DISCUSSION PAPER SERIES

No. 2519

WAITING LISTS AND PATIENT SELECTION

Pedro Pita Barros and Pau Olivella

PUBLIC POLICY

Centre for Economic Policy Research

www.cepr.org

ISSN 0265-8003

WAITING LISTS AND PATIENT SELECTION

Pedro Pita Barros, Universidade Nova de Lisboa
Pau Olivella, Universitat Autònoma de Barcelona

Discussion Paper No. 2519
August 2000

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 2519

August 2000

ABSTRACT

Waiting Lists and Patient Selection*

We develop a model of waiting lists for public hospitals when physicians deliver both private and public treatment. Public treatment is free but rationed, i.e. only cases meeting some medical criteria are admitted for treatment. Private treatment has no waiting time but entails payment of a fee. Both physicians and patients take into account that each patient treated in the private practice schedule reduces the waiting list for public treatment. We show that physicians do not necessarily select the mildest cases from the waiting list. We provide sufficient conditions on the rationing policy under which cream skimming is always partial. We show that, to a large extent, one can by-pass the analysis of doctors' behaviour in the characterization of patient selection.

JEL Classification: H51, I11 and I18
Keywords: cream skimming and waiting lists

Pedro Pita Barros
Faculdade de Economia
Universidade Nova de Lisboa
Trav. Estêvão Pinto (Campolide)
P-1099-032 Lisbon
PORTUGAL
Tel: (351 21) 3380 1600
Fax: (351 21) 388 6073
Email: ppbarros@fe.unl.pt

Pau Olivella
Department of Economics and CODE
Universitat Autònoma de Barcelona
09193 Bellaterra
SPAIN
Tel: (34 93) 581 2369
Fax: (34 93) 581 2461
Email: polivella@volcano.uab.es

* This research has partially been financed by projects SGR 98-00062, DGES PB97-0181 and HP98-0057 (Pau Olivella) and Acções Integradas Luso-Espanholas and PRAXIS XXI 13132/1998 (Pedro Pita Barros).

Submitted 22 June 2000

NON-TECHNICAL SUMMARY

The existence of waiting lists is a main feature of several health systems. This phenomenon has been studied as a means to control demand when price cannot be used as a rationing device, either for ethical or political reasons. Waiting lists and their associated waiting time are much more important in the public than in the private sector.

We focus our analysis on non-urgent treatments for a fixed speciality, e.g. elective heart surgery. It is then justified to assume that patients are treated on a first-come/first-served basis (i.e. that they are not prioritized by severity). Although there exists an extensive informal as well as theoretical literature on prioritization, a large proportion of this literature deals with prioritization across specialities. Notice also that prioritization schemes vary widely from one country's (or even region's) health system to the other. Different weights are given to distinct criteria such as expected deterioration of health status in the absence of hospital treatment, expected improvement after treatment, number of economic dependents, possibility to afford the alternative private care, current health status, age and so on. The casuistics would make it very difficult to derive general conclusions.

A common feature of countries with public health services and waiting lists, is that there is coexistence, with the same doctor, of private and public practice. In particular, the South of Europe seems to have a higher intensity of this phenomenon, although it is present also in the UK. These two features, significant waiting lists in the public sector and doctors acting in both private and public sectors, raise a basic concern. The private sector may only offer treatment to the easiest (mildest) cases. That is, patient selection (cream skimming) may exist. Notice that the actual cream skimming actions (e.g. the rejection of patients who prefer the private sector) may not really take place at all. This may be for two reasons. First, doctors may set their private offices and equipment so that only certain health problems may be treated. Second, if patients know their own health status because some previous tests have been carried out, only those patients who know they will be accepted request treatment in the private sector.

Notice that the larger the segment of patients that are admitted to (or that can be treated in) the private sector, the shorter the segment of patients that remains for the public waiting list. Therefore, our concern with patient selection leads to consideration of a supply-side model of waiting lists. It contrasts with most health economics literature, which looks essentially at demand-side determinants of the waiting list.

Another common feature of public health systems is the existence of rationing. That is, only the more severe cases are allowed access to public hospital treatment, while other patients are offered an alternative non-hospital

treatment, like medication. For instance, if the parameter measuring visual acuity runs from zero to one, public cataract surgery may be rationed by excluding patients with more than 0.5 acuity.

In general, this policy divides the population into two groups. One group is formed by all individuals who satisfy the admission requirement. The other group is formed by all individuals who are denied access to the public waiting list and who must resort to either another type of treatment, or to the private hospital treatment.

We must point out that our analysis is not devoted to finding the optimal rationing policy. We rather take the rationing policy as given and analyze how the opportunities for cream skimming are affected by changes in the rationing policy. That is, our analysis is positive in nature, rather than normative.

We limit our analysis to the segment of patients who are admitted to the public waiting list. Therefore, all our cream-skimming definitions are made in reference to this segment. Thus, we define full cream skimming as the situation where all the mildest patients out of those who were admitted in the waiting list end up being treated in the private sector. Partial cream skimming, on the other hand, is the situation in which doctors treat, in their private practice, patients with an intermediate range (again in reference to the segment of admitted patients) of illness severities. Our most important result is that full cream skimming is only compatible with intermediate rationing policies. If, on the contrary, rationing is either very lax or very stringent then cream skimming will always be partial.

Surprisingly, the only assumption on doctors' preferences needed for this result is that, if doctors are willing to treat a case with a certain severity level, then doctors are also willing to treat any case with lower severity. Therefore, our analysis is consistent with doctors who care about patients' welfare. Our restriction would mean in this case that treatment costs rise sufficiently fast with severity. That is, cost saving motives linked to financial viability may be behind the observed cream-skimming behaviour. Our model differs considerably from the Feldstein-Pauly argument, which states that doctors want to have a list from which they pick the most interesting cases. Similarly, with respect to the queueing process, the only condition we impose is that the waiting list be governed by the first-come/first-served rule.

The intuition behind our main results is the following. For a very strict rationing policy, only the most severe cases are admitted. This induces short waiting lists. Consequently, the private sector becomes less attractive. People will be willing to wait in order to save the private sector fee.

If, on the other hand, even mild cases are admitted in the waiting list, that is, rationing is quite lax, then again cream skimming will be only partial. However, the reason is quite different from the previous one. With a loose rationing

criterion, many people are admitted in the waiting list. The waiting list will be long. Nonetheless, since there are patients in the waiting list with mild conditions, these patients will be willing to wait, because their cost of doing so is small.

We also show, by means of a numerical example, that one cannot in general ascertain whether private practice serves a population with a higher or a smaller average illness severity than the population that remains in the waiting list. Only in the particular case when cream skimming is full do we know that the private sector treats a lower average severity.

1 Introduction

The existence of waiting lists is a main feature of several health systems. This phenomenon has been studied as a means to control demand when price cannot be used as a rationing device, either for ethical or political reasons.[1]

Waiting lists and their associated waiting time are much more important in the public than in the private sector. For instance, Bosanquet (1999) states, for the U.K., that "at present, there is underocupation in private hospitals, with occupancy rates at 50% or less," while Laing and Buisson (1992) report that "there is no evidence that the National Health Service (NHS) [in the UK] consultants are short of time to do their private work" and that "they do not have waiting lists for private surgery even in London, where the ratio of private to NHS work is highest."[2]

On the contrary, waiting lists in the public sector may be quite long. For instance, waiting time in the UK for hip replacement could be as long as 2 years in 1987. Due to this, some patients are willing to pay for (almost immediate) treatment in the private sector. For instance, 20% of elective hart surgery and 30% of hip replacements in the UK were conducted in the private sector in 1995.

We focus our analysis on non-urgent treatments for a fixed specialty, e.g., elective heart surgery. It is then justified to assume that patients are treated on a first-come/first-served basis (i.e., that they are not prioritized by severity). Although there exists an extensive informal as well as theoretical literature on prioritization (see again Cullis, Jones and Propper's survey), a large proportion of this literature deals with prioritization across specialties.[3] Notice also that prioritization schemes vary widely from one

[1] See the survey by Cullis, Jones, and Propper (forthcoming) for an overview on both the empirical and the theoretical literature on hospital waiting lists.

[2] On a more informal note, Richmond (1996) also reports that when John Yates, author of a book on the interaction of the NHS with the private sector (Yates, 1995), telephoned 18 orthopedic surgeons in the Fall of 1995 "to seek an appointment as a NHS patient; only 4 could see him within 3 months, and for 7 of them the wait was between 6 months and 2 years. He then called as a private patient; 2 did not see private patients but the remaining 16 offered an appointment within 1 to 7 weeks; the average wait was 3.5 weeks, even though some consultants were on holiday."

[3] See, for instance, Coast (1996) on the Oregon Plan and the New Zealand Core Services

country's (or even region's) health system to the other. Different weights are given to distinct criteria such as expected deterioration of health status in the absence of hospital treatment, expected improvement after treatment, number of economic dependents, possibility to afford the alternative private care, current health status, age, and so on. The casuistics would make it very difficult to derive general conclusions. In this respect, Goddard and Tavakoli (1994) compare three alternative regimes in terms of the associated waiting times and their consequences for equity: (i) first-come/first-served, (ii) prioritize in proportion of severity, and (iii) high priority to the more severe cases while imposing long waits to the less serious ill. These authors, however, do not consider the possibility of resorting to private practice. Notice that the choice between waiting in the public sector and paying in the private sector is very much blurred if patients in the public waiting list for a given specialty are prioritized.[4] Examples of non-urgent treatments with long waiting lists are, apart from the two examples already cited, the surgery of cataracts, herniae, varicose veins, or hemorrhoids.

A common feature of countries with public health services and waiting lists, is that there is co-existence, in the same doctor, of private and public practice. In particular, the South of Europe seems to have a higher intensity of this phenomenon, although it is present also in the UK. For instance, Richmond (1996) reports that "except in the poorest parts of Britain, there is one private hospital within a mile of every major district general hospital and this in turn means that consultant surgeons can easily gallop down the road and operate on their private patients." In fact, whenever we find some sort of public infrastructure with doctors, these same doctors also maintain some private practice. These two features, significant waiting lists in the public sector and doctors acting in both private and public sectors, raise a basic concern. The private sector may only offer treatment to the easiest (mildest) cases. That is, patient selection (cream skimming) may

List.

[4] As an illustration, in section 4.2 we discuss the effects of a simple prioritization scheme in the spirit of case (iii) in Goddard and Tavakoli (1994).

exist. Notice that the actual cream skimming actions (e.g., the rejection of patients who prefer the private sector) may not really take place at all. This may be for two reasons. First, doctors may set their private offices and equipment so that only certain health problems may be treated.[5] Second, if patients know their own health status because some previous tests have been carried out, only those patients who know they will be accepted request treatment in the private sector.[6]

Notice that the larger the segment of patients that are admitted to (or that can be treated in) the private sector, the shorter the segment of patients that remains for the public waiting list. Therefore, our concern with patient selection leads consideration of a supply-side model of waiting lists. It contrasts with most health economics literature, which looks essentially at demand-side determinants of the waiting list. There are some exceptions. Iversen (1993) analyzes the political game that hospitals and the Government play. If hospitals act as Stackelberg leaders, they may have an incentive to maintain longer waiting lists. The same author (Iversen, 1997) shows that cost savings concerns can also lead hospitals to maintain longer waiting lists in the presence of a private sector: longer waiting lists induce a shift of patients towards the private sector. Notice that, in both articles, waiting list length is never a choice variable of doctors.

Both Iversen (1997) and Goddard, Malek and Tavakoli (1995) concentrate on the patients' decisions. Both papers assume that all patients who are willing to pay for private treatment are served in the private sector. Therefore, they directly rule out the possibility of cream skimming on the part of the doctors. In fact, both authors treat doctors' decisions in a simple way, while doctor's strategic behavior plays an important role in our supply-side argument. Thus, our analysis is complementary to theirs.

More specifically, Goddard, Malek and Tavakoli (1995) develop a model

[5]Richmond (1996), when describing the private practice, reports that "many diagnostic procedures have to be done in the NHS facility, and post-operative complications, although rare, may require an ambulance ride to the NHS hospital."

[6]The issue of patient selection also arises in other circumstances. See the editorial of the Economist (1998), which addresses the criticisms directed towards health maintenance organizations in the U.S. for excluding costly cases.

with a specific functional form for the patient's utility that allows for income effects. Their emphasis is on developing comparative statics results for changes in the alternative treatment quality, the time discount rate, and other economic variables. The private supply side is determined exogenously by the number of patients that the private hospital is able to treat per period. Equilibrium determines simultaneously the price for treatment and the duration of wait in the public health service.

Another common feature of public health systems is the existence of rationing. That is, only the more severe cases are allowed access to public hospital treatment, while other patients are offered an alternative non-hospital treatment, like medication. For instance, if the parameter measuring visual acuity runs from zero to one, public cataract surgery may be rationed by excluding patients with more than 0.5 acuity (as it is the case in the public health system of the Vasc Country).

In general, this policy divides the population into two groups. One group is formed by all individuals who satisfy the admission requirement (visual acuity below 0.5 in the previous example). The other group is formed by all individuals who are denied access to the public waiting list and who must resort to either another type of treatment, or to the private hospital treatment.

We must point out that our analysis is not devoted to finding the optimal rationing policy. We rather take the rationing policy as given and analyze how the opportunities for cream skimming are affected by changes in the rationing policy. That is, our analysis is positive in nature, rather than normative.

We concentrate our analysis to the segment of patients who are admitted to the public waiting list. Therefore, all our cream-skimming definitions are made in reference to this segment. Thus, we define full cream skimming as the situation where all the mildest patients out of those who were admitted in the waiting list end up being treated in the private sector. Partial cream skimming, on the other hand, is the situation in which doctors treat, in their private practice, patients with an intermediate range (again in reference to

the segment of admitted patients) of illness severities. Our most important result is that full cream skimming is only compatible with intermediate rationing policies. If, on the contrary, rationing is either very lax or very stringent then cream skimming will always be partial.

Surprisingly, the only assumption on doctors' preferences needed for this result is that, if doctors are willing to treat a case with a certain severity level, then doctors are also willing to treat any case with lower severity. Therefore, our analysis is consistent with doctors who care about patients' welfare. Our restriction would mean in this case that treatment costs rise sufficiently fast with severity. That is, cost saving motives linked to financial viability may be behind the observed cream-skimming behaviour.[7] Similarly, with respect to the queueing process, the only condition we impose is that the waiting list be governed by the first-come/first-served rule.

The intuition behind our main results is the following. For a very strict rationing policy, only the most severe cases are admitted. This induces short waiting lists. Consequently, the private sector becomes less attractive. People will be willing to wait in order to save the private sector fee.

If, on the other hand, even mild cases are admitted in the waiting list, that is, rationing is quite lax, then again cream skimming will be only partial. However, the reason is quite different from the previous one. With a loose rationing criterion, many people are admitted in the waiting list. The waiting list will be long. Nonetheless, since there are patients in the waiting list with mild conditions, these patients will be willing to wait, because their cost of doing so is small.

We also show, by means of a numerical example, that one cannot in general ascertain whether private practice serves a population with a higher or a smaller average illness severity than the population that remains in the waiting list. Only in the particular case when cream skimming is full we know that the private sector treats a lower average severity.

[7] Our model differs considerably from the Feldstein-Pauly argument (Feldstein, 1970; Pauly, 1980) which states that doctors want to have a list from which they pick the most interesting cases.

The paper is organized as follows. In Section 2, we introduce the model. In Section 3, we provide the conditions that ensure that cream skimming is always partial, and characterize the equilibrium with cream skimming. Next, we characterize the equilibrium with cream skimming when the aforementioned conditions do not hold. In Section 4, we provide the numerical example and discusses the effects of a simple prioritization scheme on the equilibrium level of cream skimming. In Section 5, we make some final remarks. The proofs of the lemmata are relegated to the appendix while the proofs of the propositions are kept in the main text.

2 The Model

There exists a continuum of patients with mass $\overline{m}$. Each patient is characterized by an index of severity $m \in [0, \overline{m}]$ and by his initial wealth $a \in [0, \overline{a}]$. The index m and the initial wealth a are distributed independently. The density of m is $h(m) = 1$ (uniform) and the density of a is $g(a)$ with mean a_0. Hence, the joint density is given by $f(m, a) = g(a)$. The final utility of a patient is given by $\widetilde{V}(w, x, m) = x - \widetilde{c}(w, m)$, where w is the amount of time that the patient expects to wait from the moment he becomes ill until his discharge from the hospital and x is the final wealth of the individual. That is, if s is the cost of treatment, then the utility of a patient of type (m, a) is given by $a - s - \widetilde{c}(w, m)$. Since we restrict our attention to a specific illness or specialty, the assumption that the fee s is independent of severity is justified. Each severity m generates an arrival process of new patients seeking admittance to the public hospital that is also independent of initial wealth. We do not model explicitly the arrival process.[8]

Once a patient is admitted for public hospital treatment, he may be offered the possibility to resort to a private treatment. Moreover, as an outside option, an alternative non-hospital treatment exists with treatment cost equal to $s' \geq 0$. For instance, if the non-hospital treatment is public then $s' = 0$. The non-hospital treatment cures the patient after a length of

[8] See Worthington (1987, 1991) on this.

time equal to $\overline{t}$.

A single public hospital must serve two types of agents: those patients that are elligible for public treatment and prefer it over the private treatment, and those patients who, although they prefer (or are willing to pay for) the private treatment, they are not offered that option.[9] The set of all patients admitted to the public waiting list is referred to as the *total calling population*, which we denote by Ω. A patient with severity m has associated service time (that is, the time elapsed between admission into the hospital and discharge) $\tau(m)$. We do not impose any restrictions on the function τ other than it must take positive values. (For instance, if cesarean section has a shorter service time than natural birth, one could say that τ is decreasing in the severity index.)

We assume that there is no idle time during public-treatment hours. Therefore, the time t that is necessary necessary to treat all patients in the set Ω is given by

$$t = \int_0^{\overline{a}} \int_{m\in\Omega} \tau(m) f(m,a) dm = \int_{m\in\Omega} \tau(m) dm, \tag{1}$$

whenever this integral is well defined. Let $T(z) \equiv \int_0^z \tau(z')dz'$ be an auxiliary function. It denotes the time to treat all patients with severity distributed in the interval $[0, z]$.

The expected waiting time for any patient with severity m is a function of t and m, which we denote by $G(t, m)$. The form of the function G depends on the process of arrival of patients.[10] For instance, in the discrete case, if all patients in Ω arrive at the beginning of the period in a random order, i.e., independently of their severity, then $G(t, m) = \tau(m) + t/2$. As patients become atomistic, this reduces to $G(t, m) = t/2$. Other possibilities can be considered. The only requirement we impose is that G be a one-to-one mapping.

We can re-define the patients' utility function as follows. Let $c(t, m) \equiv \widetilde{c}(G(t, m), m)$, that is, the total wait in the public sector enters directly into

[9] This will be an equilibrium feature of the model. It is not an assumption.

[10] The arrival process of patients is assumed to be independent of severity.

the utility function. Similarly, let $V(t, x, m) \equiv x - c(t, m)$. Since there is a one-to-one relationship between expected waiting time and the total wait in the public sector, t, we refer to the first argument of the utility function as waiting time in the remainder. We use the following assumptions on the function c.

Assumption 1

$$\frac{\partial c}{\partial t} > 0, \qquad \frac{\partial c}{\partial m} > 0, \text{ and } c(t, m) \text{ is continuous and twice differentiable.}$$

We require that utility cost be increasing in both waiting time and severity. Also, it would be natural to think that marginal cost of waiting is increasing in the illness severity. This would ensure that the more severe the condition of a patient is, the more he is willing to pay for a reduction in the waiting time. Nonetheless, we do not need to impose this.

An important assumption is the following.

Assumption 2

$$c(0, m) = c(t, 0) = 0, \forall t \text{ and } \forall m \in [0, \overline{m}].$$

In other words, we are assuming that waiting time measures the wait until full recovery (so that $c(0, m) = 0$ for all m) and that a patient with severity $m = 0$ is not ill (so that $c(t, 0) = 0$ for all t). This also implies that the wait itself does not cause any direct disutility other than having to put up with the illness during that period. Hence, almost no disutility is suffered if a patient is almost not ill. For a discussion on the difference between the direct disutility caused by queueing and the disutility caused by waiting for a cure, see Propper (1995).

The public hospital and its equipment are owned and run by the public health system. The specific contracts that align the government's objectives with those of the hospital managers are ignored here.

The advantage of being treated in the private practice is that waiting time is smaller. We assume, for simplicity, that waiting time in the private

practice is zero. The advantage of being treated in the public practice is that the treatment is free, whereas the private practice costs $s > 0$ to the patient. Consequently, even if the patient has ex-ante contracted some private insurance, the actual payment if treatment is received is positive. This allows to accommodate in our framework diverse institutional settings. The existence of a National Health System, with an alternative private practice available, fits in our description. And it does so irrespective of supplementary health insurance contracted by patients. Also, private health plans that define preferred providers, to which the patient pays nothing (full insurance), and "outside" providers, to which some copayment is due, falls within our stylized framework.

Anyhow, for exposition purposes, we refer to s as the treatment fee. The fee s is chosen by an outside institution, in accordance to the particular setting considered.

We model physicians' behavior as that of a single representative agent. We call this agent "the doctor." She works both in the public sector, by treating patients in the public hospital (in the morning, say) and in the private hospital (in the afternoon, say). The private and public practice may even be in the same hospital, under different types of contracts.

The only assumption we need in reference to doctor's preferences is the following.

Assumption 3 *If the doctor is willing to treat, in her private practice, a patient with severity m', then she is also willing to treat any patient with severity $m \leq m'$.*

This assumption implies the following statement: "the cost of treating a patient with severity m' is higher than the cost of treating a patient with severity $m < m'$ (even if $\tau(m') < \tau(m)$)." Thus, an implicit assumption is that treating more complex problems involves the use of more sophisticated resources. The use of such more sophisticated resources is more costly to the doctor (either in time he devotes to the case or the attention and skills required). If doctors are selfish and receive a flat fee per patient treated,

then the statement and Assumption 3 are equivalent. If doctors care for patient's welfare, then Assumption 3 requires that the cost of treating a patient increases with m sufficiently fast and/or the fee per treatment increases with m sufficiently slow.[11]

More importantly, Assumption 3 implies that, on the part of the doctor, the only relevant decision is the choice of the maximum level of severity she is willing to treat privately, which we denote by $m^{\max}$. Once $m^{\max}$ is chosen, all patients with illness severity up to $m^{\max}$ are offered the private treatment.

The timing of moves is as follows. First, third-party payers and/or the doctor set the fee $s > 0$, taken exogenously here.[12] Some institution (e.g., the public payer or the college of medicine), which we refer to as the administration, fixes the criteria for admittance into the public waiting list. Namely, a cutoff value $\widehat{m} < \overline{m}$ is chosen so that all patients with severity $m \in [\widehat{m}, \overline{m}]$ are admitted to the public waiting list. The value $\widehat{m}$ fully describes the rationing policy. This divides the population into two groups, the more severe (Group A - for "accepted"- from now on) and the less severe cases (group B from now on).

The doctor then chooses $m^{\max}$. The following terminology will be used throughout.

Definition 1 *If $\widehat{m} > 0$ we say that the access to the public hospital is rationed.*

The last movers are the patients. We concentrate our analysis on group A. Since we have assumed that patients are atomistic, they take waiting time as exogenous when they decide whether to leave the public system and choose the private care option. Since patients are perfectly informed about the environment, their expectation on waiting time in the public system and

[11] Notice that if Assumption 3 was not satisfied, then the concept of "cream-skimming" would be meaningless.

[12] If s is a copayment, it is usually set by the third-party payer. If it is a full payment in private practice, it is set by the doctor.

the true value coincide in equilibrium.[13] The timing of the game implies that the doctor acts as Stackelberg leader on patients' choices, while the administration acts as a Stackelberg leader on the doctor's choice of $m^{\max}$ and all the patients' choices. The administration choice of the rationing policy is not explicitly modelled.

We simplify the analysis by assuming that the waiting time for recovery in the non-hospital treatment is so lengthy that those admitted in the public service will never consider it. Namely, we assume that $\overline{t} > T(\overline{m})$ so that $c(\overline{t}, m) + s' \geq c(t, m)$ for all t in the interval $[0, T(\overline{m})]$ and for all m in $[0, \overline{m}]$. That is, even if the public waiting time presents its ever possible maximum ($t = T(\overline{m})$) any patient in group A prefers the public service to the alternative treatment (strictly, if $m < 0$). Hence, patients in group A choose between the public treatment and the private treatment.[14]

Given waiting time t, a patient with severity m will prefer the public sector if the cost of waiting is lower than the fee s to be paid in the private sector. Since the cost of waiting is increasing in the severity level (and it is zero for $m = 0$ by Assumption 2), then it exists a unique treshold value such that for severity levels above it, the patient chooses to go to the private sector, and below it stays in the public-sector waiting list. This cutoff point is given by $c(t, m^c) = s$. We call the patient with condition m^c the "indifferent patient". Without loss of generality, we assume that the indifferent patient stays in the waiting list.

The equilibrium is computed by solving the following system of equations for m^c and t.

$$\left.\begin{array}{c} c(t, m^c) = s \\ t = T(\overline{m}) - T(m^{\max}) + \min\left\{\max\{T(m^c), T(\widehat{m})\}, T(m^{\max})\right\} - T(\widehat{m}). \end{array}\right\} \tag{2}$$

The first equation defines the indifferent patient, as explained above. The

[13] It may seem somewhat heroic to assume perfect information by the patients. However, it seems quite reasonable to say that patients do have an accurate estimate on waiting times. This information may be conveyed by friends or word of mouth. This is the sense in which we use the perfect information assumption.

[14] Note that, for fixed $\widehat{m}$, the condition $\overline{t} > T(\overline{m}) - T(\widehat{m})$ is sufficient to guarantee that patients in group A discard the non-hospital treatment.

second equation defines waiting time as the sum of the lengths of intervals served by the public sector. The apparently complex form of the second equation is due to the need to account for different cases. Its utility will be clear below.

The following definitions set up the basic terminology and help to fully understand the second equation, by considering all relevant cases. Notice first that the problem is interesting only if $\widehat{m} \leq m^{\max} \leq \overline{m}$.

Definition 2

(a) If $m^c \geq m^{\max}$*, irrespectively of whether* m^c *is above or below* $\overline{m}$*, then* $t = T(\overline{m}) - T(\widehat{m})$*. We then say that* `the private sector is inactive`*.*

(b) If $m^c < m^{\max}$*, then we say that* `the private sector is active`*; moreover,*

(b1) If $m^c < \widehat{m}$*, then* $t = T(\overline{m}) - T(m^{\max})$ *and we say that* `doctors fully cream skim patients.`

(b2) If $\widehat{m} \leq m^c \leq m^{\max} < \overline{m}$*, then* $t = T(\overline{m}) - T(m^{\max}) + T(m^c) - T(\widehat{m})$*. The solution is interior and we say that* `doctors cream skim patients only partially`*.*

(b3) If $\widehat{m} \leq m^c \leq m^{\max} = \overline{m}$*, then we say that* `doctors do not cream skim patients.`

In case (a), all patients in group A who are offered private treatment prefer the public sector. The private sector is therefore inactive. In case (b), on the other hand, the private sector is active since $m^c < m^{\max}$ and $\widehat{m} \leq m^{\max} \leq \overline{m}$ jointly imply that the interval of patients treated in this sector, i.e., $[\max\{m^c, \widehat{m}\}, m^{\max}]$ is non-empty. In particular, in case $(b1)$, there exists a non-empty interval of patients $[\widehat{m}, m^{\max}]$ who resort to the private sector. Moreover, these patients present the least severe conditions in group A. This is why we say that there is full cream skimming. However, full cream skimming does not necessarily imply that *all* patients in group A

are diverted to private practice, since $m^{\max}$ may be below $\overline{m}$. In cases ($b2$) and ($b3$) all the patients in the non-empty interval $[\widehat{m}, m^c]$ reject the doctor's offer to resort to the private sector. Note, moreover, that these patients present the least-severe conditions. Therefore, they are the ones willing to wait more time for treatment and save payment s. In case ($b2$), there is still some patients (with $m \in [m^{\max}, \overline{m}]$) who are rejected by the private sector, so there is some degree of cream skimming. In case ($b3$), on the other hand, the private sector only treats the most severe cases and looses the mildest cases, so there is no cream skimming. Notice that cases (a), ($b1$), ($b2$) and ($b3$) are mutually exclusive and exhaustive given $\widehat{m} \leq m^{\max} \leq \overline{m}$.

Since $\partial c/\partial m > 0$ and $\partial c/\partial t > 0$, the first equation in (2) defines a strictly decreasing schedule in the space (m^c, t), whereas the second equation is increasing in m^c. Thus, the system yields at most a unique solution, which we denote by $[m^c(m^{\max}), t(m^{\max})]$. To ensure existence of this solution we make the following technical assumption.

Assumption 4 *For any $\overline{c} > 0$, $c(t, m) = \overline{c}$ implies that $t \to 0$ as $m \to \infty$.*

In other words, the level curves of the waiting cost function have a horizontal asymptote at zero, Intuitively, we want to ensure that, if the waiting time is strictly positive, we can make the cost of waiting arbitrarily large by choosing a sufficiently severe condition.[15]

Figure 1 illustrates equations system (2), where we have depicted both equations in the (m^c, t) space for the case when the solution is interior (i.e., in case ($b2$), where $\widehat{m} < m^c(m^{\max}) \leq m^{\max} < \overline{m}$).

[Figure 1 about here]

Figure 1 can be used to illustrate some comparative statics on the optimal patients' decisions. Interpret the loci $c(t, m^c) = s$ (the first equation in (2)) as the demand curve for public health, where waiting time

[15] Equivalently, we want to rule out the possibility that, for a fixed and positive level of waiting cost $\overline{c}$ there exists some positive waiting length t_0 such that, no matter how bad the medical condition is, waiting costs are always lower than $\overline{c}$. Notice that this assumption is compatible with Assumption 2.

plays the role of the price for public treatment. Similarly, interpret the loci $t = T(\overline{m}) - T(m^{\max}) + \min\{\max\{T(m^c), T(\widehat{m})\}, T(m^{\max})\} - T(\widehat{m})$ as the supply of public health care (as one allows for longer waiting times, the public sector is able to treat a larger segment of patients). An increase in s (keeping the other parameters fixed) induces an upward shift in the demand schedule. Thus, a larger fee results in a larger level of severity for the indifferent consumer (so fewer patients are diverted to the private sector) and a higher waiting time (because the waiting list is longer). On the other hand, an increase in $m^{\max}$ induces a right shift of the supply curve. This results in a shorter wait and again a higher severity of the marginal consumer. This means that the segment of patients treated in the private sector shifts up.[16]

These comparative statics refer to the patient's decision only. Full equilibrium comparative statics require positing a specific functional form for the doctor's objective function. As we opted for keeping the analysis at a general level, no further comparative statics results are reported.

The doctor chooses $m^{\max} \geq \widehat{m}$ taking into account that her decision will affect the calling population for the public waiting list. The following lemmata will be useful later on. They give some properties of the solution $(m^c(m^{\max}), t(m^{\max}))$.

Lemma 1 *The equilibrium value for m^c, that is, $m^c(m^{max})$, does not decrease with $m^{\max}$.*

The lemma establishes that if doctors set a higher value for the maximum severity level that they are willing to treat in the private sector, then the indifferent patient will also have a higher severity level. As $m^{\max}$ increases, waiting time decreases, and the patients with mildest conditions will be more willing to wait for free treatment.

Lemma 2 *If $\widehat{m} < m^c(m^{\max}) < m^{\max}$, then*

[16] We address the comparative statics of changes of $m^{\max}$ on the *length* (rather than the position) of that segment later on, when we give the derivative of the equilibrium $m^c(m^{\max})$ with respect to $m^{\max}$.

$$0 < m^{c\prime}(m^{\max}) = \frac{\frac{\partial c}{\partial t}\tau(m^{\max})}{\frac{\partial c}{\partial m} + \frac{\partial c}{\partial t}\tau(m^{c}(m^{\max}))}.$$

A sufficient condition for $m^{c'}(m^{\max}) < 1$ is that $\tau(m^c(m^{\max})) \geq \tau(m^{\max})$. That $m^{c'}(m^{\max}) < 1$ means that an increase in $m^{\max}$ triggers a less than proportional change in the severity level of the indifferent patient. In other words, an increase in $m^{\max}$ enlarges the set of patients treated in private practice whenever $m^{c'} < 1$.

Let us explain, using the discrete case version of our model, why $m^{c'}$ may be larger than one. Suppose that $m^{\max}$ is already far apart from m^c (perhaps because the doctor's fee per service is quite high). Suppose now that the doctor privately attends the next patient to the right of the most severe case in her private practice (i.e., $m^{\max}$ increases). Suppose also that the service time of this additional patient is extremely lengthy. Waiting time in the public sector is so greatly reduced that not only the patient who (before the change) was indifferent between sectors now prefers to wait, but also the two patients to his right may also prefer to wait.

It turns out that we can characterize with great detail the cream-skimming conditions in the health sector without having to find the explicit solution of the doctor's decision problem (the optimal $m^{\max}$).

This feature of the model allows us to avoid a detailed discussion on either the health production technology or the preferences of doctors. In particular, the qualitative results emerging from our analysis are robust to a variety of objective functions of doctors. Our approach encompasses both fully self-centered utility functions, according to which the doctor cares only about own rewards, as well as altruistic utility functions, in which a considerable weight may be attached to patient's welfare.

In order to characterize the extent of cream-skimming in equilibrium, it is first necessary to study the attitudes of patients towards potential offers of private practice when the private sector is inactive.

Whenever the private sector does not operate, the public waiting list presents its maximum length $t = T(\overline{m}) - T(\widehat{m})$. The system of equations

(2) reduces to

$$\left.\begin{array}{c} c(T(\overline{m}) - T(\widehat{m}), m^c) = s \\ t = T(\overline{m}) - T(\widehat{m}). \end{array}\right\} \tag{3}$$

Denote by $\widetilde{m}(\widehat{m})$ the solution for m^c in this case. Thus, $\widetilde{m}(\widehat{m})$ is the severity level of the consumer indifferent between being the first to go to the private practice and staying in the waiting list. In particular, for all $m^{\max} \leq \widetilde{m}(\widehat{m})$, the private sector is inactive and $m^c(m^{\max})$ is constant and equal to $\widetilde{m}(\widehat{m})$. The question to be solved is what happens if $m^{\max} > \widetilde{m}(\widehat{m})$. The following lemma is quite intuitive and it will be useful later on.

Lemma 3 *If $m^{\max} > \widetilde{m}(\widehat{m})$, then $m^c(m^{\max}) < m^{\max}$.*

Although an increase in scope of severities treated in the private sector makes the public sector more attractive (since this reduces the public waiting list), it is obvious that it can never be the case that this effect is so large that the private sector becomes inactive.

Define now $\phi(\widehat{m}) = c(T(\overline{m}) - T(\widehat{m}), \widehat{m})$. That is, $\phi(\widehat{m})$ is the cost of waiting for the patient with the mildest condition admitted to the waiting list when the waiting list is at its maximum, that is, when the private sector is inactive.

Two cases are possible. In Case I, $\phi(\widehat{m}) \leq s$. This implies that, even when the public queue is at its maximum length, the individual with the mildest condition in Group A (weakly) prefers the public sector. We then say that the private sector is relatively unattractive. This in turn implies that the severity of the individual who is indifferent between the two sectors must be higher than or equal to $\widehat{m}$. To sum up, $\widetilde{m}(\widehat{m}) \geq \widehat{m}$.

In Case II, $\phi(\widehat{m}) > s$. This implies that, the individual with mildest condition in Group A strictly prefers the private sector when the waiting list is at its maximum. We then say that the private sector is relatively attractive. This implies that the individual who is indifferent between sectors must have a severity below $\widehat{m}$. In other words, $\widetilde{m}(\widehat{m}) < \widehat{m}$.

We now deal with each of the two cases in more detail. In the first case, the first patient indifferent between public and private practice has a severity level higher than the critical threshold for admission into the public waiting list. We will show that only partial cream-skimming is possible in this case, no matter what the optimal decision of doctors regarding divertion of patients to public practice is. In the second case, the first patient indifferent between public and private practice has severity level below the admission threshold. Cream-skimming can be full or partial, depending on doctors' choice of the maximal level of severity they are willing to treat in the private sector. We now show these two claims.

2.1 Case I – The private sector is relatively unattractive

The next lemma provides sufficient conditions ensuring that we are in Case I (which occurs for $\widetilde{m}(\widehat{m}) \geq \widehat{m}$).

Lemma 4 *There exist $\widehat{m}^*$, $\widehat{m}^{**}$ in the open interval $(0, \overline{m})$ such that $\phi(\widehat{m}) \leq s$, and therefore $\widetilde{m}(\widehat{m}) \geq \widehat{m}$ for all $\widehat{m} \in [0, \widehat{m}^*] \cup [\widehat{m}^{**}, \overline{m}]$.*

This lemma states that there is a range of values for the threshold value $\widehat{m}$ such that private practice is not attractive for the patient with the mildest condition giving access to the waiting list. Patients will never accept proposals of moving from the public list to immediate private treatment. For this to happen, either $\widehat{m}$ must be sufficiently small or sufficiently high.

The following proposition is quite straightforward and it is one of the important results in the paper.

Proposition 1 *If the admission to the public sector waiting list is either too lenient or too strict, then doctors can cream skim patients only partially.*

Technically, the proposition can be restated as: if $\widehat{m}$ is sufficiently small ($\widehat{m} \leq m^{**}$) or sufficiently high ($\widehat{m} \geq m^{**}$), then, no matter what $m^{\max}$ doctors choose in $[\widehat{m}, \overline{m}]$, either the private sector is inactive or, if the private sector is active, doctors can cream-skim patients only partially. In

particular, for all $\widehat{m} \leq m^{\max} \leq \widetilde{m}(\widehat{m})$, the private sector is inactive, so $m^c(m^{\max}) = \widetilde{m}(\widehat{m})$ (which is independent of $m^{\max}$). On the other hand, for all $\widetilde{m}(\widehat{m}) < m^{\max} \leq \overline{m}$, the private sector is active and $m^{c'}$ is given in Lemma 2.

Proof. Recall first that if $\widehat{m}$ is sufficiently small or sufficiently large, then $\widetilde{m}(\widehat{m}) \geq \widehat{m}$ by Lemma 4. Suppose first that $\widehat{m} \leq m^{\max} \leq \widetilde{m}(\widehat{m})$. Then the private sector is inactive, and $m^c(m^{\max}) = \widetilde{m}(\widehat{m})$, independent of $m^{\max}$. Suppose now that $\widetilde{m}(\widehat{m}) < m^{\max} < \overline{m}$. Then, by Lemma 1, $m^c(m^{\max}) < m^{\max}$ and the private sector is active. This implies that $t(m^{\max}) < T(\overline{m}) - T(\widehat{m})$.

Now, since by definition $c(tm^{\max}), m^c(m^{\max})) = c(T(\overline{m}) - T(\widehat{m}), \widetilde{m}(\widehat{m})) = s$, then the last inequality implies that $m^c(m^{\max}) > \widetilde{m}(\widehat{m})$. To sum up, we have $\widehat{m} < \widetilde{m}(\widehat{m}) < m^c(m^{\max}) < m^{\max} \leq \overline{m}$. That is, cream skimming is only partial and Lemma 2 applies. ■

Intuitively, if $\widehat{m}$ is sufficiently small ($\widehat{m} < m^*$), that is, the rationing into the waiting list is lenient, patients with a mild condition are admitted into the waiting list. These patients being offered the option of private treatment will choose to stay in the waiting list. Their relatively good condition ensures that cost of waiting is smaller than the fee to be paid to private practice. On the other hand, for $\widehat{m}$ sufficiently large ($\widehat{m} > m^{**}$), that is, for a strict admission rule to the waiting list, only high-severity patients will be in the public sector. Thus, waiting time will be small, which decreases the relative attractiveness of the private sector. Patients are in this case more willing to wait to save the fee s, as the wait is not long.

The intuition is illustrated in Figure 2. Lemmata 3 and 1 are reflected in the fact that $m^c(m^{\max})$ is increasing in $m^{\max}$ but it never hits the 45 degree line.

[Figure 2 about here]

2.2 Case II. The private sector is relatively attractive

At the current level of generality, we cannot provide sufficient conditions to ensure that Case II occurs ($\widetilde{m}(\widehat{m}) < \widehat{m}$). In the remainder we assume that $\phi(\widehat{m}) > s$ holds and characterize how the degree of cream skimming depends on the doctor's decision on $m^{\max}$.

In this case, $\widehat{m} > \widetilde{m}(\widehat{m})$, and it turns out that the behavior of the function $m^c(m^{\max})$ is quite different depending on whether $m^c(m^{\max})$ is above or below $\widehat{m}$, except for the following lemma.

Lemma 5 *In Case II, the equilibrium $m^c(m^{\max})$ is a strictly increasing function of $m^{\max}$ in $[\widehat{m}, \overline{m}]$. Otherwise, for $m^{\max} \leq \widehat{m}$, there is no private practice.*

The intuition in this lemma is the same as the one in Lemma 1. The difference is that, in Lemma 1, $m^{\max}$ could take any value, including the values in the segment $[0, \max\{\widehat{m}, \widetilde{m}(\widehat{m})\}]$ where there is no scope for an active private sector. Consequently, $m^c(m^{\max})$ is constant instead of increasing. In Lemma 5, on the other hand, when we restrict attention to the case when $m^{\max} > \widetilde{m}$, then $\widehat{m} > \widetilde{m}(\widehat{m})$ (since we are in Case II), and the private sector is active.

We can now state another important result in the paper.

Proposition 2 *There exists a unique threshold value for $m^{\max}$ (the maximum severity patient a doctor wants to treat) such that:*

a) if the doctor's choice is below this threshold, the private sector is active and conducts full cream skimming;

b) if the doctor's choice is above (or equal to) this threshold, the private sector is active and in equilibrium cream-skimming is only partial.

Technically, there exists a unique $m_0^{\max}$ in the open interval $(\widehat{m}, \overline{m})$, such that

(i) For all $\widehat{m} \leq m^{\max} < m_0^{\max}$, the private sector is active and conducts full cream skimming. Moreover,

$$m^{c\prime}(m^{\max}) = \frac{\partial c/\partial t}{\partial c/\partial m} \cdot \tau(m^{\max}) > 0$$

(ii) For all $m_0^{\max} \leq m^{\max} \leq \overline{m}$, the private sector is active and conducts cream skimming that is only partial. Moreover, $m^{c\prime}(m^{\max})$ is given in Lemma 1.

Proof. Suppose first that $m^{\max} = \widehat{m}$. Then no private sector exists and $m^c(m^{\max}) = m^c(\widehat{m}) = \widetilde{m}(\widehat{m}) < \widehat{m}$, since we are in Case II. Suppose now that $m^{\max} = \overline{m}$. We show now that $m^c(\overline{m}) > \widehat{m}$.

Suppose, by contradiction, that $m^c(\overline{m}) \leq \widehat{m}$. Then, using the second equation in (2), we have $t = 0$ and, by the definition of $m^c(m^{\max})$, we have that $c(0, m^c(\overline{m})) = s > 0$. This contradicts Assumption 2.

To sum up, $m^c(\widehat{m}) < \widehat{m}$ while $m^c(\overline{m}) > \widehat{m}$. Since $m^c(m^{\max})$ is continuous and strictly increasing in $[\widehat{m}, \overline{m}]$ (by Lemma 5), there exists a unique value $m_0^{\max}$ in $(\widehat{m}, \overline{m})$ satisfying

$$m^c(m^{\max}) \begin{cases} < \widehat{m} \text{ for all } m^{\max} \text{ in } [\widehat{m}, m_0^{\max}), \text{ i.e., full cream-skimming;} \\ = \widehat{m} \text{ for } m^{\max} = m_0^{\max}, \text{ i.e., partial cream-skimming;} \\ > \widehat{m} \text{ for all } m^{\max} \text{ in } (m_0^{\max}, \overline{m}] \text{ i.e., partial cream-skimming.} \end{cases}$$

Now, for $m^{\max}$ in $[\widehat{m}, m_0^{\max})$, the total waiting time is

$$t = T(\overline{m}) - T(m^{\max}),$$

since

$$m^c(m^{\max}) < \widehat{m} \leq m^{\max}.$$

Therefore, $c(T(\overline{m}) - T(m^{\max}), m^c(m^{\max})) \equiv s$. The proof of part (i) is completed by differentiating this identity with respect to $m^{\max}$.

For $m^{\max}$ in $(m_0^{\max}, \overline{m}]$, we have $\widehat{m} < m^c(m^{\max}) < m^{\max}$, and Lemma 1 applies.■

The content of this Proposition is illustrated in Figure 3.

[Figure 3 about here]

3 Extensions

In this section we address two particular questions that have not been treated in our rather general approach. The first one is the determination of the average severity treated in each sector. The second one is the effects of prioritization.

3.1 Average severity

A typical criticism to waiting lists and divertion of patients to private practice is that only the best cases are captured by the private sector. The public waiting list then keeps all the complex (and costly) situations. We have shown this is not necessarily the case. But it could be that *on average*, even if cream skimming is only partial, severity mix is lower in private practice than in public pratice.

The attractiveness of thinking in terms of average severity is that it provides simple empirically testable implications. Unfortunately, one cannot state, in general, that average severity of cases treated is higher (or lower) in the public sector. That is, when one keeps the supply-side analysis of the waiting list at a general level, no precise prediction can be made on the relative position of average case mix in private and public sectors if partial cream skimming emerges in equilibrium.

We show this claim by way of a numerical example. Take the utility function to be:

$$U = A - tm. \tag{4}$$

Let service time be constant on m and equal to 1. Take $\overline{m} = 1$. That is, $T(z) = z$. The first equation in (2), defining the indifferent patient, m^c, between staying in the waiting list or going to the private practice becomes $s = tm^c$.

First, it is easy to check that we are in the interior case.[17] Therefore,

[17] To see this, it suffices to check that we are in Case I, or equivalently, that $c(T(\overline{m}) - T(\widehat{m}), \widehat{m}) < s$. This is straightforward by substitution.

the waiting time is given by

$$t = m^c - \widehat{m} + 1 - m^{\max}. \tag{5}$$

Solving (3) and (4), the equilibrium value m^c is defined implicitly by

$$s = (m^c + 1 - \widehat{m} - m^{\max})m^c. \tag{6}$$

Take the cost of treating patient with severity m as $k(m) = \alpha m$. The doctor is assumed to care only about self-interest.[18] Hence, his problem is given by:

$$\max_{\{m^{\max}\}} V = s(m^{\max} - m^c) - \alpha \int_{m^c}^{m^{\max}} m dm. \tag{7}$$

The first-order condition for this problem is

$$\frac{\partial V}{\partial m^{\max}} = s - \alpha m^{\max} - (s - \alpha m^c)\frac{\partial m^c}{\partial m^{\max}} = 0. \tag{8}$$

Straightforward substitutions yield that the equilibrium values of m^c and $m^{\max}$ solve the following two equations:

$$(s - \alpha m^{\max})(2m^c - \widehat{m} + 1 - m^{\max}) = (s - \alpha m^c)m^c,$$

$$s = m^c(m^c - \widehat{m} + 1 - m^{\max}).$$

Take now the following set of parameters ($\widehat{m} = 0.25; s = 0.55; \alpha = 0.55$). The numerical solution yields ($m^c = 0.821; m^{\max} = 0.902$). The average severity level in the public sector is 0.555 and in the private sector is 0.861. Thus, it is not true that, in general and on average, the private practice treats the mildest cases. The reverse is not true either. Change the private sector fee to $s = 0.25$. That is, resorting to the private sector becomes cheaper. The new equilibrium values are ($m^c = 0.362; m^{\max} = 0.423$).[19] Computation of average severity level gives a value of 0.645 in the public

[18] This is more restrictive than what we assumed above. However, to obtain an explicit solution, some assumption about the utility function of doctors is needed. We take the assumption that is the least favorable for cream-skimming to be only partial.

[19] It is easy to check, again, that $c(T(\overline{m}) - T(\widehat{m}), \widehat{m}) < s$.

sector and 0.393 in the private sector. Both examples exhibit partial cream skimming.

Thus, a priori, under partial cream skimming (case b2) we cannot state whether, on average, we should find easier cases in private practice, or not. This will be essentially an empirical matter. On the other hand, we know that average severity is higher in the public sector than in the private sector under full cream skimming (case b1), while the opposite is true under no cream skimming (case b3).

However, we believe that the questions that we are able to address, i.e., what is the condition of those patients that are rejected by the private sector and what is the condition of those patients who reject the private sector, are truly interesting. After all, it is quite difficult to observe average severity. That is why most accusations of cream skimming are based on the exclusion of certain cases by the private sector. This is exactly the issue that we have addressed here.

3.2 Prioritization

We have restricted our analysis to first-come/first-served waiting lists. One of the reasons, as explained in the introduction, is that prioritization criteria vary widely form instance to instance. We can however derive some implications of our model that could be useful for the prioritization discussion, by considering a simple case of prioritization by severity, in the spirit of case (iii) in Goddard and Tavakoli (1994).

Suppose that the administration fixes a second cutoff value $\widehat{\widehat{m}}$ in the closed interval $[\widehat{m}, \overline{m}]$ such that all patients in the segment $[\widehat{m}, \widehat{\widehat{m}}]$ are treated last (while keeping the first-come/first-served rule in both $[\widehat{m}, \widehat{\widehat{m}}]$ and $[\widehat{\widehat{m}}, \overline{m}]$). Then, we can repeat the same previous analysis for the segment $[\widehat{\widehat{m}}, \overline{m}]$; just replace $\widehat{m}$ by $\widehat{\widehat{m}}$ everywhere. (Notice that if $\widehat{\widehat{m}}$ is close to $m^{\max}$ then most of the individuals in $[\widehat{\widehat{m}}, \overline{m}]$ are not offered private treatment and therefore stay in the waiting list.) The question is what happens in the segment $[\widehat{m}, \widehat{\widehat{m}}]$. Perhaps in the world without prioritization these individuals would have stayed in the waiting list by rejecting the private offer, since they were in

mild condition and were not relegated to the end of the line. With prioritization, these individuals are forced to wait for a much longer time. They will now most probably accept the private offer. Thus, we can say that prioritization may reinforce cream-skimming practices.

Again, this discussion is made just for illustrative purposes. Prioritization may take much more complicated forms and be based on other criteria other than severity, like age, economic status, or prognosis (among many others). Each form will lead to very different cream-skimming outcomes. Any productive analysis will have to be made on a case-by-case basis.

4 Final remarks

We close the paper by stating once more our main contribution. We have put together doctors' and patients' decisions, and have asked to what extent are doctors able to cream skim patients. That is, we have taken a positive rather than a normative approach. Using a quite general model (most notably, without almost no restriction on doctors' preferences or on queueing processes), we have been able to characterize the equilibrium cream-skimming outcome.

The main conclusion is that, under quite general conditions, one should not observe full cream skimming (where all the mildest cases end up being treated in the private sector, while the worst cases remain in the waiting list). Instead, one should see a partial cream-skimming regime, under which doctors treat patients with an intermediate range of illness severities in their private practice. These conditions are that the rationing policy be either sufficiently lenient or sufficiently strict, where by rationing policy we mean the admission criteria applied to the public waiting list, and that patients pay a flat fee per service (independent of the level of severity) when resorting to the private sector. The full cream-skimming outcome is only compatible (if at all) with intermediate rationing policies.

The main intuition is straightforward. Although doctors may have an incentive to offer their services (in their private practice) to the lowest seg-

ment of severity, only the more severe in this subgroup are willing to accept to paying for private treatment. The rest will reject the private sector's offer. This intuition is reinforced whenever the rationing policy is either very lenient or very strict. In the first case, there are many patients in the waiting list in mild condition. In the second case, the waiting list is so short that only the extremely severe cases will accept the private sector's offer. If, on the other hand, the rationing policy is intermediate, the fact that waiting list is quite long and the fact that patients who are admitted are in quite severe condition jointly imply that even the mildest cases admitted in the waiting list opt for private treatment. We observe full cream skimming in this case.

There are some issues that we have not addressed in this paper. First of all, our model of the demand side is fairly simple. Our analysis is to be seen as complementary to the work of several other contributors to this literature, namely, Iversen (1997), Goddard, Malek and Tavakoli (1995), and the references in the survey by Cullis, Jones and Propper (forthcoming).

In addition, we have not addressed the interesting issue of the role of different remuneration schemes and rationing policies on doctor's choices, and the impact of these choices on the size and composition of the waiting list.

This analysis would be a necessary first step to determine the optional rationing policy, which we have taken as given. Notice, however, that this endeavour would require positing a specific preference profile on the part of doctors. This would have reduced the level of generality that we have been able to maintain throughout our analysis. In fact, we have shown that one can bypass the study of doctor's choices when characterizing to a large extent the composition of the waiting list. We consider this to be an important lesson of our approach.

References

Bosanquet, N. (1999). A Successful National Health Service, London: Adam Smith Institute.

Cullis, J., P. Jones and C. Propper (forthcoming). Waiting Lists and Medical Treatment: Analysis and Policies. In A. Culyer and J. Newhouse, editors, Handbook of Health Economics.

Feldstein, M.S. (1970). The Rising Price of Physicians' Services, Review of Economics and Statistics, 52: 121–131.

Economist, The (1991), Helth Care Survey, July 6, 1991, pp 1-22.

Iversen, T. (1993). A Theory of Hospital Waiting Lists, Journal of Health Economics, 12: 55–71.

Iversen, T. (1997). The Effect of a Private Sector on the Waiting Time of a National Health Service, Journal of Health Economics, 16: 381-396.

Goddard, J.A. and M. Tavakoli (1994). Rationing and waiting list management – Some efficiency and equity considerations. In M. Malek, editor, Setting Priorities in Health Care, John Wiley & Sons Ltd: Chichester.

Goddard, J.A., M. Malek, and M. Tavakoli (1995). An Economic Model of the Market for Hospital Treatment for non-Urgent Conditions. Health Economics 4:41-55.

Laing and Buisson, Ltd. (1992). Going Private. London: King's Fund.

Pauly, M.V. (1990). Doctors and their Workshops. Chicago: University of Chicago Press.

Propper, C. (1995). The Disutility of Time Spent on the United Kingdom's National Health Service Waiting Lists. Journal of Human Resources 30(4):677-700.

Richmond, C. (1996). NHS Waiting Lists Have Been a Boon for Private Medicine in the UK, Canadian Medical Association Journal, 154: 378–381.

Worthington, D. (1987). Queueing Models for Hospital Waiting Lists, Journal of the Operational Research Society, 38: 413–422.

Worthington, D. (1991). Hospital Waiting Lists Management Models, Journal of the Operational Research Society, 42: 833–843.

Yates, J. (1995). Private Eye, Heart and Hip: Surgical Consultants, the National Health Service and Private Medicine. London: Churchil Livingstone.

Appendix

Proof of Lemma 1

To show our claim, we will use the following facts.

`Fact 1:` $\forall x, y \in R, \min\{x, 0\} < \min\{y, 0\}$ implies that $x < y$.

To see this, if $x \leq 0$ and $y \leq 0$ (case 1), the implication follows directly. If $x \leq 0$ while $y > 0$ (case 2), then $\min\{x, 0\} < \min\{y, 0\}$ implies that $x < 0$. Therefore, $x < 0 < y$. If $x > 0$ and $y \leq 0$ (case 3), then $\min\{x, 0\} < \min\{y, 0\}$ implies that $0 < y$, which contradicts $y \leq 0$. Thus, case 3 is impossible. If $x > 0$ and $y > 0$ (case 4), then $\min\{x, 0\} < \min\{y, 0\}$ implies that $0 < 0$, an impossibility. Hence, also case 4 is impossible.

`Fact 2:` $\forall x, y, r, s \in R$, if $x > r$ and $y > s$ then $\max\{x, y\} > \max\{r, s\}$.

To see this, if $x \geq y$ and $r \geq s$ (case 1) then $\max\{x, y\} = x$ and $\max\{r, s\} = r$, and the implication follows directly. If $x \leq y$ and $r \leq s$ (case 2), then again the implication follows directly. If $x \geq y$ and $r \leq s$ (case 3), then $\max\{x, y\} = x \geq y > s = \max\{r, s\}$, and the implication is proved. If $x \leq y$ and $r \geq s$ (case 4), then $\max\{x, y\} = y \geq x > r = \max\{r, s\}$, and again the implication is proved.

Let's now proceed to the proof of the lemma. Suppose, by contradiction, that there exists $0 < a < b < \widehat{m}$ such that $m^c(a) > m^c(b)$. Then the following statements hold:

(i) $T(a) < T(b)$;

(ii) $T(m^c(a)) > T(m^c(b))$

(iii) $t(a) < t(b)$, since by the first equation in (2),

$$c(t(a), m^c(a)) = c(t(b), m^c(b)) = s,$$

while $m^c(a) > m^c(b)$.

Using the second equation in (2), and rearranging terms, (iii) can be rewritten as

(iii') $\min\{\max\{T(m^c(a)) - T(a), T(\widehat{m}) - T(a)\}, 0\} <$
$$< \min\{\max\{T(m^c(b)) - T(b), T(\widehat{m}) - T(b)\}, 0\}$$

Using Fact 1, this can still be rewritten as

(iii") $\max\{T(m^c(a)) - T(a), T(\widehat{m}) - T(a)\} <$
$$< \max\{T(m^c(b)) - T(b), T(\widehat{m}) - T(b)\}.$$

Now, (i) and (ii) imply

(iv) $T(m^c(a)) - T(a) > T(m^c(b)) - T(b)$;

(v) $T(\widehat{m}) - T(a) > T(\widehat{m}) - T(b)$.

Fact 2 tell us that (iv), (v) and (iii") are incompatible. The lemma follows. ■

Proof of Lemma 2:

Substitute the expression for t given in the second equation of (2) into the first equation, and differentiate totally with respect to $m^{\max}$.■

Proof of Lemma 3:

We prove the counterpositive. Suppose that $m^c(m^{\max}) \geq m^{\max}$. Then $\min\{\max\{T(m^c(m^{\max})), T(\widehat{m})\}, T(m^{\max})\} = T(m^{\max})$, and $t = T(\overline{m}) - T(\widehat{m})$ (the private sector is not active). Therefore, by definition, $m^c(m^{\max}) = \widetilde{m}(\widehat{m})$. Therefore $m^{\max} \leq \widetilde{m}(\widehat{m})$.■

Proof of Lemma 4:

By continuity, it suffices to show that $\phi(0) \leq s$ and that $\phi(\overline{m}) \leq s$. By Assumption 2, $\phi(0) = c(T(\overline{m}) - T(0), 0) = 0$ and $\phi(\overline{m}) = c(T(\overline{m}) - T(\overline{m}), \overline{m}) = c(0, \overline{m}) = 0$. ■

Proof of Lemma 5:

Suppose, by contradiction, that there exist $\widehat{m} \leq a < b \leq \overline{m}$ such that $m^c(a) \geq m^c(b)$. Then, the following statements hold:

(i) $T(\widehat{m}) \leq T(a) < T(b) < T(\overline{m})$;

(ii) $m^c(a) \leq a, m^c(b) < b$ since $b > a \geq \widehat{m} > \widetilde{m}\widehat{m}$, so Lemma 3 applies;

(iii) $T(m^c(a)) < T(a), T(m^c(b) < T(b)$, by (ii);

(iv) $T(m^c(a)) \geq T(m^c(b))$.

(v) Since by definition $c(t(a), m^c(a)) = c(t(b), m^c(b) = s$, we have $t(a) \leq t(b)$.

Use (iv) and (v) and apply (iii) to get

(vi) $T(m^c(b)) - T(b) < T(m^c(a)) - T(a) < 0$.

Use (i) to get (vii) $T(\widehat{m}) - T(b) < T(\widehat{m}) - T(a) < 0$.

Use (vi) and (vii) to rewrite (v) as

$$\min\{\max\{T(m^c(a) - T(a), T(\widehat{m}) - T(a)\}, 0\} \leq$$
$$\leq \min\{\max\{T(m^c(b)) - T(b), T(\widehat{m}) - T(b)\}, 0\}$$

or

(v') $\max\{T(m^c(a) - T(a), T(\widehat{m}) - T(a)\} \leq \max\{T(m^c(b)) - T(b), T(\widehat{m}) - T(b)\}$

The contradiction comes from (v'), (vi) and (vii), which are incompatible, by fact 2 in the proof of Lemma 1. ■

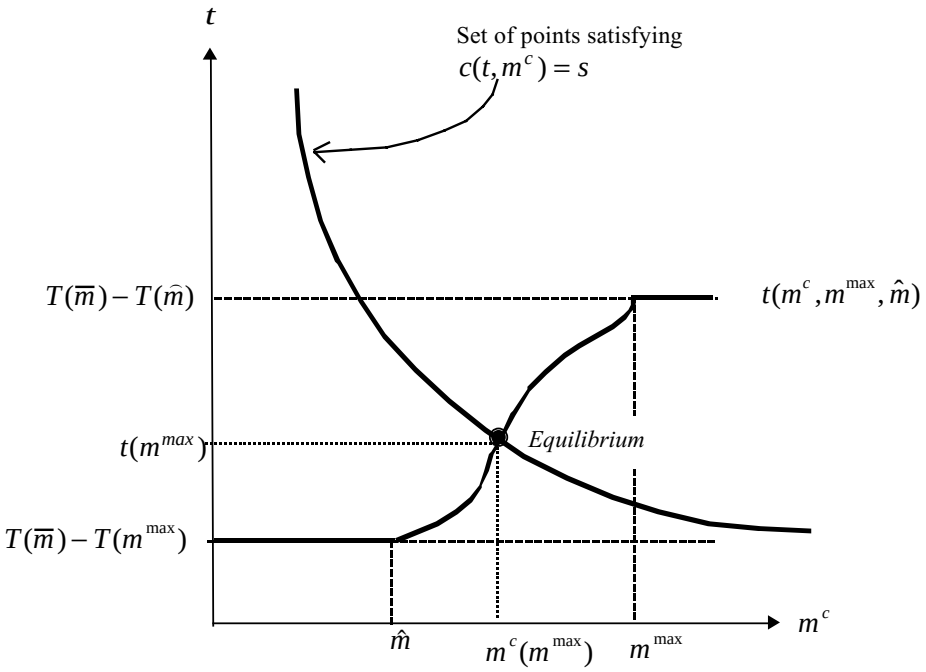

Figure 1. The equilibrium waiting time

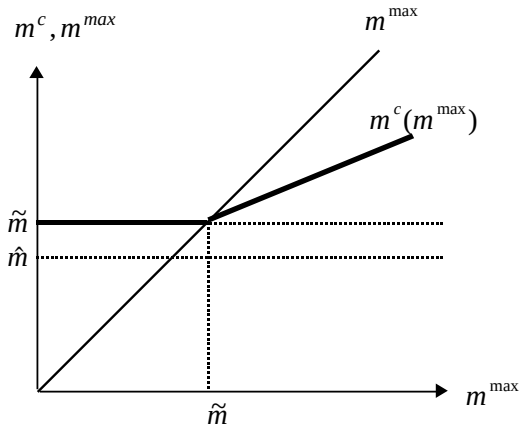

Figure 2. Case I: $\tilde{m} > \hat{m}$

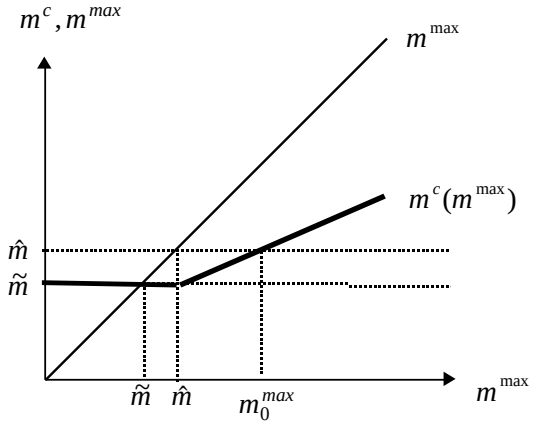

Figure 3: Case II: $\tilde{m} < \hat{m}$