Can a Mixed Health Care System be Desirable on Equity Grounds ?*

Maurice Marchand† and Fred Schroyen‡

10 February 2003

Abstract

Should health care provision be public, private, or both? We look at this question in a setting where people differ in their earnings capacity and face some illness risk. We assume that illness reduces a person's time endowment when waiting for treatment. Treatment can be obtained in a competitive private sector (through private insurance) or in the National Health Service (NHS) where it is provided free of charge but after some (endogenous) waiting time. The equilibrium in the health care sector consists of a waiting time in the NHS such that no patient wants to switch health care provider. This equilibrium is governed by two public policies: the income tax system and the size of the NHS. Our findings are threefold. First, a mixed system with a small public health care sector always gives a lower social welfare than a pure public system. Second, a mixed system with a sufficiently large NHS may improve upon a pure public system if the dispersion of earnings capacities is large enough. And finally, when health risk is negatively correlated with ability, there is an extra argument for a large NHS.

JEL classification: H21, H51, I18
Keywords: health care systems, redistribution

* This is a thoroughly revised version of the paper *Markets for public and private health care: redistribution arguments for a mixed system* (NHH-DP 20/2000, CORE-DP 2001/23). We are grateful to Tim Besley, Hugh Gravelle, Paula González, Jean Hindriks and Pierre Pestieau for their comments on that earlier version of this paper. The paper benefited as well from the comments by seminar participants at the University of Oslo and at the University of Warwick. Fred Schroyen thanks the Programme for Health Economics in Bergen (HEB) for financial support through a SNF grant. This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

†CORE (Université catholique de Louvain), 34 voie du Roman Pays, B-1348 Louvain-la-Neuve (Belgium). E-mail: mmarchand@core.ucl.ac.be

‡Dept of Economics, Norwegian School of Economics & Business Administration, Helleveien 30, N-5045 Bergen (Norway). E-mail: fred.schroyen@nhh.no

1 Introduction

In several OECD countries, health care is mainly provided publicly and financed out of tax revenue or social insurance contributions. Examples are Norway, Sweden and the United Kingdom. In these countries, there also exists a parallel private health care sector. In Norway, this private sector is still small, but experience from other countries suggests that it may grow to a significant level. In the UK, where the NHS is free of charge, the proportion of private expenditure in total expenditure on health care has grown from 9% in 1979 to 15% in 1995 (Propper, 2000). The other extreme is a health care system mainly financed by private means, as in the US (and Switzerland, up til 1995).

There exists by now a large literature collecting the arguments in favour and against public and private health care systems. The papers by Besley and Gouveia (1994), Cullis *et al* (2000), Propper and Green (2001) are examples. This literature covers many dimensions: from efficiency and equity to political sustainability and administration. The purpose of our paper is more modest in the sense that we want to construct a formal and consistent framework within which we can discuss *some*, but certainly not all, dimensions in the debate on a 'private *vs* public' health care system. Our concern in this paper is to examine whether a mixed system, in which public and private sectors coexist, can be superior, on equity grounds, to a fully public system.

It was indeed claimed by Besley and Coate (1991) that when there are limits to redistribution, such mixed system can be socially optimal for providing private goods like education or health care. In their framework, such private goods can either be acquired free of charge from the public sector or be bought at a price in the private sector. If the quality level in the public sector is lower than in the private market, some people will be willing to pay for a higher-quality good in the private sector. Furthermore, if quality is a normal good, these individuals will also have the highest incomes. Besley and Coate show that in such framework public provision of the private good can redistribute income from rich to poor when it is financed by a head tax levied on all citizens irrespective of the sector they resort to. Whereas their argument is presented in general terms, our purpose in this paper is to verify whether it still applies when some specifities of health care are accounted for.

We consider an economy where citizens differ in earning capacity and face the risk of needing a well-defined medical treatment once a year. For receiving treatment they can either resort to the NHS (i.e. the public health care sector) or sign a private health insurance policy that delivers them treatment on the spot. In the former case, health care is free of charge but rationing takes place through waiting lists. In the latter, the competitive price mechanism makes demand compatible with supply. In equilibrium, no citizen wants to change health care provider. In particular, all citizens with a earnings capacity below a certain level will resort to the NHS while the others buy a private insurance policy. Thus, in this paper we will ignore the information asymmetries between patients, physicians and health insurers, causing moral hazard, adverse selection and risk selection problems. Those are important issues in the markets for health care, and the only justification for not including them in our setting is to develop a tractable model for studying the equity issues we focus on in this paper.

We analyse the arguments that a welfare maximizing government should account for when deciding on the size of the NHS and the parameters of a linear income tax scheme. In this respect, a key question is whether a mixed health care system is desirable. There are deadweight losses in a mixed system: these are due to the waiting lists that cause discomfort, inconvenience and even more painful complications to patients. In our setting waiting lists provoke deadweight losses of resources in the form of shorter time available for leisure and labour. A mixed system is only desirable if the benefits of redistribution outweigh those deadweight losses. Our main conclusions are threefold. First, a mixed health care system with a small NHS is never desirable. If the size of the NHS is too small, the social benefits are of second order importance relative to the associated social costs. Second, it may be optimal to have a mixed system that include a large enough NHS, but a necessary condition for this is that the spread in the income distribution is sufficiently wide. Otherwise, it is optimal to have a fully public system covering the needs of the whole population. Third, when individual health risk is negatively correlated with earnings capability, there is a further argument for a large NHS: a more equitable risk pooling arrangement.

In our setting, it is noteworthy that if there were no limits to redistribution, that is if lump-sum income taxes and transfers could be differentiated by individual abilities, an indifference between the pure private and public systems would obtain in the absence of any correlation between health risk and ability while a fully public system would be superior to any other system in the presence of a negative correlation. This emphasizes that when it is socially optimal to operate a mixed system, it is because it allows to move the limits to redistribution beyond those implementable through feasible tax-and-transfer policies.

Analytical work on these issues is both recent and sparse. We mention two contributions related to our paper. Iversen (1997) lets patients differ in their income and the expected health benefit of treatment. He looks at the effect of a private sector on the waiting time for treatment in public hospitals. When patients are admitted to a waiting list without consideration of the expected health benefit of treatment, Iversen shows that the presence of a private sector results in a longer waiting time if the demand for treatment in public hospitals is sufficiently elastic with respect to waiting time. When waiting list admissions are rationed, the waiting time is shown to increase if public-sector physicians are allowed to work in the private sector in their spare time.

The model developed by Hoel and Sæther (2000) is closer to ours. They have patients differing in their willingness to wait for treatment. There is a public health sector where patients are put on a waiting list and are treated at a constant marginal cost. But patients have also the option to turn to a private sector where the marginal cost of treatment is at least as high as in the public sector. They find that it may be optimal to have an active private sector if there is sufficient inequality in patient's willingness to spend time waiting. They also discuss the optimal level of subsidy of private care and how the size of that subsidy affects the political support for a public health system with a lower waiting time.

The paper is organized as follows. First we discuss patients' choice of resorting to either the NHS or a private insurance contract (Section 2). Next, we study in our basic model how the equilibrium in the health care sector determines the waiting time in the NHS (Section 3). Thereafter, we set up the normative problem (Section 4) and analyse the optimality properties of the size of the NHS (if a mixed system is desirable) and the linear tax policy. We then provide numerical simulations to assess the desirability of a mixed system

(Section 5). Finally, we extend the basic model by allowing for discomfort of illness as well as a health risk that depends negatively on ability (Section 6). Concluding remarks are offered in Section 7.

2 The basic model

In the simplest setting developed in this paper, citizens only care about their consumption of a composite good and leisure, denoted by c and ℓ respectively. Their preferences on these two goods are described by a strictly concave utility function $u(c, \ell)$. There is some probability, denoted by π, that any individual will suffer from illness, in which case his or her labour productivity diminishes. In our setting it takes the form of a reduction in the individual's time endowment available for labour and leisure. This time endowment is equal to A for an individual in good health and to $A - \lambda(w)$ for an individual being sick, where $\lambda(w)$ is the loss of time caused by illness. This increases with w, the time a sick person has to wait before receiving medical treatment. This function satisfies the following properties: $\lambda(0) = 0$, $\lambda'(w) > 0$ and $\lambda''(w) \geq 0$.

Citizens differ in their earnings ability denoted by a. This is distributed on the support $[\underline{a}, \overline{a}]$ according to distribution function $F(a)$ with density function $f(a) > 0$ for any $a \in (\underline{a}, \overline{a})$. Let L denote an individual's labour supply. Labour earnings (aL) are subjected to a linear income tax characterized by a constant marginal tax rate, denoted by t, and a lump-sum transfer, denoted by T. Thus the available income of an individual of ability a amounts to $(1-t)aL+T$ while his or her leisure time, ℓ, is equal to either $A - L$ in case of good health or $A - \lambda(w) - L$ in case of illness.

A citizen can choose to receive medical treatment either in the NHS or in a private practice (whose fee is covered by a private insurance contract), this choice being made before the state of health is known. There is free access to the NHS that is financed out of income tax revenue. However, a sick person having opted for the NHS will be put on a waiting list before receiving medical treatment. The optimal labour supply of an individual having chosen the NHS will depend upon his or her state of health since his or her time endowment will depend upon it. Whatever the state of health, it satisfies $(1-t)a\partial u/\partial c = \partial u/\partial \ell$. Using index Ng and Nb for NHS in the good and bad health states respectively, this yields the following conditional labour supply and indirect utility functions:

$$L^{Ng} = L((1-t)a, T, A) \ \text{ and } L^{Nb} = L((1-t)a, T, A - \lambda(w)) \quad (1.1)$$

$$v^{Ng} = v((1-t)a, T, A) \ \text{ and } v^{Nb} = v((1-t)a, T, A - \lambda(w)) \quad (1.2)$$

where the three arguments of these functions are the net-of-tax wage rate, the lump-sum transfer and the time endowment depending upon the health status. The above indirect utility functions satisfy the well known Roy identities:

$$v_t^i = -aL^i v_T^i, \quad i = Ng, Nb, \quad (2.1)$$

where v_T^i is the marginal utility of income and subscripts denote partial derivatives. We also have:

$$v_\lambda^{Nb} = -(1-t)av_T^{Nb}. \quad (2.2)$$

The expected indirect utility of an individual having chosen to be treated in the NHS in case of illness is given by:

$$Ev^N = (1-\pi)v^{Ng} + \pi v^{Nb} \quad (3)$$

where (to recall) π stands for the probability of falling sick.

On the other hand, if an individual opts for a private insurance policy, he or she will be given medical treatment on the spot ($\lambda(w) = 0$) but will have to pay a fee-for-service q. We assume that competitive insurance contracts are available that provide full coverage of this risk. Therefore the labour supply and indirect utility functions do not depend upon the individual's state of health. The insurance premium being πq, they are given by:

$$L^P = L((1-t)a, T - \pi q, A) \quad (4.1)$$

and

$$v^P = v((1-t)a, T - \pi q, A) \quad (4.2)$$

where upperscript P refers to private medicine. Note that the second argument is $T - \pi q$ (instead of T) to account for the fact that individuals opting for a private insurance have their income available for consumption of the composite good reduced by the insurance premium. As earlier, the indirect utility satisfies

$$v_t^P = -aL^P v_T^P. \quad (5)$$

We now turn to the individual's choice of the health care provider. If $Ev^N \geq v^P$ the individual opts for the NHS; if $Ev^N < v^P$, he or she opts for a private health insurance policy. Since Ev^N and v^P depend upon ability a through the individual's income, this choice differs across ability types. Throughout the paper we maintain the following normality assumption.

Assumption N. *For any t, T and w, there exists some critical ability level $\hat{a}$ such that*

$$\begin{aligned} Ev^N &\geq v^P \quad \text{for any } a \leq \hat{a}, \\ Ev^N &< v^P \quad \text{for any } a > \hat{a}. \end{aligned}$$

In other words, the least able persons opt for the NHS while the most able ones opt for private medicine.

This assumption simply means that the quality of health care – here inversely related to waiting time – is a normal good. This is in line with the empirical literature that shows that the quality of care provided rises with income.

3 The comparative statics of the waiting time in the NHS

At the end of the previous section, attention was focused on the critical ability level $\hat{a}$ and so on the proportion of individuals resorting to the NHS, $F(\hat{a})$, for a given tax system (t and T) and a given waiting time in the NHS (w). This enables us to determine the size of the NHS, that is the supply of NHS services, needed to satisfy demand. With S denoting this size, we simply have $S = \pi F(\hat{a})$.

However, when formulating the government's problem in the next section, the size of the NHS will be taken as a government decision variable (together with t and T). What will then matter is how the waiting time adjusts for the demand for NHS services to clear their supply (S). This reflects the idea exposed in the introduction that in our setting, the waiting time for being treated in the NHS is used as a rationing device. Since $\hat{a} = F^{-1}(S/\pi)$, choosing S amounts to choosing $\hat{a}$. Therefore, in the rest of this section, we shall investigate the comparative statics of the waiting time in the NHS with respect to $\hat{a}, t$ and T, the results of which will be used in the next section.

To this end, let us first define $\Delta(\hat{a}, t, T, w)$ as:

$$\Delta(\hat{a}, t, T, w) \stackrel{\text{def}}{=} v((1-t)\hat{a}, T - \pi q, A) - (1-\pi)v((1-t)\hat{a}, T, A)$$
$$-\pi v((1-t)\hat{a}, T, A - \lambda(w)) = 0 \tag{6}$$

where the equality to 0 is due to the indifference of individuals of ability $\hat{a}$ between a private insurance and the NHS ($v^P - Ev^N = 0$ for these individuals). Remark that Assumption N implies:

$$\frac{d\Delta}{d\hat{a}} = (1-t)\left[\hat{L}^P \hat{v}_T^P - (1-\pi)\hat{L}^{Ng}\hat{v}_T^{Ng} - \pi\hat{L}^{Nb}\hat{v}_T^{Nb}\right] > 0, \tag{7.1}$$

where a hat on a function means that it is taken at $a = \hat{a}$. We also have:

$$\frac{\partial\Delta}{\partial w} = \pi(1-t)\hat{a}\,\hat{v}_T^{Nb}\lambda'(w) > 0, \tag{7.2}$$

$$\frac{\partial\Delta}{\partial t} = -\frac{\hat{a}}{1-t}\frac{\partial\Delta}{\partial\hat{a}} < 0 \tag{7.3}$$

and

$$\frac{\partial\Delta}{\partial T} = \hat{v}_T^P - (1-\pi)\hat{v}_T^{Ng} - \pi\,\hat{v}_T^{Nb}. \tag{7.4}$$

Using the above derivatives we obtain the following comparative static results:

$$\frac{\partial w}{\partial\hat{a}} = -\frac{\mathrm{d}\Delta/\mathrm{d}\hat{a}}{d\Delta/dw} < 0, \tag{8.1}$$

$$\frac{dw}{dt} = -\frac{\partial\Delta/\partial t}{\partial\Delta/\partial w} > 0, \tag{8.2}$$

and

$$\frac{\partial w}{\partial T} = -\frac{\partial\Delta/\partial T}{\partial\Delta/\partial w} \lesseqgtr 0. \tag{8.3}$$

A key consequence of Assumption N is that an increase in $\hat{a}$, and so in the size of the NHS, causes the equilibrium waiting time to fall.

4 The government's problem

To evaluate social welfare, we assume the following social welfare function defined over expected utilities:

$$SW \stackrel{\text{def}}{=} \int_{\underline{a}}^{\hat{a}} \psi(a)Ev^N(a)\mathrm{d}F(a) + \int_{\hat{a}}^{\overline{a}} \psi(a)v^P(a)\mathrm{d}F(a)$$

where the weight $\psi(a)$ is non-increasing in ability. The advantage of this formulation over the more standard concave transformation of (expected) utilities is that it allows for an explicit solution to the optimal tax problem in the numerical examples we present later on (see also Deaton, 1983). Besides the utilitarian case ($\psi(a) = 1, \ \forall \ a$), it contains the rank-ordered social welfare function as a special case. In the latter, $\psi(a) = 1 - F(a)$, such that the worst off agent gets unit weight, the person in the F-th percentile gets weight $1 - F$, and the best-off agent gets weight zero. It can be shown that social welfare can then be written as the product of the mean of the utility distribution and concentration measure equal to "1 − the Gini-coefficient of the utility distribution".

As already mentioned, the government ought to choose the size of the NHS, which is equivalent to choosing $\hat{a}$, and the parameters of the linear income tax system, t and T. They are the solution to the following problem:

$$\max_{\hat{a},t,T} SW \equiv \int_{\underline{a}}^{\hat{a}} \psi(a) \Big[(1-\pi)\, v((1-t)a, T, A) + \pi v((1-t)a, T, A - \lambda(w(\hat{a}, t, T)) \Big] \mathrm{d}F(a) + \int_{\hat{a}}^{\overline{a}} \psi(a) v((1-t)a, T - \pi q, A) \mathrm{d}F(a) \quad (9)$$

subject to

$$t \int_{\underline{a}}^{\hat{a}} a \Big[(1-\pi)\, L((1-t)a, T, A) + \pi L((1-t)a, T, A - \lambda(w(\hat{a}, t, T)) \Big] \mathrm{d}F(a) + t \int_{\hat{a}}^{\overline{a}} a L((1-t)a, T - \pi q, A) \mathrm{d}F(a) - T - \overline{R} - \pi q F(\hat{a}) \geq 0, \quad (10)$$

where $\overline{R}$ is the exogenously fixed amount of public expenditures for other purposes than income redistribution and the NHS. Note that the last term on the *lhs* of the budget constraint, $\pi q F(\hat{a})$, is the overall cost of the NHS. Therefore, q is assumed to be both the price of medical treatment in the competitive private market and its unit cost in the NHS.

The optimal size of the NHS can correspond to either one of two corner solutions or an interior solution. At these two corner solutions health care is provided by either only the NHS ($\hat{a} = \overline{a}$) or only private medicine ($\hat{a} = \underline{a}$). However in our basic setting the social welfare function takes the same value at these corner solutions. The reason is twofold: first, waiting lists are not needed to ration the demand for the NHS-services when citizens cannot opt for a private practice (which corresponds to the upper corner solution $\hat{a} = \overline{a}$)[1] and, second, it is equivalent for citizens to pay for their expected cost of medical treatment

[1] The function $w(\widehat{a})$, implicitly defined by (6), exhibits a discontinuity at $\widehat{a} = \overline{a}$ where its drops to zero.

through either an insurance premium (πq) at the lower corner solution ($\hat{a} = \underline{a}$) or through a reduction in T of the same magnitude at the upper corner solution ($\hat{a} = \overline{a}$). It is worth mentioning that this equivalence between the two corner solutions will not longer hold in the extension of the basic model that we will present in Section 6.

Whether the optimal size of the NHS corresponds to an interior or corner solution will be crucial in the next section where numerical results will be presented. For the time being our purpose in the remainder of this section is to characterize interior solutions to the government's problem formulated above.

The solution to this problem is formally derived in the Appendix. Here we focus on the interpretation of the results, looking first at the optimal choice of the NHS size. Denoting by $\mathcal{L}$ the Lagrangian of the government's maximization problem and by μ the multiplier of its budget constraint, the following expression is derived in the appendix:

$$\begin{aligned} \frac{1}{\mu}\frac{\partial \mathcal{L}}{\partial \hat{a}} &= -\pi \int_{\underline{a}}^{\hat{a}} \Big[(1-t)aB^{Nb}(a) + ta\Big] \mathrm{d}F(a)\lambda'(w)\frac{\partial w}{\partial \hat{a}} \\ &\quad + \Big\{ t\hat{a}\Big[\pi \hat{L}^{Nb} + (1-\pi)\hat{L}^{Ng} - \hat{L}^{P}\Big] - \pi q \Big\} f(\hat{a}) \qquad (11) \end{aligned}$$

where

$$B^{Nb}(a) \equiv \frac{1}{\mu}\psi(a)v_T^{Nb} + taL_T^{Nb} \qquad (12)$$

is the net marginal social valuation of the income of a person of ability a who has opted for the NHS and is sick. Since everything in this expression has been divided by μ, it is expressed in terms of government revenue.

The economic interpretation of the expression on the *rhs* of (11) is straightforward. The first term accounts for the fact that an increase in $\hat{a}$ (and so in the size of the NHS) causes the time endowment of the sick persons resorting to the NHS to rise by $-\lambda'(w)\frac{dw}{d\hat{a}} > 0$ since the waiting time diminishes. This has an income effect that amounts to $(1-t)a$ for a person of ability a, which is valued at $B^{Nb}(a)$, and also a direct effect on the tax revenue collected from such a person that amounts to ta. The second term in (11) reflects the budgetary implications for the government of those individuals withdrawing from the private insurance market. On the one hand, $\pi\, q$ stands for the government's additional expected health expenditures per switching individual. On the other hand, those switching individuals change their (expected) labour supply by

$\pi\hat{L}^{Nb} + (1-\pi)\hat{L}^{Ng} - \hat{L}^P < 0$. This negative sign comes from the assumption that the composite commodity and leisure are normal goods and the facts that the switching individuals have their time endowment reduced and do no longer need to pay an insurance premium.

Evaluating (11) for $a = \underline{a}$ leaves us with no benefits and only budgetary costs. Therefore, introducing a small NHS sector is harmful for social welfare. As expression (11) indicates, this result is explained by the number of NHS patients, $F(\hat{a})$, who benefit from the fall in waiting time when the size of the NHS is increased, relative to the number of patients, $f(\hat{a})$, who shift from the private sector to the NHS and so negatively affect the government's budget balance. When the NHS is of small size, there are only a few individuals benefiting from the fall in waiting time, and so the social cost of an increase in this size outweighs its social benefit.[2] The same reasoning also explains why the social benefit of an increase in the NHS size can dominate its social cost when the size of the NHS is large enough: there are then enough patients who benefit from the reduction in waiting time.

If it is optimal to have a strictly positive NHS, the welfare effects of a reduction in waiting time should at the margin balance its budgetary implications. Setting therefore (11) to zero and rearranging gives us:

$$\frac{\pi E\big[(1-t)aB^{Nb}(a) + ta \mid a \le \hat{a}\big]\lambda'(w)\left(-\dfrac{\partial w}{\partial \hat{a}}\right)}{t\big[\hat{L}_p - \pi\hat{L}^{Nb} - (1-\pi)\hat{L}^{Ng}\big] + \dfrac{\pi q}{\hat{a}}} = \frac{\hat{a}f(\hat{a})}{F(\hat{a})}\,. \tag{13}$$

The *lhs* is the ratio of the benefit per NHS patient of the rise in $\hat{a}$ (again measured in units of government revenue) to its budgetary cost per patient moving to the NHS, while the *rhs* is the elasticity of the distribution function at $\hat{a}$. For many familiar distribution functions, this elasticity falls in $\hat{a}$.

[2] This reasoning is valid when $f(\underline{a}) > 0$. When $f(\underline{a}) = 0$, $\frac{\partial \mathcal{L}}{\partial \widehat{a}}|_{\widehat{a}=\underline{a}} = 0$, and we need to investigate the sign of $\frac{\partial^2 \mathcal{L}}{\partial \widehat{a}^2}|_{\widehat{a}=\underline{a}}$. Since $\frac{\partial^2 \mathcal{L}}{\partial \widehat{a}^2}|_{\widehat{a}=\underline{a}} = \mu\{t\widehat{a}[\pi\widehat{L}^{Nb} + (1-\pi)\widehat{L}^{Ng} - \widehat{L}^P] - \pi q\}\widehat{f}|_{\widehat{a}=\underline{a}}$, and the square bracket term is negative due to Assumption N, we can claim that $\frac{\partial^2 \mathcal{L}}{\partial \widehat{a}^2}|_{\widehat{a}=\underline{a}} < 0$. Initially, social welfare is therefore a concave fucntion of $\widehat{a}$, and will never increase with $\widehat{a}$ at $\underline{a}$.

In the next section we construct numerical examples showing that in some circumstances it can be socially optimal to have a mixed health sector, involving both a NHS for the least able persons and private medicine for the most able persons. For the sake of completeness, we close the present section by characterising the optimal linear tax policy. The following expression of the marginal tax rate is derived in the appendix:

$$t = \frac{-\text{cov}(B(a), aL) - \left[\pi F(\hat{a}) E\big[(1-t)aB^{Nb} + ta \mid a \leq \hat{a}\big] \lambda'(w) \left(\dfrac{\partial w}{\partial t} - E[aL]\dfrac{\partial w}{\partial T}\right)\right]}{\left\{-E\left[a\dfrac{\partial \tilde{L}}{\partial t}\right]\right\}} \tag{14}$$

where the covariance, $\text{cov}(B(a), aL)$, is taken over the full interval $[\underline{a}, \overline{a}]$ and for the individuals resorting to the NHS over the two states of health, and $B(a)$ can be equal to $B^{Nb}(a)$, $B^{Ng}(a)$ or $B^{P}(a)$ according to the ability of the individual and his or her state of health (the last two being defined like $B^{Nb}(a)$ in (12)). In the above expression, $\dfrac{\partial \tilde{L}}{\partial t}$ stands for the income-compensated (or substitution) effect on labour supply of a change in t: $\dfrac{\partial \tilde{L}}{\partial t} = \dfrac{\partial L}{\partial t} + aL\dfrac{\partial L}{\partial T} < 0$.

Expression (14) is a modified Sheshinski (1972)-rule for the optimal marginal income tax rate. If we had only the first term in the numerator, we would have the standard ratio that trades off equity considerations (numerator) with efficiency considerations (denominator). The second term in the numerator of (14) is new and has to do with the effect on the waiting time of a change in the marginal tax rate. If this effect is negative, this second term pushes up the value of the marginal tax rate. However, it is difficult to say a priori how an increase in the marginal tax rate, accompanied by a decrease in the lum-sum transfer (T) intended to balance the government budget, will affect the waiting time.[3]

5 Numerical examples of the basic model

Throughout this section, we represent individual preferences by the following Cobb-Douglas utility function: $u(c, \ell) = [(\beta^{\beta}(1-\beta)^{1-\beta}]^{-1} c^{\beta} \ell^{1-\beta}$. This specification yields the following labour supply and indirect utility functions for an

[3] We have indeed $\dfrac{\partial w}{\partial t} - E[aL]\dfrac{\partial w}{\partial T} = \{\hat{v}_T^{P}[\hat{a}\hat{L}^{P} - E[aL]] - \pi \hat{v}_T^{Nb}[\hat{a}\hat{L}^{Nb} - E[aL]] - (1 - \pi)\hat{v}_T^{Ng}[\hat{a}\hat{L}^{Ng} - E[aL]]\}[\lambda'(w)\hat{v}_T^{Nb}(1-t)\hat{a}]^{-1}$. The expression in curly brackets is in general difficult to sign.

individual of ability a having opted for the NHS:

$$L^{Ng} = \beta A - (1-\beta)[(1-t)a]^{-1}T,$$
$$L^{Nb} = \beta[A - \lambda(w)] - (1-\beta)[(1-t)a]^{-1}T \tag{15.1}$$

and

$$v^{Ng} = [(1-t)a]^{\beta}A + [(1-t)a]^{\beta-1}T,$$
$$v^{Nb} = [(1-t)a]^{\beta}[A - \lambda(w)] + [(1-t)a]^{\beta-1}T. \tag{15.2}$$

For an individual of ability a having opted for a private insurance these functions are:

$$L^{P} = \beta A - (1-\beta)[(1-t)a]^{-1}(T - \pi q) \tag{16.1}$$

and

$$v^{P} = [(1-t)a]^{\beta}A + [(1-t)a]^{\beta-1}(T - \pi q). \tag{16.2}$$

From relation (6) it is easy to derive that the equilibrium level of waiting time satisfies the following condition:

$$\lambda(w) = [(1-t)\hat{a}]^{-1}q. \tag{17}$$

Since in the basic model w affects individuals only through $\lambda(w)$ one can dispense with choosing a particular functional form for $\lambda(w)$.

Using the above labour supply functions and substituting $\lambda(w)$ from (17), the budget constraint given in (10) yields after a few manipulations:

$$T = \frac{1-t}{1-\beta t}\left[t\beta E(a)A - \pi q F(\hat{a}) - \overline{R} + \frac{t}{1-t}\pi q\left\{(1-\beta)(1-F(\hat{a})) - \beta\int_{\underline{a}}^{\hat{a}}\frac{a}{\hat{a}}dF(a)\right\}\right]. \tag{18}$$

This is the Laffer curve, expressing the lump-sum transfer, T, in terms of the marginal tax rate, t, and the critical level of ability, $\hat{a}$ (and so the size of the NHS, $\pi F(\hat{a})$).[4]

Likewise, using the indirect utility functions derived above and substituting again $\lambda(w)$ from (17), we obtain the following expression for the social welfare

[4] The first three *rhs* terms are obvious: the first is the direct revenue effect of the marginal tax rate, while the second and third are government revenue requirements. The two terms in curly brackets take account of the fact that: (i) people with a private insurance policy tend to increase their labour supply and income tax payments due to the income effect of the insurance premium, and (ii) people on the NHS waiting list supply less labour and therefore generate less tax revenue.

given in (9):

$$SW = (1-t)^\beta E[\psi(a)a^\beta]A + (1-t)^{\beta-1}E[\psi(a)a^{\beta-1}]T$$
$$-\pi q(1-t)^{\beta-1}\left[\int_{\underline{a}}^{\hat{a}} \frac{\psi(a)a^\beta}{\hat{a}}\mathrm{d}F(a) + \int_{\hat{a}}^{\overline{a}} \psi(a)a^{\beta-1}\mathrm{d}F(a)\right], \qquad (19)$$

in which T can be substituted from (18) in order to express SW in terms of t and $\hat{a}$ alone, i.e. $SW(t,\hat{a})$.

Setting the derivative of $SW(t,\hat{a})$ w.r.t. t equal to zero results in a third-degree polynomial in t having three roots. It is the lowest root that corresponds to the optimal marginal tax rate: $t(\hat{a})$ (see Appendix). Next, we can trace out the behaviour of the function $SW(t(\hat{a}),\hat{a})$ that depicts the highest level of SW in terms of $\hat{a}$.

In all our numerical examples, β is chosen equal to .4, πq is set at 0.05, the time endowment A is normalised to one and the distribution of ability in the population is chosen such that the average ability is equal to one: $E[a] = 1$. The value for $\overline{R}$ is .3: it is 30% potential income (A times the average wage rate). The dispersion of ability in the population turns out to be a key parameter. It is captured by means of D defined as the ratio of the highest ability to the lowest ability: $D \equiv \overline{a}/\underline{a}$. We have considered three ability distributions: the uniform distribution ($F(a) = \frac{a-\underline{a}}{\overline{a}-\underline{a}}$), the log-uniform distribution ($F(a) = \frac{\log a - \log \underline{a}}{\log \overline{a} - \log \underline{a}}$), and the Beta (2,5) distribution ($F(a) = \int_{\underline{a}}^{a} \frac{1}{B(2,5)} \frac{1}{\overline{a}-\underline{a}} \frac{x-\underline{a}}{\overline{a}-\underline{a}} \left(\frac{\overline{a}-x}{\overline{a}-\underline{a}}\right) \mathrm{d}x$). The latter two are skewed to the right; the density of the log-uniform is monotonically decreasing, while that of the Beta(2,5) distribution is bell-shaped. Of these three distributions, the last one is the most relevant on empirical gounds. Figure 1 below gives the Beta(2,5) density function for $D = 100$ with support $[\underline{a},\overline{a}] = [.034, 3.41]$.

As it was analytically shown in the previous section, social welfare diminishes when $\hat{a}$ is raised from 0 to a small positive amount. It implies that a mixed system can only be optimal if the size of the NHS is large enough. It also means that the social welfare function is not everywhere concave in $\hat{a}$ when a mixed system is socially desirable. The shape of the social welfare curve is then as illustrated in Figure 2 where the local interior maximum of the curve does correspond to its global maximum. However, the curve may also be U-shaped or have its global maximum different from its local interior maximum. In both

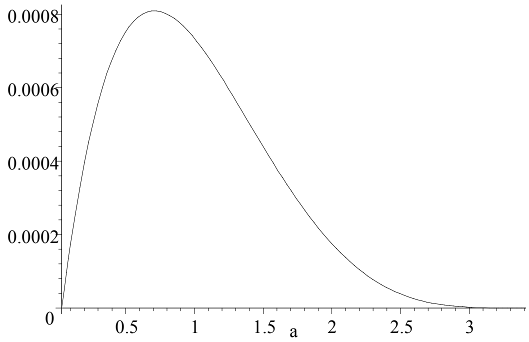

Figure 1: The Beta (2,5) density function with mean normalised to 1.

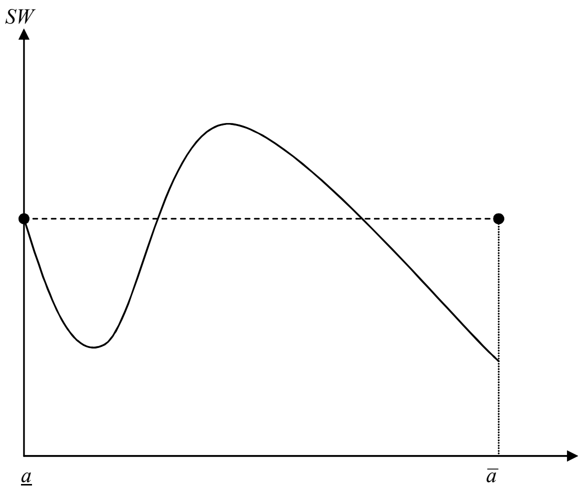

Figure 2: Shape of $SW(t(\widehat{a}), \widehat{a})$: the case of an optimal mixed system.

these cases, a mixed system is not desirable. As a matter of fact, according to our numerical examples, the dispersion parameter of the ability distribution (D) needs to be sufficiently large for a mixed system to be welfare optimal. This will be shown in the next tables.

In the tables below, we give the optimal tax policy and the NHS size for the minimal value of D for which a mixed system dominates a pure NHS or pure private system (remember that these two pure cases are in terms of resource allocation and social welfare equivalent) as well as for some higher values of D. The DWL figure (deadweight loss) indicates by how much, expressed as a percentage of actual GDP, we should decrease the value of $\overline{R}$, that is the exogenous government expenditure, for the pure (NHS or private) system to give the same amount of social welfare as the mixed system.[5]

For instance, Table 1 reads as follows. With a utilitarian objective, the lowest value of D for which a mixed system becomes optimal is 35. For this value of D, the Gini coefficient of the ability distribution is .315. A pure (NHS or private) system or a mixed system where 9.8% of the population resorts to the NHS are socially equivalent. Note that as D rises the optimal value of $F(\hat{a})$ suddenly jumps from 0 to 9,8% at $D = 35$. This reflects the non-concavity of the social welfare function as illustrated in Figure 2. However, with a rank-ordered social welfare function, a mixed system where 19.3% resorts to the NHS performs strictly better than a pure system: we would have to reduce $\overline{R}$ by 1.6% of GDP for a pure system to generate the same social welfare level as the mixed system. When D is as low as 17, a pure system performs as well as a mixed one in which 18 % of the population goes to the NHS.

		Utilitarian SW				Rank order SW			
D	Gini	$F(\hat{a})$	t	T	DWL	$F(\hat{a})$	t	T	DWL
17	.296	0 or 100%	.40	-0.8	0	18.0 %	.64	.005	0
35	.315	9.8 %	.46	-.06	0	19.3 %	.70	.01	1.6%
70	.324	11.4 %	.51	-.04	0.6%	18.1 %	.75	.02	3.2%
120	.327	11.4 %	.54	-.03	1.0%	17.5 %	.77	.02	4.4%
250	.330	11.2 %	.57	-.02	1.6%	16.7 %	.80	.03	5.8%
1000	.332	10.4 %	.61	-.01	2.3%	14.9 %	.83	.04	7.8%

Table 1: Uniform ability distribution: numerical results with constant π

		Utilitarian SW				Rank order SW			
D	Gini	$F(\hat{a})$	t	T	DWL	$F(\hat{a})$	t	T	DWL
51	.531	0 or 100%	.69	-.02	0	54.1 %	.82	.011	0
128	.603	54.2 %	.77	-.00	0	52.1 %	.91	.02	6.0%
150	.614	53.9 %	.78	-.01	0.5%	51.7 %	.93	.02	7.6%
180	.626	53.6 %	.80	-.01	1.2%	53.2%	.96	.02	10.6%
250	.645	52.8 %	.83	-.02	2.7%	57.8%	.97	.02	16.5%

[5] More precisely, $DWL \stackrel{\text{def}}{=} \dfrac{\overline{R} - R}{GDP(\text{pure NHS}; \overline{R})} 100$ where R solves SW(optimal mixed system;$\overline{R}$) $= SW$(pure NHS; R)

Table 2: Log-uniform ability distribution: numerical results with constant π

		Utilitarian SW				Rank order SW			
D	Gini	$F(\hat{a})$	t	T	DWL	$F(\hat{a})$	t	T	DWL
63.5	.298	0 or 100%	.37	-.10	0	2.6 %	.62	-.002	0
134	.306	0.5 %	.39	-.09	0	5.8 %	.64	.002	0.3%
250	.310	1.3%	.40	-.08	0.03%	6.8 %	.65	.004	0.5%
1000	.313	2.1 %	.42	-.08	0.1%	7.1 %	.67	.007	0.8%

Table 3: Beta(2,5) ability distribution: numerical results with constant π

As the tables show, whether a mixed system is superior to a pure one very much depends upon the specification of the social welfare function as well as that of the ability distribution. This is not surprising since in our setting the driving force towards a mixed system is to redistribute real income across individuals of different abilities. As expected a mixed system is more likely to be optimal when the more redistribution-oriented rank order social welfare function represent the preferences of society. As already mentioned the dispersion of the ability distribution also plays an important role: it must be large enough for a mixed system to be optimal. However, the shape of the ability distribution matters as well in the determination of the optimal NHS size when a mixed system is superior to a pure one. A comparison of our results for the log-uniform distribution and the beta one shows this very clearly.

Another interesting conclusion we infer from our numerical examples is that while the optimal marginal tax rate (t) is always monotonically increasing with the dispersion of the ability distribution (D), the optimal NHS size in case of a mixed system is not monotonic in D. Therefore, the marginal tax rate and the NHS size can be either complements or substitutes in redistributing real income across individuals of different abilities.

6 Extension of the basic model

In this section we modify in two respects the basic setting developed in the previous sections. First, in addition to its effect on time endowments the presence of waiting lists in the NHS directly affects the utility of individuals having opted for this health system. Let $r(w)$ denote their non-pecuniary and additive loss of utility in case of illness with $r'(w) > 0$ and $r''(w) > 0$. Second, a negative correlation of ability and illness risk is accounted for. There is indeed empirical evidence of this negative correlation. To simplify the presentation we assume that the probability of illness depends upon a in a deterministic way: $\pi(a)$ with $\pi'(a) < 0$. For the sake of simplicity, we also rule out any adverse selection in the insurance markets. For an individual of ability a the private insurance premium is thus $\pi(a)q$.

Let the indirect utility, $v(\cdot)$, be defined in the same way as in Section 2, therefore not including the non-pecuniary loss of utility $r(w)$. Then, $\Delta(\hat{a}, t, T, w)$ given in (6) in the basic model is now defined by:

$$\Delta(\hat{a}, t, T, w) \equiv v((1-t)\hat{a}, T-\pi(\hat{a})q, A) - (1-\pi(\hat{a}))v((1-t)\hat{a}, T, A) \\ -\pi(\hat{a})\Big[v((1-t)\hat{a}, T, A-\lambda(w)) - r(w)\Big] = 0. \tag{20}$$

Differentiating this expression yields

$$\begin{aligned} \frac{\partial \Delta}{\partial \hat{a}} &= (1-t)\Big[(\hat{L}^P \hat{v}_T^P - (1-\pi(\hat{a}))\hat{L}^{Ng}\hat{v}_T^{Ng}\Big] \\ & -\pi(\hat{a})\hat{L}^{Nb}\hat{v}_T^{Nb} + \pi'(\hat{a})\Big[-q\hat{v}_T^P + (\hat{v}^{Ng} - \hat{v}^{Nb}) + r(w)\Big], \\ & > 0 \qquad \text{by assumption,} \end{aligned} \tag{21.1}$$

$$\frac{\partial \Delta}{\partial w} = \pi(\hat{a})\Big[(1-t)\hat{a}\lambda'(w)\hat{v}_T^{Nb} + r'(w)\Big] > 0, \tag{21.2}$$

with $\partial\Delta/\partial t$ and $d\Delta/dT$ being defined in the same way as in (7.3) and (7.4). The effects of $\hat{a}, t$ and T on w are obtained as in (8).

The government's optimization problem is now written as:

$$\begin{aligned} \max_{\hat{a},t,T} SW &= \int_{\underline{a}}^{\hat{a}} \Big[(1-\pi(a))v((1-t)a, T, A) + \pi(a)v((1-t)a, T, A \\ &- \lambda(w(\hat{a}, t, T))) + \pi(a)r(w(\hat{a}, t, T))\Big] \mathrm{d}F(a) \\ &+ \int_{\hat{a}}^{\overline{a}} v((1-t)a, T - \pi(a)q, A)\mathrm{d}F(a) \end{aligned} \tag{22}$$

subject to:

$$t\int_{\underline{a}}^{\hat{a}} a\Big[(1-\pi(a))L((1-t)a,T,A)+\pi(a)L((1-t)a,T,A$$

$$-\quad \lambda(w(\hat{a},t,T)))\Big]\mathrm{d}F(a)+t\int_{\hat{a}}^{\overline{a}} aL((1-t)a,T-\pi(a)q,A)\mathrm{d}F(a)$$

$$-T-\overline{R}-q\int_{\underline{a}}^{\hat{a}}\pi(a)\mathrm{d}F(a)\geq 0. \tag{23}$$

Proceeding in the same way as in the appendix for the basic setting we obtain the following condition for the optimal choice of $\hat{a}$ (and so of the NHS size: $S = q\int_{\underline{a}}^{\hat{a}} \pi(a)dF(a)$):

$$\frac{E\Big[\pi(a)\big\{(1-t)aB^{Nb}(a)+ta\big\}\lambda'(w)+\pi(a)r'(w)\mid a\leq\hat{a}\Big]\left(-\dfrac{dw}{d\hat{a}}\right)}{t\Big[\hat{L}_p-\pi(\hat{a})\hat{L}^{Nb}-(1-\pi(\hat{a}))\hat{L}^{Ng}\Big]+\dfrac{\pi(\hat{a})q}{\hat{a}}}=\frac{\hat{a}f(\hat{a})}{F(\hat{a})}, \tag{24}$$

which is the equivalent of (13) in the basic setting. Likewise, the optimal choice of the marginal tax rate t satisfies:

$$t = \tag{25}$$

$$\frac{-\mathrm{cov}(\beta,aL)-F(\hat{a})E\Big[\pi(a)\big\{(1-t)aB^{Nb}+ta\big\}\lambda'(w)+\pi(a)r'(w)\,|\,a\leq\hat{a}\Big]\Big(\dfrac{\partial w}{\partial t}-E[aL]\dfrac{\partial w}{\partial T}\Big)}{E\left[a\dfrac{\partial\tilde{L}}{\partial t}\right]}$$

which is the equivalent of (14) in the basic model.

Except for the introduction of $r(w)$ and of $\pi(a)$ instead of π, conditions (24) and (25) are very close to their counterparts in the basic model. However the main difference lies in that the pure NHS (upper corner solution: $\hat{a} = \overline{a}$) now yields a *higher* value for social welfare than the pure private system (lower corner solution: $\hat{a} = \underline{a}$). This is because the former operates a redistribution from low- to high-risk individuals. While the premium in the private insurance market depends upon risk and is thus higher for a low-ability individual than for a high-ability one (because of the negative correlation between ability and risk), all health expenditures are financed by means of progressive income taxes in the pure NHS. This redistributive consideration also makes it, *ceteris paribus*, more attractive to have a larger NHS sector in any mixed system.

For the numerical simulations below, we assume that

$$\pi(a) = \pi(\overline{a})\frac{a - \underline{a}}{\overline{a} - \underline{a}} + \pi(\underline{a})\frac{\overline{a} - a}{\overline{a} - \underline{a}}$$

with $\pi(\underline{a}) = .55$ and $\pi(\overline{a}) = .45$. In other words, we took a 10% difference in health risk between the most and the least able person. Except for this change we use the same specifications and parameter values as in the previous section. In particular, we assume $r(w) = 0$, $\forall\ w$ and $\frac{E[\pi(a)]q}{E[a]A} = .05$.

		Utilitarian SW				Rank order SW			
D	Gini	$F(\hat{a})$	t	T	DWL	$F(\hat{a})$	t	T	DWL
28	.310	100%	.44	-.06	0	20.0 %	.68	.00	0
60	.322	11.7 %	.50	-.05	0	19.5 %	.73	.01	1.6 %
120	.327	11.3 %	.54	-.03	0.5%	18.0 %	.76	.02	3.1 %
250	.330	11.2 %	.57	-.03	1.1%	16.7 %	.79	.03	4.4 %
1000	.332	10.4%	.60	-.01	1.8%	15.4 %	.82	.03	6.2 %

Table 4: Numerical results: uniform ability distribution with variable π

		Utilitarian SW				Rank order SW			
D	Gini	$F(\hat{a})$	t	T	DWL	$F(\hat{a})$	t	T	DWL
69	.557	100%	.72	.03	0	54.7 %	.84	.01	0
164	.620	54.7 %	.79	.01	0	51.8 %	.92	.02	6.4 %
180	.626	54.4 %	.80	.01	0.3 %	51.5 %	.94	.02	7.4 %

Table 5: Numerical results: log-uniform ability distribution with variable π

		Utilitarian SW				Rank order SW			
D	Gini	$F(\hat{a})$	t	T	DWL	$F(\hat{a})$	t	T	DWL
280	.311	100%	.65	-.04	0	6.8 %	.65	.00	0
1000	.313	100%	.66	-.04	0	7.1%	.66	.006	0.2 %
10000	.315	100%	.66	-.04	0	7.2 %	.66	.007	0.3 %

Table 6: Numerical results: Beta(2,5) ability distribution with variable π

In interpreting the numerical results given in the above tables, it is important to keep in mind that as stated above, the pure NHS system performs better than the fully private system when $\pi(a)$ falls as a increases. Therefore, when assessing whether a mixed system is desirable, it is with this pure NHS system that the mixed system needs to be compared. However, the numerical results we obtain here are not very different from the ones we have obtained in the previous section except for one case. This occurs with the Beta ability distribution and the utilitarian objective. In this case the pure NHS system always performs better than any mixed system. This is due to the fact that the pure NHS system achieves some redistribution of real income among individuals of different risks. However, the shape of the ability distribution plays a crucial role in that result since it does not occur with the other ability distribution functions.

7 Conclusions

This paper has investigated whether Besley and Coate's argument in favor of a mixed (private/public) system for the provision of some private services applies to the health care sector when some of its specificities are taken into account. Besley and Coate (1991) claim indeed that if there are limits to redistribution (due to the impossibility of implementing lump-sum taxes) it may be socially optimal to make a private and a public sector coexist, the service provided in the latter being free of charge but of lower quality. In countries with mixed health care systems, this lower quality is caused by waiting lists for elective treatments. In our set-up these waiting lists act as a rationing device to equate demand and supply in the public sector. However the waiting time for being treated in the public sector is a pure deadweight loss, that could be avoided if a benevolent social planner had control over which individuals had to resort to public health care. The issue is then to know when the welfare gains from redistribution outweighs this deadweight loss in a mixed system.

To address this question, we have investigated how the level of social welfare evolves with the size of the public sector. It turns out that the maximisation problem we face is not concave (see Figure 2). Accordingly, since the two pure health care systems (either entirely private or entirely public) are characterized by the same levels of social welfare in our basic model (sections 2–5), the optimal system can be either a mixed system (interior solution) or a pure one (corner solution) in this model. A key finding of our paper is that the dispersion of the ability distribution in the population needs to be large enough for a mixed health care system to be socially optimal. Furthermore it is never desirable to have a small public sector. In our extension to the basic model (section 6), in which a.o. the probability of being sick falls with ability, the level of social welfare achieved with an entirely public system outweighs the one achieved with an entirely private one in which the premium of private insurance is adjusted to individual illness risk. It is due to the fact that the former system redistributes across risk classes. Therefore it is with the pure public system that mixed systems must be compared in terms of social welfare.

Even though in both our basic and extended models a mixed health care system is socially desirable with a sufficient dispersion of abilities in the population, our numerical results show that the welfare gains that may be achieved by a mixed system relative to the best pure system are quiet low; especially for the Beta(2,5)-distribution which is from an empirical point of view the most relevant one. The reason is the importance of the deadweight losses caused by the waiting lists in the public sector, that sort individuals with different incomes between the public and private health care providers.

In this paper, some features of the health care sector have been swept under the carpet: distinction between diagnosis exams and therapeutic treatments, uncertainty about the outcome of treatments, information asymmetry between patients and doctors, adverse selection in the insurance market ... Furthermore our set-up has concentrated on elective care for which waiting lists occur in actual mixed systems; clearly, it does not apply to emergency care. Including these features would complicate the analysis, but we believe would not affect our qualitative results.

References

[1] Besley, Tim and Stephen Coate (1991). Public provision of private goods and the redistribution of income. *American Economic Review*, 81, 979–984.

[2] Besley, Tim and Miguel Gouveia (1994). Alternative systems of health care provision. *Economic Policy*, 19, 199–258.

[3] Cullis, John, Philip Jones and Carol Propper (2000). Waiting lists and medical treatment: analysis and policies, ch 23 in: Newhouse J P & A J Culyer, *Handbook of Health Economics* (Amsterdam: North-Holland).

[4] Deaton Angus (1983). An explicit solution to an optimal tax problem. *Journal of Public Economics*, 20, 333-346.

[5] Hoel Michael and Erik M Sæther (2000). Private health care as a supplement to a public health system with waiting time for treatment. Mimeo, University of Olso.

[6] Iversen, Tor (1997). The effect of a private sector on the waiting time in a national health service. *Journal of Health Economics*, 16, 381–396.

[7] Propper, Carol (2000). The demand for private health care in the UK. *Journal of Health Economics*, 19, 855–876.

[8] Propper, Carol and Katherine Green (2001). A larger role for the private sector in financing UK health care: the arguments and the evidence. *Journal of Social Policy*, 30, 685-704.

[9] Sheshinski, Eytan (1972). The optimal linear income tax. *Review of Economic Studies*, 39, 297–302.

[10] Sydsæter, Knut (1981). *Topics in mathematical analysis for economists.* London, Academic Press.

Appendix

7.1 Derivation of the first order conditions in Section 4

The Lagrange function to the planning problem is

$$\begin{aligned}\mathcal{L} &= \int_{\underline{a}}^{\hat{a}} \psi(a)\Big[(1-\pi)\, v((1-t)a, T, A) + \pi v((1-t)a, T, A \\ &\quad -\lambda(\hat{a}, t, T)\Big] dF(a) + \int_{\hat{a}}^{\overline{a}} \psi(a) v((1-t)a, T-\pi q, A) \mathrm{d}F(a) \\ &\quad +\mu \Big\{ t \int_{\underline{a}}^{\hat{a}} a \Big[(1-\pi)\, L((1-t)a, T, A) + \pi L((1-t)a, T, A-\lambda(\hat{a}, t, T)\Big] \mathrm{d}F(a) \\ &\qquad +t \int_{\hat{a}}^{\overline{a}} aL((1-t)a, T-\pi q, A) dF(a) - T - \overline{R} - \pi q F(\hat{a}) \geq 0, \Big\}.\end{aligned}$$

The derivative w.r.t. $\hat{a}$ is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \hat{a}} &= -\pi \int_{\underline{a}}^{\hat{a}} \psi(a) v_T^{Nb}(1-t) a \mathrm{d}F(a) \frac{\partial \lambda}{\partial \hat{a}} + \mu \pi t \int_{\underline{a}}^{\hat{a}} a \frac{\partial L^{Nb}}{\partial \lambda} \mathrm{d}F(a) \frac{\partial \lambda}{\partial \hat{a}} \\ &\quad +\mu \Big\{ t\hat{a} \Big[\pi \widehat{L}^{Nb} + (1-\pi) \widehat{L}^{Ng} - \widehat{L}^{P}\Big] f(\widehat{a}) - \pi q f(\widehat{a}) \Big\}.\end{aligned}$$

Defining

$$B^{Nb}(a) \equiv \frac{1}{\mu} \psi(a) v_T^{Nb} + ta \frac{\partial L^{Nb}}{\partial T},$$

using the fact that

$$\frac{\partial L^{Nb}}{\partial \lambda} = -\left[\frac{\partial L^{Nb}}{\partial T}(1-t)a + 1\right] \;\; (= -\frac{\partial c^{Nb}}{\partial T}),$$

and collecting terms gives expression (11) in the text.

The derivative w.r.t. t is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial t} &= -\int_{\underline{a}}^{\hat{a}} \psi(a)[\pi v_T^{Nb} a L^{Nb} + (1-\pi) v_T^{Ng} a L^{Ng}] \mathrm{d}F(a) - \int_{\hat{a}}^{\overline{a}} \psi(a) v_T^{P} a L^{P} \mathrm{d}F(a) \\ &\quad -\pi \int_{\underline{a}}^{\hat{a}} \psi(a) v_T^{Nb}(1-t) a \mathrm{d}F(a) \frac{\partial \lambda}{\partial t} + \mu \pi t \int_{\underline{a}}^{\hat{a}} a \frac{\partial L^{Nb}}{\partial \lambda} \mathrm{d}F(a) \frac{\partial \lambda}{\partial t} \\ &\quad +\mu \Big\{ \int_{\underline{a}}^{\hat{a}} [\pi a L^{Nb} + (1-\pi) L^{Ng}] \mathrm{d}F(a) + \int_{\hat{a}}^{\overline{a}} a L^{P} \mathrm{d}F(a) \\ &\qquad +t \int_{\underline{a}}^{\hat{a}} [\pi a \frac{\partial L^{Nb}}{\partial t} + (1-\pi) a \frac{\partial L^{Ng}}{\partial t}] \mathrm{d}F(a) + t \int_{\hat{a}}^{\overline{a}} a \frac{\partial L^{P}}{\partial t} \mathrm{d}F(a) \Big\},\end{aligned}$$

and that w.r.t. T is

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial T} = & \int_{\underline{a}}^{\hat{a}} \psi(a)[\pi v_T^{Nb} + (1-\pi)v_T^{Ng}]\mathrm{d}F(a) + \int_{\hat{a}}^{\overline{a}} \psi(a)v_T^P \mathrm{d}F(a) \\
& -\pi \int_{\underline{a}}^{\hat{a}} \psi(a)v_T^{Nb}(1-t)a\mathrm{d}F(a)\frac{\partial \lambda}{\partial T} + \mu\pi t \int_{\underline{a}}^{\hat{a}} a\frac{\partial L^{Nb}}{\partial \lambda}\mathrm{d}F(a)\frac{\partial \lambda}{\partial t} \\
& +\mu\left\{ t\int_{\underline{a}}^{\hat{a}} [\pi a\frac{\partial L^{Nb}}{\partial T} + (1-\pi)a\frac{\partial L^{Ng}}{\partial T}]\mathrm{d}F(a) + t\int_{\hat{a}}^{\overline{a}} a\frac{\partial L^P}{\partial T}\mathrm{d}F(a) - 1\right\}.
\end{aligned}$$

Performing $\frac{\partial \mathcal{L}}{\partial t} + \frac{\partial \mathcal{L}}{\partial T}E[aL]$ gives

$$\begin{aligned}
& -E[\psi(a)v_T aL] + E[\psi(a)v_T]E[aL] \\
& -\pi\left\{\int_{\underline{a}}^{\hat{a}} [\psi(a)v_T^{Nb}(1-t)a + \mu ta\frac{\partial L^{Nb}}{\partial \lambda}]\mathrm{d}F(a)\right\}\left(\frac{\partial \lambda}{\partial t} + \frac{\partial \lambda}{\partial T}E[aL]\right) \\
& +t\mu\left\{E[a\frac{\partial L}{\partial t}] + E[a\frac{\partial L}{\partial T}]E[aL]\right\}
\end{aligned}$$

Using the Slutsky decomposition $\frac{\partial L}{\partial t} = \frac{\partial \widetilde{L}}{\partial t} - \frac{\partial L}{\partial T}aL$ and the definition $B(a) = \frac{1}{\mu}\psi(a)v_T + ta\frac{\partial L}{\partial T}$ then gives expression (14) in the text.

7.2 An explicit solution for the optimal tax policy

Making use of the indirect utility functions (15.2) and (16.2), the expression for social welfare can be written as

$$\frac{SW}{E[a]AE[\psi(a)a^{\beta-1}]} = (1-t)^\beta \kappa + \tau(1-t)^{\beta-1} - \rho(1-t)^{\beta-1}\gamma(\widehat{a})$$

where

$$\tau \stackrel{\text{def}}{=} \frac{T}{E[a]A}, \; \kappa \stackrel{\text{def}}{=} \frac{1}{E[a]}\frac{\int_{\underline{a}}^{\overline{a}} a\; \psi(a)a^{\beta-1}\mathrm{d}F(a)}{\int_{\underline{a}}^{\overline{a}} \psi(a)a^{\beta-1}\mathrm{d}F(a)}, \; \rho \stackrel{\text{def}}{=} \frac{E[\pi(a)]q}{E[a]A}$$

and

$$\gamma(\widehat{a}) \stackrel{\text{def}}{=} \frac{\int_{\underline{a}}^{\widehat{a}} \pi(a)\frac{a}{\widehat{a}}\psi(a)a^{\beta-1}\mathrm{d}F(a) + \int_{\widehat{a}}^{\overline{a}} \pi(a)\psi(a)a^{\beta-1}\mathrm{d}F(a)}{E[\pi(a)]\int_{\underline{a}}^{\overline{a}} \psi(a)a^{\beta-1}\mathrm{d}F(a)}.$$

The parameter κ is the ratio of a weighted average wage rate to the arithmetic average; for $\psi'(a) \leq 0$, it is smaller than 1. For a given value of $\widehat{a}$, the welfare trade-off between t and τ is given by

$$-\left.\frac{\mathrm{d}\tau}{\mathrm{d}t}\right|_{\mathrm{d}SW=0} = \beta\kappa - \tau\frac{1-\beta}{1-t} + \rho\frac{1-\beta}{1-t}\gamma(\widehat{a}),$$

which is the labour supply of a representative agent with ability κ and lump sum income $\tau - \rho\gamma(\widehat{a})$.

Using the labour supply functions (15.1) and (16.1), the government budget constraint can be written as

$$\tau = \frac{1-t}{1-\beta t}\left[t\beta - \rho\varphi(\widehat{a}) - s + \frac{t}{1-t}\rho\delta(\widehat{a})\right] \tag{26}$$

where

$$\varphi(\widehat{a}) \stackrel{\text{def}}{=} \frac{\int_{\underline{a}}^{\widehat{a}} \pi(a)\mathrm{d}F(a)}{E[\pi(a)]}, \quad s \stackrel{\text{def}}{=} \frac{\overline{R}}{E[a]A}$$

and

$$\delta(\widehat{a}) \stackrel{\text{def}}{=} (1-\beta)\frac{\int_{\widehat{a}}^{\overline{a}} \pi(a)\mathrm{d}F(a)}{E[\pi(a)]} - \beta\frac{\int_{\underline{a}}^{\widehat{a}} \pi(a)a\mathrm{d}F(a)}{E[\pi(a)]\hat{a}}.$$

Note that with constant π, $\varphi(\widehat{a})$ reduces to $F(\widehat{a})$.

The budgetary trade-off between t and τ is then

$$-\left.\frac{\mathrm{d}\tau}{\mathrm{d}t}\right|_{d\overline{R}=0} = -\frac{1}{(1-\beta t)^2}\left[(1-\beta)(\rho\varphi(\widehat{a}) + s) + (1 - 2t + \beta t^2)\beta + \rho\delta(\widehat{a})\right]. \tag{27}$$

Replacing in this trade-off τ by the rhs of (26) and equating it to the welfare trade-off yields the following third degree polynomial in t:

$$\begin{gathered}\left\{\beta^3(1-\kappa)\right\}t^3 + \\ \left\{\beta[\rho\varphi(\widehat{a}) + s + \rho\delta(\widehat{a}) - 1] - \beta^2[\rho\varphi(\widehat{a}) + s - \kappa] - (\beta^2+\beta^3)(1-\kappa) + \beta^2[\gamma(\widehat{a}) - \delta(\widehat{a})]\rho\right\}t^2 + \\ \left\{(\beta^2+\beta)(1-\kappa) + \beta^2[\rho\varphi(\widehat{a}) + s - \kappa] - \beta[\rho\varphi(\widehat{a}) + s - 1] - (\beta^3 - 2\beta^2 + 2\beta)\rho\gamma(\widehat{a}) + \beta\rho\delta(\widehat{a})\right\}t + \\ \left\{-\beta(1-\kappa) + (1-\beta)\rho\gamma(\widehat{a}) - \rho\delta(\widehat{a})\right\} = 0\end{gathered}$$

For the general third degree polynomial

$$a_1x^3 + a_2x^2 + a_3x + a_4 = 0,$$

let us define $Q \stackrel{\text{def}}{=} \frac{1}{3}\frac{a_3}{a_1} - (\frac{1}{3}\frac{a_2}{a_1})^2$, $R \stackrel{\text{def}}{=} \frac{1}{6}\frac{a_3}{a_1}\frac{a_2}{a_1} - \frac{1}{2}\frac{a_4}{a_1} - (\frac{1}{3}\frac{a_2}{a_1})^3$, $S \stackrel{\text{def}}{=} (R+\sqrt{P})^{\frac{1}{3}}$ and $T \stackrel{\text{def}}{=} (R-\sqrt{P})^{\frac{1}{3}}$, with $P \stackrel{\text{def}}{=} Q^3 + R^2$.

Then if $P < 0$, the polynomial has three different real roots; if $P = 0$, it has three real roots of which at least two are identical; and if $P > 0$, it has one real and two complex roots (see Sydsæter, 1981, p 54). Cardano's formulae for the roots are then as follows:

$$\begin{aligned} root_1 &= -\frac{1}{3}\frac{a_2}{a_1} + (S+T) \\ root_2 &= -\frac{1}{3}\frac{a_2}{a_1} - \frac{1}{2}(S+T) + \frac{1}{2}i\sqrt{3}(S-T), \text{ and} \\ root_3 &= -\frac{1}{3}\frac{a_2}{a_1} - \frac{1}{2}(S+T) - \frac{1}{2}i\sqrt{3}(S-T), \end{aligned}$$

where $i = \sqrt{-1}$. It turns out that for our model

$$0 < root_2 < root_3 < 1 < root_1.$$

The optimal tax rate is therefore $root_2$.

In two simulations (rank-order SW, loguniform distribution, $D = 180$ and $D = 250$) does the highest SW-contour *not* have a tangency point on the upward sloping part of the Laffer curve (26). Since the SW contours are never negatively sloped (the slope measures the labour supply of a representative agent), the optimal marginal tax rate must be the rate corresponding to the maximum of the Laffer curve. It is found by solving the second degree polynomial defined by setting the *rhs* of (27) equal to zero.