Working Paper No. 14/01

Markets for public and private health care: redistribution arguments for a mixed system

by

Maurice Marchand and Fred Schroyen

SNF Project No. 2704
Markeder for helsetjenester

The project is financed by the Research Council of Norway

FOUNDATION FOR RESEARCH IN ECONOMICS AND BUSINESS ADMINISTRATION
BERGEN, APRIL 2001

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, for instance, 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.

There exists by now a large literature collecting the arguments in favour and against one type of health care or the other. The papers by Besley and Gouveia (1994), Cullis *et al* (1999), Propper and Green (1999) 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 main concern in this paper is to examine some equity arguments in favour of a mixed system.

Our approach will therefore abstract from a number of aspects. Some of these are of minor importance but others are not, and in so far as they are not present our model should be considered as a benchmark. Thus, in this paper we will abstract from the uncertain incidence of illness, the informational asymmetries between patients and physicians, and the market power of physicians. Obviously, these are all important characteristics in the market for health care, and the only reason for not having them in the picture is that we want to develop a simple setting that we can handle in a formal way.

We consider an economy where citizens need a well-defined medical treatment once a year. These citizens differ in earnings capacity. For receiving treatment they can resort either to the NHS or to the pri-

vate market for health care. In the former, 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. The total number of physicians is allowed to depend on their living standard. Private physicians determine themselves how many consultations to perform, while the work load and salary of NHS doctors is specified in a labour contract with the government. In equilibrium, physicians are indifferent where to work and no patient wants to change physician. In particular, all citizens with a earnings capacity below a certain level will resort to the NHS while the others buy a private treatment. We analyse the arguments that a welfare maximising government should account for when deciding on (i) the size of the NHS, (ii) the subsidy of private health care, (iii) the terms of the contract it offers to NHS physicians, and (iv) the parameters of a linear income tax. Our main conclusions are threefold. First, a small NHS system is not desirable. The social benefits are of second order importance relative to the associated social costs. Second, it may be optimal to have an extensive public health care system, but a necessary condition for this is that the spread in the income distribution is sufficiently wide. Third, there is a good reason for having an NHS labour contract such that public physicians work harder than their private colleagues. In this way the government can use its monopsony power on the market for physicians to improve social welfare. It means that there is some optimal departure from production efficiency as the marginal reservation wage is higher in the NHS than in the private market.

It is noteworthy that if there were no limit to redistribution, that is if lump-sum income taxes and transfers could be differentiated by individual abilities, an indifference about the mix of NHS and private practice would result in our setting. This emphasizes that when it is socially optimal to operate a mixed system, it is because it allows to move forward 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.[1] Iversen (1997) lets patients

[1]Brekke and Sørgard's (2000) approach is more distant from ours because they

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 practice (Section 2). Next, we build a simple model of occupational choice in the medical profession. When this is combined with the demand side, we are able to discuss how the equilibrium in the health care sector depends on the contract offered to NHS physicians (Section 3) and in particular how the waiting time in the NHS is determined (Section 4). Thereafter, we set up the normative problem (Section 5) and analyse the optimality properties of the three above mentioned policies: first under a production efficiency requirement

consider an imperfect competition setting. Physicians in a local market for health care divide their time between working in the public health sector (where they earn a wage set by the government and cater for patients that they otherwise would have seen in their private practice) and working in private. Brekke and Sørgard characterize the equilibria both when the two physicians choose their labour supplies in the two sectors non-cooperatively and when they coordinate. They look at the wage rate and subsidy rate that maximise the sum of consumer and producer surplus, taking into account a marginal cost of public funds.

(Section 6) and next in a more general context where the government can also offer contracts that depart from production efficiency (Section 7). Concluding remarks are offered in Section 8.

2 The demand for private and public health care

Citizens care about their health status (h), consumption of a composite good (c) and leisure (ℓ). Their preferences on these three 'goods' can be described by an additive and strictly concave utility function $r(h) + u(c, \ell)$. We thus abstract from the possibility that better health may enhance the enjoyment of consumption and leisure and that bad health may affect the disutility of labour. Throughout the paper, leisure and the composite good are assumed to be normal goods.

All citizens are endowed with some earnings capacity (or ability) a, and a time endowment normalised to 1. People are distributed over the ability interval $[\underline{a}, \overline{a}]$ according to the distribution function $F(a)$ (with density $f(a)$). We have in mind an ability distribution that is skewed to the right, but we will also look at a numerical example with a uniform distribution. Labour earnings are taxed at a constant marginal tax rate t, and a per capita tax T is levied. Thus when working during L hours, the disposable income of an individual of ability a amounts to $(1-t)aL - T$, and leisure time to $1 - L$. His or her optimal labour supply satisfies $a(1-t)\partial u/\partial c = \partial u/\partial \ell$.

Each year, every person needs one unit of medical treatment (one consultation), and this treatment can be obtained either in the NHS or in a private practice. There is free access to the NHS which is financed out of income tax revenue. However, a patient will be put on a waiting list before receiving medical treatment in the NHS. If the patient has to wait for w weeks, the discomfort in terms of reduced health status is taken to be δw. Hence, utility when resorting to an NHS doctor is

$$r(\overline{h} - \delta w) + v\big((1-t)a, -T\big) \tag{1}$$

where $\overline{h}$ is the after-treatment health status in the absence of any waiting time and $v\big((1-t)a, m\big)$ is the indirect utility function giving the non-

health utility for a net wage rate $(1-t)a$ and exogenous income m when labour supply is chosen at its optimal level, $L((1-t)a, m)$. For further reference, we denote the marginal utility of income by α and remind the reader of Roy's identity: $\partial v/\partial t = -\alpha a L$.

If a patient goes to see a private doctor, she is given medical care on the spot but will have to pay a consultation fee q. This out-of-pocket fee can be smaller than what the physician receives for carrying out the treatment if private medical care is subsidised. Overall utility is then equal to

$$r(\overline{h}) + v\big((1-t)a, -T-q\big). \tag{2}$$

Because health status is a normal good, there exists a critical ability level $\widehat{a}$ such that citizens with a higher ability make use of private health care, while all those with a lower ability seek help with the NHS. Formally,

$$\exists\, \widehat{a} \in [\underline{a}, \overline{a}] : r(\overline{h} - \delta w) + v\big((1-t)a, -T\big) \gtreqless r(\overline{h}) + v\big((1-t)a, -T-q\big)$$
$$\iff a \lesseqgtr \widehat{a}. \tag{3}$$

When we normalise the population size to unity, the demand for NHS and private consultations is thus given by $F(\widehat{a})$ and $1-F(\widehat{a})$, respectively.

3 The equilibrium conditions in the physician market

The total number of physicians will depend on the living standard one can obtain in this occupation, denoted by $\widetilde{U}$, compared with the one achieved in one of the alternative occupations. Thus, one can expect that a higher utility level $\widetilde{U}$ will convince more students to study for a medical degree and join the medical profession. The supply of physicians is then modelled as

$$M(\widetilde{U}), \text{ with } M' > 0, M'' < 0. \tag{4}$$

We denote the elasticity of M with respect to $\widetilde{U}$ by ε^M_U.[2] The size of this elasticity will depend on the time perspective one takes. In the short run,

[2]Throughout the paper, we will denote the elasticity of x w.r.t. z as ε^x_z. Thus $\varepsilon^x_z = \frac{\mathrm{d}\ln x}{\mathrm{d}\ln z}$.

it is probably close to zero, while in the long run it can be significantly higher.

Physicians have preferences over consumption (c) and leisure (ℓ) that can be represented by a strictly quasi-concave utility function $U(c, \ell)$. They are endowed with one unit of time and have the opportunity to work either for the NHS or in a private practice. An NHS contract specifies the number of consultations a physician is required to perform per period (s_N) as well as the salary (y). We do not consider the possibility for NHS doctors to supplement their income with private consultations. This is because we show later on that the optimal NHS contract would not give them any incentive to do so.

We are assuming that the market for private health care is perfectly competitive and abstract from any informational asymmetries that may lead to agency problems and local monopoly power on behalf of the physician. The consultation fee that a private physician charges is denoted by π. As mentioned earlier, this may differ from the patient's out-of-pocket payment q owing to some (ad valorem) subsidy rate σ: $q = (1 - \sigma)\pi$. Thus a private doctor takes the market price for private consultations, π, as given and determines himself how many consultations to perform per period by solving

$$\max_{s_P} U(\pi\, s_P, 1 - s_P). \tag{5}$$

Let the solution to the private physician's problem be denoted as $s_P(\pi)$, and the corresponding utility level as $V(\pi)$.

Denoting the number of NHS and private physicians as M_N and M_P, respectively, we are now ready to state the equilibrium conditions for the two markets:

$$M_N \geq 0, M_P \geq 0, \tag{6}$$

$$M_N + M_P = M(\widetilde{U}), \tag{7}$$

$$s_N\, M_N + s_P(\pi)\, M_P = F(\overline{a}), \tag{8}$$

$$M_N\, [U(y, 1 - s_N) - \widetilde{U}] = 0 \;\text{ and }\; U(y, 1 - s_N) \leq \widetilde{U}, \tag{9}$$

$$M_P\, [V(\pi) - \widetilde{U}] = 0 \text{ and } \; V(\pi) \leq \widetilde{U}. \tag{10}$$

Eq (7) states that the number of physicians working in the two sectors should match the number willing to work in the medical profession. Eq (8) ensures that the total number of consultations delivered in the two sectors satisfies the total demand for consultations, equal to $F(\overline{a}) = 1$. Finally, eqs (9) and (10) are complementary slackness conditions that make sure that a sector is active only if the living standard enjoyed by working there equals the living standard of the medical profession.

By defining the equilibrium in the doctor market in this way, we are formally ruling out any rationing of doctor positions (implemented for example through a limitation on the number of students admitted in medical schools). Such rationing in the doctor market would be required if the NHS contract (y, s_N) was making the medical profession too attractive. However, it is clear that rationing is never optimal in our model since if there were rationing, total public health expenditure could be reduced by lowering the salary of NHS physicians, y.

The NHS contract will govern the type of equilibrium that obtains. *If* there is coexistence of an NHS and a private sector, then the equilibrium consultation fee in a private practice should not give any incentive for a physician to switch sector. We denote the no-arbitrage fee when the NHS contract is (y, s_N) by $\pi(y, s_N)$. From (9) and (10), it is the value for π that solves

$$U(y, 1 - s_N) = V(\pi). \tag{11}$$

Therefore, when the NHS contract (y, s_N) is chosen so as to make a private and a public sector coexist (mixed system), the above system of equations defining the equilibrium in the doctor market simplifies to

$$M_N + M_P = M\big(V(\pi(y, s_N))\big) \tag{12}$$

and

$$s_N\, M_N + s_P\big(\pi(y, s_N)\big)\, M_P = F(\overline{a}). \tag{13}$$

If s_N is equal to $s_P(\pi(y, s_N))$, these two equations do not provide a unique solution for M_N and M_P. In that case (which we will study in Section 6), the capacity of the NHS sector – let us label it K_N $(= s_N\, M_N)$

– has necessarily to be decided by the government. This determines then the allocation of physicians across the two sectors.

On the contrary, if the NHS contract (y, s_N) is such that $s_p(\pi(y, s_N))$ differs from s_N (as it will be the case in Section 7) and an NHS coexists at equilibrium with a private sector, equations (12) and (13) can uniquely be solved for M_N and M_P. From this solution one can infer the NHS size that results from the NHS contract (by using $K_N = M_N s_N$). Alternatively, one can however think of the government as choosing NHS capacity K_N and work load s_N. Then, the following system of equations

$$\frac{K_N}{s_N} + M_P = M\big(V(\pi(y, s_N))\big) \tag{14}$$

and

$$K_N + s_P(\pi(y, s_N))\, M_P = F(\overline{a}). \tag{15}$$

yields the salary of NHS doctors (y) and the amount of private physicians (M_P) that ensure an equilibrium without rationing in the doctor market: $y = y(s_N, K_N)$ and $M_P = M_P(s_N, K_N)$.

In line with what has just been said, the government is, in the following, seen as choosing NHS capacity K_N. Equivalently, it can be seen as deciding on $\widehat{a}$ since the demand for NHS consultations must equate the NHS capacity $(F(\widehat{a}) = K_N)$.

4 The equilibrium waiting time for NHS services

As we have set up the model, all citizens need medical care once a year and total demand for health care is therefore entirely inelastic. What is left to patients' discretion is whether they resort to the NHS or the private sector. As we have seen in the previous section, when an NHS and a private sector coexist, the equilibrium fee for a private consultation is determined by the no-arbitrage condition (11). This means that for a given capacity of the NHS sector, $K_N = s_N M_N$, it is up to the waiting

time w to allocate patients across the two sectors: w will adjust such that the demand for NHS consultations matches the NHS capacity, i.e. $F(\widehat{a}) = K_N$ (see also Lindsay and Feigenbaum, 1984 and Martin and Smith, 1999). The equilibrium waiting time, that acts here as a rationing device[3], is thus defined by a no-arbitrage condition for patients:

$$r(\overline{h} - \delta w) + v\big((1-t)\widehat{a}, -T\big) = r(\overline{h}) + v\big((1-t)\widehat{a}, -T - q\big) \tag{16}$$

with $\widehat{a}$ satisfying $F(\widehat{a}) = K_N$.

Straightforward differentiation of (16) gives the following insights:

$$\frac{\mathrm{d}r(\overline{h} - \delta w)}{\mathrm{d}\widehat{a}} = (1-t)[\widehat{\alpha}_P \widehat{L}_P - \widehat{\alpha}_N \widehat{L}_N] > 0, \tag{17}$$

$$\frac{\mathrm{d}r(\overline{h} - \delta w)}{\mathrm{d}t} = -\widehat{a}[\widehat{\alpha}_P \widehat{L}_P - \widehat{\alpha}_N \widehat{L}_N] < 0, \tag{18}$$

$$\frac{\mathrm{d}r(\overline{h} - \delta w)}{\mathrm{d}T} = -[\widehat{\alpha}_P - \widehat{\alpha}_N] < 0, \tag{19}$$

$$\frac{\mathrm{d}r(\overline{h} - \delta w)}{\mathrm{d}q} = -\widehat{\alpha}_P < 0, \tag{20}$$

where a hat on a variable denotes its value for an individual with ability $\widehat{a}$ i.e. indifferent between the two sectors, and subscripts N and P mean evaluation of the variable at lump-sum income of $m = -T$ for an NHS patient and $m = -T - q$ for a private patient, respectively. Recall that α stands for the marginal utility of income. The signs of the derivatives are motivated by the facts that marginal utility of income is falling in consumption and that leisure is a normal good. Along with the separability assumption that health status does not affect labour supply, these facts imply that $\widehat{\alpha}_P > \widehat{\alpha}_N$ and $\widehat{L}_P > \widehat{L}_N$. The marginal patient resorting to the private market works more and has a lower consumption of the composite good than the one resorting to the NHS. Therefore, we may conclude that the equilibrium waiting time is

[3]In our model, the waiting time does not result from the stochastic arrival of patients facing a NHS capacity constraint. See e.g. Worthington (1987) for a model where it does.

- reduced by a larger NHS capacity;
- increased by a higher marginal tax rate;
- increased by a higher lump sum tax; and
- increased by a higher out-of-pocket payment to private physicians.

This is for a mixed system. If we have a pure NHS system, the waiting time has no longer a role as an allocative device and thus falls to zero. The model is thus discontinuous in the waiting time: with a small parallel private sector, the waiting time is strictly positive, but it falls to zero when health care is exclusively provided by the NHS.

5 Welfare evaluation

The government disposes of the following policy instruments: K_N (or equivalently $\widehat{a}$), y, s_N, σ, t and T, some of them being linked by the equilibrium conditions (14) and (15). In the remainder of the paper, we want to characterise the optimal selection of these instruments. We take as the criterion for evaluation an additive social welfare function that is defined as follows:

$$SW = \int_{\underline{a}}^{\widehat{a}} \psi\left(r(\overline{h} - \delta w) + v\big((1-t)a, -T\big)\right) \mathrm{d}F(a) + \int_{\widehat{a}}^{\overline{a}} \psi\left(r(\overline{h}) + v\big((1-t)a, -T-q\big)\right) \mathrm{d}F(a), \tag{21}$$

with $\psi(\cdot)$ being a concave transformation of individual utility functions: $\psi' > 0, \psi'' \leq 0$. We close the model by including a budget constraint for the government. It states that labour income tax revenue should cover the wage bill of the NHS, the subsidies to private health care, and possibly some exogenous level of government expenditure R:

$$\int_{\underline{a}}^{\widehat{a}} taL\big((1-t)a,-T\big)\mathrm{d}F(a) + \int_{\widehat{a}}^{\overline{a}} taL\big((1-t)a,-T-(1-\sigma)\pi\big)\mathrm{d}F(a) + T \;\geq\; yM_N + \sigma\pi[1-F(\widehat{a})] + R. \tag{22}$$

The problem of a social planner is then to maximise social welfare by choosing appropriately the above policy instruments under the government's budget constraint, equilibrium conditions (14) and (15) in the physician market, and condition (16) that determines the equilibrium waiting time.

To analyse this problem in Section 6, we first consider the case where we impose the NHS contract to be production efficient. As shown later, production efficiency is optimal when the supply elasticity of physicians ε_U^M is infinite. Next, we will relax this assumption in Section 7 and allow for a finite supply elasticity and so for the possibility that the government exerts some monopsony power in the physician market.

6 Optimal NHS size under a production-efficient NHS contract

As said in the introduction, production efficiency in providing health care requires that the marginal reservation wage of physicians be equated across the two sectors. In this section, the NHS contract is chosen so as to satisfy production efficiency. In a system where both sectors coexist, we have from (9) and (10) that $U(y, 1 - s_N) = V(\pi)$. As illustrated by

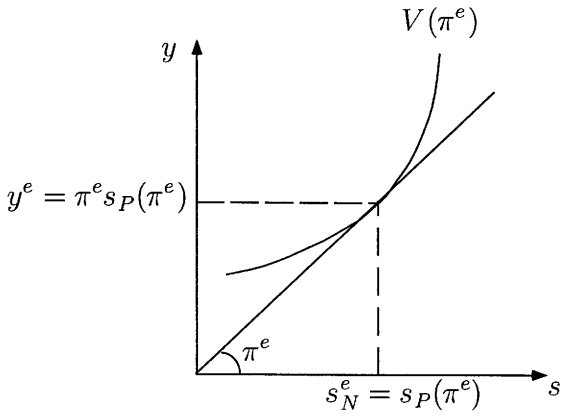

Figure 1: Work load and income for physicians under production efficiency

Figure 1, production efficiency then requires that $s_N^e = s_P(\pi^e)$ and $y^e = \pi^e\, s_P(\pi^e)$, where π^e denotes the market clearing consultation fee that satisfies:

$$M\big(V(\pi^e)\big)\; s_P(\pi^e) \equiv F(\overline{a}) = 1. \tag{23}$$

Therefore, with the NHS contract satisfying production efficiency, the equilibrium fee, π^e, does not depend upon the policy instruments controlled by the government, and the terms of the NHS contract, y and s_N, can be determined in a straightforward way, which simplifies the analysis.

As pointed out in Section 3, in this production efficiency case the market shares of the two sectors are determined through the choice by the government of the capacity of the NHS or equivalently $\widehat{a}$ (since $F(\widehat{a}) = K_N$). The marginal patient resorting to the NHS must be indifferent between receiving care there and getting private treatment. And for a given out-of-pocket payment $q = (1-\sigma)\,\pi^e$, it is the waiting time that will ensure indifference. In what follows, we regard $\widehat{a}, t, T$,and σ as the decision variables of the government.

The Lagrangian function of the maximisation problem then becomes:

$$\begin{aligned}
\mathcal{L} = & \int_{\underline{a}}^{\widehat{a}} \psi\big[r(\overline{h} - \delta w^e) + v\big((1-t)a, -T\big)\big]\mathrm{d}F(a) \\
& + \int_{\widehat{a}}^{\overline{a}} \psi\big[r(\overline{h}) + v\big((1-t)a, -T - (1-\sigma)\,\pi^e\big)\big]\mathrm{d}F(a) \\
& + \lambda\Bigg\{\int_{\underline{a}}^{\widehat{a}} taL\big((1-t)a, -T\big)\mathrm{d}F(a) \\
& + \int_{\widehat{a}}^{\overline{a}} taL\big((1-t)a, -T - (1-\sigma)\,\pi^e\big)\mathrm{d}F(a) \\
& + T - F(\widehat{a})\frac{y^e}{s_P(\pi^e)} - \sigma\pi^e[1 - F(\hat{a})] - R\Bigg\}
\end{aligned} \tag{24}$$

where w^e is obtained from

$$r(\overline{h} - \delta w^e) + v\big((1-t)\widehat{a}, -T\big) = r(\overline{h}) + v\big((1-t)\widehat{a}, -T - (1-\sigma)\,\pi^e\big).$$

Using (17), the derivative of this Lagrangian w.r.t. $\widehat{a}$ is given by

$$\frac{d\mathcal{L}}{d\widehat{a}} = (1-t)\left(\widehat{\alpha}_P\widehat{L}_P - \widehat{\alpha}_N\widehat{L}_N\right)E[\psi'(\cdot)|a \leq \widehat{a}]\,F(\widehat{a}) \\ -\lambda\{t\widehat{a}(\widehat{L}_P - \widehat{L}_N) + (1-\sigma)\pi^e\}f(\widehat{a}). \qquad (25)$$

The first term is the effect on the welfare of the NHS patients of an increase in $\widehat{a}$ through a reduction in the waiting time. There are $F(\widehat{a})$ patients enjoying this rise of welfare. The second term stands for the budgetary implications of the increase in $\hat{a}$. These consists of (i) a reduction in income tax revenue because the labour supply of the patients switching sector is negatively affected, and (ii) a net increase in the health care budget (each new NHS patient receives a treatment that costs π^e, but saves the government refunding $\sigma\pi^e$). There are $f(\widehat{a})$ marginal patients switching sector.

Evaluating expression (25) for $a = \underline{a}$ leaves us with no benefits and only budgetary costs. Introducing a small NHS sector is therefore welfare deteriorating. We summarise this as

Proposition 1 *In the production efficiency case, the introduction of a small NHS sector is harmful for social welfare.*

As expression (25) indicates, this result is explained by the number of NHS patients, $F(\widehat{a})$, who benefit from the fall in waiting time when the size of the NHS is increased, relative to the number of patients, $f(\widehat{a})$, 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. The same reasoning also explains why when the size of the NHS is large enough, the social benefit of an increases in the NHS size can dominate its social cost. 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 with the budgetary implications. Setting therefore (25) to zero and rearranging

gives us

$$\frac{1}{\lambda}\frac{\widehat{a}(1-t)(\widehat{\alpha}_P\widehat{L}_P-\widehat{\alpha}_N\widehat{L}_N)\,E[\psi'(\cdot)|a\leq\widehat{a}]}{t\widehat{a}(\widehat{L}_P-\widehat{L}_N)+(1-\sigma)\pi^e}=\frac{\widehat{f}\widehat{a}}{\widehat{F}}. \tag{26}$$

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

Our purpose is now to construct examples that show that it can be optimal to have an NHS sector if its size is sufficiently large. In these examples, we adopt the following specification of the utility function: $\log c - L/\gamma$. If $(1-t)a \geq m/\gamma$, an agent of ability a will want to participate in the labour market, and the optimal levels of his consumption, labour supply and non-health indirect utility are respectively given by $c=\gamma(1-t)a$, $L=\gamma-m/(1-t)a$, and $v=\log((1-t)a\gamma)-1+m/\gamma(1-t)a$. Recall that m is equal to either $-T$ for NHS patients or $-T-q$ for private ones.

Let us assume for the moment that the tax policy does not drive any NHS patient out of the labour force. Replacing the optimal labour supplies in (22) by $\gamma+T/(1-t)a$ for NHS patients and $\gamma+(T+q)/(1-t)a$ for private patients, the government budget constraint can then be rewritten as

$$T=(1-t)[R+\pi^e F(\widehat{a})-t\,E(a)\gamma]-(t-\sigma)\pi^e[1-F(\widehat{a})] \tag{27}$$

where $E(a)\equiv\int_{\underline{a}}^{\overline{a}} a\,\mathrm{d}F(a)$. Assuming a utilitarian social welfare function ($\psi'\equiv 1$), condition (26) can then be filled in to give:[4]

$$\frac{a^H}{\widehat{a}}=\frac{\widehat{f}\widehat{a}}{\widehat{F}} \tag{28}$$

[4]The relevant individual variables are as follows: $\widehat{L}_P=\gamma+(T+(1-\sigma)\pi^e)/(1-t)\,\widehat{a}$, $\widehat{L}_N=\gamma+T/(1-t)\,\widehat{a}$ and $\widehat{\alpha}_P=\widehat{\alpha}_N=1/\gamma(1-t)\,\widehat{a}$. Since $\mathrm{d}T/\mathrm{d}R=1-t$ (cf (27)), the social marginal utility of income, λ, is: $\int_{\underline{a}}^{\overline{a}}(1/\gamma a)\mathrm{d}F(a)$, or just $1/\gamma a_H$.

where a^H is defined as the harmonic mean of the ability distribution, that is

$$a^H \equiv \left(\int_{\underline{a}}^{\overline{a}} \frac{1}{a} \mathrm{d}F(a) \right)^{-1}.$$

Therefore with this specification of the utility function the tax and subsidy parameters no longer enter the *lhs* of first-order condition (26): the optimal size of the NHS depends only on the shape of the ability distribution. Let us now take a closer look at two examples of distribution functions:

Example 1 *Uniform ability distribution.*

If $F(a) = \frac{a - \underline{a}}{\overline{a} - \underline{a}}$ on $[\underline{a}, \overline{a}]$, then $a^H = \frac{\overline{a} - \underline{a}}{\log \overline{a} - \log \underline{a}}$ and (28) becomes

$$\frac{1}{\widehat{a}} \frac{\overline{a} - \underline{a}}{\log \overline{a} - \log \underline{a}} = \frac{\widehat{a}}{\widehat{a} - \underline{a}}.$$

This expression may be written as a second degree equation in $\widehat{a}/\underline{a}$:

$$\left(\frac{\widehat{a}}{\underline{a}} \right)^2 + \left(1 - \frac{\widehat{a}}{\underline{a}} \right) \frac{\frac{\overline{a}}{\underline{a}} - 1}{\log \frac{\overline{a}}{\underline{a}}} = 0$$

which has real roots if and only if $\log \frac{\overline{a}}{\underline{a}} \leq \frac{\overline{a}}{\underline{a}} - 1$, or $\overline{a} \geq 10.35\underline{a}$. In that case there are two positive real roots, and since social welfare is initially decreasing in $\widehat{a}$, it is the second root that is a local maximum.

Figure 2 gives an example where this local maximum is also the global one.[5] It makes use of the following parameters: $\underline{a} = 10, \overline{a} = 160, \gamma = .25, \pi^e = 5$ and $R = 10$. With σ set equal to 0, the optimal value for $\widehat{a}$ is 40.78, meaning that the NHS serves almost 25% of the population. The optimal tax policy is $t = .375$ and $T = .428$. Under this policy nobody withdraws from the labour market.

[5] As mentioned at the end of Section 4, the waiting time discontinuously falls to zero when the mixed system becomes a pure private one ($\hat{a} = \overline{a} = 160$). Furthermore as the figure shows, social welfare with a pure NHS system is at the same level as with a pure private system ($\hat{a} = 0$). The latter result holds only if $s_N = s_P(\pi^e)$ as it is assumed in the present section.

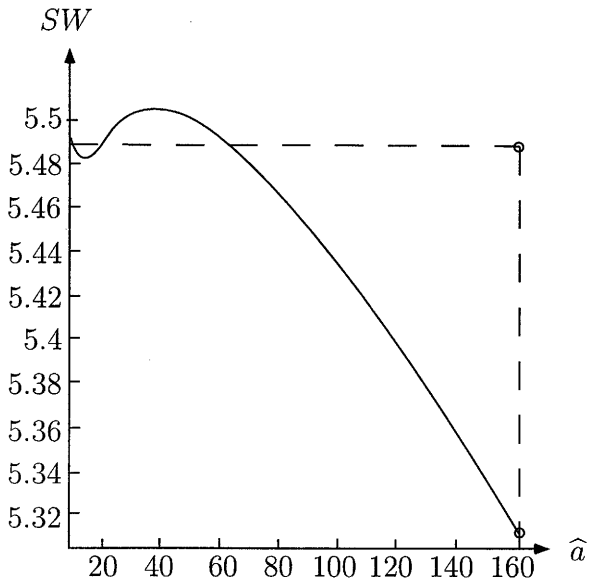

Figure 2: Social welfare as a function of $\widehat{a}$ (the uniform case)

Example 2 *Log-uniform ability distribution.*

With $F(a) = \dfrac{\log a - \log \underline{a}}{\log \overline{a} - \log \underline{a}}$, the ability distribution is skewed to the right on $[\underline{a}, \overline{a}]$ and a^H now equals $\left(\dfrac{\log \overline{a} - \log \underline{a}}{\overline{a} - \underline{a}}\right) \overline{a}\underline{a}$. The $\hat{a}$-rule may now be written as

$$\frac{\frac{\overline{a}}{\underline{a}} \log \frac{\overline{a}}{\underline{a}}}{\frac{\overline{a}}{\underline{a}} - 1} = \frac{\frac{\hat{a}}{\underline{a}}}{\log \frac{\hat{a}}{\underline{a}}}$$

which has more than one real root for $\hat{a}/\underline{a}$ if $\overline{a} \geq 12.2\underline{a}$. Again, the spread between the lowest and largest ability level must be sufficiently high for a mixed health care system to be optimal.

Figure 3 provides an example with an interior global maximum for a log-uniform ability distribution. The parameters used in this example are: $\underline{a} = 10, \overline{a} = 300, \gamma = .25, \pi^e = 5$, and $R = 20$. Average ability is then 85.26. Setting again $\sigma = 0$, the optimal value for $\widehat{a}$ is 67.22 with

$F(\hat{a}) = .56$ and the optimal tax policy is $t = .603$ and $T = 2.615$. Once again, under this policy nobody withdraws from the labour market.

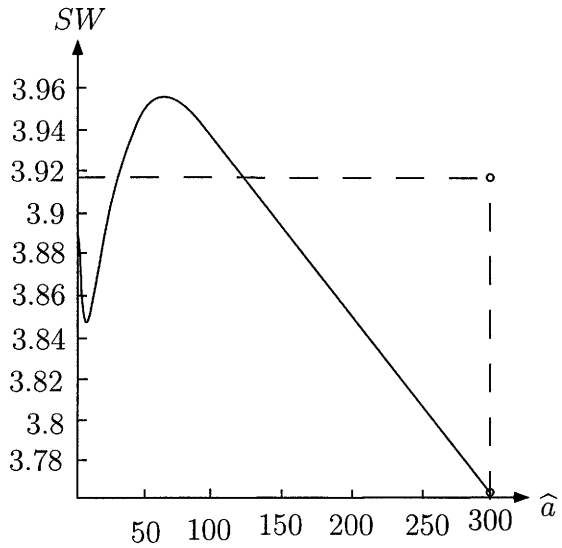

Figure 3: Social welfare as a function of $\hat{a}$ (the log-uniform case)

These two examples point to a more general result, namely that social welfare is not a nicely concave function in the size of the NHS. The fact that the introduction of a small NHS reduces welfare does not preclude that a larger public health care system may improve upon a purely private solution. But the necessary condition for this to occur is that there is a sufficiently unequal distribution of earnings capacities.

We summarize our findings in the following proposition:

Proposition 2 *A mixed health care system can be socially optimal. The likelihood for this to happen rises with increased dispersion of the ability distribution.*

Our result about the desirability of an NHS is akin to the one obtained by Besley and Coate (1991). Like in their model, individuals

consume here one unit of a private good, which they can either acquire free of charge by resorting to the public sector or buy at some price in the private sector. If the quality level in the public sector is lower than in the private market (due to the waiting time in our model), some individuals will prefer to pay for a higher-quality good in the private sector. 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 the whole population. The same argument can be used here to explain why a mixed health care system can be optimal when the NHS is financed by a linear income tax that is imposed to all citizens irrespective of the sector they resort to for receiving care.[6]

To characterise the optimal linear tax policy, we proceed in the standard way, substracting the first-order condition w.r.t. T multiplied by $E(aL)$ from the first-order condition w.r.t. t. After rearranging the result of this operation, we obtain:

$$t = \frac{-\text{cov}(\beta, aL)}{E(a\dfrac{\partial L^c}{\partial \omega}a)} + [E(aL)(\widehat{\alpha}_P - \widehat{\alpha}_N) - \widehat{a}(\widehat{\alpha}_P \widehat{L}_P - \widehat{\alpha}_N \widehat{L}_N)] \frac{F(\widehat{a})\, E[\psi'(\cdot)|a \leq \widehat{a}]}{\lambda\, \text{E}(a\dfrac{\partial L^c}{\partial \omega}a)} \qquad (29)$$

where $\partial L^c/\partial\omega$ is the compensated wage effect on labour supply, and $\beta \equiv \psi'(\cdot)\,\alpha/\lambda + ta\partial L/\partial m$ is Diamond's net social marginal utility of a one Euro income transfer to agent a.

Expression (29) is a modified rule à la Sheshinski (1972) for the optimal marginal income tax rate. The first *rhs* term is the standard term

[6]Hoel and Sæther (2000) derive a similar result. Their setting is a partial equilibrium model of the health care sector with an exogenous marginal cost of public funds and with patients differing in their waiting time cost. They construct an example with a two class population, and find that if there is a sufficiently wide gap between the waiting costs for the two classes, it is optimal to have a positive waiting time that makes the high waitng cost class resort to the private market.

trading off equity considerations (numerator) with efficiency considerations (denominator). The second *rhs* term is new and has to do with the effect on the waiting time of a change in the marginal tax rate; from (18) and (19) the square bracket is in fact $(\mathrm{d}r/\mathrm{d}t) - E(aL)(\mathrm{d}r/\mathrm{d}T)$. It is difficult to say a priori how an increase in the marginal tax rate, accompanied by a decrease in the lump sum rate to balance the budget, will affect the waiting time. What we can say is that[7]

$$\begin{array}{rll} \dfrac{\mathrm{d}r}{\mathrm{d}t} - E(aL)\dfrac{\mathrm{d}r}{\mathrm{d}T} & < 0 & \text{if } \widehat{a}\widehat{L}_N < E(al) < \widehat{a}\widehat{L}_P, \\ & \lesseqgtr 0 & \widehat{a}\widehat{L}_N < \widehat{a}\widehat{L}_P < E(al), \\ & < 0 & E(al) < \widehat{a}\widehat{L}_N < \widehat{a}\widehat{L}_P. \end{array}$$

Therefore if the income level earned by the marginal patient resorting to the private sector lies above the average income level, then a small tax reform boosts the waiting time. This is in itself a reason for setting the marginal tax rate below the standard level.

In the standard model of income redistribution, the lump-sum tax T is chosen to set the average value of Diamond's β equal to 1. In the present model, this rule is modified because the lump-sum tax has an impact on the equilibrium waiting time:

$$E\beta = 1 - \frac{1}{\lambda}(\widehat{a}_P - \widehat{a}_N)\; E[\psi'(\cdot)|a \leq \widehat{a}]\, F(\widehat{a}). \tag{30}$$

The benefit for the government's budget of a marginal increase in T is reduced by the undesirable consequence of a longer waiting time for all NHS patients.

The first-order condition with respect to σ is

$$E(\beta|a \geq \widehat{a}) = 1 - \frac{\widehat{\alpha}_P}{\lambda}\, E[\psi'(\cdot)|a \leq \widehat{a}]\frac{F(\widehat{a})}{1 - F(\widehat{a})} \tag{31}$$

where the *lhs* is the average β for individuals with $a > \widehat{a}$. The refund rate σ acts as a lump-sum instrument for all patients visiting a private doctor. However, when private visits are made more attractive through

[7]We make use of the facts that $\widehat{\alpha}_P - \widehat{\alpha}_N > 0$ and that $\widehat{L}_P - \widehat{L}_N > 0$.

a larger refund, the equilibrating waiting time will fall and this benefits the part of the population that visits an NHS physician. This explains why the direct cost of a one Euro refund on the *rhs* of (31) is corrected downwards. This is a formalisation of an argument used in public debates in favour of a parallel private health care system: it relieves pressure on the public system and allows lower income classes being treated there to get faster treatment. Inspection of the second *rhs* term in (31) shows that this argument is stronger, the bigger the NHS (for the obvious reason that more people benefit from the shorter waiting time, *and* because the impact on the equilibrium waiting time is then stronger) and the higher the social evaluation of the living standard of lower income classes.[8]

Condition (31) can be combined with (30) to yield:

$$E(\beta|a \leq \widehat{a}) = 1 + \frac{\widehat{\alpha}_N}{\lambda} E[\psi'(\cdot)|a \leq \widehat{a}]. \tag{32}$$

This expression would be the first-order condition with respect to a copayment for NHS consultations if the government were to charge one. This copayment would act as a lum-sum tax on NHS patients, the marginal net social cost of which appears on the *lhs* of the expression. There would be also benefits associated with the copayment: they consist of additional government revenue and reduction in waiting time for NHS patients, that are accounted for the by the first and second terms respectively on the *rhs*. It is however worthwhile recognizing that with both σ and T available as policy instruments, a copayment for NHS services becomes a redundant instrument.

7 The optimal NHS contract

In the previous section, we assumed that the government offers NHS physicians a contract with a work load identical to the one chosen by private physicians. This achieved production efficiency in the provision of consultations, i.e. NHS and private doctors have the same marginal reservation fee for a consultation, and as Figure 1 illustrates, this common reservation fee is equal to y/s_N. We now look at the desirability of

[8]In our examples, both the marginal utility of income, the marginal cost of public funds, and the income effects on labour supply are constant so that condition (31) can only hold by accident in which case σ is indeterminate.

offering a different contract. Of course, the utility that physicians derive from an NHS contract must equal that of a private practice. But by manipulating the terms of the contract and departing from production efficiency, the government can influence this common equilibrium utility level and in this way reduce the cost of running the NHS. The reason is that the government has monopsony power on the market for physicians.

We proceed in two steps. First, we investigate the optimal choice of the NHS contract when the government is only concerned with minimizing public expenditure on health care. This allows us to better understand the optimal contract choice when the government instead maximises social welfare, an issue we turn to in the second step.

Let us therefore assume in a first stage that the government only cares about public expenditure on health care, subject to the constraint that $F(\widehat{a})$ patients are treated in the NHS. Expenditure on health care consists of the NHS wage bill, $M_N\, y$, and subsidies to private health care, $\sigma\pi(y, s_N)[1-F(\widehat{a})]$. We start from eqs (14) and (15) that give $y(s_N, K_N)$ as explained at the end of Section 3, and find in the appendix that

$$\frac{\partial y}{\partial s_N} = MRF - \frac{M_N}{M}\frac{y}{s_N}\frac{1}{\varepsilon_y^M + \dfrac{M_P}{M}\varepsilon_\pi^{s_P}\varepsilon_y^\pi} \tag{33}$$

where $MRF \equiv U_\ell/U_y = -U_s/U_y$ is an NHS physician's marginal reservation fee for a consultation and ε_y^π and ε_y^M are elasticities that are defined by $\varepsilon_y^\pi = (\partial\pi(y, s_n)/\partial y)(y/\pi) = (U_y^N/s_P U_y^P)(y/\pi)$ and $\varepsilon_y^M = \varepsilon_U^M\, \varepsilon_y^U$.

Since $K_N = F(\widehat{a})$, the expenditure minimisation problem is

$$\min_{s_N} \Omega[s_N, \sigma, K_N] \equiv y(s_N, K_N)\frac{K_N}{s_N} + \sigma\pi(y(s_N, K_N), s_N)\,[1 - K_N].$$

In the appendix, we show that the first order condition that s_N must satisfy to minimize expenditure can be written as:

$$\frac{MRF - \frac{y}{s_N}}{\frac{y}{s_N}} = \frac{M_N}{M}\frac{1}{\varepsilon_y^M + \dfrac{M_P}{M}\varepsilon_\pi^{s_P}\varepsilon_y^\pi}\left(1 + \varepsilon_y^\pi\frac{\sigma\pi}{\dfrac{y}{s_N}}\frac{1 - K_N}{K_N}\right). \tag{34}$$

We therefore depart from the standard rule that the work load of NHS physicians be chosen so as to equate their marginal reservation fee for a consultation with the cost of an additional consultation (i.e. y/s_N) (which is in our setting equivalent to satisfy production efficiency). Condition (34) provides the optimal markup between these two quantities. It is noteworthy that if the elasticity of doctor supply ε_U^M, and therefore ε_y^M, were infinite, this markup would be nil. In this case, production efficiency would be optimal.

Condition (34) is an inverse elasticity rule where three elasticities are important: the elasticity of total doctor supply w.r.t. the NHS salary, ε_y^M, the elasticity of private consultations supplied w.r.t. the consultation fee, $\varepsilon_\pi^{s_P}$, and the elasticity of the equilibrium consultation fee w.r.t. the NHS salary, ε_y^π. Assuming for a moment that $\varepsilon_\pi^{s_P} \equiv 0$ (e.g. Cobb-Douglas physician preferences) and $\sigma = 0$ (no refunding of private care), it is easy to see why the government wants to give NHS physicians a higher work load than the one their private colleagues choose ($s_N > s_P$). If the government's objective is to care for $F(\widehat{a})$ patients through the NHS, the higher work load per NHS physician means that fewer such physicians need to be hired. If ε_U^M is finite, this allows the NHS to reduce its overall wage bill since the reservation utility of the marginal physician ($\widetilde{U}$) falls. So the monopsony power of the government on the market for doctors gives an incentive for the government to depart from the standard rule, and this will be more pronounced the higher the share of NHS doctors and the smaller the elasticity of the supply of doctors. The smaller this elasticity, the more 'stingy' the government can be in remunerating its personnel for a higher work load.

Let us now assume that $\varepsilon_\pi^{s_P} \neq 0$, still keeping $\sigma = 0$. The elasticity $\varepsilon_\pi^{s_P}$ can be either positive or negative according to whether the substitution effect dominates the income effect or not. If positive, the second term in the denominator of the second factor on the *rhs* of (34) pushes the markup and so the work load s_N downwards. This can be explained as follows. As mentioned earlier, a rise in s_N makes $\widetilde{U}$ fall; this in turn lowers π and therefore s_P since $\varepsilon_\pi^{s_P} > 0$. However, a lower work load in the private sector means that more private physicians are needed to take care of the $F(\overline{a}) - F(\widehat{a})$ patients there, and this acts on M and so $\widetilde{U}$ in

opposite direction to the rise in s_N. By contrast, if $\varepsilon_\pi^{s_P} < 0$, the effect on the required number of physicians of an increase in s_N is reinforced by that of an increase in s_P.

Also the presence of a health care subsidy at rate σ pushes the work load upwards. The reason here is that a tougher work load in the NHS has a negative impact on the equilibrium fee for private consultations and, to the extent these are subsidised, on government spending.

We may thus conclude that, in the short run, when the supply of doctors is rather inelastic, the government should demand a heavier work load from NHS physicians than the one private doctors perform. This is in stark contrast with what we observe in countries with an NHS: there public doctors work less than their private colleagues, but often the former also have a private practice on the side. This will not happen when salaries are set according to (34) because at the margin, an NHS doctor reservation fee exceeds market fee π.

How does minimal public expenditure vary when the NHS capacity $K_N = F(\widehat{a})$ increases? Plugging the optimal NHS work load back into the expenditure function gives the minimal expenditure as a function of the size of the NHS:

$$\Theta(\sigma, K_N) \equiv \min_{s_N} \Omega[s_N, \sigma, K_N].$$

It is shown in the appendix that

$$\frac{\partial \Theta}{\partial K_N} = (1 - \frac{s_N}{s_P})(MRF - \frac{y}{s_N}) + \frac{y}{s_N} - \sigma\pi. \tag{35}$$

First, notice that for $s_N = s_P$, this expression reduces to $(1-\sigma)\,\pi^e$ so that public expenditure is proportional to K_N. But as we just observed, it will only be optimal under special circumstances to have the same number of consultations per physician in the two sectors. The government can in general reduce expenditure by choosing s_N above s_P. And this means that public expenditure will rise less than proportionally to K_N, and that for small values of K_N the relationship will be concave. In fact, since the first round bracket term in (35) is negative, public expenditure on health

care may even *fall* as the NHS treats more patients. This will occur when the markup in (34) is very large, which in turn is optimal when e.g. the elasticity of the supply of physicians is very small. Figure 4 illustrates minimal expenditure as a function of the number of patients treated in the NHS in an example for which ε_U^M is equal to 2.[9]

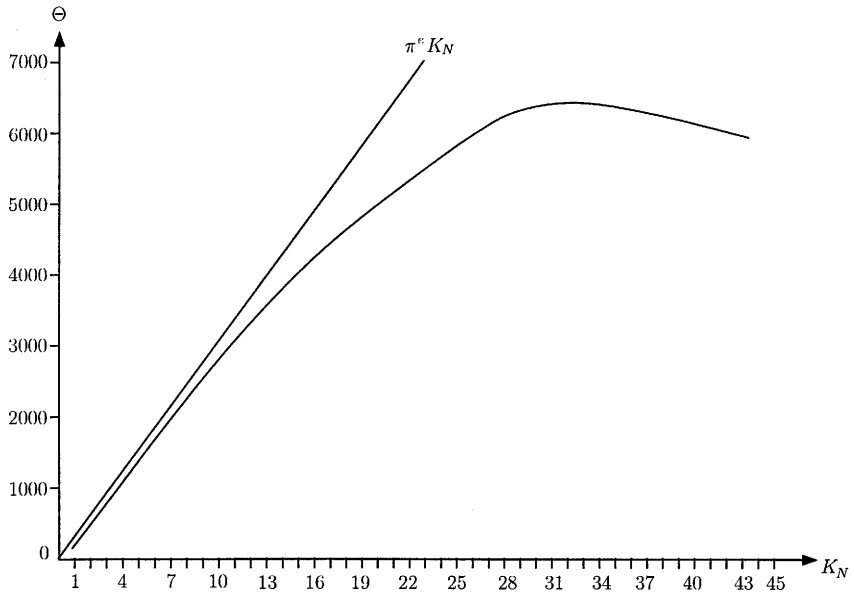

Figure 4: Minimal public expenditure on health care with and without the use of monopsony power for $\varepsilon_y^M = 1/2$ (the curve and the straight line respectively)

In this figure the curve and the straight line exhibit how public health expenditure varies when the government does and does not exploit respectively its monopsony power in the physician market. According to

[9]The figure is drawn under the following assumptions: $U = c^{\frac{1}{4}}\ell^{\frac{3}{4}}$, a physician's time endowment is 200, $M(U) = 900\left(\frac{U}{1500}\right)^{\gamma}$, and there are 45000 citizens. Therefore we depart from our above normalization rules concerning a physician's time endowment and population size. Then, under a pure private market equilibrium (or, alternatively, when $s_N = s_P$), $\pi^e = 30000$, $s_P(\pi^e) = 50$, $V(\pi^e) = 1500$ and $M = 900$. Figure 4 is for $\gamma = \varepsilon_U^M = 2$. Since $\varepsilon_y^U = 1/4$, this gives $\varepsilon_y^M = 1/2$. The units on the vertical axis are in 100 Euros and on the horizontal axis in 1000 citizens.

Figure 4, the minimal public expenditure on health care, $\Theta(\sigma, K_N)$, will not be a convex function of the number of NHS patients. This means that we have identified a second reason in addition to the one exhibited in Section 6 for which social welfare is not necessarily a concave function of the NHS size.

Our assumption in the first step that the government selects the terms of the NHS contract to minimise public expenditure on health care, for a given size of the NHS, is however *ad hoc* because this contract will determine the equilibrium consultation fee of private physicians and thus affect the welfare of private patients and of NHS patients through the waiting time. We therefore turn in the second step to the choice of s_N when social welfare is to be maximised. If anything, the concern for patient welfare reinforces the argument for pushing s_N upwards, because that will reduce the equilibrium consultation fee and the waiting time. There is however a third effect which we have ignored so far: the reduction in the consultation fee will have a negative income effect on private patients' labour supply and thus on income tax revenue.

The three side effects just mentioned – on waiting time for NHS patients, on disposable income of private patients, and on tax revenue – are also present when the government increases the subsidy rate σ. Optimizing the Lagrangian (24) with respect to s_N and σ, we show in the appendix that the following inverse elasticity rule holds:

$$\frac{MRF - \dfrac{y}{s_N}}{\dfrac{y}{s_N}} = \frac{M_N}{M} \frac{1}{(\varepsilon_y^M + \dfrac{M_P}{M} \varepsilon_\pi^{s_P} \varepsilon_y^\pi)} \left(1 + \varepsilon_y^\pi \frac{\pi}{\dfrac{y}{s_N}} \frac{1 - K_N}{K_N} \right). \qquad (36)$$

It is according to this rule that the government should set the terms of the NHS contract if it is concerned with social welfare maximisation. Three things are worth noting about this equation. First, except for the non-appearance of σ on the *rhs*, it is exactly the same as the inverse elasticity rule for expenditure minimisation. Second, this markup rule is obtained without using any argument about the optimal choice of $\widehat{a}$, and it therefore holds for any positive size of the NHS system. Finally, it is derived without using the optimality conditions for the tax policy.

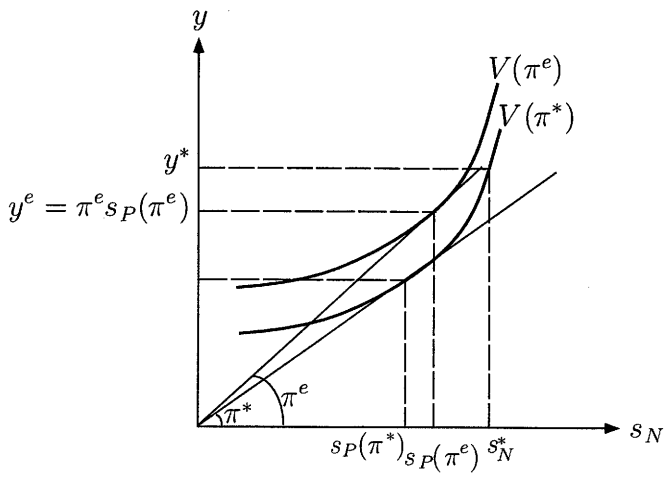

Figure 5: Work load and income for physicians when the government chooses NHS contract (y^*, s_N^*) according to (36)

The optimal contract is illustrated in Figure 5 as (y^*, s_N^*), and we summarise our discussion of this contract in

Proposition 3 *The optimal NHS contract should specify a work load and a salary for physicians that satisfies the inverse elasticity rule (36). According to this rule, the marginal reservation fee for an NHS physician should exceed the average fee, which in turn should exceed the equilibrium fee for private consultations. Equivalently, an NHS physician should perform more consultations than a private physician. This optimal contract takes away any incentive for NHS physicians to top up their income by private consultations.*

Finally, we mention that we have treated physicians as absentees by not including their welfare into the social welfare function. It is clear that if we had done so, the reasons for an upward distortion in the work load of NHS doctors would have been mitigated.

8 Concluding remarks

The organisation of the health care sector is a theme for public debate in many countries. One of the main issues in this discussion is who should be responsible for the provision of health care. Should it be provided privately, should the production be in public hands, or should patients have both options available? We have been looking at this issue in a setting where people differ in their earnings capacity and display an inelastic demand for health care. We assumed that illness reduces a person's living standard when treatment is postponed, but does not affect the person's earnings capacity. Treatment can be obtained either in a competitive private sector or in the NHS where it is provided free of charge but after some (endogenous) waiting time. The health care sector is in equilibrium when the fee for private consultations is such that no physician has an incentive to switch place of work, and in addition waiting time in the NHS is such that no patients wants to switch health care provider. This equilibrium, and in particular the size of each sector, is governed by three public policies: the income tax system that should ensure a balanced budget, the subsidisation of private health care, and the terms of the contract offered to NHS physicians. Assuming that the government has welfaristic objectives, we analysed the optimal policies. Our analysis highlighted three findings. First, a small public health care sector gives a lower social welfare level than a pure private system. Second, a mixed system with a sufficiently large NHS may improve upon a pure private system. A necessary condition for this to happen is that the distribution of earnings capacities in society is sufficiently wide. And finally, the doctors working there should be given a contract that specifies a heavier work load than the one their private colleagues choose.This is the way the government can exert its monopsony power on the market for physicians to improve social welfare.

Our assumptions also delineate the limitations of our model. The actual market for private health care is not likely to be perfectly competitive. Physicians may exert monopoly power in setting the consultation fee because it is costly for their patients to switch to a new physician who is not informed about the patient's health history. Satterthwaite (1985) has argued that a monopolistic structure for the health care market is

more reasonable for many countries.

A second issue is the cost of having to wait for a treatment. We have assumed that waiting incurs a direct cost because it inflicts discomfort on the patient. But on top of that, a reduced health status will have consequences for one's earnings capacity. Even though a social insurance policy can protect the individual patient from the loss in income, it cannot protect the productive possibilities in the economy as a whole.

References

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

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

Brekke, Kurt and Lars Sørgard, (2000). Where are all the physicians? Private versus public health care. Mimeo, University of Bergen.

Cullis, John, Philip Jones and Carol Propper, (1999). Waiting lists and medical treatment: analysis and policies, mimeo, University of Bristol (forthcoming in *Handbook of Health Economics*).

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 Oslo.

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.

Lindsay, Cotton and Bernard Feigenbaum, (1984). Rationing by waiting lists. *American Economic Review* **74**, 404–417.

Martin, Stephen and Peter Smith, (1999). Rationing by waiting lists: an empirical investigation. *Journal of Public Economics* **71**, 141–164.

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

Propper, Carol and Katherine Green, (2000). A larger role for the private sector in health care? A review of the arguments. Mimeo, University of Bristol.

Satterthwaite, Mark, (1985). Competition and equilibrium as a driving force in the health services sector. In: Robert Inman (ed) *Managing the service economy* (Cambridge: Cambridge University Press).

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

Worthington, David (1987). Queueing models for hospital waiting lists. *Journal of the Operations Research Society* **38**, 413–22.

Appendix

A. Derivation of (33)

Eliminating M_P between (14) and (15) gives

$$K_N + \left[M(U(y, 1 - s_N)) - \frac{K_N}{s_N}\right] s_P(\pi(y, s_N)) = F(\overline{a}). \qquad \text{(A.1)}$$

Differentiating this expression and accounting for the fact that from (11), $\partial\pi/\partial s_N = (-U_\ell/U_y)\partial\pi/\partial y$ yields:

$$dy\left[M'U_y s_P + M_p \frac{\partial s_P}{\partial \pi}\frac{\pi}{y}\right] + ds_N\left[-M'U_\ell s_P - M_P \frac{\partial s_P}{\partial \pi}\frac{U_\ell}{U_y}\frac{\partial \pi}{\partial y} + \frac{K_N s_P}{s_N^2} s_P\right] = 0, \qquad \text{(A.2)}$$

from which (33) can be obtained after some easy manipulations.

B. Derivation of (34)

The first-order condition that s_N must satisfy for a minimum of public expenditure Ω is obtained by equating to zero:

$$\frac{\partial \Omega}{\partial s_N} = \left(\frac{\partial y}{\partial s_N} - \frac{y}{s_N}\right)\frac{K_N}{s_N} + \sigma(1 - K_N)\left(\frac{\partial \pi}{\partial y}\frac{\partial y}{\partial s_N} - \frac{\partial \pi}{\partial s_N}\right). \qquad \text{(A.3)}$$

Using relations $\partial\pi/\partial y = U_y/V_\pi$ and $\partial\pi/\partial s_N = -U_\ell/V_\pi$, recalling that $MRF = U_\ell/U_y$ and substituting $\partial y/\partial s_N$ from (33) into this first-order condition yields relation (34) in a straightforward way.

C. Derivation of (35) and (36)

Differentiating (A.1) yields:

$$dy\left[M'U_y s_P + M_P \frac{\partial s_P}{\partial \pi}\frac{\partial \pi}{\partial y}\right] + dK_N\left[1 - \frac{s_P}{s_N}\right] = 0, \qquad \text{(A.4)}$$

which we rearrange to obtain

$$\frac{\partial y}{\partial K_N} = \frac{s_P(\pi) - s_N}{M\, s_P(\pi) s_N} \frac{1}{(\varepsilon_y^M + \frac{M_P}{M}\varepsilon_\pi^{s_P}\varepsilon_y^\pi)}. \tag{A.5}$$

Using the envelope theorem leads to:

$$\frac{\partial \Theta}{\partial K_N} = \frac{y}{s_N} - \sigma\pi + \left[M_N + \sigma\frac{\partial \pi}{\partial y}(1 - K_N)\right]\frac{\partial y}{\partial K_N}. \tag{A.6}$$

Sustituting $\partial y/\partial K_N$ from (35) into this derivative and using (34) then gives expression (36) after some manipulations.

D. Derivation of (37)

The Lagrangian in (24) can be written as:

$$\mathcal{L} = \Sigma + \lambda\Omega \tag{A.7}$$

where Σ is the sum of the government's objective and the part of its budget constraint that is related to tax revenue (multiplied by λ). Recall that Ω stands for public expenditure. Thus the derivatives of the Lagrangian with respect to s_N and σ are :

$$\frac{d\mathcal{L}}{ds_N} = \frac{\partial \Sigma}{\partial q}(1-\sigma)\left[\frac{\partial \pi}{\partial y}\frac{dy}{ds_N} + \frac{\partial \pi}{\partial s_N}\right] + \lambda\frac{\partial \Omega}{\partial s_N} = 0, \tag{A.8}$$

$$\frac{d\mathcal{L}}{d\sigma} = -\frac{\partial \Sigma}{\partial q}\pi + \lambda\frac{\partial \Omega}{\partial \sigma} = 0. \tag{A.9}$$

Combining (A.8) and (A.9) yields

$$(1-\sigma)\frac{\partial \Omega}{\partial \sigma}\left[\frac{\partial \pi}{\partial y}\frac{dy}{ds_N} + \frac{\partial \pi}{\partial s_N}\right] + \pi\frac{\partial \Omega}{\partial s_N} = 0. \tag{A.10}$$

Using (A.3) and the fact that $\partial\Omega/\partial\sigma = \pi(1 - K_N)$, this expression can be rewritten as:

$$(1-\sigma)(1+K_N)\left[\frac{\partial \pi}{\partial y}\frac{dy}{ds_N} + \frac{\partial \pi}{\partial s_N}\right] + \left(\frac{\partial y}{\partial s_N} - \frac{y}{s_N}\right)$$

$$+\sigma(1-K_N)\left[\frac{\partial \pi}{\partial y}\frac{\partial y}{\partial s_N} - \frac{\partial \pi}{\partial s_N}\right] = 0 \tag{A.11}$$

which simplifies to

$$\left(\frac{\partial y}{\partial s_N} - \frac{y}{s_N}\right) + (1 - K_N)\left[\frac{\partial \pi}{\partial y}\frac{\partial y}{\partial s_N} - \frac{\partial \pi}{\partial s_N}\right] = 0. \tag{A.12}$$

This is identical to the *rhs* of (A.3) except for the absence of σ in front of the second term. Proceeding in the same way as (34) was obtained from (A.3) yields (37).