THE UNIVERSITY of York

Discussion Papers in Economics

No. 2001/01

Hospital Specialists' Private Practice and its Impact on the Number of NHS Patients Treated and on the Delay for Elective Surgery

by

Antonia Morga and Ana Xavier

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

Hospital specialists' private practice and its impact on the number of NHS patients treated and on the delay for elective surgery.*

Antonia Morga† and Ana Xavier‡

March 9, 2001

Abstract

This paper analyses UK NHS waiting times and waiting lists for elective surgery looking at the hospital specialists' behaviour and the conflict of interest these may face when allowed to practice privately. We look at the relationship between the government as the health care purchaser and principal of a two-tier hierarchy, and two hospital specialists, the agents, that deal with elective and emergency treatement. Specialists are organised in a separated structure, each responsible for only one type of surgery (either elective or emergency). We formalise specialists' preferences when dealing with the two activities. We see how specialists' interest in the income obtained with private practice (and altruism) affects negatively (positively) the optimal NHS numbers treated and increases the waiting time for elective surgery. Asymmetry of information also has a negative impact on the NHS leading to fewer patients treated or higher transfers paid. If remuneration is based on performance, transfers have to take private practice into account. As a result, there may be benefits from extra investment so as to improve information systems as well as seeking out instruments for nurturing more altruistic behaviour on the part of the specialists

Key words: waiting times and lists, elective surgery, hospital specialists

JEL Nos: H0, I1, L2, L5

*We would like to thank Alan Williams, Aki Tsuchiya, Peter Smith, Peter Simmons, Diane Dawson, Michael Kuhn, Andrew Street and those at the IOH group at York for their valuable comments. We also thank those attending the Oxford Young Economists Meeting, SIHCM St. Andrews, EARIE - Lausanne, 2000. All remaining errors are ours.

†Antonia Morga is funded by the E.S.R.C. research grant no. R00429824574, under the theme priority "Governance, Regulation and Accountability", and by the Dipartimento di Scienze Economiche, Universitá di Bari, Italy. Address: Department of Economics and Related Studies, University of York, York YO10 5DD, UK. Email: amcm103@york.ac.uk.

‡Ana Xavier was funded by the Fundação para a Ciência e Tecnologia, Portugal and the Centre for Health Economics, University of York. Corresponding author: LICOS - Centre for Transition Economics, Katholieke Universiteit Leuven, Deberiostraat 34, 3000 Leuven, Belgium. Email: anna.xavier@econ.kuleuven.ac.be

1 Introduction

This paper looks at the role that hospital specialists play in managing the UK National Health Service (NHS) waiting lists. We look at specialists' activity and the possible conflict of interest they face between the public and the private sector, and investigate how this affects waiting times for elective surgery.

In the NHS a patient has a first appointment with his General Practitioner (GP) after which, if the GP finds it necessary, the patient may be referred to a hospital specialist. The specialist then decides whether the patient should be admitted to hospital. If so, a patient is either admitted for emergency treatment or put on a waiting list depending on whether the waiting is critical. The patients whose names are put on a waiting list wait for non-urgent or elective surgery.

Within the UK, the NHS is the main provider of medical care. It is financed by general taxation with user charges applied to certain services (*e.g.* dental services). This implies that provision is mostly free of charge at the moment of consumption. The private sector provides an increasing amount of non-urgent surgery: around 17% of all elective procedures in the UK are provided from private sources, an increase from the 4% registered in the 50s (Beeby *et al.*, 1989; Laing's, 1994, 1999). The general public spent about £3,287 million on private health care in 1997, a two-fold increase in real terms since 1975 (Office of Health Economics, 1999). The increase in private health care expenditure is intimately related to an increase in the uptaking of private medical insurance (Propper, 1998; Besley *et al.*, 1999; Office of Health Economics, 1999, Cullis *et al.*, 2000). According to Besley *et al.* (1999), around 14-17% of the population (mainly the middle class and the rich) purchases private health care insurance. Long waits have been seen by the general population as an unsatisfactory characteristic of the NHS (Bosanquet, 1988) and appear to affect the purchasing of private insurance (Besley *et al.*, 1999). Indeed, the length of the waits or the lists have been used by insurance companies to make insurance look more attractive.

Waiting lists are a common feature of markets where the allocation of resources is achieved by a non competitive money price mechanism. It is sometimes felt that a competitive price mechanism is not the correct one in terms of equity, for ethical or political reasons, and when the good is considered a merit good. Hence, several commodities are delivered at a zero or low price at the point of consumption (Barzel, 1974). That is the case of public health care systems such as the UK NHS. It is the aim of the UK NHS that inability to pay should not deter access to medical care. Moreover, a stochastic demand coupled with constrained capacity implies that not all the patients can be admitted immediately at all times, and waiting lists are used for planning and smoothing hospital activity if hospitals are working at full capacity.

Large waiting lists, and the associated long waits for many procedures, have been a persistent feature of the UK NHS since its inception in 1948. During the first twenty five years, there were about half a million people waiting for hospital treatment. Since then, the number of people on the list has increased to around 0.7 million in the 1980s and, after

a decrease, the number started rising again to around 1 million people at the end of 2000, despite an increase in capacity (more than 60%) and an 87% increase in the number of cases treated (Martin and Smith, 1999; Hamblin *et al.*, 1998; Department of Health Web Site, 2000; Cullis *et al.* 2000). The number of people waiting less than one year has been fairly constant over time. The number of patients waiting more than one year increased in the 1980s but in the 1990s (due to governmental policies) the number of those waiting more than a year has seen a great reduction. Patients were waiting an average of 111 days for elective surgery at the start of 1999 (Martin and Smith, 1999; Department of Health Web Site, 2000; Cullis *et al.*, 2000).

Interpreted as unmet demand for the services both times and lists have received the interest of politicians and of the general public. The political debate has led to the implementation of a number of initiatives in order to decrease waiting lists or waiting times including: the "Waiting List Fund" in 1986, the "Waiting Time Fund" in 1991, which aimed to reduce a large number of patients waiting a long time, and the "Patients' Charter" in 1992, specifying the maximum time a patient should wait. More recently, the government introduced a "Performance Fund" in 1998, rewarding the Health Authorities (HAs) that would do most to reduce lists and times, as well as an extra funding of £500 million directed at the reduction of waiting lists. The "Performance Fund" also provided a threat of dismissal to all non-executive directors of HAs and Trusts if the targets concerning waiting lists were not attained.

Importantly, waiting lists and waiting times for elective care do not appear to be independent from the hospital specialist and the way he manages his list as well as his relation to private practice. About 70% of senior NHS specialists engage in private practice (Monopolies and Mergers Commission, 1993). Private practice accounts for about 10% of working time for full time NHS specialists but the percentage is higher for part-time specialists. Yates (1995) estimates that on average an NHS surgeon undertakes two operations per week in the private sector. There is some concern that the pursuit of private practice by NHS surgeons leads to a conflict of interest (Yates, 1987, 1995). For example, an NHS specialist may use the long waits of the NHS to advise a patient to undertake quicker private treatment which that specialist may be able to offer. Hence, NHS specialists may have a perverse incentive to maintain long NHS waiting lists in order to make their private health care appear more attractive and thus increase their earnings (Yates, 1987, 1995). This perverse incentive is strengthened by the existence of private insurance coupled with the idea that a long list may be a sign of a doctor's prestige and capabilities.

We believe that the analysis that follows contributes to the debate over NHS waiting times and lists. It does so by using a Principal - Multi Agent approach to study the influence of specialists' private practice on the NHS activity and waiting times, thus adding some more insights to the political and social debate over specialists' conflict of interest. We discuss and formalise the motivation and behaviour of hospital specialists within an NHS hospital defining possible utility functions when in the presence of different activities (elective or

emergency care).

We look at the context where the government, as the health care purchaser, acts as a principal of a two-tier hierarchy and contracts with two hospital specialists (the agents) so as to have two tasks accomplished: emergency and non-emergency (elective) patients treated within an NHS hospital. Hospital specialists are organised in a separated horizontal structure, that is, each specialist is responsible for only one type of surgery and the government is able to observe the number of patients treated by each specialist.

We wish to determine the optimal contract that provides the incentives for the specialists to maximise the government's expected utility in the context of asymmetric information when hospital specialists may have an incentive in keeping a certain list length so as to increase demand for and income from their private practice. We analyse the impact of having altruistic and/or selfish specialists on the optimal number of patients treated and on the optimal waiting time.

The next section describes the remuneration system of hospital specialists. We then proceed by discussing the utility functions of the government and the hospital specialists in section (3) and solve for the optimal contract in the context where tasks are separated and information can be symmetric or asymmetric (correlated or non-correlated) in section (4). We finalise with a summary and conclusions (section 5).

2 Hospital Specialists' Remuneration System

The specialists' remuneration is important because specialists' decisions affect the use of hospital and community health services. Hospital specialists are remunerated on a salary basis using recommended fixed salary scales. They may also be given a distinction award and engage in private practice. Salaries and wages are the largest single component of NHS costs (around 62% of the gross annual revenue expenditure with the NHS Hospital and Community Services) and the cost of awards can be up to tens of millions of pounds. Private practice may increase specialists' income by tens of thousands and may create a conflict with the public interest. The possibility of engaging in private practice may provide a perverse incentive to treat fewer patients in the NHS so as to increase demand for NHS specialists' private practice (Bloor and Maynard, 1992, Bloor and Maynard, 1993; Office of Health Economics, 1999).

It has also been argued that the system is not efficient because payment is not in any way related to performance, workload, quality, or health outcome and thus should be reformed to include an explicit performance related element (Bloor and Maynard, 1992; Bloor *et al.*, 1992; Frankel and West, 1993). Salaries are fixed and the system of awards is rather secretive. Awards appear to be based on prestige, international recognition, and on academic distinction, varying widely with the specialty and region. As such, the current remuneration system fails to provide a financial incentive for specialists to work in areas or specialties of, perhaps, lower scientific interest but of interest for the public, and fail to motivate doctors

who then delegate the non-interesting duties to junior doctors (Bloor and Maynard, 1992). A salary system has the advantage that it does not encourage unnecessary treatment but a system of awards could be used to make specialists treat more patients, work more hours, decrease waiting lists or waiting times. The NHS reforms allowed the Trusts to appoint their own specialists and set their payments. This perhaps would give an opportunity to introduce an element of performance related payment. The issues of performance related payment and the relationship between the state and private sector are in fact the object of the governmental document "Agenda for Change" (NHS Executive, 1999).

Analysing a possible reform of the system that bases remuneration on performance (patients treated and waiting times) should be done having in mind that health care is a service (*e.g.* diagnosis and advice), thus heterogeneous and non-retradable, and subject to information asymmetries (Gaynor, 1994). In the market for health care, there are several participants that establish different relationships between them and have different degrees of information: the state (politicians, government), the general public as citizens and patients, insurance providers, purchasers (HAs and GP fundholders in the UK case), and health care providers including hospitals, doctors and other health professionals (Jones and Zanola, 1995). When one of the parties in an exchange has more information than the other, an agency problem is said to exist (Arrow, 1985). The use of performance related reimbursement schemes whereby an optimal contract is established so as to create financial incentives to attain a certain target may thus look reasonable.

Summarising, hospital specialists may be able to influence waiting lists and waiting times for elective surgery as they own and manage their lists. They may have conflicting objectives when in the presence of private practice and private insurance: long NHS lists may increase the demand for their private practice and consequently their earnings. The presence of asymmetric information may emphasize the perverse behaviour of specialists.

It is our aim to analyse the behaviour and the possible reimbursement of specialists when these may not act purely on medical grounds and, as a consequence, may affect the number of patients treated and the waiting time for elective surgery. With the analysis we hope to shed some light on how, in the context of elective care, specialists preferences for private earnings affect NHS activity as well as public transfers when these are performance related (as appears to be the aim of the English government). Doing so, it is our hope that this study by may help a more informed policy making process in what concerns elective care.

3 The Utilities of the Government and of the Hospital Specialists

The research concerning the utility functions of the various participants in the health care market is ongoing.[1] Information on this respect would influence the type of incentives to give

[1]Papers include Mooney and Ryan (1993), Lerner and Claxton (1994), and Scott (1996, 1997).

agents. Given that knowledge on specialists' preferences is limited, when constructing the utility functions of hospital specialists we make use of agency theory as well as information from previous studies and governmental documents concerning hospital specialists' behaviour and their private practice.

We start by assuming that there are a number of individuals in need for health care. Of these potential patients, some need emergency care (and thus constitute the demand for emergency care, D^y) and some need elective or non-urgent treatment (thus constituting the demand for elective care, D^x). The demand for elective care is assumed to depend on morbidity factors, technology, and social attitudes and perceived measures of waiting time or list (see Hamblin *et al.* 1998, Iversen, 1997, Gravelle *et al.*, 2000).[2]

Those treated as elective, x, and those treated as emergency cases, y, obtain a net benefit from treatment (*e.g.* QALYs) defined respectively as Q^x and Q^y. All patients are assumed to obtain equal benefits from treatment.

Waiting time for emergency treatment is zero (*i.e.* an individual cannot wait). Hence, when not all the emergency cases can be treated in a hospital they have to be sent somewhere else or patients die. In either case, we assume there is a cost present. All those needing emergency care but not treated are the difference between total demand and total numbers treated, $(D^y - y)$, and the unit opportunity cost of not treating these is V^y (*e.g.* the social loss due to death).

Concerning elective cases, there are those who wait and receive treatment, x, and those that remain on the list without receiving treatment, $(D^x - x)$. All patients, treated and non-treated, face an average waiting time on a list, t^x, time that delays the reception of that benefit and decreases the present value of the benefit when received (Lindsay and Feigenbaum, 1984; Goddard *et al.* 1995; Martin and Smith, 1999). Hence, delay implies a cost/loss defined here as V^x.

Waiting time for elective care, t^x, is defined as the result of the interaction between those demanding treatment, D^x, and those being treated, x, given a certain capacity attributed to elective surgery, K^x. Hence, we can write a "technological relationship" that expresses waiting time as a function of capacity, number of patients treated and total level of demand: $t^x = \frac{D^x - x}{K^x}$. K^x is fixed and determined *a priori* (Iversen, 1997).[3] The more cases are treated,

[2]Possible measures of perceived waiting time include the time to clear the list, the proportion waiting more than three months and one year, and total list, computed using past information (Gravelle et al., 2000). We believe that it is realistic to think that the actual time in each period is not known to the demanders of care who have only limited access to information such as past information. As such, the demand for elective surgery is dependent on a measure of perceived waiting time which is different from the measure of actual waiting time that we look at in this analysis. Therefore, demand can be considered exogenous to the problem investigated.

[3]The analysis we conduct in this paper can be seen as one where the government wants to treat both types of patients, and wishes to achieve a balance in the numbers treated, therefore maximising health gains for both elective and emergency treatment. An alternative context would be that of having K^{ele} related to the total capacity of the hospital and thus to the capacity allocated to emergency: $K^{ele} = K - K^{eme}$. This formalisation corresponds to the case where the regulator has some prioritising rule treating first all the

the lower is the delay for elective treatment.

The government is assumed to wish to maximise the population's health (White Paper 1999). With this in mind, we use a broad definition of a health production function, $H(HC, Z)$, relating level of health, H, to health care, HC, and to a set of other factors (*e.g.* income), Z. For simplicity reasons, the health production function is assumed to be additively separable in health care and in the set of other factors: $H(HC, Z) = H(HC) + H(Z)$. The government pays transfers to the specialists, T^A and T^B, respectively the transfer to hospital specialist A and the transfer given to hospital specialist B, for the treatment of both types of patients. Government's surplus (which corresponds to patients' surplus)[4] obtained with the health care received is given by the difference between the net benefit of the treatment and the social cost of taxes, $(1+\lambda)$, paid to finance the production of the services.

In order to obtain explicit solutions in the case of this model, we adopt the following formulation for the government's utility function U^P:[5]

$$\begin{aligned} U^P &= \sum_N H(HC_N, Z) - (1+\lambda)\left(T^A + T^B\right) = \\ &= \left(Q^x - V^x t_i^x\right) x_i - V^x\left(D^x - x_i\right) t_i^x - \frac{1}{2}{x_i}^2 + Q^y y_j - \frac{1}{2}{y_j}^2 - V^y\left(D^y - y_j\right) \\ &+ H(Z) - (1+\lambda)\left(T_i^A + T_j^B\right) \end{aligned} \tag{1}$$

with $i = h, l$ and $j = h, l$, that is, when specialists can be of two different types: high cost, h, or low cost, l (see below). $Q^x x_i^A + Q^y y_j^B$ represents the total benefit that patients obtain from both elective and emergency hospital treatment whilst $V^x\left(D^x - x_i\right) t_i^x$ is the total cost imposed on those not treated, $V^x t_i^x x_i$ is the total waiting cost imposed on those treated and $V^y\left(D^y - y_j\right)$ is the total cost of not treating emergency patients. Simplifying,

$$\begin{aligned} U^P &= \sum_N H(HC_N, Z) - (1+\lambda)\left(T^A + T^B\right) = \\ &= Q^x x_i - \frac{1}{2}{x_i}^2 - V^x D^x t_i^x + Q^y y_j - \frac{1}{2}{y_j}^2 - V^y\left(D^y - y_j\right) + H(Z) - (1+\lambda)\left(T_i^A + T_j^B\right) \end{aligned} \tag{2}$$

where $V^x D^x t_i^x$ is the total cost of waiting. It can be seen that total benefit increases with numbers treated but in a decreasing way. The function just presented corresponds to having

emergency cases and then allocating the spare capacity to elective surgery. In that case there is the implicit assumption that health gains obtained with elective are always lower than those obtained with emergency.

[4]The government is not maximising a social welfare function which is the sum of the utilities of patients and specialists, although the government considers the cost of transfers paid for the production of care. This does not change the results qualitatively (Bös, 1994). We follow the aim expressed by the English government in the White Paper 1999.

[5]According to Tirole (1988) and Grossman and Hart (1983) very few general results can be obtained in the absence of specific assumptions concerning the utility functions of the principal and the agents.

specialists organised in a separated structure with specialist A treating elective cases and specialist B doing emergency surgery.

With respect to hospital specialists, it is our intention to consider the possibility that the length of the waiting list may have a positive impact on the utility of the hospital specialists because long lists may increase the demand for their private practice. We wish to see the impact of those preferences on patients treated in the NHS and on the NHS waiting time.

Hospital specialists, when operating in an emergency room, treat the cases as they appear without any discretionary power to postpone treatment for patients face a life threatening situation. As such, the treatment of emergency cases does not depend on the will of a hospital specialist to treat, although the specialist's knowledge and effort may affect the number of emergency cases treated. With elective surgery the situation is quite different. Waiting lists for elective surgery are owned by hospital specialists who are directly responsible for treating the patients on their lists. Given that elective cases are not life threatening hospital specialists have discretionary power over whom and when to treat and, hence, their preferences have a direct effect on the waiting time for elective surgery. We assume that hospital specialists may nevertheless have altruistic attitudes: they may care about the negative impact that waiting time imposes on patients (if elective) and they care about the loss due to not treating emergency patients.

Thus, using the insights of Tirole (1988) and Laffont and Tirole (1993), we build the following utility functions with specialist A dealing with elective care, x, and specialist B dealing with emergency surgery, y, and with the preferences of specialist A having both a selfish and an altruistic component:[6]

$$\begin{aligned} U_i^A &= T_i^A - \left(\theta_i^A - e_i^A\right) x_i^A - \left(e_i^A\right)^2 + \tau\left(-V^x t_i^x\right) - (1-\tau) r x_i^A = \\ &= T_i^A - \left(\theta_i^A - e_i^A\right) x_i^A - \left(e_i^A\right)^2 - \tau V^x \left(\frac{D^x - x_i^A}{K^x}\right) - (1-\tau) r x_i^A \end{aligned} \tag{3}$$

and

$$U_j^B = T_j^B - \left(\theta_j^B - e_j^B\right) y_j^B - \left(e_j^B\right)^2 - \omega V^y \left(D^y - y_j^B\right) \tag{4}$$

with $i = h, l$ and $j = h, l$ (where i is the type of specialist A and j is the type of specialist B). T_i^A and T_j^B are the transfers received by each specialist and paid by the government, e_i^A and e_j^B stand for the effort specialist A and B, respectively, put into the services they provide to patients, and θ_i^A and θ_i^B are random exogenous variables affecting each agent's cost (*e.g.* knowledge or skills and hence, the less skillful or knowledgeable the specialist is, the higher the cost of producing x_i). θ assumes discrete values (high or low) and defines whether the specialist has a low, θ_l, or high cost, θ_h. It is assumed to follow a known probability distribution. x_i^A and y_j^B represent respectively the number of patients treated as elective and as emergency cases.

[6]The way we modelled the specialists' utility functions considers a cost function rather than a production function. The reason to do this is that we want to address the issue of cost minimisation and namely the role of effort in the number of patients treated as elective or emergency cases.

$\left(\theta_i^A - e_i^A\right) x_i^A$ and $\left(\theta_j^B - e_j^B\right) y_j^B$ are the cost functions related to the production of output x_i and y_j, whilst $\left(e_{x_i}^A\right)^2$ and $\left(e_j^B\right)^2$ measure the disutility of effort measured in monetary terms. τ and ω, with $0 \leq \tau \leq 1$ and $0 \leq \omega \leq 1$, represent respectively the degree of altruism presented by a specialist towards the elective wait and towards those still in need for emergency treatment. $V^y\left(D^y - y_j^B\right)$ measures the disutility specialist B attaches to those not treated as emergency whereas rx_i^A can be seen as the disutility hospital specialist A derives from having that number of patients treated in the public. If r is an average per patient payment received by the specialist in the private sector, then rx_i^A is a proxy for what hospital specialist A may forego as income were he to treat those patients in the private sector, that is, the opportunity cost of treating the patients in the NHS.

Another way of looking at the possible 'selfish' behaviour of hospital specialists is to consider that a longer NHS list increases the specialist's utility because it increases the potential demand for his private practice:

$$\begin{aligned} U_i^A &= T_i^A - \left(\theta_i^A - e_i^A\right) x_i^A - \left(e_i^A\right)^2 + \tau\left(-V^x t_i^x\right)\varepsilon + (1-\tau)r\left(D^x - x_i^A\right)(1-\varepsilon) \\ &= T_i^A - \left(\theta_i^A - e_i^A\right) x_i^A - \left(e_i^A\right)^2 - \tau V^x\left(\frac{D^x - x_i^A}{K^x}\right) + (1-\tau)r\left(D^x - x_i^A\right)(1-\varepsilon) \end{aligned} \quad (5)$$

In this context, the specialist perceives that a potential number of those potentially not treated in the NHS will go private, $\left(D^x - x_i^A\right)(1-\varepsilon)$, number that is affected by the number of patients he treats, x_i^A. In (5) the more patients a specialist treats in the NHS, the smaller is the number of those left to treat and potentially demanding private surgery $\left(D^x - x_i^A\right)$. ε measures patients preferences concerning NHS treatment and how this is preferred over private treatment.[7] $r\left(D^x - x_i^A\right)(1-\varepsilon)$ is the potential income that the specialist may derive from his private practice.

Note that when dealing with emergency the specialist does not face any conflict of interest or perverse incentives. As such, emergency is here to allow us to establish the comparison between different and possible activities specialists conduct within the hospital. It allows us to show that perverse incentives may develop in some but not in other contexts and certain motivations may arise from external factors, in our case, the possibility of conducting private surgery.

We assume risk-neutral agents and a risk-neutral principal (where risk neutrality refers to income).

We now investigate the consequences of having the different functions on numbers treated and waiting times.[8]

[7] $(1-\varepsilon)$ is assumed to be exogenous to the model. The literature appears to be ambiguous with respect to whether the wait affects the demand for private care. Besley et al. (1999) finds that the waiting list is positively associated with the purchasing of health insurance whereas Propper (1998) does not find evidence that the demand for private health care is affected by the waiting list. The demand for private care is found to be related to income and political attitudes as well as to past use of private care.

[8] In the analysis that follows we take private practice and the number of patients treated in the private

4 Analysing the Impact of Private Practice

We derive the optimal contract for each specialists in three different scenarios: first, the symmetric information case, where the government has complete information about specialists types, θ, and is aware of their effort, e (subsection 4.1); second, in subsection (4.2), the situation where hospital specialists have superior information about their type and effort but information is not correlated (*e.g.* specialists do not contact with each other in their activities); third, the context where there is correlation of information (subsection 4.3).

4.1 When Information is Symmetric

In this section, as a benchmark case, we derive the optimal contract between the government and the two hospital specialists for the separated treatment of emergency and elective patients when the government perfectly observes θ^A and θ^B. Specialist A is responsible for elective surgery, x, while specialist B is responsible for emergency surgery, y. They can be of two cost types measured by the random parameter θ: low cost, θ_l,or high cost, θ_h. The level of outcome achieved by the specialists depends on how costly it is for them to produce due to their knowledge. Moreover, specialist are required to put effort, e, in the treatment of patients, and their choice of effort is conditional on the their private information of θ (*i.e.* on their cost parameter). We solve for the optimal contract by first deriving each specialists' choice of effort (Laffont and Tirole, 1993; Tirole, 1988). Each specialist chooses the level of effort that maximises his utility function given its type.[9]

For hospital specialist B when he is of type j and treats emergency cases only, we have:

$$\max_{e_j^B} U_j^B = T_j^B - \left(\theta_j^B - e_j^B\right) y_j^B - \left(e_j^B\right)^2 - \omega V^y \left(D^y - y_j^B\right)$$

$$\Longrightarrow e_j^{B*} = \frac{y_j^B}{2}$$

sector as something contrary to the government' objectives. In our context, the government achieves greater health gains the greater the number of patients treated publicly. One could think that patients that go private also obtain health gains that the government could take into consideration. It may be the case, however, that the gains obtained are outweighed by the costs of an increase in the waiting time for NHS patients so as to make patients go private (Iversen, 1997). Also, the government may not wish to consider those gains if it is the rich and upper medium class that benefits from going private (White Paper 1999). Moreover, in an efficiency perspective the government should maximise the number of patients treated with the resources allocate to the NHS.

[9]We are not considering the choice of effort to be put in the NHS and in the specialist's private practice given a total effort possible. Our aim is to look at the effort put in the NHS and see how it affects numbers treated and the waiting time. The private practice impacts on the NHS not via effort but as an extra element in the utility function namely as income foregone. This corresponds to the context where more effort in the public allows for a better use of the existing fixed capacity whereas in the private sector there is no capacity or effort constraint.

with $j = h, l$. T_j^B, θ_j^B, e_j^B, and y_j^B are as defined previously. Similarly, for hospital specialist A, when A is of type i, with $i = h, l$:

$$\max_{e_i^A} U_i^A = T_i^A - \left(\theta_i^A - e_i^A\right) x_i^A - \left(e_i^A\right)^2 - \tau V^x \left(\frac{D^x - x_i^A}{K^x}\right) - (1-\tau) r x_i^A$$

$$\Longrightarrow e_i^{A*} = \frac{x_i^A}{2}$$

where T_i^A, θ_i^A, e_i^A, and x_i^A are as before. These expressions for the optimal efforts give us the first best effort.

When the non-altruistic term is $(1-\tau) r \left(D^x - x_i^A\right)(1-\varepsilon)$, the optimal effort is also $e_i^{A*} = \frac{x_i^A}{2}$.

Given the assumption of full information, the government maximises his utility knowing that specialist B is of type j and specialist A is of type i, and subject to having specialists' utility greater or equal to zero, $U_i^A, U_j^B \geqslant 0$, that is, the government has to ensure that specialists remain in activity in every state of the world - limited liability case. Principal - Agent theory assumes that, in a competitive market for agents, the principal has to make sure that the utility the agent obtains by working in the activities she wishes is greater than the agent's reservation utility, that is, the utility he can obtain by going somewhere else (Tirole, 1988). In our case, the utility of working in the NHS must be greater than that of dropping the activity which is assumed to be zero. Thus, we can write the government's maximisation problem as follows

$$\max_{y_j^B, x_i^A, T_i^A, T_j^B} U^P = Q^x x_i^A - \frac{1}{2}(x_i^A)^2 - V^x D^x t_i^x + Q^y y_j^B - \frac{1}{2}(y_j^B)^2 - V^y \left(D^y - y_j^B\right)$$
$$+ H(Z) - (1+\lambda)\left(T_i^A + T_j^B\right)$$
$$\text{subject to } U_j^B \geqslant 0 \text{ and } U_i^A \geqslant 0, \text{ and to } t_i^x = \frac{D^x - x_i^A}{K^x}$$

with $i = h, l$ and $j = h, l$. $U_j^B \geqslant 0$ and $U_i^A \geqslant 0$ are the individual rationality or participation constraints (PCs) reflecting the fact that the government has to ensure that specialists participate in production.

Using the expressions of the optimal levels of effort of the specialists as well as the expression for the waiting time we have with $i = h, l$ and $j = h, l$:

$$\max_{y_j^B, x_i^A, T_i^A, T_j^B} U^P = Q^x x_i^A - \frac{1}{2}(x_i^A)^2 - V^x D^x \left(\frac{D^x - x_i^A}{K^x}\right) + Q^y y_j^B - \frac{1}{2}(y_j^B)^2 - V^y \left(D^y - y_j^B\right)$$
$$+ H(Z) - (1+\lambda)\left(T_i^A + T_j^B\right)$$

$$\text{s.t. } T_i^A - \theta_i^A x_i^A + \frac{\left(x_i^A\right)^2}{4} - \tau V^x \left(\frac{D^x - x_i^A}{K^x}\right) - (1-\tau) r x_i^A \geqslant 0$$

$$\text{and } T_j^B - \theta_j^B y_j^B + \frac{\left(y_j^B\right)^2}{4} - \omega V^y \left(D^y - y_j^B\right) \geqslant 0$$

Given the hypothesis of complete information about the cost types, and given that the transfers are costly, the participation constraints are binding for all types and for both specialists. Hence, with $i = h, l$ and $j = h, l$:

$$T_i^A = \theta_i^A x_i^A - \frac{\left(x_i^A\right)^2}{4} + V^x \left(\frac{D^x - x_i^A}{K^x}\right) + \tau V^x \left(\frac{D^x - x_i^A}{K^x}\right) + (1-\tau) r x_i^A \tag{6}$$

$$T_j^B = \theta_j^B y_j^B - \frac{\left(y_j^B\right)^2}{4} + \omega V^y \left(D^y - y_j^B\right) \tag{7}$$

Substituting (6) and (7) into the utility function of the government, as in (2) we have

$$\begin{aligned}
\max_{y_j^B, x_i^A} U^P = {} & Q^x x_i^A - \frac{1}{2}(x_i^A)^2 - V^x D^x \left(\frac{D^x - x_i^A}{K^x}\right) + Q^y y_j^B - \frac{1}{2}(y_j^B)^2 - V^y \left(D^y - y_j^B\right) + H(Z) \\
& - (1+\lambda) \left[\theta_i^A x_i^A - \frac{\left(x_i^A\right)^2}{4} \tau V^x \left(\frac{D^x - x_i^A}{K^x}\right) + (1-\tau) r x_i^A\right] \\
& - (1+\lambda) \left[\theta_j^B y_j^B - \frac{\left(y_j^B\right)^2}{4} + \omega V^y \left(D^y - y_j^B\right)\right]
\end{aligned}$$

and the first order conditions are:

$$\begin{aligned}
\frac{\partial U^P}{\partial x_i^x} &= Q^x - x_i^A + \frac{V^x D^x}{K^x} - (1+\lambda) \left[\theta_i^A - \frac{x_i^A}{2} - \tau \frac{V^x}{K^x} + (1-\tau) r\right] = 0 \\
\frac{\partial U^P}{\partial y_j^y} &= Q^y - y_j^B + V^y - (1+\lambda) \left(\theta_j^B - \frac{y_j^B}{2} - \omega V^y\right) = 0
\end{aligned}$$

where $\frac{\partial U^P}{\partial x_i^x}$ is the first derivative of the government's utility function with respect to x_i^A and $\frac{\partial U^P}{\partial y_j^B}$ is the first derivative with respect to y_j^B. From these expressions we can obtain the optimal number of patients treated as elective and emergency cases, which are respectively, with $i = h, l$ and $j = h, l$:

$$x_i^{A*} = \frac{2}{(1-\lambda)} \left[Q^x + \frac{V^x D^x}{K^x} + (1+\lambda) \frac{V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)} \left[\theta_i^A + (1-\tau) r\right] \tag{8}$$

$$y_j^{B*} = \frac{2}{(1-\lambda)} \left[Q^y + V^y + (1+\lambda) \omega V^y\right] - \frac{2(1+\lambda)}{(1-\lambda)} \theta_j^B \tag{9}$$

When the expression concerning 'selfish' behaviour is $(1-\tau)r\left(D^x - x_i^A\right)(1-\varepsilon)$, the optimal number of elective patients treated is:

$$x_i^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + (1+\lambda)\frac{V^x}{K^x}\right] - \left[\frac{2(1+\lambda)}{(1-\lambda)}\theta_i^A + (1-\tau)(1-\varepsilon)r\right] \qquad (10)$$

Thus, it can be conclude by looking at (8), (10), and (9) that (see also appendix A.1):

Conclusion 1 *The optimal number of emergency patients treated increases with the benefit from treatment, Q^y, and the disutility generated by turning emergency patients down, V^y. It decreases with the marginal cost of production, θ_j^B.*

Conclusion 2 *Similarly the number of elective patients treated increases with the benefit from treatment, Q^x, and with the disutility generated by the time spent waiting, V^x. The number decreases with the marginal cost of production, θ_i^A.*

Conclusion 3 *The more altruistic the specialists' preferences, in so far as specialists care about the cost of waiting or the cost of turning down patients, the higher the number of patients treated in the NHS.*

Conclusion 4 *The greater the private income foregone, the smaller the number of NHS patients treated as elective cases.*

Conclusion 5 *The greater the preferences of the public for private practice the greatest the impact of specialists perverse behaviour, that is, the smaller the number of NHS patients treated*

Hence, when the specialist dealing with elective surgery has an incentive to keep lists long in order to increase demand for private practice the number of NHS patients treated is reduced. Moreover, a shift in population preferences towards private care may enhance this perverse behaviour.

The waiting time for elective surgery is defined as the technological relationship $t^x = \frac{D^x - x_i^A}{K^x}$. Using the expression for the optimal number of elective patients treated, x_i^{A*}, and substituting it into the above expression we obtain the resulting waiting time, $t^x = \frac{D^x - x_i^{A*}}{K^x}$. It can be seen that:

Conclusion 6 *Waiting time is lower the more altruistic specialists are, and it is higher the greater the impact on specialists' utility of the income foregone from private surgery or the greater the preferences for private practice.*

This adds to the claims by Iversen (1997) that private practice leads to a higher waiting time in the NHS.

The optimal transfers, when specialist A has both 'selfish' and altruistic preferences, are:

$$T_j^{B*} = \theta_j^B y_j^{B*} - \frac{\left(y_j^{B*}\right)^2}{4} + \omega V^y \left(D^y - y_j^{B*}\right) \tag{11}$$

$$T_i^{A*} = \theta_i^A x_i^{A*} - \frac{\left(x_i^{A*}\right)^2}{4} + \tau V^x \left(\frac{D^x - x_i^{A*}}{K^x}\right) + (1-\tau) r x_i^{A*} \tag{12}$$

or:

$$T_i^{A*} = \theta_i^A x_i^{A*} - \frac{\left(x_i^{A*}\right)^2}{4} + \tau V^x \left(\frac{D^x - x_i^{A*}}{K^x}\right) - (1-\tau) r \left(D^x - x_i^{A*}\right)(1-\varepsilon) \tag{13}$$

It can be seen that the transfers paid in the NHS depend on the existence of private practice and on how specialists relate to it.

When specialists care about the private income foregone in the private sector due to treating the patients in the NHS, as in equation (12), the specialist obtains an increase in the rent in the public sector. Thus, the transfers paid in the NHS increase when specialists see in the possibility of treating patients privately an opportunity to increase their earnings. The rationale is the following: to give an incentive (financial) for the specialist to treat those patients in the NHS and not in the private sector, that is, to control the possible perverse behaviour of specialists, the government has to increase public transfers so as to compensate the specialist for what he foregoes due to not treating them in the private sector. To adjust for the possible perverse behaviour of specialists the government increases the transfers. It can be concluded that the existence of private practice increases the transfer the government has to pay to specialists in the NHS and hence increases health care costs.

The case of transfer (13) is one that reflects the fact that specialists perceive that a potential number of those not treated in the NHS will go private, $\left(D^x - x_i^{A*}\right)(1-\varepsilon)$, and this number is affected by how many they treat in the NHS, x_i^{A*}, with $r\left(D^x - x_i^{A*}\right)(1-\varepsilon)$ being the potential income obtained by treating the potential demand for private practice. In this context, the optimal transfer scheme suggests that the government should reduce the state transfers paid to hospital specialists proportionally to what is perceived specialists may receive outside the NHS by increasing the demand for their private practice. The optimal transfer is thus a decreasing function of the income received from their private practice.

An alternative system would be to establish a regulatory rule that obliges specialists to choose the sector (private or public) of their activity. The analysis of this regulatory tool lies outside the scope of this chapter and is left for future research. We restrict the analysis to the context where NHS specialists are allowed to practice privately and performance related remuneration is put in place (which appears to be the intention of the English government).

With a simple setting it can be shown that, independently of the way specialists dealing with elective surgery internalises their private practice, the latter affects the number of patients treated in the NHS and thus the waiting time for elective surgery. We formalise

the different motivations (measured by parameters τ and ω) faced by hospital specialists depending on the task developed, and how these together with their skills/knowledge (θ) and effort (e) impact on the number of patients treated and on the resulting waiting time for elective surgery. If the government is to create performance related remuneration and still allow for private practice it has to be aware that the optimal contract has to take into account all the elements of the utility of the specialist including the possible income he could obtain with his private practice. In either context of preferences, we obtain important results with important policy implications, suggesting that the government may have something to gain by taking into account private income sources when setting public transfers to hospital specialists.

4.2 When Information is Asymmetric and Uncorrelated

As said, the health care market is one where asymmetry of information is an important feature. Therefore, it is realistic to assume that the government has less information than the specialists: only the specialists know if they have a low or a high production cost, θ_l or θ_h. Consequently, specialists' compensation can only be based on output and the government holds expectations over θ. In this situation, besides having to ensure that specialists participate in the production, the government has to provide the incentives to induce the specialists to reveal their type (*i.e.* the incentive compatibility constraints - ICCs). In other words, the government must design an ex-ante contract that induces the specialists to produce according to their type. Limited liability is still assumed in place.

When there is no correlation between the agents' private information, the principal cannot do better than contract with each agent separately (Demski and Sappington, 1983 and 1984). In our case, this implies setting a contract with specialist A for the treatment of elective cases, x_i^A, and a contract with specialist B for the treatment of emergency patients, y_j^B.

The timing of the contract is as follows: first, the realisation of θ occurs and is observed by the specialists alone; second, the government offers a contract to the specialists specifying the state-contingent transfers for each possible level of production; third, conditional on θ, the specialists choose their optimal level of effort; and, fourth, production takes place. Conditional on the observed numbers treated the specialists receive the agreed transfers from the government. We further assume that specialists play a non-cooperative Nash game, that is, they do not collude in their choice of effort or in the revelation of private information.

Specialists' choice of effort takes place in the same manner as in complete information as technology is separable in effort and the cost θ. Thus:

$$e_j^{B*} = \frac{y_j^B}{2} \text{ and } e_i^{A*} = \frac{x_i^A}{2}$$

The government maximises health gains establishing a separate contract with each specialist. Thus, when contracting with specialist A for the treatment of elective patients and

knowing the probabilities, p_i, that the specialist is of a low or a high cost, the government:

$$\max_{x_i^A, T_i^A} U^P = \sum_{i=h,l} p_i \left[Q^x x_i^A - \frac{1}{2}(x_i^A)^2 - V^x D^x \left(\frac{D^x - x_i^A}{K^x} \right) - (1+\lambda) T_i^A + H(Z) \right] \tag{14}$$

subject to the individual rationality constraints that guarantee that the specialist receives his reservation utility if he truthfully reveals his type:

$$T_i^A - \theta_i^A x_i^A + \frac{\left(x_i^A\right)^2}{4} - \tau V^x \left(\frac{D^x - x_i^A}{K^x} \right) - (1-\tau) r x_i^A \geqslant 0, \forall i = h, l$$

and subject to the incentive compatibility constraints which ensure that the specialist prefers to tell the truth rather than lie about his actual productivity, $\forall i, q = h, l$ and $i \neq q$,[10]

$$\begin{aligned} & T_i^A - \theta_i^A x_i^A + \frac{\left(x_i^A\right)^2}{4} - \tau V^x \left(\frac{D^x - x_i^A}{K^x} \right) - (1-\tau) r x_i^A \\ & \geqslant T_q^A - \theta_i^A x_q^A + \frac{\left(x_q^A\right)^2}{4} - \tau V^x \left(\frac{D^x - x_q^A}{K^x} \right) - (1-\tau) r x_q^A \end{aligned}$$

Note that according to the Revelation Principle (Myerson, 1979), among all possible games an agent may choose to play when he privately observes his type, one can restrict the attention to the class of direct mechanisms in which the agent is induced to truthfully reveal his own type. Therefore, in our context we have:

Conclusion 7 *Each specialist is kept at his reservation level of utility in the high cost state of the world, $\theta = \theta_h$, while he gains a rent due to his informational monopoly when he is of a low cost, $\theta = \theta_l$.* [11]

This corresponds to having for specialist A, with $i = h, l$:

a) the following optimal payment functions for each state of the world

$$T_l^A = \theta_l^A x_l^A - \frac{\left(x_l^A\right)^2}{4} + (\theta_h^A - \theta_l^A) x_h^A + \tau V^x \left(\frac{D^x - x_l^A}{K^x} \right) + (1-\tau) r x_l^A \tag{15}$$

$$T_h^A = \theta_h^A x_h^A - \frac{\left(x_h^A\right)^2}{4} + \tau V^x \left(\frac{D^x - x_h^A}{K^x} \right) + (1-\tau) r x_h^A \tag{16}$$

[10] Here we just present one type of specialist A preferences. The results with the second type of preferences are similar and are presented in appendix A.1.

[11] This result follows from the fact that in the optimal incentive scheme if the agent is in the most productive state he is indifferent between telling the truth and claiming to be of lower productivity, and in such a context he chooses the action most preferred by the principal. We check whether these results are verified using Laffont and Tirole (1993) standard approach.

b) and substituting (15) and (16) into (14) the first-order conditions with respect to the levels of output x_h^A and x_l^A:

$$\frac{\partial U^P}{\partial x_l^A} = Q^x - x_l^A + \frac{V^x D^x}{K^x} - (1+\lambda)\left[\theta_l^A - \frac{x_l^A}{2} - \frac{\tau V^x}{K^x} + (1-\tau)r\right] = 0$$
$$\frac{\partial U^P}{\partial x_h^A} = Q^x - x_h^A + \frac{V^x D^x}{K^x} - (1+\lambda)\left[\theta_h^A - \frac{x_h^A}{2} - \frac{\tau V^x}{K^x} + (1-\tau)r + \frac{p_l\left(\theta_h^A - \theta_l^A\right)}{(1-p_l)}\right] = 0$$

where p_l is the probability that we are in the low cost state of the world, $\theta_i^A = \theta_l^A$.

The optimal number of elective patients treated by hospital specialist A, and for each type of cost, is therefore:

$$x_l^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + \frac{(1+\lambda)\tau V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[(1-\tau)r + \theta_l^A\right] \tag{17}$$
$$x_h^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + \frac{(1+\lambda)\tau V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[\theta_h^A + (1-\tau)r + \frac{p_l(\theta_h^A - \theta_l^A)}{(1-p_l)}\right] \tag{18}$$

and similarly for hospital specialist B the optimal number of emergency patients treated is:

$$y_l^{B*} = \frac{2}{(1-\lambda)}\left[Q^y + V^y + (1+\lambda)\omega V^y\right] - \frac{2(1+\lambda)}{(1-\lambda)}\theta_l^B \tag{19}$$
$$y_h^{B*} = \frac{2}{(1-\lambda)}\left[Q^y + V^y + (1+\lambda)\omega V^y\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[\theta_h^B + \frac{p_l}{(1-p_l)}(\theta_h^B - \theta_l^B)\right] \tag{20}$$

Using (17), (18), (19), and (20) one obtains the expressions for the optimal transfers:

$$T_l^A = \theta_l^A x_l^{A*} - \frac{\left(x_l^{A*}\right)^2}{4} + \left(\theta_h^A - \theta_l^A\right)x_h^{A*} + \tau V^x\left(\frac{D^x - x_l^{A*}}{K^x}\right) + (1-\tau)r x_l^{A*} \tag{21}$$
$$T_h^A = \theta_h^A x_h^{A*} - \frac{\left(x_h^{A*}\right)^2}{4} + \tau V^x\left(\frac{D^x - x_h^{A*}}{K^x}\right) + (1-\tau)r x_h^{A*} \tag{22}$$

It can be seen that:

Conclusion 8 *The first best number of patients treated is achieved when specialists are of a low cost, $\theta = \theta_l$.*

Conclusion 9 *Due to asymmetry of information the specialist gains a rent if in a low cost state of the world (e.g. high skills or knowledge when $\theta = \theta_l$). When there is asymmetry of information and the specialist is of a high cost (e.g. low skills or knowledge) fewer patients are treated.*

Conclusion 10 *Altruistic behaviour has a positive impact on the number of patients treated and decreases waiting time, whereas the interest on increasing private practice decreases numbers treated and increases the time spent waiting.*

Therefore, asymmetry of information leads to fewer patients being treated or bigger transfers being paid to the specialist, effects that enhance the negative impact already brought up by private practice.

4.3 When Information is Asymmetric and Correlated

As before, the government has less information than the hospital specialists over their cost type and effort. However, specialists' private information, the value of θ, is now correlated with the joint probability distribution of θ for the two specialists, ϕ_{ij}, known. When the sign of the correlation is known, the principal can use some information on one agent to derive the type of the other agent and design a contract accordingly. Specialists are still assumed to play a non-cooperative Nash game.

The choice of effort takes place in the same manner as in the complete information case so that the optimal levels of efforts chosen by the specialists are, with $i, j = h, l$:

$$e_{ij}^{A*} = \frac{x_{ij}^A}{2} \text{ and } e_{ji}^{B*} = \frac{y_{ji}^B}{2}$$

As previously, the government maximises health gains and has to ensure that the specialists participate in the production and provide the incentives to induce them to reveal their cost type. Following the formulation proposed by Demski and Sappington (1984), the problem of the government is, with $i, j = h, l$:

$$\begin{aligned}\max_{x_{ij}^A, x_{ji}^B, T_{ij}^A, T_{ji}^B} U^P = \sum_{i=h,l}\sum_{j=h,l} \phi_{ij} \Bigg[& Q^x x_{ij}^A - \frac{1}{2}\left(x_{ij}^A\right)^2 - V^x D^x \left(\frac{D^x - x_{ij}^A}{K^x}\right) \\ & + Q^y y_{ji}^B - \frac{1}{2}\left(y_{ji}^B\right)^2 - V^y\left(D^y - y_{ji}^B\right) + H(Z) - (1+\lambda)\left(T_{ij}^A + T_{ji}^B\right)\Bigg] \qquad (23) \\ \text{s.t. PCs and to ICCs}\end{aligned}$$

where ϕ_{ij} indicates the joint probability that $\theta_i^A = h, l$ and $\theta_j^B = h, l$, and takes four values: ϕ_{ll} when θ_l^A and θ_l^B; ϕ_{lh} when θ_l^A and θ_h^B; ϕ_{hl} when θ_h^A and θ_l^B; and ϕ_{hh} when θ_h^A and θ_h^B.

Using the optimal effort levels we have:

a) the PCs for specialist A and specialist B and for each state of the world (*i.e.* a limited liability case):

$$T_{ij}^A - \theta_i^A x_{ij}^A + \frac{(x_{ij}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{ij}^A}{K^x}\right) - (1-\tau) r x_{ij}^A \geqslant 0$$

$$T_{ji}^B - \theta_j^B y_{ji}^B + \frac{(y_{ji}^B)^2}{4} - \omega V^y\left(D^y - y_{ji}^B\right) \geqslant 0$$

b) the ICCs for specialist B:

$$T_{jl}^B - \theta_j^B y_{jl}^B + \frac{(y_{jl}^B)^2}{4} - \omega V^y \left(D^y - y_{jl}^B\right)$$
$$\geq T_{kl}^B - \theta_j^B y_{kl}^B + \frac{(y_{kl}^B)^2}{4} - \omega V^y \left(D^y - y_{kl}^B\right)$$

c) the ICCs for specialist A:

$$T_{il}^A - \theta_i^A x_{il}^A + \frac{(x_{il}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{il}^A}{K^x}\right) - (1-\tau) r x_{il}^A$$
$$\geq T_{ql}^A - \theta_i^A x_{ql}^A + \frac{(x_{ql}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{ql}^A}{K^x}\right) - (1-\tau) r x_{ql}^A$$

$\forall j, k = h, l$ and $j \neq k$,and $\forall i, q = h, l$ and $i \neq q$ and $j \neq k$.

As in Mas Colell *et al.* (1995), we define the above expressions for the incentive compatibility constraints as "ex-post" since the contract is offered by the Principal after each agent has learned his type and that of the other agent as well. Yet, in the organisation we are studying, we assume that there is no form of collusion between the two agents, so that the fact that they observe each other's type before private information becomes publicly revealed does not have an impact on the form of the incentive compatibility constraints that each agent faces in the optimal contract.[12] As a consequence,

Conclusion 11 *The optimal contract for each specialist has the participation constraints binding in the high cost state of the world, for every realisation of the other specialist's private information. In the low cost state, the incentive compatibility constraints are binding at the optimum, for each specialist and for every realisation of the other specialist's private information. Numbers treated are decreasing in the cost.* [13]

The optimal transfer functions are, for specialists A and B respectively with $i, j = h, l$

$$T_{lj}^A = \theta_l^A x_{lj}^A - \frac{(x_{lj}^A)^2}{4} + \tau V^x \left(\frac{D^x - x_{lj}^A}{K^x}\right) + (1-\tau) r x_{lj}^A + \left(\theta_h^A - \theta_l^A\right) x_{hj}^A \tag{24}$$

$$T_{hj}^A = \theta_h^A x_{hj}^A - \frac{(x_{hj}^A)^2}{4} + \tau V^x \left(\frac{D^x - x_{hj}^A}{K^x}\right) + (1-\tau) r x_{hj}^A \tag{25}$$

[12] Alternatively, if the agents could not observe each other's type, we would have used "interim incentive compatibility constraints" (Holmstrom and Myerson, 1983). Interim ICCs imply that the optimal contract is derived after each agent "has learned his type but before the agents' types are publicly revealed" (Mas-Colell *et al.*, 1995) so that each agent guesses the other agent's type. They capture the fact that for every expected type of the other agent, each agent is at least as better off by telling the truth about his type than lying. Therefore, the ICCs hold in expectation for every value that the other agent's private information may take.

[13] See Appendix B for the proofs.

$$T_{li}^B = \theta_l^B y_{li}^B - \frac{(y_{li}^B)^2}{4} + \omega V^y \left(D^y - y_{li}^B\right) + \left(\theta_h^B - \theta_l^B\right) y_{hi}^B \tag{26}$$

$$T_{hi}^B = \theta_l^B y_{hi}^B - \frac{(y_{hi}^B)^2}{4} + \omega V^y \left(D^y - y_{hi}^B\right) \tag{27}$$

Hence, both specialist B and specialist A are kept at their reservation level in the high cost state, whatever the value of the other specialist's cost state. In the low cost state of the world, specialist B experiences an informational rent, given by the expression $(\theta_h^B - \theta_l^B)y_{hi}^B$. Similarly, specialist A obtains a rent given by $(\theta_h^A - \theta_l^A)x_{hj}^A$.

Substituting (24), (25), (26), and (27) into the utility function of the government as in (23) and computing the first-order conditions we have:

a) for hospital specialist A, with $i = h, l$ and $j = h, l$:

$$\frac{\partial U^P}{\partial x_{lj}^A} = Q^x - x_{lj}^A + \frac{V^x D^x}{K^x} - (1+\lambda)\left[\theta_l^A - \frac{x_{lj}^A}{2} - \frac{\tau V^x}{K^x} - (1-\tau)r\right] = 0$$

$$\frac{\partial U^P}{\partial x_{hj}^A} = Q^x - x_{hj}^A + \frac{V^x D^x}{K^x} - (1+\lambda)\left[\theta_h^A - \frac{x_{hj}^A}{2} - \frac{\tau V^x}{K^x} - (1-\tau)r + \frac{\phi_{lj}\left(\theta_h^A - \theta_l^A\right)}{\phi_{hj}}\right] = 0$$

b) for hospital specialist B, with $i = h, l$ and $j = h, l$:

$$\frac{\partial U^P}{\partial y_{li}^B} = Q^y - y_{li}^B + V^y - (1+\lambda)\left(\theta_l^B - \frac{y_{li}^B}{2} - \omega V^y\right) = 0$$

$$\frac{\partial U^P}{\partial y_{hi}^B} = Q^y - y_{hi}^B + V^y - (1+\lambda)\left(\theta_l^B - \frac{y_{hi}^B}{2} - \omega V^y + \frac{\phi_{li}\left(\theta_h^B - \theta_l^B\right)}{\phi_{hi}}\right) = 0$$

And the optimal numbers treated are, with $i = h, l$ and $j = h, l$:

$$x_{lj}^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + \frac{(1+\lambda)\tau V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[\theta_l^A + (1-\tau)r\right] \tag{28}$$

$$x_{hj}^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + \frac{(1+\lambda)\tau V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[\theta_h^A + (1-\tau)r + \frac{\phi_{lj}\left(\theta_h^A - \theta_l^A\right)}{\phi_{hj}}\right] \tag{29}$$

and

$$y_{li}^{B*} = \frac{2}{(1-\lambda)}\left\{Q^y + V^y\left[1 + \omega(1+\lambda)\right]\right\} - \frac{2(1+\lambda)}{(1-\lambda)}\theta_l^B \tag{30}$$

$$y_{hi}^{B*} = \frac{2}{(1-\lambda)}\left\{Q^y + V^y\left[1 + \omega(1+\lambda)\right]\right\} - \frac{2(1+\lambda)}{(1-\lambda)}\left[\theta_h^B + \frac{\phi_{li}}{\phi_{hi}}(\theta_h^B - \theta_l^B)\right] \tag{31}$$

Conclusion 12 *In the low cost state of the world, each specialist treats the first-best level of patients, for every realisation of the other specialist's cost level.*

Conclusion 13 *In the high cost state of the world, each specialist treats a sub-optimal number of patients. The departure of the level of production from the first best depends however on the correlation between the two random variables, θ_j^B and θ_i^A.*

In particular, we would have that:

$y_{hh}^{B*} > y_{hl}^{B*}$ when $\frac{\phi_{ll}}{\phi_{hl}} > \frac{\phi_{lh}}{\phi_{hh}}$
$y_{hh}^{B*} = y_{hl}^{B*}$ when $\frac{\phi_{ll}}{\phi_{hl}} = \frac{\phi_{lh}}{\phi_{hh}}$
$y_{hh}^{B*} < y_{hl}^{B*}$ when $\frac{\phi_{ll}}{\phi_{hl}} < \frac{\phi_{lh}}{\phi_{hh}}$

These findings depend on the covariance between θ_j^B and θ_i^A, since:

$\frac{\phi_{ll}}{\phi_{hl}} > \frac{\phi_{lh}}{\phi_{hh}}$ when $\sigma > 0$
$\frac{\phi_{ll}}{\phi_{hl}} = \frac{\phi_{lh}}{\phi_{hh}}$ when $\sigma = 0$
$\frac{\phi_{ll}}{\phi_{hl}} < \frac{\phi_{lh}}{\phi_{hh}}$ when $\sigma < 0$

where σ indicates the covariance between the two random variables.

It is important to note that when specialists private information is correlated, the well known trade-off between incentives and efficiency (or between incentives and rent-extraction) in the optimal contract assumes different features. For example, when specialist B is a high cost type, and specialist A a low cost one, then specialist B will produce more output when the covariance between θ_j^B and θ_i^A is positive then when it is negative. As a consequence, in the former case the Principal obtains a higher number of patients treated (i.e. closer to the first-best level), while in the latter one he obtains a lower level of patients treated (i.e. more distortion from the first-best level). On the other hand, since the informational rent that the Principal leaves to the specialist in the *low cost state* depends on the level of patients the specialist is asked to treat in the *high cost state*, it follows that when the optimal contract implements a higher level of patients to be treated in the high cost state, the Principal has to leave a higher informational rent to the specialist in the low cost state. Hence, in the contract for specialist B, the conflict between optimal provision of incentives, on the one hand, and efficient rent extraction, on the other, are exacerbated when the two random variables, θ_j^B and θ_i^A, are positively correlated. Similar considerations can be drawn for specialist A.

Note that the second-best levels of both elective and emergency services depend on the ratio $\frac{\phi_{li}}{\phi_{hi}}$, and that the smaller the ratio $\frac{\phi_{li}}{\phi_{hi}}$, the smaller will be the distortions from the first-best levels. In fact, if strong positive correlation is in place so that $\lim \phi_{hh}, \phi_{ll} \to 1$, and $\lim \phi_{lh}, \phi_{hl} \to 0$, it follows that the departures from the first-best level will be lower.

Using the insights of Sappington and Demski (1983), it can be seen that when correlation between agents' private information is perfect the first best numbers of patients treated are obtained. For example in the case of known perfect positive correlation which implies that $\phi_{hh}, \phi_{ll} = 1$ (i.e. both specialists are either of a high cost or of a low cost) and $\phi_{lh}, \phi_{hl} = 0$,

the optimal number of patients treated in this case is equal to the optimal number of patients in the symmetric information. Hence:

$$x_{ll}^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + (1+\lambda)\,\tau\frac{V^x}{K^x}\right] - \frac{2\,(1+\lambda)}{(1-\lambda)}\left[\theta_l^A + (1-\tau)r\right] \tag{32}$$

$$x_{hh}^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + (1+\lambda)\,\tau\frac{V^x}{K^x}\right] - \frac{2\,(1+\lambda)}{(1-\lambda)}\left[\theta_h^A + (1-\tau)r\right] \tag{33}$$

and

$$y_{ll}^{B*} = \frac{2}{(1-\lambda)}\left\{Q^y + V^y\left[1+\omega\,(1+\lambda)\right]\right\} - \frac{2\,(1+\lambda)}{(1-\lambda)}\theta_l^B \tag{34}$$

$$y_{hh}^{B*} = \frac{2}{(1-\lambda)}\left\{Q^y + V^y\left[1+\omega\,(1+\lambda)\right]\right\} - \frac{2\,(1+\lambda)}{(1-\lambda)}\theta_h^B \tag{35}$$

This is due to the fact that, when private information is known to be perfectly positively correlated, the number of PCs and ICCs, and thus the number of optimal transfers and first order conditions, is reduced from four to two cases per specialist so that only the above cases are obtained. Similarly for when information is perfectly negatively correlated so that

Conclusion 14 *When information is asymmetric but the specialists costs variables are perfectly correlated, and the government knows how the types relate, then he can devise a contract that extracts any possible rents due to the asymmetry of information and can implement the first-best levels of outputs.*

Summarising, information asymmetries can enhance the negative behaviour of specialists leading to higher transfer being paid by the government or fewer patients treated. Altruistic and 'selfish' behaviour have respectively a positive and negative impact on the number of patients treated and 'selfish' behaviour increases waiting time for elective surgery. If it is possible for the government to obtain better information on how specialists' capabilities relate to each other then the negative impact of asymmetry of information may be reduced.

5 Conclusions and Policy Discussion

This paper looked at the role hospital specialists play in managing the UK NHS waiting lists, and the effect of their behaviour in determining numbers treated and waiting times for elective surgery when perverse incentives may be in place.

In the NHS UK specialists may be able to influence waiting lists and waiting times for elective surgery as they own and manage those lists. In such a context, the possibility of conducting private elective activity, which often more than duplicates their earnings, may create perverse incentives not observed in other activities such as emergency surgery which

in the paper serves as a term of comparison. When dealing with elective surgery specialists may have conflicting objectives in that a longer NHS waiting list and time may contribute to increase the demand for their private practice, and thus their earnings. They may have the incentive to treat fewer NHS patients. It is also argued that the current remuneration system does not contribute to increase production and/decrease NHS waiting times or lists, and the introduction of some element of performance related payment in the wage structure is currently being discussed by the English government. Hence, the relevance of this study.

Using a Principal - Multi Agent framework, we analysed the relationship between the government, as the health care purchaser and the principal of a two-tier hierarchy, and two hospital specialists that accomplish two tasks separately: elective and emergency surgery. We explored the motivation of both the government and the specialists and analysed the impact of altruistic and 'selfish' behaviour by specialists on the number of patients treated and waiting times.

The presence of altruism increases the number of patients treated because more attention is paid to the disutility generated by turning emergency patients down or to the disutility generated by the time spent waiting by elective patients. The presence of 'selfish' behaviour, that is, the perverse interest specialists have in the income they can obtain from their private practice was found to lead to: a) a decrease in the optimal number of patients treated as NHS elective surgery cases; and b) an increase the waiting time NHS elective care patients face. Moreover, the presence of asymmetric information helped emphasizing the perverse behaviour. Asymmetry of information (which a reasonable hypothesis in the context where the government contracts with specialists) implied that a greater transfer had to be paid to the high skills specialist so as to have the first best number of patients treated, or led to fewer patients being treated.

Moreover, 'selfish' behaviour when dealing with elective care led to an increase in health care costs due to an increase in the state transfers that the government had to pay so as to provide a financial incentive for specialists to treat more patients within the NHS. Otherwise, and if information on private practice is available a reduction in the transfers paid to specialists proportional to what they obtain with their private practice must be put in place.

We believe these are very important results with important policy implications, namely that if remuneration of NHS specialists is to be in some way dependent on their levels of NHS activity, then

a) Remuneration should take into account the level of income the specialist might receive from private practice. Thus, the government should created the means to obtain information on the specialists' private earnings.

b) The existence of imperfect information on the part of the government implies the need for higher payments to specialists. Therefore, there may be benefits for investing in improved information systems on specialists' workload within the NHS.

c) The presence of altruism on the part of specialists leads to increased levels of production and decreased waiting times under all types of organisation. There is therefore a case for

seeking out instruments for nurturing more altruistic behaviour on the part of specialists.

We believe that this analysis contributed to the debate over waiting times and waiting lists within the NHS. It used a Principal - Multi Agent approach to study the influence of specialists' private practice upon the number of patients treated in the NHS and NHS waiting times for elective surgery. We formalised and analysed how different motivations on the part of specialists when dealing with different activities, as well as their skills/knowledge and their effort influenced the number of NHS patients treated and the waiting time patients faced. We believe that with this study we added some more insights into the political and social debate over the presence of perverse incentives in specialists behaviour which is at the centre of recent governmental discussion on the reform of NHS specialists' remuneration. We hope to have contributed to the development of a better informed policy making process.

A number of issues are left for future research and shall be tackled in time. They include: having a hospital manager as the principal (which may have more information then the government); comparing the separation of tasks with a multitask organisation; analysing the case of having a vertical organisation where the government contracts with a hospital manager that then contracts with a specialist; and finally considering specialists cooperation or collusion in their activities.

6 References

Arrow, K. J., (1985), "The economics of agency.", In J. Pratt and R. J. Zeckhauser (Eds.), *Principals and Agents: The Structure of Business.*, pp. 37–51, Harvard Business School Press.

Barzel, Y., (1974), "A theory of rationing by waiting", *The Journal of Law and Economics*, 17, 73-95.

Beeby, N. R., Nicholl, J. P., and Williams, B. T., (1989), "Role of the private sector in elective surgery in England and Wales.", *BMJ*, 298, 243–247.

Besley, T., Hall, J. and Preston, I., (1999), "The demand for private health insurance: Do waiting lists matter?", *Journal of Public Economics*, 72, 155-181.

Bloor, K. and Maynard, A., (1992), "Rewarding excellence? Consultants' distinction awards and the nedd for reform", *DP 100, Centre for Health Economics*, University of York.

Bloor, K. and Maynard, A., (1993), "Expenditure in the NHS during and after the Tatcher's years: Its growth and utilisation.", *DP 113, Centre for Health Economics*, University of York.

Bloor, K., Maynard, A., and Street, A., (1992), "How much is a doctor worth?", *DP 98, Centre for Health Economics*, University of York.

Bosanquet, N., (1988), "An ailing state of National Health", in : Jowell, R., Brook, L. and Witherspoon, S. (eds), *British Social Attitudes Survey.*

Bös, D., (1994), "Pricing and price regulation. An economic theory for public enterprises and public utilities.", In C. J. Bliss and M. D. Intrilligator (Eds.), *Advanced Textbooks in*

Economics, volume 34. Amsterdam: Elsevier.

Cullis, J. G., Jones, P. R., and Propper, C., (2000), "Waiting lists and medical treatment: Analysis and policies.", In *Handbook of Health Economics*, volume 1, chapter 23, pp. 1202–1249, Elsevier Science B. V.

Demski, J.S and Sappington, D.E.M., (1984), "Optimal incentive contracts with multiple agents", *Journal of Economic Theory*, 44, 156-167.

Department of Health, (1992), "*The Patient's Charter and You - A Charter for England.*"

Department of Health, (1999), "*Saving Lives: Our Healthier Nation.*", White Paper

Department of Health, (2000), www.doh.gov.uk., Web Site.

Frankel, S. and West, R., (1993), "What is to be done?", In S. Frankel and R. West (Eds.), *Rationing and Rationality in the National Health Service: The Persistence of Waiting Lists*, chapter 7, pp. 115–131, London: MacMillan.

Gaynor, M., (1994), "Issues in the industrial organisation of the market for physician services.", *Journal of Economics and Management Strategy*, 3(1), 211–255.

Goddard, J. A., Malek, M. and Tavakoli, M., (1995), "An economic model of the market for hospital treatment for non-urgent conditions", *Health Economics*, 4(1), 41-55.

Gravelle, H., Smith, P. and Xavier, A., 2000, Waiting times and waiting lists: A model of the market for elective surgery., *DP 2000/27, DERS, University of York.*

Grossman, S. J. and Hart, O. D., (1983), "An analysis of the principal-agent problem.", *Econometrica*, 51(1), 7–45.

Hamblin, R., Harrison, A., and Sean, B., (1998), "*Access to Elective Care: Why Waiting Lists Grow.*", King's Fund Institute.

Holmstrom, B. and Myerson, R., (1983), "Efficient and durable decision rules with incomplete information", *Econometrica,* 51, 1799-1819

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

Jones, A. and Zanola, R., (1995), "Agency in health care", paper presented at a meeting on "*Public decision-making processes and asymmetry of information*", University of Catania.

Laffont, J.J. and Tirole, J., (1993), "*A Theory of Incentives in Procurement and Regulation*", Cambridge and London: MIT Press.

Laing and Buisson, (1994), "*Laing's Review of Private Health Care.*", Laing and Buisson Publications.

Laing and Buisson, (1999), "*Laing's Review of Health Care Market.*", Laing and Buisson Publications.

Lerner, C. and Claxton, K., (1994), "Modelling the behaviour of general practitioners.", *DP 116 Centre for Health Economics*, University of York.

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

Martin, S. and Smith, P., (1999), "Rationing by waiting lists: An empirical investigation.", *Journal of Public Economics*, 71, 141–164.

Mas-Colell, A., Whinston, M. D., and Green, J. R., (1995), "*Microeconomic Theory.*", Oxford University Press, 1st edition.

Monopolies and Mergers Commission, (1993), "*Private Medical Services: A Report on Agreements and Practices Relating to Charges for the Supply of Private Medical Services*", London: HMSO.

Mooney, G. and Ryan, M., (1993), "Agency in health care: Getting beyond first principles.", *Journal of Health Economics*, 12(2), 125–135.

Myerson, R., (1979), "Incentive compatibility and the bargaining problem", *Econometrica,* 47, 61-73

NHS Executive, (1999), "*Agenda for Change - Modernising the NHS Pay System.*".

Office of Health Economics, (1999), "*Compendium of Health Statistics*", 11th Edition.

Propper, C., (1998), "The demand for private health care in the UK.", *Working Paper 98/004, CMPO*, University of Bristol.

Sappington, D. and Demski, J. S., (1983), "Multi-agent control in perfectly correlated environments.", *Economics Letters*, 13, 325–330.

Scott, A., (1996), "Agency, incentives and the behaviour of general practitioners: The relevance of principal agent theory in designing incentives for GPs in the UK.", *DP 03.96. HERU*, University of Aberdeen.

Scott, A., (1997), "Designing incentives for GPs. a review of the literature on their preferences for pecuniary and nonpecuniary job characteristics.", *DP 01.97 HERU* , University of Aberdeen.

Tirole, J., (1988), "*The Theory of Industrial Organisation*", London: MIT Press.

Yates, J., (1987), "*Why Are We Waiting? An Analysis of Hospital Waiting Lists*", Oxford: Oxford University Press.

Yates, J., (1995), "*Private Eye, Heart and Hip*", London: Churchill Livingstone.

A Appendix

A.1 Some comparative statics

We can derive some comparative statics on the optimal numbers treated that are as follows, with $i = h, l$ and $j = h, l$:

in the case of hospital specialist B:

$$\frac{\partial y_j^{B*}}{\partial \theta_j^B} = -\frac{2\left(1+\lambda\right)}{\left(1-\lambda\right)} < 0; \quad \frac{\partial y_j^{B*}}{\partial V^y} = \frac{2}{\left(1-\lambda\right)}\left[1+\omega\left(1+\lambda\right)\right] > 0;$$

$$\frac{\partial y_j^{B*}}{\partial Q^y} = \frac{2}{\left(1-\lambda\right)} > 0; \quad \frac{\partial y_j^{B*}}{\partial \omega} = \frac{2\left(1+\lambda\right)}{\left(1-\lambda\right)} V^y > 0;$$

$$\frac{\partial y_j^{B*}}{\partial \lambda} = \frac{2}{\left(1-\lambda\right)^2}\left(Q^y + V^y + 2\omega V^y - 2\theta_j^B\right) = +/-$$

in the case of hospital specialist A:

$$\frac{\partial x_i^{A*}}{\partial Q^x} = \frac{2}{(1-\lambda)} > 0; \quad \frac{\partial x_i^{A*}}{\partial V^x} = \frac{2}{(1-\lambda)}(\frac{D^x}{K^x} + (1+\lambda)\frac{\tau}{K^x}) > 0;$$
$$\frac{\partial x_i^{A*}}{\partial \theta_i^A} = -\frac{2(1+\lambda)}{(1-\lambda)} < 0; \quad \frac{\partial x_i^{A*}}{\partial K^x} = -\frac{2}{(1-\lambda)}\left[\frac{V^x D^x}{(K^x)^2} + \tau\frac{V^x}{(K^x)^2}\right] < 0;$$
$$\frac{\partial x_i^{A*}}{\partial \lambda} = \frac{2}{(1-\lambda)^2}\left(Q^x + \frac{V^x D^x}{K^x} + 2\alpha\frac{V^x}{K^x} - 2\theta_i^A - (1-\tau)r\right) < 0;$$
$$\frac{\partial x_i^{A*}}{\partial r} = -\frac{2(1+\lambda)}{(1-\lambda)}(1-\tau) < 0; \quad \frac{\partial x_i^{A*}}{\partial \alpha} = \frac{2(1+\lambda)}{(1-\lambda)}\left(\frac{V^x}{K^x} + r\right) > 0;$$
$$\frac{\partial x_i^{A*}}{\partial D^x} = \frac{2}{(1-\lambda)}\frac{V^x}{K^x} > 0; \quad \frac{\partial x_i^{A*}}{\partial(1-\tau)} = -\frac{2(1+\lambda)r}{(1-\lambda)} < 0;$$
$$\frac{\partial x_i^{A*}}{\partial(1-\varepsilon)} = -\frac{2(1+\lambda)(1-\tau)r}{(1-\lambda)} < 0;$$

And the signs are as expected.

A.2 The alternative interpretation of specialist A preferences

When in the context of asymmetry of information and with the 'selfish' behaviour defined as $(1-\tau)r\left(D^x - x_i^A\right)(1-\varepsilon)$, the PCs and the ICCs are:

$$T_i^A - \theta_i^A x_i^A + \frac{\left(x_i^A\right)^2}{4} - \tau V^x\left(\frac{D^x - x_i^A}{K^x}\right) + (1-\tau)r\left(D^x - x_i^A\right)(1-\varepsilon) \geqslant 0$$

and

$$T_i^A - \theta_i^A x_i^A + \frac{\left(x_i^A\right)^2}{4} - \tau V^x\left(\frac{D^x - x_i^A}{K^x}\right) + (1-\tau)r\left(D^x - x_i^A\right)(1-\varepsilon)$$
$$\geqslant T_q^A - \theta_i^A x_q^A + \frac{\left(x_q^A\right)^2}{4} - \tau V^x\left(\frac{D^x - x_q^A}{K^x}\right) + (1-\tau)r\left(D^x - x_q^A\right)(1-\varepsilon)$$

and the optimal number of elective cases treated is:

$$x_l^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + \frac{(1+\lambda)\tau V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[(1-\tau)r(1-\varepsilon) + \theta_l^A\right] \tag{36}$$

$$x_h^{A*} = \frac{2}{(1-\lambda)}\left[Q^x + \frac{V^x D^x}{K^x} + \frac{(1+\lambda)\tau V^x}{K^x}\right] - \frac{2(1+\lambda)}{(1-\lambda)}\left[(1-\tau)r(1-\varepsilon) + \theta_h^A + \frac{p_l(\theta_h^A - \theta_l^A)}{(1-p_l)}\right] \tag{37}$$

B Appendix

B.1 Proofs of Binding Constraints in the Optimal Contract

We solve the problem for the case in which specialist A observes that specialist B is of a low cost type. Looking at the constraints relevant to the transfers to specialist A, for the case in which the specialist B is in the efficient state of the world, i.e. when $\theta^B = \theta_l^B$, the ICC for A in the low cost state of the world is:

$$\begin{aligned} &T_{ll}^A - \theta_l^A x_{ll}^A + \frac{(x_{ll}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{ll}^A}{K^x}\right) - (1-\tau) r x_{ll}^A \\ &\geq T_{hl}^A - \theta_l^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x}\right) - (1-\tau) r x_{hl}^A + \end{aligned}$$

And the PC in the high cost state of the world is:

$$T_{hl}^A - \theta_h^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x}\right) - (1-\tau) r x_{hl}^A \geq 0$$

The ICC in the low cost state of the world and the PC in the high cost state case imply (rearranging the terms) that the PC for the low cost state of the world would not be binding at the optimum:

$$\begin{aligned} &T_{ll}^A - \theta_l^A x_{ll}^A + \frac{(x_{ll}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{ll}^A}{K^x}\right) - (1-\tau) r x_{ll}^A \\ &\geq T_{hl}^A - \theta_l^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x}\right) - (1-\tau) r x_{hl}^A \\ &\geq T_{hl}^A - \theta_h^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x}\right) - (1-\tau) r x_{hl}^A \geq 0 \end{aligned}$$

Moreover, it can also be concluded that the PC for the high cost state of the world must be binding. Indeed, if the last term is not equal to zero the government could decreased the transfer to A by the same amount in both states (T_{ll}^A and T_{hl}^A) thus decreasing costs and increasing its utility. But if this is the case then it means that the transfers and the mechanism design was not the optimal. Hence, the PC must bind in the optimum, that is,

$$T_{hl}^A - \theta_h^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x}\right) - (1-\tau) r x_{hl}^A = 0$$

Finally, the ICC for the low cost type must also bind in the optimum. If it was not binding then the government could decrease the transfer paid when in low cost state of the world (T_{ll}^A) thus decreasing government costs and increasing its utility without breaking the PC of

the low cost type or the ICC for the low cost type or the ICC for the high cost type which still hold. Thus, in the optimum we must have:

$$T_{ll}^A - \theta_l^A x_{ll}^A + \frac{(x_{ll}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{ll}^A}{K^x} \right) - (1-\tau) r x_{ll}^A$$
$$= T_{hl}^A - \theta_l^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x} \right) - (1-\tau) r x_{hl}^A$$

The ICC for the high cost state of the world is:

$$T_{hl}^A - \theta_h^A x_{hl}^A + \frac{(x_{hl}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{hl}^A}{K^x} \right) - (1-\tau) r x_{hl}^A$$
$$\geq T_{ll}^A - \theta_h^A x_{ll}^A + \frac{(x_{ll}^A)^2}{4} - \tau V^x \left(\frac{D^x - x_{ll}^A}{K^x} \right) - (1-\tau) r x_{ll}^A$$

Summing up the ICCs we obtain the following condition:

$$0 \geq (x_{ll}^A - x_{hl}^A)\left(\theta_l^A - \theta_h^A\right)$$

Given that $\theta_h^A > \theta_l^A$ the expression holds for $x_{ll}^A \geqslant x_{hl}^A$, that is, output is decreasing in the cost. Moreover, if ICC for low cost is binding as is the PC for the high cost of the world it can also be shown (by substituting the expressions for the transfers T_{hl}^A and T_{ll}^A obtained from the ICC in the low cost state and the PC in the high cost state) that the ICC for the high cost state of the world will not be binding at the optimum, and can be neglected.

The same procedure can be followed to determine the binding constraints for the case in which A observes that B is of a high cost type and when applying the analysis to specialist B.