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

Maurice Marchand* and Fred Schroyen†

30 January 2004

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. We find that: (i) a mixed system with a small NHS is never desirable; (ii) actuarially fair sickness insurance is never desirable either; (iii) 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 (iv) the welfare gains from such a mixed system are not likely to be significant.

*CORE (Université catholique de Louvain)
†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 (current) expenditure on health care has grown from 10.5% in 1980 to 17.8% in 2000 (OECD, 2003). Around 15% of the population was in 1991 covered by a supplementary private health insurance (Besley et al, 1999). 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 that we want to construct a formal and consistent framework within which we can discuss and assess the equity grounds for a mixed health care system to be superior to a fully public one.

Public health care services are a specific instance of public provision of a private good. In a first best economy, there is no good justification for public provision of such goods, unless the government can deliver it at a lower marginal cost. But first best reasoning can be misleading in a second best setting. When there are limits to redistributional policy, instruments that help to sort income classes become useful. Public provision can work as such a sorting device if low income citizens choose the publicly provided good, while high income citizens go private. The latter contribute to public revenue through income taxation, which is then used to finance the delivery of the private good to the former at a price below the marginal cost. Indirectly, the rich thus finance the consumption of the poor. Crucial for sorting to take place is that the publicly and privately provided goods are good substitutes for low income people, but poor substitutes for high income people. In other words, that the goods differ in one or more attributes for which the preferences are positively correlated with income. This argument, has been formalised by Besley and Coate (1991).

They and others\(^1\) mention health care, education, child care as potential candidates for in-kind redistribution. In health care, there are indeed several at-

---
tributes that may be thought of sharing the above property. The quality of the treatment itself, the quality of the environment in which it is provided, and the speed with which the treatment is delivered, are all attributes that come to mind as ‘normal’, i.e. for which richer people have a higher willingness to pay than poorer people.

Interestingly, an econometric study of the demand for private health insurance in the UK (Besley et al, 1999) reveals that such insurance is a normal good and that among the six regional public health authority quality indicators, only the size of the long term waiting lists shows up as a significant explanatory variable.\(^2\) In the light of the aforementioned second best arguments, this suggests that the coexistence of a public and private health sector, with waiting times in the former, enacts redistribution. It would also mean that negative redistributional side effects are part of the price tag for policy measures aimed at reducing waiting times/lists.

In this paper, we embed the choice of health care provider in a model where citizens differ in (unobservable) ability and taxed (linearly) according to their observable income. People without private insurance who fall ill are admitted free of charge to the NHS, but the time spent on the waiting list crowds out their market and non-market activities. A mixed system is then desirable if the benefits of redistribution outweigh those deadweight losses. Our main conclusions can be summarised as follows:

1. a mixed health care system with a small NHS is never desirable because the redistributional benefits are of second order relative to deadweight losses;
2. actuarially fair sickness insurance that protects people without private health care insurance against the waiting time risk—though desirable from the citizen’s point of view—is detrimental for the in-kind redistribution argument;
3. a mixed system can be optimal but a necessary condition is that the spread of the ability distribution is sufficiently large; and
4. the welfare gains implied by such a system are not significant if the ability distribution is bell-shaped and skewed to the right.

While the first two results are of a theoretical nature, the latter two follow from numerical simulations on the model. We think they are important, because the literature exposing the potential second best argument for public provision has never attempted to gauge to what extent that argument is empirically justified.

\(^2\) The effect of a 1% increase in a household’s position in the income distribution on the probability of purchasing private health insurance is around 3.7 percentage points. The effect of an increase in the long term waiting list with one person per thousand is around 2.2 percentage points.
The focus on waiting times/lists as a sorting device needs justification. Quality of treatment, and the quality of the treatment environment (rooms, food, staff friendliness) could also be used as a screening device if set low enough. They have the advantage over waiting lists that their reduction leads to costs savings, and thus a lower social cost. One should therefore compare the trade-offs of different attributes. In this paper, we focus exclusively on waiting times. Lack of space is one reason, but another is the above mentioned empirical finding that proxies for these other attributes do not help to explain the demand for private insurance very much.

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 assume heterogeneity in the willingness to wait for treatment. In the public health sector patients are put on a waiting list and are treated at a constant marginal cost. Patients can also choose to be treated privately at a marginal cost at least as high as in the public sector. Hoel and Sæther 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. They do not pay attention, as we do, to the uncertain incidence of illness, to the crowding out effects of waiting on labour market decisions, to the question of sickness insurance. Nor do they make any attempt at measuring the required degree of inequality.

The paper is organized as follows. First we discuss citizens’ choice of resorting to either the NHS or a private insurance contract (Section 2). Next, we study how the equilibrium in the health care sector determines the waiting time in the NHS (Section 3). Thereafter, we set up the normative problem and analyse the optimality properties of the size of the NHS (if a mixed system is desirable) and the linear tax policy both under the absence (Section 4) and presence (Section 5) of a competitive market for sickness insurance. We then provide numerical simulations to assess the desirability of a mixed system (Section 6). Concluding remarks are offered in Section 7.
2 Choice of health care provider

In the simplest setting developed in this paper, citizens only care about their consumption of a composite good and leisure, denoted by $c$ and $\ell$ respectively. Their preferences on these two goods are described by a strictly concave utility function $u(c, \ell)$. There is some probability, denoted by $\pi$, that any individual will suffer from illness, in which case her time endowment available for labour and leisure diminishes. In our setting it takes the form of a reduction in the individual’s time endowment which is equal to $A$ for an individual in good health and to $A - w$ for an individual being sick, where $w$ is the loss of time caused by illness. We assume $w$ is related in a one to one way to the time a sick person has to wait before receiving medical treatment, and will in the sequel be interpreted like that.

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 her leisure time, $\ell$, is equal to either $A - L$ in case of good health or $A - w - L$ 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 her state of health since 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:

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

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

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:

\[
\frac{\partial v^i}{\partial L} = -aL^i, \quad i = Ng, Nb,
\]

\[
(2a)
\]
where \( v^i_T \) is the marginal utility of income and subscripts denote partial derivatives. We also have:

\[
v^N_b = -(1 - t)a v^N_T. \tag{2b}
\]

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

\[
Ev^N = (1 - \pi)v^Ng + \pi v^Nb, \tag{3}
\]

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

On the other hand, if an individual opts for a private insurance policy, she will be given medical treatment on the spot \((w = 0)\) 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 \( \pi q \), they are given by:

\[
L^P = L((1 - t)a, T - \pi q, A) \tag{4a}
\]

and

\[
v^P = v((1 - t)a, T - \pi q, A), \tag{4b}
\]

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

\[
v^P_T = -a L^P v^P_T. \tag{5}
\]

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

\[
Ev^N \geq v^P \quad \text{for any } a \leq \hat{a},
\]

\[
Ev^N < v^P \quad \text{for any } a > \hat{a}.
\]

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 \((t, T)\) 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 \( \Delta \) as \( v^P - Ev^N \):

\[
\Delta(a, t, T, w) \overset{\text{def}}{=} v((1 - t)a, T - \pi q, A) - (1 - \pi)v((1 - t)a, T, A) - \pi v((1 - t)a, T, A - w). \tag{6}
\]

The critical ability level \( \hat{a} \) is then defined as \( \Delta(\hat{a}, t, T, w) = 0 \). Note that Assumption N implies:

\[
\frac{\partial \Delta}{\partial a} \bigg|_{\hat{a}} = (1 - t) \left[ \hat{L}^P \hat{v}_T^P - (1 - \pi)\hat{L}^N g \hat{v}_T^N - \pi \hat{L}^N b \hat{v}_T^N \right] > 0, \tag{7a}
\]

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

\[
\frac{\partial \Delta}{\partial w} \bigg|_{\hat{a}} = \pi(1 - t)\hat{a} \hat{v}_T^{N_b} > 0, \tag{7b}
\]

\[
\frac{\partial \Delta}{\partial t} \bigg|_{\hat{a}} = -\frac{\hat{a}}{1 - t} \frac{\partial \Delta}{\partial \hat{a}} < 0, \tag{7c}
\]

and

\[
\frac{\partial \Delta}{\partial T} \bigg|_{\hat{a}} = \hat{v}_T^P - (1 - \pi)\hat{v}_T^N - \pi \hat{v}_T^{N_b}. \tag{7d}
\]

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

\[
\frac{\partial w}{\partial \hat{a}} = -\frac{\partial \Delta/\partial \hat{a}}{\partial \Delta/\partial w} < 0, \tag{8a}
\]
\[
\frac{\partial w}{\partial t} = -\frac{\partial \Delta}{\partial t} \frac{\partial \Delta}{\partial w} > 0, \quad (8b)
\]
and
\[
\frac{\partial w}{\partial T} = -\frac{\partial \Delta}{\partial T} \frac{\partial \Delta}{\partial w} \leq 0. \quad (8c)
\]
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:

\[
SW \overset{\text{def}}{=}= \int_{\hat{a}}^{a} \psi(a) E\nu^{N}(a) dF(a) + \int_{\hat{a}}^{\pi} \psi(a) \nu^{P}(a) dF(a), 
\]
where the weight \( \psi(a) \) is non-increasing in ability.\(^3\) 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.

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:

\[
\max_{\hat{a}, t, T} \int_{\hat{a}}^{\hat{a}} \psi(a) \left[ (1 - \pi) v ((1 - t)a, T, A) + \pi v ((1 - t)a, T, A - w(\hat{a}, t, T)) \right] dF(a) \\
+ \int_{\hat{a}}^{\pi} \psi(a) v ((1 - t)a, T - \pi q, A) dF(a)
\]
subject to

\[
t \int_{\hat{a}}^{\hat{a}} a \left[ (1 - \pi) L ((1 - t)a, T, A) + \pi L ((1 - t)a, T, A - w(\hat{a}, t, T)) \right] dF(a) \\
+ t \int_{\hat{a}}^{\pi} aL ((1 - t)a, T - \pi q, A) dF(a) - T - \bar{R} - \pi q F(\hat{a}) \geq 0, \quad (10)
\]

where \( \bar{R} \) is the exogenously fixed amount of public expenditures for other purposes than income redistribution and the NHS. Note that the last term on the \( lth \)s of

\(^3\)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).
the budget constraint, $\pi qF(\hat{a})$, 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 either to one of two corner solutions or to an interior solution. At these two corner solutions health care is exclusively provided by either the NHS ($\hat{a} = \pi$) or private medicine ($\hat{a} = a$). 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 $\hat{a} = \pi$) and, second, it is equivalent for citizens to pay for their expected cost of medical treatment through either an insurance premium ($\pi q$) at the lower corner solution ($\hat{a} = a$) or through a reduction in $T$ of the same magnitude at the upper corner solution ($\hat{a} = \pi$). Whether the optimal size of the NHS corresponds to an interior or corner solution depends on the parameters of the model, as will be illustrated in section 6. In the remainder of this section, we characterise the interior solution.

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 $\mu$ the multiplier of its budget constraint, the following expression is derived in the appendix:

$$\frac{1}{\mu} \frac{\partial \mathcal{L}}{\partial \hat{a}} = -\pi \int_a^{\hat{a}} \left[ (1 - t)a B^{N_b}(a) + ta \right] dF(a) \frac{\partial w}{\partial \hat{a}}$$

$$- \left\{ ta \left[ \hat{L}^P - \pi \hat{L}^{N_b} - (1 - \pi) \hat{L}^{N_g} \right] + \pi q \right\} f(\hat{a})$$

(11)

where

$$B^{N_b}(a) \overset{\text{def}}{=} \frac{1}{\mu} \psi(a) v_T^{N_b} + ta L_T^{N_b}$$

(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 $\mu$, 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 sick persons resorting to the NHS to rise by $-\frac{\partial w}{\partial \hat{a}} > 0$ 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 $B^{N_b}(a)$, and also a direct effect on tax revenue collected form such a person that amounts to $ta$.

---

4The function $w(\hat{a}, t, T)$, implicitly defined by (8a)–(8c), exhibits a discontinuity at $\hat{a} = \pi$ where it drops to zero.
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 reduce their (expected) labour supply by \( \hat{L}^P - \pi \hat{L}^{Nb} - (1 - \pi)\hat{L}^{Ng} \).\(^5\)

Evaluating (11) for \( a = \bar{a} \) leaves us with no benefits and only budgetary costs. We thus have

**Result 1.** *The introduction of 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.\(^6\) 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:

\[
\frac{\hat{a}f(\hat{a})}{F(\hat{a})} = \frac{\pi E \left[ (1 - t)aB^{Nb}(a) + ta \mid a \leq \hat{a} \right] \left( -\frac{\partial w}{\partial a} \right) \hat{a}}{t\hat{a} \left[ \hat{L}^P - \pi \hat{L}^{Nb} - (1 - \pi)\hat{L}^{Ng} \right] + \pi q}.
\]

The rhs 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 lhs is the elasticity of the distribution function at \( \hat{a} \). For many familiar distribution functions, this elasticity falls in \( \hat{a} \).

In section 6, we construct numerical examples showing in which 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. We close

\(^5\)\( \hat{L}^P - \pi \hat{L}^{Nb} - (1 - \pi)\hat{L}^{Ng} > 0 \) by the assumption that the composite commodity and leisure are normal goods and the facts that the switching individuals have their time endowment reduced and no longer pay an insurance premium.

\(^6\)This reasoning is valid when \( f(\bar{a}) > 0 \). When \( f(\bar{a}) = 0, \frac{\partial^{2}E}{\partial a^{2}} \bigg|_{a=\bar{a}} = 0 \), and we need to investigate the sign of \( \frac{\partial^{2}E}{\partial a^{2}} \bigg|_{a=\bar{a}} \). Since \( \frac{\partial^{2}L}{\partial a^{2}} \bigg|_{a=\bar{a}} = \mu \{t\hat{a}[\pi L^{Nb} + (1 - \pi)\hat{L}^{Ng} - \hat{L}^P] - \pi q \} \hat{a} \bigg|_{a=\bar{a}}, \) and the square bracket term is negative due to Assumption N, we can claim that \( \frac{\partial^{2}E}{\partial a^{2}} \bigg|_{a=\bar{a}} < 0 \). Initially, social welfare is therefore a concave function of \( \hat{a} \), and will never increase with \( \hat{a} \) at \( \bar{a} \).
the present section by characterising the optimal linear tax policy. The following expression of the marginal tax rate is derived in the appendix:

\[
\begin{align*}
t &= -\text{cov}(B(a), aL) - \left[ \pi F(\hat{a})E \left[ (1 - t)aB^N + ta \mid a \leq \hat{a} \right] \left( \frac{\partial w}{\partial t} - E[aL]\frac{\partial w}{\partial T} \right) \right] \\
&\quad - E \left( a\frac{\partial t}{\partial t} \right),
\end{align*}
\]

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

Expression (14) is a modified Sheshinski (1972)-rule for the optimal marginal income tax rate and holds for an arbitrary size if the NHS (\(\hat{a}\)). 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.

Conditional on an interior solution for the NHS (i.e. (13)), (14) can be rewritten as

\[
\begin{align*}
t &= -\text{cov}(B(a), aL) - \left( t\hat{a}[\hat{L}P - E\hat{L}^N] + \pi q \right) \left( -\frac{\partial w}{\partial a} \right)^{-1} \left( \frac{\partial w}{\partial t} - E(aL)\frac{\partial w}{\partial T} \right) f(\hat{a}) \\
&\quad - E \left( a\frac{\partial t}{\partial t} \right).
\end{align*}
\]

(15)

If \(\partial w/\partial t - E[aL]\partial w/\partial T\) is positive, the optimal marginal tax rate will, to a first approximation, be lower than without a mixed health care system: the \(T\)-compensated increase in \(t\) increases the waiting time, making \(f(\hat{a})\) people going private who increase their labour supply and thus their tax payments and reduce the claim on the NHS budget. In general, it is difficult to establish the sign of \(\partial w/\partial t - E[aL]\partial w/\partial T\). However, in the institutional setting discussed in the next section, this waiting time effect simplifies straightforwardly.

**Result 2.** If a mixed health care sector is optimal, its size should satisfy (13). The corresponding optimal marginal tax rate is implicitly given by (15).

### 5 Actuarially fair insurance against waiting

A citizen resorting to the NHS receives treatment at a double cost: (i) she has to wait, leading to an ex post negative time endowment effect; and (ii) she face ex

\[\text{The reason is the indeterminate sign of } \frac{\partial w}{\partial T} - \text{see (8c).}\]
ante uncertainty regarding to the size of the time endowment, leading to a risk premium. A risk averse person will want to insure herself against the endowment effect of the waiting time risk. Fair insurance would entitle her to an indemnity $(1-t)aw$ in the case of illness in return for paying ex ante a premium $(1-t)aπw$. This type of insurance will be provided on a competitive insurance market even when insurance firms do not observe a consumer’s ability ex ante. All that is required is that NHS patients can produce evidence of being on a waiting list. The result is a fully insured NHS consumer who supplies

$$L^N = L((1-t)a, T - (1-t)aπw, A)$$

(16a)

hours of labour and ends up with utility

$$v^N = v((1-t)a, T - (1-t)aπw, A),$$

(16b)

irrespective of her health status.

Comparing (16b) with the utility level of a citizen opting for private health care (4b) shows that the equilibrium waiting time is straightforwardly given by $w = [(1-t)\hat{a}]^{-1}q$, a figure that is independent of the lump sum transfer $T$ and the probability $π$. As a result, a citizen with critical ability $\hat{a}$, who switches from the private to the public system, will not change her labour supply:

$$\hat{L}^N = \hat{L}^P = L((1-t)a, T - πq, A).$$

and the optimality conditions (13) and (15) reduce to:

$$\hat{a}f(\hat{a}) = \frac{π\mathbb{E}[aB^{N_b}(a) | a \leq \hat{a}]}{\hat{a}} + \frac{t}{1-t} \frac{\mathbb{E}[a | a \leq \hat{a}]}{\hat{a}},$$

(17)

$$t = \frac{-\text{cov}(B(a), aL) - \frac{π\hat{a}f(\hat{a})}{1-t}}{-\mathbb{E}\left(\frac{\partial aL}{\partial t}\right)}.$$  

(18)

To a first approximation, the optimal marginal tax rate is thus lower than in the absence of a mixed health care system.

**Result 3.** With competitive waiting time insurance, the optimal interior size of the NHS is governed by (17), while the optimal marginal income tax rate satisfies (18), and is thus—to a first approximation—lower than with a pure health care system.

Though it is clear that, ceteris paribus, the availability of waiting time insurance increases the ex ante welfare of any person relying on the NHS, this will at

---

8Competition on the insurance market leads to an efficient separating equilibrium, since the profitability of each (premium, indemnity)-contract does not depend on the identity of the consumer—unlike in the case where $π$ is private information.
the same time make the NHS more attractive to richer citizens who so far relied on private insurance, making them switch, reducing their labour supply and tax payments, and increasing the burden on the NHS budget. It is thus not obvious that waiting time insurance is desirable. In the appendix we show that if the premium for such insurance has a loading factor \( \lambda \), i.e. \((1 + \lambda)(1 - t)aw\), then

\[
\frac{1}{\mu} \frac{\partial L}{\partial \lambda} |_{\lambda=0} = t \int_a^b adF > 0.
\]

Result 4. From a social welfare point of view, it is not optimal to have a perfectly competitive market for insurance against the waiting time risk.

Result 4 is an instance that markets may act as constraints on redistribution policy. This was earlier shown to be the case when people can side trade commodities within coalitions (Hammond, 1987, Schroyen, 1997) or on second hand markets (Guesnerie, 1995).

6 Numerical results from the model

In section 4, we asserted that the government’s problem is not a concave in terms of the size of the NHS, meaning that a size satisfying (13) be dominated by a pure (private or public) health care system. Whether a mixed system is optimal, and if so how large the should be NHS, depends both on positive parameters (the shape of the ability distribution, the income elasticity of labour supply) of the model, as well as normative ones (the degree of inequality aversion). It is the purpose of this section to assess some of these dependencies.

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

\[
L^N_g = \beta A - (1 - \beta)[(1 - t)a]^{-1}T,
L^N_b = \beta(A - w) - (1 - \beta)[(1 - t)a]^{-1}T,
\]

and

\[
\nu^N_g = [(1 - t)a]^{\beta}A + [(1 - t)a]^{\beta-1}T,
\nu^N_b = [(1 - t)a]^{\beta}(A - w) + [(1 - t)a]^{\beta-1}T.
\]

For an individual of ability \( a \) having opted for a private insurance these functions are

\[
L^p = \beta A - (1 - \beta)[(1 - t)a]^{-1}(T - \pi q)
\]
and
\[ v^P = [(1 - t)a]^\beta A + [(1 - t)a]^{\beta - 1} (T - \pi q). \]  

(21b)

Having chosen for the particular cardinalisation of the consumer’s preferences, she is risk averse w.r.t temporal uncertain consumption and leisure prospects, but risk neutral w.r.t timeless uncertain endowment prospects. This risk neutrality significantly simplifies the analysis, because it allows us to proceed in two stages. It also means that an NHS citizen will behave as if fully insured against the waiting time risk. The equilibrium level of waiting time is thus

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

(22)

In the appendix, we use the above labour supply and waiting time functions in the budget constraint (10) to obtain a Laffer curve \( T(t, \hat{a}) \). Likewise, using the indirect utility functions derived above and substituting again \( w \) from (22), we obtain an expression for social welfare given in (9): \( SW(t, \hat{a}) \). 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, \( \beta \) is chosen equal to .4, 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: \( \mathbb{E}[a] = 1 \). The value for \( \pi \) is set at .3: it is 30% potential income (\( A \) times the average wage rate); while that for \( \pi q \) is set at 0.05. We have considered three ability distributions: uniform \( (F(a) = \frac{a - \alpha}{\pi - \alpha}) \), log-uniform \( (F(a) = \frac{\log a - \log \alpha}{\log \pi - \log \alpha}) \), and Beta(2,5) \( (F(a) = \int_{\alpha}^{\pi} \frac{1}{B(2,5)} \left( \frac{a - \alpha}{\pi - \alpha} \right)^4 \frac{1}{\pi - \alpha} dx) \). 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 grounds. For each distribution, we consider different mean preserving spreads. These spreads are modeled by the means of the dispersion parameter \( D \overset{\text{def}}{=} \pi/\alpha \). Figure 1 below gives the Beta(2,5) density function for \( D = 100 \) with support \([\alpha, \pi] = [.034, 3.41]\).
The Beta (2,5) density function with mean normalised to 1.

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 curve I 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 (case III). In case II, a mixed system performs equally well w.r.t. a fully private system. In the tables below, we show that the dispersion parameter $D$ must be sufficiently large to justify a mixed system.

Shapes for $SW(t(\hat{a}), \hat{a})$.

In these tables, we give the optimal tax policy and the NHS size for different values of $D$. Because this support size is hard to interpret, we provide two other characteristics of the ability distribution: the decile ratio $P90/P10$ and the Gini coefficient. To measure how much better the optimal mixed system performs, compared with a pure system, we computed a deadweight loss figure ($DWL$).
This figure tells by how much, expressed as a percentage of actual GDP, we should decrease the value of $\mathcal{R}$, 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.\footnote{More precisely, $DWL = \frac{\mathcal{R} - R}{GDP(\text{pure NHS}; R)} \times 100$ where $R$ solves $SW(\text{optimal mixed system}; R) = SW(\text{pure NHS}; R)$}

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 and $F^{-1}(.9)/F^{-1}(.1) = 7.2$. 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 $\mathcal{R}$ 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.

### Table 1: Mean preserving spreads of the uniform ability distribution.

<table>
<thead>
<tr>
<th>Summary statistics</th>
<th>Utilitarian SW</th>
<th>Rank order SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$P_{90}/P_{10}$</td>
<td>Gini</td>
</tr>
<tr>
<td>17</td>
<td>5.9</td>
<td>.296</td>
</tr>
<tr>
<td>35</td>
<td>7.2</td>
<td>.315</td>
</tr>
<tr>
<td>70</td>
<td>8.0</td>
<td>.324</td>
</tr>
<tr>
<td>120</td>
<td>8.4</td>
<td>.327</td>
</tr>
<tr>
<td>250</td>
<td>8.7</td>
<td>.330</td>
</tr>
<tr>
<td>1000</td>
<td>8.9</td>
<td>.332</td>
</tr>
</tbody>
</table>

### Table 2: Mean preserving spreads of the log-uniform ability distribution.

<table>
<thead>
<tr>
<th>Summary statistics</th>
<th>Utilitarian SW</th>
<th>Rank order SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$P_{90}/P_{10}$</td>
<td>Gini</td>
</tr>
<tr>
<td>51</td>
<td>23.3</td>
<td>.531</td>
</tr>
<tr>
<td>128</td>
<td>48.5</td>
<td>.603</td>
</tr>
<tr>
<td>150</td>
<td>55.2</td>
<td>.614</td>
</tr>
<tr>
<td>180</td>
<td>63.7</td>
<td>.626</td>
</tr>
<tr>
<td>250</td>
<td>82.6</td>
<td>.645</td>
</tr>
</tbody>
</table>

\footnote{More precisely, $DWL = \frac{\mathcal{R} - R}{GDP(\text{pure NHS}; R)} \times 100$ where $R$ solves $SW(\text{optimal mixed system}; R) = SW(\text{pure NHS}; R)$}
Table 3: Mean preserving spreads of the Beta(2,5) ability distribution.

<table>
<thead>
<tr>
<th>Summary statistics</th>
<th>Utilitarian SW</th>
<th>Rank order SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$P90/P10$</td>
<td>Gini</td>
</tr>
<tr>
<td>63.5</td>
<td>4.8</td>
<td>0.298</td>
</tr>
<tr>
<td>134</td>
<td>5.2</td>
<td>0.306</td>
</tr>
<tr>
<td>250</td>
<td>5.3</td>
<td>0.310</td>
</tr>
<tr>
<td>1000</td>
<td>5.5</td>
<td>0.313</td>
</tr>
</tbody>
</table>

These tables invite for the following remarks.

1. If it is optimal to have a mixed health care system with utilitarian social preferences, then such a system (though not necessarily of the same size) is also optimal under the rank order social welfare function. This is not surprising, the driving force towards a mixed system being to redistribute real income across individuals of different abilities.

2. The size of the NHS under an optimal mixed system hinges very much upon the shape of the ability distribution: the fatter the left hand tail, the larger the NHS.

3. For a given distribution, a mixed system only becomes superior if the dispersion ($D$ or $P90/P10$) is large enough.

4. The summary statistics indicate that the Beta (2,5) distribution is probably the closest description of a real ability distribution. Our simulations then show that (i) a mixed system should have a very small NHS, and (ii) that the welfare gain w.r.t. a pure (public or private) system is insignificant.

5. 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 $D$. Marginal tax rate and NHS size can be either complements or substitutes in redistributing real income across individuals.

7 Concluding remarks

It is often argued that the government may stretch the limits to redistribution by supplying private goods at sufficiently low quality free of charge. In countries with mixed health care systems, this lower quality is, among others, caused by waiting lists for elective treatments. In our set-up such 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 constructed a model that embeds the choice
of health care provider in an otherwise standard optimal income taxation prob-
lem. We found that (i) a small NHS is never desirable, (ii) the dispersion of the
ability distribution in the population needs to be large enough for a mixed health
care system to be socially optimal, (iii) conditional on our parameter choices, the
welfare gains to be accomplished by a mixed system relative to the best pure
system are quiet low – especially for bell shaped distributions that are skewed to
the right. This last conclusion has to implications. First, that policy measures
aimed at reducing waiting times in the NHS need not necessarily override any
redistribution in society. Second, it invites for investigating whether cost reduc-
ing measures, such as lower quality of amenities in the NHS, are cheaper sorting
devices.

In this paper, some features of the health care sector have been swept un-
der 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 ac-
tual mixed systems; clearly, it does not apply to emergency care. Including these
features would complicate the analysis, but we believe would not affect our qual-
itative results.

References


insurance: do waiting lists matter? Journal of Public Economics, 72, 155-
181.


as a redistributive device in an optimum income tax model. Scandinavian
Journal of Economics, 97, 547-567.


Appendix

7.1 Derivation of the first order conditions in Section 4

The Lagrange function to the planning problem is

\[ L = \int_{\hat{a}}^{\bar{a}} \psi(a) \left[ (1 - \pi) v((1 - t)a, T, A) + \pi v((1 - t)a, T, A) \right. \]

\[ - w(\hat{a}, t, T)) \int_{\hat{a}}^{\pi} \psi(a) v((1 - t)a, T - \pi q, A) \] \[ dF(a) + \mu \left\{ t \int_{\hat{a}}^{\bar{a}} a \left[ (1 - \pi) L((1 - t)a, T, A) + \pi L((1 - t)a, T, A - w(\hat{a}, t, T)) \right] dF(a) \right. \]

\[ + t \int_{\hat{a}}^{\pi} aL((1 - t)a, T - \pi q, A) dF(a) - T - \bar{F} - \pi q F(\hat{a}) \} . \]

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

\[ \frac{\partial L}{\partial \hat{a}} = -\pi \int_{\hat{a}}^{\bar{a}} \psi(a) \nu_T^{N_b} (1 - t) a dF(a) \frac{\partial w}{\partial \hat{a}} + \mu \pi t \int_{\hat{a}}^{\bar{a}} a \frac{\partial L^{N_b}}{\partial w} dF(a) \frac{\partial w}{\partial \hat{a}} \]

\[ + \mu \left\{ t \hat{a} \left[ \pi L^{N_b} + (1 - \pi) \hat{L}^{N_g} - \hat{L}^P \right] f(\hat{a}) - \pi q f(\hat{a}) \} . \]

Defining \( B^{N_b}(a) \overset{\text{def}}{=} \frac{1}{\mu} \psi(a) \nu_T^{N_b} + t a \frac{\partial L^{N_b}}{\partial T}, \) using the fact that

\[ \frac{\partial L^{N_b}}{\partial w} = - \left[ \frac{\partial L^{N_b}}{\partial T} (1 - t) a + 1 \right] (= - \frac{\partial c^{N_b}}{\partial T}), \]

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

The derivative w.r.t. \( t \) is

\[ \frac{\partial L}{\partial t} = - \int_{\hat{a}}^{\bar{a}} \psi(a) [\pi v_T^{N_b} a L^{N_b} + (1 - \pi) v_T^{N_g} a L^{N_g}] dF(a) - \int_{\hat{a}}^{\pi} \psi(a) v_T^{P} a L^{P} dF(a) \]

\[ - \pi \int_{\hat{a}}^{\bar{a}} \psi(a) v_T^{N_b} (1 - t) a dF(a) \frac{\partial w}{\partial t} + \mu \pi t \int_{\hat{a}}^{\bar{a}} a \frac{\partial L^{N_b}}{\partial w} dF(a) \frac{\partial w}{\partial t} \]

\[ + \mu \left\{ \int_{\hat{a}}^{\bar{a}} [\pi a L^{N_b} + (1 - \pi) L^{N_g}] dF(a) + \int_{\hat{a}}^{\pi} a L^{P} dF(a) \right. \]

\[ + t \int_{\hat{a}}^{\bar{a}} [\pi a \frac{\partial L^{N_b}}{\partial t} + (1 - \pi) a \frac{\partial L^{N_g}}{\partial t}] dF(a) + t \int_{\hat{a}}^{\pi} a \frac{\partial L^{P}}{\partial t} dF(a) \} . \]
and that w.r.t. $T$ is

\[
\frac{\partial L}{\partial T} = \int_0^\lambda \psi(a)[\pi v_T^{Nb} + (1 - \pi)v_T^{Ng}]dF(a) + \int_\lambda^\infty \psi(a)v_T^P dF(a) \\
- \pi \int_0^\lambda \psi(a)v_T^{Nb}(1 - t)adF(a)\frac{\partial w}{\partial T} + \mu \pi t \int_0^\lambda a \frac{\partial L_T^{Nb}}{\partial w} dF(a)\frac{\partial w}{\partial T} \\
+ \mu \left\{ t \int_0^\lambda \left[ \pi a \frac{\partial L_T^{Nb}}{\partial T} + (1 - \pi)a \frac{\partial L_T^{Ng}}{\partial T} \right] dF(a) + t \int_\lambda^\infty a \frac{\partial L_T^P}{\partial T} dF(a) - 1 \right\}.
\]

Performing $\frac{\partial \lambda}{\partial t} + \frac{\partial \psi}{\partial T} E[aL]$ gives

\[
- E[\psi(a)v_TaL] + E[\psi(a)v_T]E[aL] \\
- \pi \left\{ \int_0^\lambda [\psi(a)v_T^{Nb}(1 - t)a + \mu ta \frac{\partial L_T^{Nb}}{\partial w}] dF(a) \right\} \left( \frac{\partial w}{\partial t} + \frac{\partial w}{\partial t} E[aL] \right) \\
+ t\mu \left\{ E[a \frac{\partial L}{\partial t}] + E[a \frac{\partial L}{\partial T}] E[aL] \right\}.
\]

Using the Slutsky decomposition $\frac{\partial \lambda}{\partial t} = \frac{\partial L}{\partial t} - \frac{\partial L}{\partial T} aL$ 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 Derivation of expression (19)

Assume that a consumer with ability $a$ can buy an insurance indemnity $I$ to be paid out in case of waiting at a premium that is fair up to a loading factor $\lambda$: $(1 + \lambda)\pi I$. If $\lambda = 0$, the optimal indemnity, $I^o(a)$, will be $(1 - t)aw$, but if $\lambda > 0, I^o(a) < (1 - t)aw$. The person’s indirect utility is then

\[
Ev^N = (1 - \pi)v((1 - ta), T - (1 + \lambda)\pi I^0(a), A) + \pi v((1 - ta), T - (1 + \lambda)\pi I^0(a) + I^o(a), A - w),
\]

and it follows that $\frac{\partial Ev^N}{\partial \lambda} = -Ev_T^N \pi I^o(a)$. Since $\frac{\partial Ev^N}{\partial w} = -\pi v_T^{Nb}(1 - t)\hat{\lambda}$, the effect of a change in $\lambda$ on the equilibrium waiting time is

\[
\frac{\partial w}{\partial \lambda} = -\frac{\partial Ev^N/\partial \lambda}{\partial Ev^N/\partial w} = -\frac{Ev_T^N I^o(\hat{\lambda})}{v_T^{Nb}(1 - t)\hat{\lambda}},
\]

which reduces to $-w$ under full insurance ($\lambda = 0$).

We then have

\[
\frac{1}{\mu} \frac{\partial L}{\partial \lambda} = -\pi \int_0^\lambda EB_N(a)I^o(a)dF(a) - \pi \int_0^\lambda [(1 - t)aB_T^{Nb}(a) + ta] dF(a) \frac{\partial w}{\partial \lambda}.
\]

Evaluated at $\lambda = 0$ gives expression (19) in the text.
7.3 An explicit solution for the optimal tax policy

Using the indirect utility functions (20b), (21b) and substituting \( w \) from (22), we obtain the following expression for the social welfare given in (9):

\[
SW = (1 - t)^\beta E[\psi(a)a^\beta]A + (1 - t)^{\beta-1}E[\psi(a)a^{\beta-1}]T
- \pi q(1 - t)^{\beta-1} \left[ \int_a^\pi \frac{\psi(a)a^\beta}{\hat{a}}dF(a) + \int_a^\pi \psi(a)a^{\beta-1}dF(a) \right],
\]

Dividing through by \( E[a]A\) \( E[\psi(a)a^{\beta-1}] \), this can also be written as

\[
\frac{SW}{E[a]A} = (1 - t)^{\beta} \kappa + \tau(1 - t)^{\beta-1} - \rho(1 - t)^{\beta-1} \gamma(\hat{a}),
\]

where

\[
\tau = \frac{T}{E[a]A}, \quad \kappa = \frac{1}{E[a]} \int_a^\pi \psi(a)a^{\beta-1}dF(a), \quad \rho = \frac{\pi q}{E[a]A},
\]

and

\[
\gamma(\hat{a}) = \frac{\int_a^\pi \frac{\psi(a)a^\beta}{\hat{a}}dF(a) + \int_a^\pi \psi(a)a^{\beta-1}dF(a)}{\int_a^\pi \psi(a)a^{\beta-1}dF(a)}.
\]

The parameter \( \kappa \) 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 \( \hat{a} \), the welfare trade-off between \( t \) and \( \tau \) is given by

\[
- \frac{d\tau}{dt} \bigg|_{dSW=0} = \beta \kappa - \tau \frac{1 - \beta}{1 - t} + \rho \frac{1 - \beta}{1 - t} \gamma(\hat{a}), \tag{A1}
\]

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

Likewise, using the labour supply functions (20a), (21a) and substituting \( w \) from (22), the budget constraint given in (10) yields the Laffer curve:\(^{10}\)

\[
T = \frac{1 - t}{1 - \beta t} \left[ t\beta E(a)A - \pi q F(\hat{a}) - \pi F(\hat{a}) + \frac{t}{1 - t} \pi q \right] \left\{ (1 - \beta)(1 - F(\hat{a})) - \beta \int_a^\pi \frac{a}{\hat{a}}dF(a) \right\},
\]

or

\[
\tau = \frac{1 - t}{1 - \beta t} \left[ t\beta - \rho F(\hat{a}) - s + \frac{t}{1 - t} \rho \delta(\hat{a}) \right], \tag{A2}
\]

\(^{10}\)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.
The optimal tax rate is therefore 

\[ T = \frac{\delta(a)}{1 - \beta} \left( 1 - F(a) \right) - \beta \int_{a}^{\hat{a}} \frac{a dF(a)}{\hat{a}}. \]

The budgetary trade-off between \( t \) and \( \tau \) is then

\[ \frac{d\tau}{dt} \bigg|_{d\tau=0} = \frac{1}{(1-\beta t)^2} \left[ (1-\beta)\rho F(\hat{a}) + s + (1-2t+\beta^2)\beta + \rho \delta(\hat{a}) \right]. \quad (A3) \]

Replacing in the welfare trade-off (A1) \( \tau \) by the rhs of (A2) and equating it to the budget trade-off (A3) yields a third degree polynomial in \( t \):

\[ \beta^3(1-\kappa)t^3 + \{\beta[\rho \varphi(\hat{a}) + s + \rho \delta(\hat{a}) - 1] - \beta^2[\rho \varphi(\hat{a}) + s - \kappa] - (\beta^2 + \beta^3)(1-\kappa) + \beta^2[\gamma(\hat{a}) - \delta(\hat{a})] \rho \} t^2 + \{ \beta^2 + \beta)(1-\kappa) + \beta^2[\rho \varphi(\hat{a}) + s - \kappa] - \beta[\rho \varphi(\hat{a}) + s - 1] - (\beta^2 - 2\beta^2 + 2\beta)\rho \gamma(\hat{a}) + \beta \rho \delta(\hat{a}) \} t + \left\{ -\beta(1-\kappa) + (1-\beta)\rho \gamma(\hat{a}) - \rho \delta(\hat{a}) \right\} = 0. \]

For the general third degree polynomial

\[ a_1x^3 + a_2x^2 + a_3x + a_4 = 0, \]

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

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:

\[
\begin{align*}
\text{root}_1 &= -\frac{1}{3} a_1 + (S + T) \\
\text{root}_2 &= -\frac{1}{3} a_1 - \frac{1}{2} (S + T) + \frac{1}{2} i\sqrt{3}(S - T), \quad \text{and} \\
\text{root}_3 &= -\frac{1}{3} a_1 - \frac{1}{2} (S + T) - \frac{1}{2} i\sqrt{3}(S - T),
\end{align*}
\]

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

\[ 0 < \text{root}_2 < \text{root}_3 < 1 < \text{root}_1. \]

The optimal tax rate is therefore \( \text{root}_2 \).

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 (A2). 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 (A3) equal to zero.