/pubeco/v71y1999i1p141-164.html

Schnellenbach, J., 2012. Nudges and norms: On the political economy of soft paternalism. European Journal of Political Economy 28 (2), 266 – 277.
URL http://www.sciencedirect.com/science/article/pii/S0176268011001480

Zweifel, P., 2015. Rationing of health care: is there an economic rationality to it? The European Journal of Health Economics 16 (8), 797–800.

A Utility comparison

In the next analysis we will use the following notation:

$$G(z) = \int_b^z g(y)\,\mathrm{d}y, \quad F(c) = \int_0^c \frac{1}{1-\delta}\,\mathrm{d}\theta, \quad \theta_a(c) = \int_0^c \frac{\theta}{1-\delta}\,\mathrm{d}\theta$$

along with the already defined

$$y_m = \int_b^a yg(y)\,\mathrm{d}y, \quad \theta_m = \int_0^{1-\delta} \frac{1}{1-\delta}\,\mathrm{d}\theta.$$

A.1 Individual utility for an implicit restriction

In the presence of waiting lists, from equation (2) and recalling that people with income higher than the threshold y_T in equation (3) will opt out of the public sector, the ex-post individual utility can be written as

$$U_T = y(1-t_T) + \begin{cases} \varphi H & \text{if healthy and } b \leq y < y_T, \\ -\pi p + \varphi H & \text{if healthy and } y \geq y_T, \\ -\beta y \theta H T + \varphi H(\delta + \theta) & \text{after treat. and } b \leq y < y_T, \\ -\pi p + \varphi H(\delta + \theta) & \text{after treat. and } y \geq y_T. \end{cases} \tag{11}$$

In this case the number of individuals treated in the public sector is $N_T = G(y_T)$, public expenditure is equal to $\pi p\, G(y_T) = \pi p\, N_T$ and the tax rate is $t_T = \frac{\pi p\, N_T}{y_m}$. Therefore the expected (on falling ill or not and over θ) utility is:

$$\begin{aligned} E(U_T)(y) = {} & y\left(1 - \frac{\pi p\, N_T}{y_m}\right) + (1-\pi)\varphi H + \pi\varphi H(\delta + \theta_m) \\ & - \begin{cases} \pi\beta y H\theta_m T, & b \le y < y_T, \\ \pi\, p & y_T \le y \le a. \end{cases} \end{aligned} \tag{12}$$

A.2 Individual utility for an explicit restriction

If an explicit restriction $\theta \ge c$ is implemented, people with income over the limit y_c in equation (5) will opt out of the public sector and from equation (2) the ex-post individual utility can be written as:

$$U_c = y(1-t_c) + \begin{cases} \varphi H & \text{if healthy and } b \le y < y_c \\ -\pi p + \varphi H & \text{if healthy and } y \ge y_c \\ \varphi H(\delta + \theta) & \text{if } b \le y < y_c \text{ and } \theta \ge c \\ -\beta y\theta H + \varphi H(\delta+\theta) & \text{if } b \le y < y_c \text{ and } \theta < c \\ -\pi p + \varphi H(\delta + \theta) & \text{after treat. and } y \ge y_c. \end{cases}$$

The number of individuals treated in the public sector is $N_c = (1 - F(c))G(y_c)$, therefore public expenditure is given by $\pi p\,(1-F(c))G(y_c) = \pi p\, N_c$ and the tax rate is equal to $t_c = \frac{\pi p\, N_c}{y_m}$.
Thus the expected utility is:

$$\begin{aligned} E(U_c)(y) = {} & y\left(1 - \frac{\pi p\, N_c}{y_m}\right) + (1-\pi)\varphi H + \pi\varphi H(\delta + \theta_m) \\ & - \begin{cases} \pi H\theta_a(c)(\varphi + \beta y) & b \le y \le y_c, \\ \pi\, p & y_c < y \le a. \end{cases} \end{aligned} \tag{13}$$

A.3 Comparison of individual utilities

Let us first compare the "no restriction" case with both implicit and explicit restrictions. When access to public care is unrestricted the utility coincides with U_0 (i.e. either c or T equal to zero). In this case $t_0 = \frac{\pi p}{y_m}$, $y_0 = a$ and the difference of the expected utilities for the implicit case is:

$$E(U_T) - E(U_0) = y\frac{\pi p\,(1-N_T)}{y_m} - \begin{cases} \pi\,\beta y H\theta_m T & b \le y < y_T \\ \pi\, p & y \ge y_T \end{cases} \tag{14}$$

In this case the analysis is quite simple: the existence of a restriction makes richer people opt out of the public sector, thus public expenditure decreases and so the tax bill. The first term in (14) is the gain deriving from this effect;

the second term represents the loss due to either the waiting time or the cost of buying a private insurance policy.
For people opting out of the public sector, the difference in (14) increases linearly in y and is positive only if $y > \dfrac{y_m}{1-N_T}$. Since

$$y_T\,(1-N_T) = y_T \int_{y_T}^{a} g(y)\,\mathrm{d}y \le \int_{y_T}^{a} y\,g(y)\,\mathrm{d}y \le y_m \tag{15}$$

only a fraction of the people opting out may have a gain from the presence of a restriction.
Moreover, since $y_T = \dfrac{p}{\beta H \theta_m T}$ the difference in (14) for people treated by the public sector can be rewritten as:

$$E(U_T) - E(U_0) = \pi p\, y \left(\frac{1-N_T}{y_m} - \frac{1}{y_T} \right)$$

and from (15) it is always negative and decreasing in y.

If we instead compare an explicit restriction with no restriction we have:

$$\begin{aligned} E(U_c) - E(U_0) = {} & y\frac{\pi p\,(1-N_c)}{y_m} \\ & - \begin{cases} \pi\, H\theta_a(c)(\varphi + \beta y) & b \le y < y_c \\ \pi\, p & y \ge y_c \end{cases} \end{aligned}$$

As for the implicit case, the first term represents the gain everyone gets from a reduction in the own tax bill, the second is the loss deriving either from non being treated or from the cost of the insurance. The analysis of gainers and losers is not as straightforward as before.
Although the derivative for y is positive for all those opting out (and it increases with c), the gain in utility is positive only when

$$y > \frac{y_m}{1-N_c} = y^*. \tag{16}$$

The threshold y^* can be higher or lower than y_c, depending both on the distribution of the income, the parameters and the value of c.

- For c is sufficiently high $y_c < y_m$, then $y^* > y_c$, because y^* is always higher than y_m.
- Depending on the parameters, as c decreases, the inequality $y^* > y_c$ may still be valid, or there may exist low values of c for which instead $y^* < y_c$. In fact, examining the situation for the minimum value $c_{\min}$ defined in (6), since in this case $y_c = a$ the corresponding value of y^* is

$$y^* = \frac{y_m}{1-(1-F(c))\cdot G(y_c)} = \frac{y_m}{F(c_{\min})} = \frac{y_m\,(1-\delta)}{c_{\min}},$$

which from (6) is equal to

$$y_m \sqrt{\frac{(1-\delta)H(\beta a+\varphi)}{2p}} = y_m \sqrt{\frac{\theta_m H(\beta a+\varphi)}{p}}.$$

If the above expression is lower than a, since y^* and y_c are continuous in c, for low values of c the inequality $y^* < y_c$ is be valid.

In general, both y_c and y^* are decreasing in c, but, while the first is always convex, the behaviour of the second depends on the income distribution. In the uniform case where $g(y) = \frac{1}{a-b}$ both functions are convex, thus if an intersection between y_c and y^* exists it is unique and this only happens if $y^*(c_{\min}) < a$, as described above. In more general cases the presence of multiple intersections cannot be excluded.

Let us now compare the individual expected utilities for implicit and explicit restrictions when the same number of individuals opt out, i.e. $y_c = y_T = L$. The expressions of y_c and y_T are equal if and only if

$$T = T(c) = \frac{p}{\left(\frac{p}{\theta_a(c)} - \varphi H\right)\theta_m T}. \tag{17}$$

For each individual it holds

$$\begin{aligned} E(U_c) - E(U_T) = {} & y\frac{\pi p\, F(c)\, G(L)}{y_m} \\ & - \begin{cases} \pi\,\varphi\, H\theta_a(c) + \pi\,\beta\, y\, H(T(c)\theta_m - \theta_a(c)) & b \le y < L \\ 0 & y \ge L \end{cases} \end{aligned}$$

which from (17) can also be written as

$$\begin{aligned} E(U_c) - E(U_T) = {} & y\frac{\pi p\, F(c)\, G(L)}{y_m} \\ & + \begin{cases} \pi\,\varphi\, H\theta_a(c)\left(\frac{y}{L} - 1\right) & b \le y < L \\ 0 & y \ge L \end{cases} \end{aligned} \tag{18}$$

This function is increasing in y and positive for any $y \ge L$, that is anyone opting out is better off with an explicit restriction, since taxes will be lower. Part of the remaining population with relative high income will still prefer the explicit restriction and, at least for small values of c, depending on the size of b, all the population has a positive differential. On the other hand, the limit for b tending to zero is negative for all values of c, thus there exist values of the minimum income b for which an implicit rationing will be preferred by people with low income.

Note that the gain caused by the taxation difference depends on the quantity $F(c)\, G(L(c))$, expressing the number of people that will not be cured. The behaviour with respect to c depends on the income distribution. In the uniform case the derivative with respect to c is negative, i.e. the differential for the individuals opting out is higher when c is minimal. If the distribution is a Gaussian it is first increasing and then decreasing.

B Welfare

The welfare of the community is defined as the aggregation of individual preferences allows us to define the welfare of the community. The policymaker might assign a different weight to the utility of individuals with different incomes. We introduce into the model a non negative weight function $w(y)$ such that $\int_b^a w(y)g(y)\,\mathrm{d}y = 1$ and define the welfare function as

$$\int_b^a \left(\int_0^{1-\delta} U(y,\theta)\,\frac{1}{1-\delta}\,\mathrm{d}\theta \right) g(y)\,w(y)\mathrm{d}y = \int_b^a E(U)(y)\,g(y)\,w(y)\mathrm{d}y$$

To simplify equations let us define

$$G_w(z) = \int_b^z g(y)w(y)\,\mathrm{d}y, \quad y_w = \int_b^a y\,g(y)w(y)\,\mathrm{d}y.$$

From equation (12) the welfare for an implicit restriction is

$$\begin{aligned} W_T = {} & y_w(1-t_T) - \pi p(1-G_w(y_T)) + \varphi H(1-\pi+\pi\delta) + \pi\varphi H\theta_m \\ & - \pi\beta HT\theta_m \left(\int_b^{y_T} yg(y)w(y)\,\mathrm{d}y \right) \end{aligned}$$

and the welfare difference between waiting lists and no restriction to access is

$$W_T - W_0 = \pi p\,\frac{y_w}{y_m}(1-N_T) - \pi p(1-G_w(y_T)) - \pi p \int_b^{y_T} \frac{y}{y_T} g(y)w(y)\,\mathrm{d}y.$$

The first term two terms are related respectively to the difference in public and private expenditure, the last to the cost of the delay. For $w(y) = 1$ the difference in total expenditure is zero, therefore $W_T - W_0 < 0$ for any $T > 0$. Also, from the analysis in Appendix A.3 if the weight w is nonzero only for incomes below y_T the welfare difference is trivially negative.

From equation (13) for an explicit restriction the welfare function is

$$\begin{aligned} W_c = {} & y_w(1-t_c) - \pi p(1-G_w(y_c)) + \varphi H(1-\pi+\pi\delta) + \pi\varphi H\theta_m \\ & - \pi H\theta_a(c) \int_b^{y_c} (\varphi+\beta y)g(y)w(y)\,\mathrm{d}y \end{aligned}$$

and the welfare difference between an explicit restriction and no restriction to access is

$$\begin{aligned}W_c - W_0 &= \pi p\,\frac{y_w}{y_m}(1-N_c) - \pi p(1-G_w(y_c)) - \pi H\varphi\theta_a(c)G_w(y_c)\\&\quad - \pi\beta H\theta_a(c)\int_b^{y_c} y\,g(y)w(y)\,\mathrm{d}y\\&= \frac{\pi p}{y_m}\,(y_w(1-N_c)-y_m) + \pi\beta H\theta_a(c)\int_b^{y_c}(y_c-y)\,g(y)w(y)\,\mathrm{d}y.\end{aligned}$$

The first term is negative and increasing in c, while the second is positive and decreasing in c and the sign of the difference depends on the relative size of the two summands. As already discussed in Appendix A.3, the number of gainers and losers in the comparison depends on the parameters, the income distribution and the magnitude of c. In all cases where only a fraction of the people that opt out has a gain, the welfare difference is obviously negative if positive weight is assigned only to those remaining on the public market. If instead $w(y) = 1$ for all y the formula simplifies to:

$$W_c - W_0 = -\pi p\,N_c + \pi\beta H\theta_a(c)\int_b^{y_c}(y_c-y)\,g(y)\,\mathrm{d}y.$$

The behaviour depends on the parameters and on the shape of the income distribution. In case the latter is uniform the difference $W_c - W_0$ is increasing in c and, depending on the parameters can be either negative for all values of c or only up to a certain level.

Let us now compare the welfare functions for implicit and explicit restrictions when the same number of individuals opt out, i.e. $y_c = y_T = L$. From equation (18) the welfare difference is

$$\Delta W_{c,T(c)} = \pi p\,F(c)\,G(L)\,\frac{y_w}{y_m} - \pi\,\varphi\,H\theta_a(c)\frac{1}{L}\int_b^L (L-y)\;g(y)w(y)\,\mathrm{d}y$$

The first summand is the difference in taxation, which is always positive; the second is instead negative and comprises both loss in health stock and the disutility caused by delays. When $w(y) = 1$ for any y the formula simplifies to

$$\Delta W_{c,T(c)} = \frac{\pi}{1-\delta}\left(\int_b^L\int_0^c \left(p - \varphi\,H\theta + \varphi\,H\theta\frac{y}{L}\right)\,\mathrm{d}\theta\;g(y)\mathrm{d}y\right)$$

which is of course positive for any c satisfying $p - \varphi\,H\frac{c}{2} \geq 0$, i.e. if $c \leq \frac{2p}{\varphi\,H}$. If this condition is satisfied, whatever the income distribution, the welfare for an explicit restriction is higher. If moreover the distribution is uniform, from the discussion in Appendix A.3 it follows that the difference is decreasing in c