Esta tesis doctoral contiene un índice que enlaza a cada uno de los capítulos de la misma.
Existen asimismo botones de retorno al índice al principio y final de cada uno de los capítulos.

Ir directamente al índice

Para una correcta visualización del texto es necesaria la versión de Adobe Acrobat Reader 7.0 o posteriores

Aquesta tesi doctoral conté un índex que enllaça a cadascun dels capítols. Existeixen així mateix botons de retorn a l'index al principi i final de cadascun deis capítols .

Anar directament a l'index

Per a una correcta visualització del text és necessària la versió d' Adobe Acrobat Reader 7.0 o posteriors.

Aspectos Económicos de la Provisión Pública de Servicios Sanitarios

Una Tesis presentada

por

Paula González Rodríguez

al

Doctorado en Economía Cuantitativa

bajo la dirección de

Dra. Carmen Herrero Blanco

en cumplimiento de los requisitos para la

obtención del título de

Doctor en Economía

Departamento de Fundamentos del Análisis Económico
Universidad de Alicante

Octubre 2002

A Nico, por todo

Acknowledgments

There are many persons who have helped me in the process of doing this doctoral thesis. With these pages I would like to thank all of them. I am aware that probably I forget not only names but also reasons and that, in many cases, it is difficult to put into words how grateful I am.

I would like to start with a special mention to my supervisor Carmen Herrero. I am truly grateful to her, not only for her encouragement, guidance and advice during the elaboration of this thesis, but also for the confidence she has shown in me and in my work.

I am also deeply indebted to Javier López-Cuñat for the time and effort he has devoted to my work. He has carefully revised and improved the drafts of the articles included in this dissertation. I am really grateful to him.

I would like to mention the faculty members and the administrative staff of the Departamento de Fundamentos del Análisis Económico at the Universidad de Alicante. From the faculty members, I would like to thank my professors of the doctorate courses: Subir Chattopadhyay, Alfonsa Denia, Paco Marhuenda, Juan Mora, Martin Peich, Felipe Pérez, Fernando Vega and Antonio Villar. I also want to acknowledge Pepe Silva and all those who were my partners in my teaching duties. From the administrative staff, I want to thank Carlos Belando, Lourdes Garrido, Julio Giménez, Reyes Gironés, Mercedes Mateo and Mariló Rufete for their willingness to help me whenever I needed it.

I also want to thank ECARES at the Université Libre de Bruxelles, for the opportunity they have given me to spend the final months of this work there. I really appreciate the hospitality they have offered me.

This dissertation has also benefited from comments on earlier versions of the chap-

ters. Among the people who contributed to this work with their advice and suggestions, I would like to highlight Pedro Pita Barros, Ignacio García-Jurado, Tor Iversen, Izabella Jelovac, Inés Macho-Stadler, Pau Olivella and Nicolás Porteiro.

Financial support from the Instituto Valenciano de Investigaciones Económicas and from the Generalitat Valenciana (project GV01-371) is gratefully acknowledged.

My acknowledgment is extensive to those friends who have accompanied me during the elaboration of this thesis. First, I have to mention my classmates: Jose Antonio García, Jorge Guillén, Marc Escrihuela, Trino Ñiguez, Arnold Polanski, Ana Rabadán and Anaís Tarragó. I thank all of them for the good moments we have spent together and, specially, to Marc, for his calmness and constant good humor and to Anaís, for being my confidant.

I can not forget the people with whom I shared flat during my years in Alicante. They have been my second family. I give my warmest thanks to Chony Andina, for the long conversations we had, to Rebeca Jiménez, for being always ready to help me and to make me laugh, and to Dunia López, for the great cakes she cooked for us.

I also would like to mention other friends with whom I have shared unforgettable talks, drinks and parties. Among them, undoubtedly, I should remember Santi Budría, Laura Crespo, Carlos Gutiérrez, Marisa Hidalgo, Antonio Jiménez, Renatas Kizys, Tate Lacomba, Francis Lagos, Miguel Ángel Meléndez, Leonora Millán, Juan de Dios Moreno, Lorena Mullor, Alicia Pérez, Guadalupe Valera and Arantxa Valero.

Quisiera, por último, expresar mi agradecimiento a todas las personas que han seguido desde fuera mi trabajo. Me gustaría agradecerle a mi familia, y en especial a mi madre, su apoyo incondicional y sus palabras de ánimo.

A Nico le agradezco enormemente el cariño que me ha dado durante todos estos años, su paciencia y su constante apoyo en los momentos difíciles de este trabajo.

Contenidos

Resumen

Esta tesis doctoral consta de tres capítulos independientes en los que estudio distintos problemas que surgen en la provisión de servicios sanitarios. Aunque las interacciones económicas bajo estudio son diferentes, hay un marco común en este trabajo. Este marco viene dado por la utilización de la Teoría de Juegos (tanto desde una perspectiva cooperativa como no cooperativa) y de herramientas de Organización Industrial para realizar el análisis.

La importancia del sector sanitario es evidente en las sociedades modernas. Por un lado, por el creciente peso que ha adquirido como sector de actividad económica. Un dato que puede ilustrar esta relevancia es el hecho de que en la mayoria de los países desarrollados el sector sanitario contribuye al Producto Nacional Bruto en más de un 10%. Por otro lado, no debemos obviar el impacto directo que tiene sobre el bienestar de la población.

El problema del creciente gasto sanitario a nivel mundial ha sido objeto de debates públicos durante años. Sin embargo, tener en cuenta factores económicos a la hora de tomar decisiones sobre los servicios sanitarios proporcionados, era apenas concebible hace pocos años. Recientemente, ha surgido una rama de investigación que estudia, desde una perspectiva económica, diversos aspectos de la provisión de servicios sanitarios. Esta Tesis Doctoral

proporciona una contribución a esta creciente línea de investigación, centrándose en dos aspectos particulares.

El Capítulo 2 está dedicado a estudiar el reparto de los costes derivados de las intervenciones quirúrgicas, en un contexto con diferentes especialidades médicas y listas de espera.[1] Los problemas de asignación de costes surgen en numerosas situaciones de la vida real, donde los individuos, todos con sus propios propósitos, deciden trabajar juntos. En estas situaciones, surge el problema de como dividir los costes conjuntos resultantes de la cooperación (e implícitamente el ahorro en costes) entre los participantes. Esta descripción general de los problemas de asignación de costes puede ser aplicada directamente a nuestro análisis. Utilizando un enfoque de Teoría de Colas, modelizamos el problema de reparto de costes que surge cuando distintas especialidades médicas comparten la utilización, tanto de los quirófanos, como las camas de un hospital.

Los otros dos capítulos se centran en algunas de las implicaciones ecónomicas del hecho de que numerosos médicos presten sus servicios tanto en el sector público como en el privado. A pesar de que esta doble actividad del médico es claramente relevante en aquellos países con sistemas mixtos de salud, este tema ha recibido escaso interés por parte de la literatura. De todas las posibles implicaciones que puedan surgir, me centro en dos.

En el capítulo 3 estudio la optimalidad de políticas, recientemente puestas en marcha en algunos países, en las cuales el sector público transfiere una proporción de sus pacientes a hospitales privados para aliviar los problemas que la excesiva longitud de las listas de espera en el sector público genera. Me centro en los incentivos estrategicos que el médico tiene debido a su condición de proveedor dual, y en las implicaciones que esto puede tener sobre los costes de la autoridad sanitaria si se lleva a cabo la política. Caracterizo un posible problema de

[1] Este primer capítulo es un trabajo conjunto con Carmen Herrero.

selección de pacientes y estudio bajo que condiciones la política debería ser implementada.

Finalmente, el Capítulo 4 está dedicado a analizar los incentivos de los médicos a realizar diagnósticos costosos y a elegir determinados tipos de tratamiento en un contexto de diseño de contratos. La doble actividad del médico es crucial para el análisis, ya que considero que el médico utiliza su trabajo en el sector público para aumentar su "prestigio" como proveedor sanitario y aumentar así sus ingresos privados. Muestro que esta búsqueda de reputación tiene consecuencias relevantes. En particular, genera una tendencia por parte del médico a sobreproveer servicios. Pero, al mismo tiempo, puede tener un efecto positivo, ya que induce al médico a realizar un mayor esfuerzo en su actividad diagnóstica. El análisis realizado me permite evaluar algunos de los marcos regulatorios existentes en diversos países europeos en relación con la doble práctica del médico.

El resto de la Introducción está dedicada a presentar, con más detalle, el análisis realizado en los tres capítulos que forman esta Tesis Doctoral.

Capítulo 2: "Reparto Optimo de Costes Quirúrgicos en Presencia de Colas"

Los servicios públicos de salud sufren una enorme saturación y unas enormes listas de espera a nivel mundial. Su impacto directo sobre el bienestar social los han convertido en una importante línea de investigación. La literatura existente sobre listas de espera para acceder a tratamiento quirúrgico sigue dos tradiciones separadas: por un lado, la tradición de teoría de colas, que considera las llegadas de pacientes y los tiempos de servicio elementos estocásticos y, por otro lado, la literatura de la economía del bienestar, donde las colas son entendidas como un sistema para la distribución y asignación de los recursos.

En este capítulo, modelizamos el problema de las listas de espera para acceder

a tratamiento quirúrgico, haciendo uso de teoría de colas. Como consideramos que tanto la llegada de nuevos pacientes a las listas, como el proceso de tratamiento tienen una componente aleatoria, la aplicación de los resultados de teoría de colas surge de un modo natural.

Nos centramos en los costes generados por las operaciones, teniendo en cuenta que cuanto mayores sean los recursos utilizados por los hospitales, menores serán las listas de espera.

Para modelizar las listas de espera quirúrgicas como un sistema de colas debemos tener en cuenta algunas características particulares de las mismas. En primer lugar, la existencia de dos fuentes para la formación de las listas de espera: tanto la capacidad del quirófano como la disponibilidad de camas en el hospital. En segundo lugar, la presencia de distintos procedimientos médicos que comparten ambos tipos de servidores. En tercer lugar, las distintas tasas de llegada de pacientes según los procedimientos. Por último, no todos los procedimientos son considerados igualmente urgentes y, por tanto, pacientes de distintas especialidades pueden tener distintas prioridades.

La gestión de las listas de espera quirúrgicas genera problemas de asignación de costes. El objetivo principal de este capítulo es, precisamente, proponer una regla de asignación de costes para el reparto de los costes conjuntos. Construimos dicha regla utilizando una perpectiva de juegos teóricos, mediante el diseño de un juego de asignación de costes. La primera parte del análisis está dedicada al estudio de los costes asociados directamente al quirófano. Para ello procedemos en dos etapas. En primer lugar, enfrentamos una situación en la que cada procedimiento médico tiene su propio quirófano, con otra en la cual hay un único quirófano para cubrir todos los tipos de operaciones. Esto nos permite mostrar que compartir el quirófano para tratar a pacientes de distintas especialidades, conduce a reducciones en los costes. En segundo lugar, construímos un juego de reparto de costes y ofrecemos

la regla de reparto de costes que recomienda la asignación del valor de Shapley del juego. La tarifa que proponemos tiene, por tanto, todas las propiedades positivas del valor de Shapley. Además, el juego de reparto de costes que surge entre los tratamientos es la suma de un juego aditivo más un "juego del aeropuerto", lo cual facilita los cálculos de la solución cooperativa.

En la segunda parte del trabajo, extendemos nuestro análisis al tiempo de hospitalización de los pacientes. Para tratarlos a todos en un determinado periodo de tiempo, no sólo es necesario que el quirófano funcione adecuadamente, sino también que haya una oferta adecuada de camas para acomodar a los pacientes en el hospital. Una cuestión natural que surge, entonces, es analizar el impacto que la cooperación entre los procedimientos médicos tiene sobre los costes postoperatorios. Modelizando también la etapa de hospitalización como un sistema de colas, podemos calcular el número de servidores (camas) necesarios para garantizar el servicio, bajo distintos escenarios de cooperación. Como el tiempo de hospitalización tiene también una componente aleatoria, no hay posibilidad de resolver analíticamente el modelo y llegar a resultados generales. Sin embargo, si nos centramos en el número de camas necesarias en términos esperados podemos obtener resultados. El signo del efecto de compartir el quirófano sobre los costes postoperatorios esperados depende de las características específicas de los tratamientos. Calculamos dos condiciones suficientes que nos permiten asegurar ahorros también en esta segunda etapa del proceso.

Finalmente, proporcionamos un ejemplo numérico que ilustra los principales resultados de nuestro modelo. Lo hacemos sobre la base de datos reales de la tasa media de llegada de pacientes y la tasa media de hospitalización de los mismos. Aplicamos nuestro análisis teórico a este caso particular e interpretamos los resultados que surgen. En particular, mostramos que cuando los procedimientos médicos cooperan también en la utilización de las camas se obtienen importantes ahorros.

Capítulo 3: "Implicaciones Políticas de Transferir Pacientes a la Práctica Privada"

Existe un consenso generalizado sobre los problemas de congestión que sufren los servicios públicos de salud a nivel mundial. La población es sensible a esta congestión del sistema, ya que es la que sufre directamente los efectos de las largas listas de espera. Este malestar social ha llevado a distintas autoridades sanitarias a recurrir a hospitales privados para que colaboren en la reducción de las listas de espera públicas, a través de costosos programas temporales. Encontrar el balance correcto entre la contención de los costes y la mejora en la provisión de los servicios sanitarios, se ha convertido en un gran reto para la mayoría de los gobiernos europeos.

En este capítulo analizo las consecuencias de transferir pacientes públicos a clínicas privadas, en un intento por reducir las listas de espera. Estudio también bajo qué circunstancias la autoridad sanitaria debería llevar a cabo una política de estas características. Un elemento crucial para el análisis es el hecho de que el médico sea proveedor dual, es decir, que preste sus servicios tanto en el sector público como en el privado.

El punto de partida es un modelo sencillo en el cual la autoridad sanitaria contrata a un médico especialista para tratar pacientes con distintos niveles de gravedad, y establece acuerdos con hospitales privados para que ellos traten al resto de los pacientes. Hay dos objetivos principales en este capítulo. En primer lugar, caracterizar el comportamiento de un médico proveedor dual cuando dicha política es implementada. Muestro que cuando el gobierno no es capaz de controlar el comportamiento del médico en relación a la gravedad de los pacientes que trata, surge un problema de "cream-skimming". Debido a la distinta estructura remunerativa existente en ambos sistemas, los médicos prefieren tratar los casos

más leves en sus propias clínicas privadas.

El segundo objetivo del capítulo es estudiar como la aparición de este fenómeno afecta a la decisión de la autoridad sanitaria de implementar la política, y de la cantidad de pacientes que deben transferirse al sector privado. Encuentro que el valor crucial que determina la relevancia del problema es la dispersión relativa de las gravedades de los pacientes. Cuanto mayor es dicha dispersión, más gana el médico al seleccionar pacientes y, al mismo tiempo, mayor es el impacto sobre los costes soportados finalmente por la autoridad sanitaria.

En aquellas situaciones en las que es óptimo para la autoridad sanitaria llevar a cabo la política, estudio el efecto de la selección de pacientes sobre la cantidad de operaciones que finalmente se desvían al sector privado. Respecto a esto último, encuentro que cuando la dispersión relativa de las gravedades es suficientemente alta, más pacientes son desviados a hospitales privados. Cuando me enfrento a disciplinas médicas en las que la dispersión de las gravedades es baja, el resultado final viene determinado por la cuantía del pago acordado con el sector privado por cada operación desviada.

El análisis desarrollado proporciona algunas recomendaciones políticas sobre la optimalidad de aplicar este tipo de medidas. Al diseñar una política de desvío de pacientes públicos al sector privado, el decisor social debería considerar el hecho de que la diferencia que existe entre los sistemas de remuneración en ambos sectores puede crear incentivos perversos en el comportamiento de los médicos. Además, los resultados sugieren que la decisión de llevar a cabo o no la política está influida no sólo por el pago por operación pactado con el sector privado, sino también por el tipo de enfermedad. En particular, la dispersión de las gravedades de los pacientes juega un papel muy importante, ya que cuanto mayor sea el rango de gravedades más grave se vuelve el problema de selección de pacientes.

Capítulo 4 "¿Debería Limitarse la Práctica Dual de los Médicos?. Un Enfoque de Incentivos"

En este capítulo examino, en un contexto de riesgo moral, otra de las implicaciones que la práctica dual de los médicos tiene para las autoridades sanitarias públicas. La relevancia del análisis viene dada por el hecho de que, como ya fue mencionado en el capítulo anterior, en países con sistemas de salud mixtos es frecuente que muchos médicos trabajen en ambos sectores al mismo tiempo. A pesar de ello, hay pocos estudios que analizen este tema y, en particular, que se centren en los conflictos de intereses que puedan surgir debido a la doble actividad del médico.

El objetivo principal de este capítulo es analizar las consecuencias que tiene, para la autoridad sanitaria, la práctica dual del médico. El análisis desarrollado proporciona un marco teórico que nos permite evaluar algunas de las médidas introducidas por algunos países europeos para regular la doble actividad del médico.

Por ejemplo, en España la legislación actual permite a los médicos públicos ofrecer sus servicios también como proveedores privados. Sin embargo, se paga un bono fijo adicional a todos aquellos médicos que renuncien a su actividad privada. Es decir, se les ofrece a los médicos la posibilidad de firmar contratos de exclusividad.

En otros países europeos con sistemas mixtos de salud la doble actividad del médico está también permitida, aunque bajo distintos tipos de regulación. En el Reino Unido o en Francia, por ejemplo, a los médicos públicos se les permite trabajar en el sector privado pero se les limita sus ingresos privados a una determinada cantidad máxima.

Los conflictos de interés que surgen entre la actividad pública y privada del médico pueden afectar a ambos servicios en numerosas dimensiones. En este trabajo, me centro en

una en particular que no ha sido tratada por la literatura hasta ahora. Esta tiene que ver con el hecho de que hay muchos médicos cuyo trabajo en hospitales públicos les ha generado la reputación de ser "buenos doctores" y este prestigio, obviamente, influye positivamente sobre su práctica privada. Aunque esto no tiene impacto sobre sus ingresos públicos, si tiene efectos sobre sus ingresos privados ya que éstos dependen directamente de la demanda que reciben de la población de pacientes.

Construyo un modelo con un paciente, un médico especialista y la autoridad sanitaria. El paciente sufre una enfermedad de gravedad indeterminada y requiere atención sanitaria. Al atender al paciente, el doctor realizar dos tipos de tareas distintas: primero, diagnosticar la gravedad de la enfermedad y, en segundo lugar, proporcionar el tratamiento adecuado. La actuación del médico se ve afectada por su condición de proveedor dual en el sentido de que si puede curar al paciente en una única ronda de tratamiento mejora su "prestigio profesional".

La autoridad sanitaria diseña el contrato que minimiza los costes sociales, en un contexto en el que el médico tiene ventajas informacionales. Por tanto, supongo que ni el proceso de diagnóstico ni el tratamiento que el médico prescribe pueden ser verificables o contratables.

En este marco, estudio las implicaciones que tiene para la autoridad sanitaria la doble actividad del médico. El análisis me permite comprobar si la autoridad sanitaria estaría interesada en prohibir o limitar la práctica privada de los médicos. Encuentro que la práctica dual del médico tiene efectos contrapuestos. Por un lado, su interés en curar al paciente y ganar prestigio genera una sobreprovisión de servicios sanitarios. Por otro lado, si la autoridad sanitaria logra controlar estos incentivos a sobreproveer servicios, entonces se puede beneficiar del interés del médico en realizar una diagnóstico muy preciso y curar al

paciente.

Con respecto a recomendaciones políticas, nuestro análisis sugiere que el impacto general de la doble actividad del médico sobre los costes de la autoridad sanitaria depende básicamente de la estrategia de tratamiento que ésta decida seguir. Si la prioridad de la autoridad sanitaria es la contención de costes, entonces la doble actividad del médico es perjudicial. Si la prioridad es, sin embargo, minimizar la pérdida de salud de la población, entonces la práctica dual del médico permite conseguir el objetivo a un menor coste.

Este trabajo proporciona un marco teórico para poder analizar la optimalidad tanto de los contratos de exclusividad, como de los límites a los ingresos privados del médico. Si se considera la primera de estas medidas, observamos como cuando la autoridad sanitaria diseña un contrato de incentivos no hay necesidad de ofrecer al médico un contrato de exclusividad. Si, por otro lado, la autoridad sanitaria paga un salario, entonces proponer al médico que firme contratos de exclusividad puede suponer un ahorro en costes. Estos resultados permiten explicar la existencia de los contratos de exclusividad como una elección de "second best".

Si se considera el segundo tipo de regulación observamos que, bajo contratos de incentivos, limitar los ingresos privados del médico es beneficioso socialmente, excepto en los casos en los que la autoridad sanitaria está altamente preocupada por la precisión del diagnóstico médico. En estos casos, la regulación sería perjudicial socialmente.

Contents

Chapter 1

Introduction

This Doctoral Dissertation consists of three independent chapters in which I study problems that emerge from the provision of health care services. Even if the economic interactions under study differ, there is a common framework for this work. It is given by the use of Game Theory (both from a co-operative and a non co-operative perspective) and of analytical tools from Industrial Organization to perform the analysis.

The importance of the Health Care Sector is clear in modern societies. On the one hand, by the increasing weight that it has acquired as a sector of economic activity. An illustrative feature is the fact that, in the majority of the developed countries, the Health Care Sector contributes to the GNP with more than a 10%. On the other hand, we can not underrate its direct impact on the welfare of the population.

The problem of increasing expenditure on health throughout the world has been the focus of public debates for years. However, even a few years ago, taking economic factors into account was barely conceivable when making decisions about the services provided by health authorities. A branch of research that studies, from an economic perspective, different aspects of the provision of health care services has recently emerged. This Doctoral Dissertation

provides a contribution to this increasing line of research, focusing on two particular issues.

Chapter 2 is devoted to study the sharing of the costs derived from surgical operations, in a framework with different medical specialities and waiting-lists.[1] Cost allocation problems arise in many real life situations, where individuals, all with their own purposes, decide to work together. In this situations, the problem of how to divide among the participants the joint costs (and implicitly the costs savings) which result from the co-operation arises. This general description of cost allocation problems can be directly applied to our analysis. Using a queuing theory approach, we model the cost-sharing problem that emerges when different medical specialities share both operating-theatres and hospital beds.

The other two chapters focus on some economic implications of the fact that many physicians work both in the public and in the private sector. This dual activity is clearly relevant in those countries with mixed health care systems, but has received little interest by the literature. From all the possible implications I study two.

In Chapter 3 I analyze the optimality of policies, recently implemented in some countries, in which the Public Sector transfers a proportion of its patients to private hospitals to alleviate the problems that the excessive length of waiting-lists generate. I concentrate on the strategic incentives that a physician may have due to his position as dual supplier, and the implications that this can have over the costs borne by the Health Authority when undertaking the policy. I spot a potential problem of patient selection and characterize in which cases the policy should be implemented.

Finally, Chapter 4 is devoted to analyze physicians' incentives to perform a costly diagnosis and choose a particular type of treatment in a framework of contract design. The doctor's dual activity is crucial for the analysis, since he uses his work in the Public Sector

[1] This first chapter is a joint work with Carmen Herrero.

as a way to increase his "prestige" as a provider and, hence, increase his revenues as a private practitioner. I show that this reputation-seeking effect has important implications. In particular, it generates a tendency to over-provide services by the physician, but it can also have a positive effect since it induces the physician to perform a higher effort in the diagnosis. The analysis performed allows me to evaluate some alternative regulatory frameworks that exist in several European countries concerning physicians' dual practice.

The remaining of this Introduction is devoted to present, more in detail, the analysis performed in these three chapters that form this Doctoral Dissertation.

Chapter 2: "Optimal Sharing of Surgical Costs in the Presence of Queues"

Public health services, worldwide, are plagued by over-crowding and lengthy waiting-lists. Their impact on the social welfare has made them become an important issue of research. The existing literature on waiting-lists for surgical treatment follows two separate traditions: on the one hand, the queueing theory tradition, which considers arrivals and service times to be stochastic events, and on the other hand, the welfare economics literature, where queues are considered as a system for the distribution and allocation of resources.

In this chapter, we model the problem of the waiting-lists for surgical treatment, making use of queueing theory. Since we consider that both, the arrival of new patients to the waiting-list and the process of treatment have random components, the application of the queueing theory results arises naturally.

We concentrate on the costs generated by the operations, taking into account that the higher the resources spent by the hospital, the shorter its resulting waiting-list.

To model surgical waiting-lists as a queueing system several features should be

considered. First, the existence of two sources for the formation of waiting-lists: both the capacity of the operating-theatre, and the bed-capacity of the hospital. Second, the presence of several different medical procedures sharing both sets of servers. Third, the difference in the rate of arrival across procedures. Finally, not all medical procedures are considered equally urgent and, hence, patients from different specialities should have different priorities.

The management of surgical waiting-lists certainly generates cost-allocation problems. The main aim of this chapter is, precisely, to propose a cost-allocation rule for the sharing of joint costs. To construct such cost-allocation rule, we use a game theoretical perspective, designing a cost-allocation game.

The first part of the analysis is devoted to study the costs associated with the operating-theatre. We proceed, then, in two steps. First, we confront a situation in which each medical procedure has its own operating-theatre, with another in which there is just one theatre for all the operations. This allows us to show that sharing the operating-theatre to treat patients from the different medical procedures, leads to cost reductions. Secondly, we construct a cost-sharing game and we offer a cost-sharing rule that recommends the Shapley value allocation of the game. The tariff we propose has, therefore, all the properties of the Shapley value. Moreover, the cost-sharing game emerging among the treatments is the sum of an additive game plus an "airport game", what eases the computation of the co-operative solution.

In the second part of the work, we extend our analysis to the patients' recovery time. In order to treat the patients in a given time, it is not only necessary for the operating-theatre to work properly, but also for there to be an adequate supply of beds to accommodate the patients in hospital. A natural concern is, therefore, to analyze the impact of co-operation among medical procedures on the post-operational costs.

Modelling the hospitalization stage also as a queueing system, we can compute the number of servers (beds) required to guarantee the service in different scenarios of co-operation. As the recovery time has also a random component, there is no possibility to analytically solve the model and arrive at any general results.

However, we can obtain some results if we focus on the expected number of beds required. In this case, the sign of the effect that the sharing of the operating-theatre has on the average post-operational costs depends on the specific characteristics of the treatments and cannot be stated in general. We compute two sufficient conditions for making a saving at this second stage of the process as well.

Finally, we provide a numerical example that illustrates the main features of our model. We perform it on the basis of real data concerning the expected number of patients' arrivals and the expected length of their recovery time. We apply our theoretical analysis to this particular case and interpret the results that arise. In particular, we show that major savings are obtained when the different procedures cooperate in the managing of beds as well.

Chapter 3: "Policy Implications of Transferring Patients to Private Practice"

There is a general consensus in society concerning the congestion problems that bear public health services worldwide. The population is sensitive to the congestion within the system as they suffer the direct effects of long waiting-lists for urgently needed operations. This social discomfort has lead several national health authorities to turn to private hospitals for assistance in reducing their waiting-lists, through temporary and costly programs.

Finding the correct balance between cost-containment and improvements in the provision of health care services has, therefore, become a major endeavor in most European

economies.

In this chapter, I analyze the consequences of transferring public sector patients to private clinics as an attempt to reduce waiting-lists, and the circumstances under which the Health Authority should implement it. A crucial element for the analysis is the fact that the physician is a dual supplier, i.e., he works both for the public and the private sector.

The starting point is a simple model in which the Health Authority contracts a hospital specialist for treating patients with different severities, and reaches agreements with private hospitals to have the remaining patients treated there.

There are two main objectives in this chapter. First, to characterize the behavior of a dual-supplier physician when such a policy is undertaken. I show that when the government is not able to monitor the physician's behavior with regard to which severities he treats, a problem of cream-skimming arises. Due to the different structure of the physician's remuneration in either system, specialists prefer to treat only the mildest cases in their own private practices. We, then, show how this problem makes the Health Authority be more reluctant to implement this policy.

The second objective of the chapter is to study how this feature affects the decision of the Health Authority concerning two issues: when to carry out the policy and the amount of patients that should be transferred to the private sector. I find that the crucial value to determine the relevance of the problem the Health Authority bears is the relative dispersion of the patients' severities. The higher the dispersion is, the more the physician earns from selecting patients and, at the same time, the greater the impact on the costs borne by the Health Authority is.

In those situations in which it is optimal for the Health Authority to undertake the policy, I study the effects of patient-selection on the amount of patients that are finally

transferred to the private sector. In this respect, I find that when the relative dispersion of the severities is high enough, more patients are sent to private hospitals. When we deal with medical disciplines in which the dispersion of the severities is low, the result is determined by the value of the fee per operation that the Health Authority agreed with the private sector.

The analysis performed provides some policy recommendations concerning the optimality of this kind of measures. When designing a policy to transfer patients from the public to the private sector, the policy-maker should consider the fact that the difference that exists between the reimbursement systems in the two sectors can create perverse incentives for the physicians.

Moreover, the results suggest that the decision concerning whether to undertake the policy or not is not only influenced by the fee per operation agreed with the private sector, but also by the type of illness. In particular, how disperse the severities of the patients are is shown to be very important, since the wider the range of severities, the more serious the problem of patient-selection becomes.

Chapter 4 "Should Physicians' Dual Practice Be Limited? An Incentive Approach"

In this chapter I examine, in a moral-hazard environment, some implications that physicians' dual activity has for public health authorities. The relevance of the analysis is given by the fact that, as mentioned before, it is common in countries with mixed health care systems that many doctors work in both sectors at the same time. In spite of this, there are few studies that analyze this issue, and in particular that focus on the conflicting interests that arise from the doctors' dual activity.

The main objective of this chapter is to analyze the consequences for health author-

ities of doctors' dual practice. The analyses performed provides a theoretical benchmark for the evaluation of several measures that different European health authorities have introduced to deal with this issue.

For instance, in Spain the labor legislation in force allows public doctors to offer their services as private providers as well. However, an additional fixed monthly bonus is paid to those practitioners that agree to forego their private practice. That is, an exclusive contract is offered to them.

In the majority of the other European countries with mixed health care systems, physicians' dual activity is also allowed, although under different types of regulation. In the UK or in France, for example, physicians are allowed to operate in the private sector but their private income is restricted to not exceed a certain threshold.

The conflicting interests that arise between the doctor's public and private practices can affect both services in many different dimensions. In this work, I focus on a particular one that has not been analyzed by the literature so far. This one has to do with the fact that there are many physicians whose service in public hospitals have won them the reputation of being "good doctors" and such prestige obviously has a positive influence on their private practice. Even if this has no impact on their public earnings, it certainly has effects on their private revenues as such revenues depend directly on the demand they receive from the patient population.

I construct a model with one patient, one specialist, and the health authority. The patient suffers from an illness whose severity is unclear and requires medical attention. In attending to the patient, the doctor has to perform two different tasks: First, to diagnose the severity of the illness and then, to provide the required treatment. The doctor's performance is affected by his condition of dual supplier, since if he can cure the patient in a single

treatment his "professional prestige" is improved.

The health authority designs the contract that minimizes the social costs, in a framework in which the physician has informational advantages. I suppose, therefore, that neither his diagnostic process nor the treatment he prescribes can be either verifiable or contractible.

In this set-up I study the implications that the physicians' dual activity has for the health authority. These analyses allow me to verify whether the health authority would like to either prohibit or limit the doctor's private practice. I find that the doctor's double role provides certain conflicting effects. On the one hand, his keen interest in curing the patient and gaining prestige, generates an over-provision of health services. On the other hand, if the health authority manages to control these incentives to over-provide services, then it can benefit from the physician's interest in doing a more accurate diagnosis and curing the patient.

Regarding policy recommendations, our analysis suggests that the overall impact of the doctor's dual practice on the health authority's costs, depends basically on the treatment strategy that the health authority decides to follow. If the priority of the health authority is to contain costs, then the doctor's dual activity is welfare decreasing. If the priority is to minimize patients' health losses, physician's dual practice affords the objective at a lower cost.

This work provides a theoretical framework in which the optimality of exclusive contracts and of limits on physicians' private incomes can be addressed. Considering the former, we show that when the health authority is able to design an incentive contract there is no need for exclusiveness. If, on the other hand, the health authority is forced to pay a flat salary, the health authority can successfully contain costs by signing exclusive contracts. Such

results may well explain the very existence of exclusive contracts, which can be considered a "second best" choice.

Considering the later type of regulation we show that, under an incentive contract, limiting physicians's private income is beneficial from a social point of view, except in those cases in which the health authority is highly concerned about the accuracy of the diagnosis the physician performs. In such a case, this kind of regulation is socially harmful.

Chapter 2

Optimal Sharing of Surgical Costs in the Presence of Queues

2.1 Introduction

The widespread access to Public Health Care in Western European countries is placing such stress on the system that it has arrived to the point at which optimal allocation of resources is becoming a major management problem. On the one hand, and since health services are among the critical aspects in controlling the quality of public services, the regularity and adequacy of hospital services has now become crucial for the prestige of a government. On the other hand, management errors could have a tremendous impact on the Health Administration budget.

Citizens are particularly sensitive to some of the phenomena related to health services. One of such phenomena is the persistence of waiting-lists for surgical treatment. The anxiety and discomfort caused by this phenomenon forces the government to devise especial programs to alleviate the problem temporarily. Such temporary programs, however, cannot

solve the problem, and can turn out to be extremely costly.

Judging from the existing literature on waiting-lists for surgical treatment, there seems to be two separate traditions: 1) the queueing theory tradition, which considers arrivals and service times to be stochastic events, and 2) that of the welfare economics literature, where queues are considered as a system for the distribution and allocation of resources.

Queueing theory addresses these sort of problems from a statistical or operational research point of view (for an overview of this topic see Gross and Harris (1997), Hillier and Lieberman (1995), Kleinrock (1975) and Prabhu (1997)). Any system in which arrivals make excessive demands on a finite-capacity resource may be termed a queueing system. In particular, if the arrival times of the demands are unpredictable, conflicts over the use of the resource may well arise and queues of waiting customers will obviously be formed. The main idea behind the prediction of the behavior of the system, is, nonetheless, extremely simple: The length of the queue depends on the expected rate of arrivals, and on its statistical fluctuations. If the average rate of arrivals exceeds the capacity, the system automatically breaks down, and unbounded queues will arise. When the average rate is less than the system's capacity, however, we also find queues, due to statistical fluctuations and spurts of arrivals that may occur at any given moment.

The forming of waiting-lists for elective surgery can be considered a queueing system. Queueing theory predicts several characteristics of waiting-lists, such as the expected waiting-time of the agents or the expected length of the queue. When agents are attended to on a first-come-first-served basis, the only control variable is the system's capacity. Consequently, the theory can help us to make decisions concerning such capacity, considering the fact that the higher its capacity, the higher the associated costs, but the shorter the queues.

The queueing system for surgical treatment has some rather peculiar characteristics:

(1) There are two sources for the formation of waiting-lists. On the one hand, the capacity of the operating-theatre, and, on the other hand, the bed-capacity of the hospital; (2) Several different medical procedures share both servers. In other words, customers from different treatments need to use both the operating-theatre and the beds; (3) Each of these medical procedures has its own rate of arrival; (4) Not all medical procedures are considered equally urgent, so that the average waiting-time politically considered as adequate differs among procedures.

In managing such a situation, cost-allocation problems arise. Furthermore, since different procedures share both the operating-theatre and the hospital beds, we must first design a cost-allocation rule for the sharing of joint costs. This, in fact, is the main purpose of this paper. To construct a cost-allocation rule, we use a game theoretical perspective, designing a cost-allocation game. In the first part of the paper, we concentrate on the costs associated with the operating-theatre. We construct, then, a game by confronting two situations: one in which each medical procedure has its own operating-theatre, and another in which there is just one theatre for all the operations. We show that sharing the operating-theatre to treat patients from the different medical procedures, leads to cost reductions. We then construct a cost-sharing game and, given its peculiar characteristics, we offer a cost-sharing rule that recommends the Shapley value allocation of the cost-sharing game. Our optimal tariff, therefore, has all the nice properties of the Shapley value (see Shapley (1953), Tijs and Driesen (1986), Young (1994) and Moulin and Shenker (1996)). The fact that this co-operative solution can be computed easily, is certainly an important property from a practical point of view.

The cost-sharing game emerging among the treatments is the sum of an additive game plus an "airport game" (see Littlechild and Owen (1973), and Littlechild and Thompson

(1977)), where the different landing-track capacities are translated, in our model, into the capacity required by the operating-theatre to satisfy the demands placed on it, according to the maximum average waiting-time guarantee. A similar idea has been applied in Fragnelli et al. (2000), to the construction of a railway line, as a proposal for the reorganization of the European railway system. In their case, as in ours, the proposed solution to the cost-sharing game is the Shapley value.

Up to this point, only the direct costs derived from surgical operations have been considered. We must remember, however, that an operation also generates other costs, more precisely, the costs incurred during the patients' recovery. We, therefore, introduce the post-operational costs in the model and we study how they are affected by the co-operation among different medical procedures. Considering the beds as servers, we may also model the hospitalization stage as a queueing system. As such, the number of servers (beds) required to guarantee the service can be computed in different scenarios. Nonetheless, there is no possibility of our arriving at any general results, due to analytical unsolvability of the model.

In spite of this, however, something general can be said about the average number of beds. By so doing, we show that sharing the operating-theatre has an ambiguous effect on the average post-operational costs. If the medical procedure with the highest priority level has a higher recovery time than the average recovery time for the other medical disciplines, the co-operation leads to post-operational cost savings in average terms.

We then analyze a numerical example with real data. In this example, we compute the distribution of surgical costs, applying the theoretical results obtained previously. The number of beds required is also computed, under different scenarios. We, then, estimate the distribution of bed costs among the different procedures, provided that an upper bound of .1 is set on the probability of waiting after the operation.

Most of the literature on hospital waiting-lists has focused on the demand side. Culyer and Cullis (1976), and Cullis and Jones (1985), identify demand factors as the ones that most affect waiting-lists.

There are some papers, however, that address the problem from the supply side. Iversen (1993), for instance, shows that the non-cooperative character of resource allocation in Public Health Services may contribute to excessive waiting-lists. Our work must be considered among this supply-side branch of the literature, since we study the costs derived from increasing the capacity of the operating-theatre in order to decrease the time spent by the patients on the waiting-list.

Recent papers focus mainly on the effects of such waiting-lists on the patients' welfare and on the purchase of private health insurance. Johannesson (1998) develops a model of the benefits and costs of being on a waiting-list. Since changes in the duration of the waiting-time causes complex shifts in utility streams, shorter waiting-time is not necessarily preferable to a longer one. Besley (1999) shows that longer waiting-lists for public treatment are associated with greater purchases of private health insurance. In our analysis, however, neither the patients' welfare nor private provision are considered.

The problem of the hospital-bed supply has also been addressed in the literature. Joskow (1980), and Worthington (1987), use a queueing model to analyze the characteristics of the hospital-bed supply. They both consider the beds to be the servers of the system. Hence, the waiting-lists are determined by the interaction between two factors: On the one hand, the arrival of new patients and their lengths of stay and, on the other hand, the amount of beds available. Instead, we consider that the queue is formed in the previous stage. Then, when a patient leaves the queue and enters the operating-theatre, we put a small upper bound to the probability that he will not find an available hospital bed after the operation.

The rest of the paper is organized as follows: Section 2 presents the model. Section 3 studies the operating-theatre costs. Section 4 computes the optimal cost-sharing. Section 5 introduces the post-operational time in the model. Section 6 provides a numerical example. Finally, Section 7 offers some concluding remarks.

2.2 The Model

We consider the basic queueing process: customers requiring a service are generated over time by an input source. These customers arrive to the system and join a queue. At different moments, a customer is selected to receive the service by means of a queue discipline. The mechanism of service, then, provides the service and the customer leaves the system.

In our problem, the customers are patients who require surgical treatment and the mechanism of the service is a hospital. In fact, there are two different sorts of *servers* in our model: (1) the operating-theatre, and (2) the hospital beds. Any individual entering the system should first go through the operating-theatre, and once released from this server, a bed should be available for him/her. The patient only leaves the system once he is discharged from the hospital.

Let us consider a situation in which we have n kinds of medical procedures and a certain number of patients requiring service from each of the different procedures. Let $N = \{1, 2, ..., n\}$ denote the set of fields of treatment.

We assume that the number of potential patients is infinite. This is a standard assumption in queueing theory, which simply makes the model analytically more tractable. The main implication of this assumption is that the number of individuals in the queue does not affect the amount of potential entrants. This seems reasonable in our framework, since the

probability of needing some medical treatment is, in principle, independent of the amount of people who require it. In our study, we completely ignore the physicians' strategic behavior. If we considered this, the assumption would be difficult to sustain, since the length of the waiting-lists could affect the incentives of the General Practitioners to send either more or fewer patients for elective surgery.

We put no restrictions on the length that the queue can reach, which is also standard in the literature even when dealing with situations in which a finite upper bound actually exists, but it is large enough.

It is assumed that the patients' arrivals to the medical system follow a Poisson process. This means that every period of a certain length, has the same probability of receiving a patient. We can define $\lambda_i \in \mathbb{R}_{++}$ as the expected number of arrivals per unit of time, from the i_{th} medical procedure. This is equivalent to say that the time between arrivals of patients of the same type is given by an exponential distribution with mean $\frac{1}{\lambda_i}$.

The work an arriving patient brings into the operating-theatre, equals the time of service he requires. We consider this service time to follow a random process. Even if in principle the length assigned to a surgical operation from a given medical procedure is fixed, an operation may either require less time or unexpectedly become more complicated, and hence, require extra time. It is therefore necessary to consider some randomness in the service process.

We measure the service time not in absolute terms (length of the operation) but rather, in relative terms, using as a reference the total time that the operating-theatre is in use on a given day. For example, if an operation lasts 2 hours (in expected terms) and the service is open 8 hours a day, the expected service time for a patient would be $\frac{1}{4}$. In other words, each patient occupies one-fourth of the total working time of the server.

We consider, then, the service time of the i_{th} medical procedure follows an exponential distribution with mean $\frac{1}{\mu_i}$, and $\mu_i \in \mathbb{R}_{++}$, where $\frac{1}{\mu_i}$ represents the fraction of the total working time of the server employed in one patient. This choice has been made mainly for analytical convenience, since most of the exact results of queuing theory apply to queuing systems in which the service times are exponential. On top of this, we consider that the characteristics of this distribution fit well with the problem we are modelling. As we are only dealing with scheduled non-urgent surgery, the probability of an operation becoming more complicated should be relatively low.

Analogously, μ_i would stand for the maximum expected rate (capacity) at which the system can perform (different among treatments), i.e., the potential average rate of type i patients' departures per unit of time.

We require that $\lambda_i < \mu_i \ \forall i \in \{1, 2, ..., n\}$, otherwise the queue would "explode" and the system would break down.

We also assume that the queue discipline imposed is that of "first come, first served", i.e., patients are chosen to receive the service according their order of arrival. By assuming this, we are not ruling out the possibility that there is some sort of priority ordering among patients. One may think, for instance, that when a patient initially demands medical services, he is screened by a physician and forwarded to the appropriate server, according to the type of service he requests. The only restriction in our model is that, once the patients have joined a queue, they are served according their order of arrival.

As in many queueing theory models, we assume that the arrivals and departures from the system behave as a "birth and death process". Hence, we require that at any given instant, only one "birth" (arrival of a patient to the queue) and one "death" (departure of a patient from the operating-theatre) can occur.

Finally, we perform our analysis considering that the steady state of the system has been reached.

2.3 Operating-theatre Costs

In order to characterize the costs associated with giving the operating-theatre enough capacity to provide the service within the legal maximum time, we consider two alternative scenarios. We should remember that, in this Section, the *system* is simply the operating-theatre.

In the first scenario, each surgical procedure has its own operating-theatre for treating its patients. Let us denote the expected waiting-time of an individual of type i in the system by W_i. Note that this time includes not only waiting in the queue, but also the time spent in the operating-theatre. If we considered only the queue time, which in principle might seem more reasonable, the qualitative results would not change and the model would become analytically less tractable. Moreover, the time spent in the operating-theatre is negligible with respect to the total time in the system. Under all the previous assumptions, it is well-known that the expected-time of an individual in the system is given by:

$$W_i = \frac{1}{\mu_i - \lambda_i}. \tag{2.1}$$

In the second scenario, the different medical procedures share a single operating-theatre. Proceeding analogously, and denoting the expected-time of an individual in the system by W, independently of his type, we have:

$$W = \frac{1}{\mu - \sum_{i=1}^{n} \lambda_i}, \tag{2.2}$$

where μ is the capacity of the operating-theatre to treat patients coming from any of the different specialities and $\sum_{i=1}^{n} \lambda_i$ is the total expected number of arrivals per unit of time.

Since the arrivals of patients are independent events across specialities, the total number of arrivals also follows a Poisson process and its mean is computed as the sum of the mean of the arrivals of the patients coming from the n medical procedures.

Finally, we consider that the government stipulates that, on average, the maximum waiting-time in the system for the i_{th} medical procedure may not exceed t_i, i.e., a maximum average waiting-time guarantee is provided.[1] Moreover, these times differ across treatments (applying, for example, an urgency criterion) and we suppose, without loss of generality, that $t_1 \geq t_2 \geq ... \geq t_n$. There is evidence of the implementation of this kind of measure by certain European Public Health Administrations. For instance, a maximum waiting-time guarantee was introduced in Sweden in 1992, to shorten waiting-times (see Hanning and Wimblad Spånberg (2000)). In Spain, the Ministry of Health and Consumption has recently designed a program (Programa Avance INSALUD) to ensure an average waiting-time guarantee to patients requiring elective surgery.

In the following sub-sections, we study the costs of fulfilling the government's objective under the two scenarios mentioned above. To do so, we assume that the costs are proportional to the amount of patients treated per unit of time. We interpret this as having costs that are linear in the capacity of the server. In our case, an increase in capacity could be understood as having the operating-theatre open for more hours, to be able to serve more patients.

2.3.1 Different Operating-theatres

Since the operating costs are proportional to the amount of patients treated, the overall costs arising from n operating-theatres are merely the sum of the individual costs.

[1] It can be shown that this measure is analytically equivalent to another in which we set a maximum probability that the patients' waiting-time exceeds a fixed limit.

The explicit form of the individual costs is as follows:

$$C_i = k\mu_i(t_i),$$

with $k \in \mathbb{R}_{++}$.

We must set μ_i (the potential amount of patients from medical procedure i that can be treated per unit of time) in order to guarantee the corresponding legal maximum waiting-time (t_i). What we are setting indirectly is the number of hours that the theatre should stay open each day. Formally:

$$W_i = \frac{1}{\mu_i - \lambda_i} = t_i \Longleftrightarrow \mu_i(t_i) = \frac{1}{t_i} + \lambda_i. \tag{2.3}$$

Hence, the costs per medical procedure are:

$$C_i = k\left[\frac{1}{t_i} + \lambda_i\right], \ \forall i \in \{1, 2, ..., n\}. \tag{2.4}$$

As we can see, these costs are increasing in the mean rate arrival (λ_i), and decreasing in the maximum average waiting-time guarantee (t_i). The two features are reasonable: the more patients that arrive, and the lower the average-time we can keep them waiting, the higher the cost will be .

The overall cost of keeping n operating-theatres open are, therefore:

$$C^N = \sum_{i=1}^{n} C_i = k\left(\sum_{i=1}^{n} \frac{1}{t_i} + \sum_{i=1}^{n} \lambda_i\right). \tag{2.5}$$

2.3.2 A single Operating-theatre

In this scenario, as there is only one operating-theatre, the total costs of which will be given by:

$$C^1 = k\mu(T),$$

where $\mu(T)$ is the potential number of patients, coming from any medical procedure, that can be treated per unit of time, and T is the lowest value that the maximum average-time guarantee takes across treatments. Formally:

$$T = \min \{t_i / i \in N\}. \tag{2.6}$$

This means that if the system has enough capacity to guarantee the legal average-time t_i for the i_{th} medical procedure, then it also has to be able to serve any medical procedure j with $j < i$, according to its legal maximum average-time (recall that if $j < i$ then $t_j > t_i$).

Proceeding analogously as in the previous subsection, we compute $\mu(T)$:

$$W = \frac{1}{\mu - \sum_{i=1}^{n} \lambda_i} = T \Longleftrightarrow \mu(T) = \frac{1}{T} + \sum_{i=1}^{n} \lambda_i. \tag{2.7}$$

Therefore, the overall costs are:

$$C^1 = k\left(\frac{1}{T} + \sum_{i=1}^{n} \lambda_i\right). \tag{2.8}$$

2.3.3 Comparing the Costs of the Two Scenarios

We now proceed by comparing the costs of the two situations analyzed. The aim is to verify whether there is any kind of saving, understood as lower aggregate costs, in the scenario in which the medical procedures share the operating-theatre.

Proposition 1 *Sharing the operating-theatre leads to cost reduction.*

Proof. Using Equations (2.5) and (2.8) and taking (2.6) into account, it is straightforward to verify the sign of the difference between the costs of keeping n operating-theatres open and the costs of maintaining only one, which serves all the medical procedures. We obtain:

$$C^N - C^1 = k\left(\sum_{i=1}^{n-1} \frac{1}{t_i}\right) > 0,$$

since both k and t_i $\forall i \in \{1, 2, ..., n\}$ are strictly positive. ■

Hence, it is shown that a saving is possible if the different surgical procedures cooperate and share a single operating-theatre. Let us explain the reason for this. When each medical procedure maintains its own server, it has to suffer not only a cost that is proportional to the expected number of patients demanding surgical care ($k\lambda_i$), but also a fixed cost, depending on the maximum average waiting-time guarantee for the given medical procedure ($\frac{1}{t_i}$). This is due to the randomness of the process we are dealing with. Both the number of arrivals and the number of discharges are measured in expected terms, since we are working with variables that are distributed according to random processes. Each operating-theatre, therefore, would have to maintain some additional capacity to prevent a situation in which a greater number of patients than expected arrive at a given moment, or when an operation becomes complicated, requiring some extra time. If there is co-operation among the medical procedures, they can maintain the necessary additional capacity by just supporting the fixed extra cost of the procedure with the highest priority level together. The degree of priority is understood in our model as the average waiting-time guarantee, and the shorter the guarantee, the higher the priority degree.

We can interpret this smaller necessary capacity in terms of optimal risk-sharing among treatments. When a medical procedure is on its own, it has to cover all the risks of an excessive arrival of patients or a lengthy operation. This means that it has to ensure supplementary capacity to guarantee the legal average waiting-time, even when circumstances are more difficult than expected for a certain period of time.

When this medical procedure shares the operating-theatre with others, however, the "bad luck"experience in one treatment on a given day, may well be off-set by good luck in

another. In other words, it is not too costly if one medical procedure performs badly one day, as it can use some of the extra time saved by another which has been more lucky in the realization of the uncertainty. This phenomenon is similar to risk spreading. As we can compensate for the results among different procedures, we can cover the demand within the legal average-time with less installed capacity (and therefore, at a lower cost).

Since the co-operative scenario is cheaper than the other one, our main interest now is to distribute the benefits from working together among the different medical procedures. In other words, we must calculate the optimal tariff that each treatment should pay for the use of the service. In the following section, we model this problem as a cost-sharing co-operative game and we compute the optimal fee.

2.4 Optimal Cost-Sharing

To summarize, then, the problem we face is as follows: There is one operating-theatre which is being used by different surgical procedures and the costs must be divided among them. We have to decide, therefore, how the operating-theatre costs should be allocated to the medical procedures through an optimal tariff.

We shall now construct a cost-sharing game. Let us consider the *players* to be the different surgical procedures, $N = \{1, ..., n\}$. The cost-sharing game is defined as follows: $c : 2^N \rightarrow \mathbb{R}$, assigns the minimum cost $c(S)$, under which the time guarantee is fulfilled *for all the surgical procedures in* S, to any non-empty coalition S of medical procedures. For the empty coalition we have $c(\emptyset) = 0$. Since sharing the operating-theatre always affords cost reduction, our minimal cost will be the cost required to maintain a single operating-theatre

shared by all the different medical disciplines included in set S. Namely,

$$c(S) = k\left(\frac{1}{T_S} + \sum_{i \in S} \lambda_i\right),$$

where $T_S = \min\{t_i : i \in S\}$ is the shortest average-time guarantee within S, namely, the average-time guarantee established for the most urgent procedure in set S. Note that the above cost function can be divided into two parts:

$$c(S) = \left(\frac{k}{T_S}\right) + \left(k \sum_{i \in S} \lambda_i\right),$$

there is a variable expense, which is proportional to the number of patients of each medical procedure who demand the service $c^v(S) = \left(k \sum_{i \in S} \lambda_i\right)$, and a fixed cost $c^f(S) = \left(\frac{k}{T_S}\right)$, which is independent of the medical procedure that the agents belong to. Our cost-sharing game, therefore, is the sum of two other games, $c = c^v + c^f$.

We shall adopt the recommendation of the *Shapley value* of the game as a way of distributing the costs among the different surgical procedures. This solution has the following properties:

1. *It is optimal,* as it recommends the greatest possible cost reduction (in our case, for the use of a single operating-theatre). It divides the total cost $c(N)$ among the different medical disciplines.

2. *It is linear,* since we solve a game that is the sum of two other games by simply solving them separately, and then adding them up. In our case, $Sh(c) = Sh(c^v) + Sh(c^f)$.

3. *It is symmetric,* as whenever two procedures are indistinguishable in cost, they contribute the same amount to the total cost.

4. *It is fair,* as we cannot manipulate the outcome by introducing artificial procedures with a zero cost.

Our cost-sharing game, therefore, is the sum of two other cost-sharing games: the variable cost-sharing game, c^v, and the fixed cost-sharing game, c^f. Because of property (2), $Sh(c) = Sh(c^v) + Sh(c^f)$.

It turns out that the variable cost-sharing game, c^v, is a linear one, in which there are no cost reductions from co-operation among the different procedures. Consequently, $Sh_i(c^v) = k\lambda_i$, for all $i \in N$.

Note that the fixed cost-sharing game, c^f, is analogous to the one appearing in the "airport game". In our case, instead of requiring different landing-track capacities for different types of planes, we require different operating-theatre capacities, depending on the maximum average waiting-time guarantee set by the government for the different surgical operations. This is a concave game. Consequently, the Core of the game is not empty, and the Shapley value allocation belongs to it. On taking (2.6) into account, we know that:

$$\frac{k}{t_1} \leq \frac{k}{t_2} \leq .... \leq \frac{k}{t_n} = \frac{k}{T}.$$

The fixed cost of an operating-theatre with the sufficient capacity to serve all of the procedures depends, essentially, on the time guarantee of the procedure with the highest priority level (the n_{th} medical treatment in our case).

Baker (1965) and Thompson (1971), proposed a simple cost-allocation rule for solving these kinds of cost-sharing problems. Littlechild and Owen (1973), showed that the above-mentioned cost-allocation coincides with the cost-allocation recommended by the Shapley value. We can express this rule as follows: each procedure contributes equally to the cost of maintaining an operating-theatre open for the medical treatment with the least priority; the contribution of the procedure with the least priority level is, then, completely computed. All of the remaining procedures also contribute equally to the additional cost of keeping the

theatre open for the next treatment in the finite order. This way, the second procedure's contribution is completed, and so we move on to the next one.

Formally, the cost that should be charged to the j_{th} procedure is given by:

$$\begin{aligned} Sh_1(c^f) &= \tfrac{k}{nt_1} \\ Sh_2(c^f) &= k\left[\tfrac{1}{nt_1} + \tfrac{1}{(n-1)}\left(\tfrac{1}{t_2} - \tfrac{1}{t_1}\right)\right] \\ Sh_3(c^f) &= k\left[\tfrac{1}{nt_1} + \tfrac{1}{(n-1)}\left(\tfrac{1}{t_2} - \tfrac{1}{t_1}\right) + \tfrac{1}{(n-2)}\left(\tfrac{1}{t_3} - \tfrac{1}{t_2}\right)\right] \\ &\cdots \\ Sh_n(c^f) &= k\left[\tfrac{1}{nt_1} + \tfrac{1}{(n-1)}\left(\tfrac{1}{t_2} - \tfrac{1}{t_1}\right) + \tfrac{1}{(n-2)}\left(\tfrac{1}{t_3} - \tfrac{1}{t_2}\right) + \cdots + \left(\tfrac{1}{t_n} - \tfrac{1}{t_{n-1}}\right)\right]. \end{aligned} \tag{2.9}$$

Consequently, if in a certain period of time, we receive a set of patients M, where $M = M_1 \cup \cdots \cup M_n$ and M_i stands for the set of patients for procedure i, $m_i = \#M_i$, we have the following result:

Proposition 2 *An optimal schedule of fees for any user $j \in M$ of the operating-theatre $\left(\sigma_j^*(c)\right)$ is given by:*

$$\begin{aligned} \sigma_j^*(c) &= \tfrac{k}{m_1}\left[\lambda_1 + \tfrac{1}{nt_1}\right] && \text{if } j \in M_1 \\ \sigma_j^*(c) &= \tfrac{k}{m_2}\left[\lambda_2 + \tfrac{1}{nt_1} + \tfrac{1}{(n-1)}\left(\tfrac{1}{t_2} - \tfrac{1}{t_1}\right)\right] && \text{if } j \in M_2 \\ &\cdots\cdots && \cdots\cdots \\ \sigma_j^*(c) &= \tfrac{k}{m_n}\left[\lambda_n + \tfrac{1}{nt_1} + \tfrac{1}{(n-1)}\left(\tfrac{1}{t_2} - \tfrac{1}{t_1}\right) + \cdots + \left(\tfrac{1}{t_n} - \tfrac{1}{t_{n-1}}\right)\right] && \text{if } j \in M_n. \end{aligned} \tag{2.10}$$

2.5 Effects on the Post-operational Costs

We must remember that surgical operations generate more costs than just the ones directly derived from the operation. In almost every case, the patient must spend some time in hospital recovering, what we call "Recovery Time". There are obviously exceptions, like the operations for myopia, after which the patient leaves the hospital immediately. Furthermore,

this recovery time differs across medical procedures, and has a random component, since all patients do not react equally to an operation.

The possibilities of fulfilling a certain maximum average waiting-time guarantee do not depend merely on the capacity of the operating-theatre, but also on the availability of beds for the patients during their recovery time. Actually, there is a second *set of servers* in the system: the beds for recovering patients. We may think of the operating-theatre as the source of patients for this second set of servers. Once a patient leaves the theatre, he should be assigned a bed. Thus, the system only works properly if there are enough beds available for the patients leaving the theatre.

To compute the number of servers (beds) that we need for the adequate functioning of the hospital, we again make use of queueing theory. A classic result from queueing-theory, sometimes referred to as the "equivalence property", ensures that in the steady state the departures from a queueing system with Poisson arrivals at a rate λ is also a Poisson of parameter λ. If patients arrive at hospital beds at a rate λ, their stay in the server is exponentially distributed with mean d, and there are a total of b servers (beds) in the system, the probability that a patient will have to wait for a server is given by Erlang's C formula. Denoting the number of patients in the system by $\mathcal{N}$, we have:

$$P(queueing) = P(\mathcal{N} \geq b) = \frac{\left(\frac{(b\rho)^b}{b!}\right)\left(\frac{1}{1-\rho}\right)}{\left[\sum_{k=0}^{b} \frac{(b\rho)^k}{k!} + \left(\frac{(b\rho)^b}{b!}\right)\left(\frac{1}{1-\rho}\right)\right]}, \tag{2.11}$$

where $\rho = \frac{\lambda d}{b} < 1$ is the necessary and sufficient condition for convergence to the steady-state in the system.

We can set a maximum value to this probability, and then, by solving the previous equation, estimate the number of beds needed, b.

Erlang's formula can be used to compute the number of beds required for the differ-

ent scenarios: (1) If the procedures share neither the theatre nor the beds; (2) If they share the theatre but not the beds; (3) If they do not share the theatre but they do share the beds; and (4) If they share both theatre and beds.

Since we cannot analytically solve the above values with Erlang's formula, it will be used merely for computational purposes. In Section 6, we deal with a numerical example, and we illustrate the method for comparing the different scenarios.

What we can do, however, is to introduce the impact of sharing the operating-theatre on the *average* costs derived from the recovery period, assuming that the different procedures do not share the beds. That is, we can compute the average number of beds required in scenarios 1 and 2 described previously.

2.5.1 Average Post-operational Costs

Sharing or not the operating-theatre alters the way in which the different specialities provide services to their patients. The procedures have a different capacity to perform operations when sharing the operating theatre than when they do not and, hence, they can require a different number of beds. In this subsection, we are interested in studying the impact of sharing the operating-theatre on the post-operational costs.

Consider n fields of treatment, inversely ordered according to their grade of urgency (defined by t_i), so that the one with the highest priority (the one with the lowest average waiting-time guarantee) is the n^{th} medical procedure.

Let d_i with $i = 1, 2, ..., n$, denote the average number of units of time that a patient of type i spends recovering in hospital after an operation. Once again, the service time is exponentially distributed.

To perform the analysis, we need first to define the variable that measures the

number of patients requiring recovery treatment. In this work, we take the operating-theatre rate capacity (μ) as the relevant measure. Although we have already seen that the expected rate of arrivals at hospital beds is λ, by using this variable we would be treating the two set of servers (operating-theatres and hospital beds) as completely independent systems. By considering the capacity of the operating-theatre, we establish a link between the two stages of the process and, hence, we can study how they interact. Moreover, μ is the rate at which the operating-theatre can perform the operations and, hence, it is a measure of the number of operations scheduled per unit of time.

Therefore, we compute the average number of beds required for medical procedure i in the absence of co-operation ($\bar{b}_i$) as a function of the capacity fixed by its operating-theatre and of its patients' expected recovery time.[2] Formally:

$$\bar{b}_i = \mu_i(t_i)d_i = \left(\frac{1}{t_i} + \lambda_i\right) d_i, \; i = 1, 2, ..., n. \tag{2.12}$$

Analogously, the expected number of beds required by the system in the absence of co-operation in the use of the operating-theatre is:

$$\sum_{i=1}^{n} \bar{b}_i = \sum_{i=1}^{n} \frac{d_i}{t_i} + \sum_{i=1}^{n} \lambda_i d_i.$$

When the different medical procedures share the operating-theatre, as all the patients are treated in the same theatre, when we compute the expected number of beds, we need to take the proportion of individuals of each type treated in the considered period of time (p_i), and their corresponding average recovery times (d_i) into account. The number of patients assigned to each medical procedure is expressed as a fraction of the total capacity of the

[2]Throughout this sub-section we implicitly assume that there exists independence between the capacity of the operating-theatre and the length of the recovery time in hospital.

operating-theatre. Formally:

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \mu(T).$$

Therefore, the expected amount of required beds is given by:

$$\bar{b}^N = \sum_{i=1}^{n} p_i d_i = \sum_{i=1}^{n} \frac{\lambda_i \mu(T)}{\sum_{j=1}^{n} \lambda_j} d_i = \left(\frac{1}{T} + \sum_{i=1}^{n} \lambda_i\right) \left(\sum_{i=1}^{n} \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} d_i\right). \quad (2.13)$$

Finally, we assume that the post-operational costs are linear in the number of beds. This reduces the analysis of the costs to the computation of the amount of beds required .

Let us define $\Lambda = \sum_{j=1}^{n} \lambda_j$. In comparing scenarios 1 and 2, therefore, we obtain the following result:

Proposition 3 *Sharing the operating-theatre reduces the average post-operational costs if and only if:*

$$\sum_{i=1}^{n} \lambda_i d_i < \Lambda T \sum_{i=1}^{n} \frac{d_i}{t_i}$$

Proof. We have to compute the difference $\sum_{i=1}^{n} \bar{b}_i - \bar{b}^N$.

We can rewrite (2.13) as:

$$\bar{b}^N = \frac{1}{T} \sum_{i=1}^{n} \frac{\lambda_i}{\Lambda} d_i + \sum_{i=1}^{n} \lambda_i d_i.$$

Thus, $\sum_{i=1}^{n} \bar{b}_i - \bar{b}^N = \sum_{i=1}^{n} \frac{d_i}{t_i} + \sum_{i=1}^{n} \lambda_i d_i - \left(\frac{1}{T} \sum_{i=1}^{n} \frac{\lambda_i}{\Lambda} d_i + \sum_{i=1}^{n} \lambda_i d_i\right)$.

$$\bar{b}^N < \sum_{i=1}^{n} \bar{b}_i \Leftrightarrow \frac{1}{T} \sum_{i=1}^{n} \frac{\lambda_i}{\Lambda} d_i < \sum_{i=1}^{n} \frac{d_i}{t_i}.$$

Namely,

$$\bar{b}^N < \sum_{i=1}^{n} \bar{b}_i \Leftrightarrow \sum_{i=1}^{n} \lambda_i d_i < \Lambda T \sum_{i=1}^{n} \frac{d_i}{t_i}.$$

This completes the proof. ■

This result can be expressed as follows: It is not always true that sharing an operating-theatre leads to lower average costs in the second stage of the treatment process. Formula $\sum_{i=1}^{n} \lambda_i d_i < \Lambda T \sum_{i=1}^{n} \frac{d_i}{t_i}$ can be explained as follows: On the left-hand side, we have the aggregate expected time of recovery in a certain period of time, for all individuals entering the queue (if they are going to be served). On the right-hand side, we can think of a similar recovery time, in which the average stay is given by $d = \sum_{i=1}^{n} \frac{T d_i}{t_i}$. There are cost savings in the recovery period if and only if $\sum_{i=1}^{n} \lambda_i d_i < \Lambda d$.

The explanation for this result has to do with the fact that the number of beds is proportional to the capacity that was fixed at the previous stage. When the different medical procedures cooperate, they agree to set a common capacity that comprises a variable factor that depends on each individual medical procedure (λ_i), and a constant additional term determined by the most demanding one ($\frac{1}{T}$). We have already proven, in Section 3, that this co-operation ensures a saving in the direct costs of the operations. However, the fraction of the total capacity assigned to each procedure may not always be smaller than the one it would set when it is working on its own (we only can ensure that the sum across procedures is always lower). For instance, under co-operation, the less urgent field may perform more operations per unit of time than if it were working on its own (since it is guaranteeing a lower average waiting-time for its patients). Depending on the rate of its patients' arrivals, it may or may not be assigned a higher capacity, which determines the expected amount of beds it requires. Following this reasoning, we can see how the smaller the ratio $\frac{T}{t_i}$ for each and every medical treatment is, the more difficult the condition found in Proposition 3 to guarantee savings to be fulfilled. This ratio provides a measure of how much the capacity of the medical disciplines, which are not first in the priority order, increases with respect to the reference non-cooperative situation.

We will now provide two sufficient conditions that ensure savings in the average post-operational costs from sharing the operating-theatre, and which have a clearer intuition. The following one is an immediate consequence of Proposition 3:

Corollary 1 *If* $\Lambda T > \lambda_i t_i$, $\forall i = 1, 2, ..., n$, *sharing the operating-theatre reduces the average post-operational costs.*

This condition is sufficient for a saving in the required number of beds, since it ensures that each field of specialization (it is trivial for the one with the highest priority), uses in expected terms a fraction of the total capacity that is smaller than the one it would set if it worked alone. Whereas the capacity of medical procedure i when working alone was $\mu_i(t_i) = \frac{1}{t_i} + \lambda_i$, when it cooperates, its share of the global is given by $\frac{\lambda_i}{\Lambda}\frac{1}{T} + \lambda_i$. Therefore, the smaller expected capacity required implies a reduction in the average amount of beds required for post-operational treatment.

When some of the conditions of Corollary 1 do not hold, it means that some medical procedures would demand more beds. If this is the case, the effect over the total average costs is ambiguous, since the extra demand of some medical procedures could be compensated by the smaller requirement of others. This possibility is analyzed in the following corollary.

Corollary 2 *If* $d_n > \frac{\sum_{i=1}^{n-1} \lambda_i d_i}{\Lambda - \lambda_n}$, *sharing the operating-theatre reduces the average post-operational costs.*

Proof. If:

$$d_n > \frac{\sum_{i=1}^{n-1} \lambda_i d_i}{\Lambda - \lambda_n}, \text{ then } d_n > \frac{\sum_{i=1}^{n-1} \lambda_i d_i \left(1 - \frac{t_n \Lambda}{t_i \lambda_i}\right)}{\Lambda - \lambda_n}, \text{ and thus,}$$

$$d_n > \sum_{i=1}^{n-1} d_i \left(\frac{\lambda_i}{\Lambda - \lambda_n} - \frac{t_n}{t_i}\frac{\Lambda}{\Lambda - \lambda_n}\right) \Leftrightarrow d_n \left(\frac{\Lambda - \lambda_n}{\Lambda}\right) >$$

$$> \sum_{i=1}^{n-1} d_i \left(\frac{\lambda_i}{\Lambda} - \frac{t_n}{t_i}\right) \Leftrightarrow \sum_{i=1}^{n-1} d_i \left(\frac{\lambda_i}{t_n \Lambda} - \frac{1}{t_i}\right) < \frac{d_n}{t_n}\left(1 - \frac{\lambda_n}{\Lambda}\right) \Leftrightarrow$$

$$\Longleftrightarrow \frac{1}{t_n}\frac{\lambda_n}{\Lambda}d_n + \frac{1}{t_n}\sum_{i=1}^{n-1}\frac{\lambda_i}{\Lambda}d_i < \frac{d_n}{t_n} + \sum_{i=1}^{n-1}\frac{d_i}{t_i} \Leftrightarrow$$

$$\Longleftrightarrow \frac{1}{T}\sum_{i=1}^{n}\lambda_i d_i < \Lambda\sum_{i=1}^{n}\frac{d_i}{t_i}$$

that is, the hypothesis of Proposition 3 holds. ■

Corollary 2 states that if the medical procedure with the highest priority has a longer expected recovery time than the average recovery time of the rest, by cooperating in the operating-theatre we reduce the expected amount of required beds.

The longer the recovery time in hospital is, the higher the impact of an increase in the capacity of a medical procedure on the amount of beds will be. This makes that when d_n is sufficiently high, the decrease in the expected number of beds needed for medical procedure n definitely exceeds the possible increase in the requirements of the others.

2.6 A Numerical Example

In this section, we provide a numerical example to illustrate, for a particular case, the main features of our analysis. The basis of our example is real data obtained from a small hospital, on the expected number of arrivals of patients and their expected post-operational time.

To set the maximum average-time guarantee for the different procedures, we have taken the actual time spent by the patients in the waiting-lists of these medical disciplines into account. We will consider that the priority of a procedure corresponds to the time guarantee, in such a way that, given two treatments, the one with a shorter waiting-time has priority over the other.

We analyze six medical procedures ($n = 6$), all of an elective nature and with a very short recovery time (even null for some cases). The procedures employed are: cataract

surgery, inguinal hernia operations, varicose veins, arthroscopies, hysterectomies and knee replacements. In the following table, we provide the information required for the construction of the example; all of the variables are measured in monthly terms:

Medical Procedures	λ_i	t_i	d_i
Knee Replacements (r)	12	4	0.266
Cataract Surgery (c)	129	2	0.043
Hysterectomies (hy)	19	1	0.243
Arthroscopies (a)	39	$\frac{1}{2}$	0.083
Inguinal Hernias (h)	33	$\frac{1}{3}$	0.074
Varicose Veins (v)	15	$\frac{1}{3}$	0.083

2.6.1 Operational Costs

We first compute the optimal capacity that each surgical procedure should install on its own, using Equation (2.3):

μ_r	μ_c	μ_{hy}	μ_a	μ_h	μ_v
$\frac{49}{4}$	$\frac{259}{2}$	20	41	36	18

Hence the overall capacity set, in the absence of co-operation is:

$$\sum_{i\in\{r,c,hy,a,h,v\}} \mu_i = 256.75.$$

When all the procedures share the operating-theatre, the capacity ($\mu(T)$) is set according to Equation (2.7), where T is given by $t_v = t_h = \frac{1}{3}$.

$$\mu(T) = 247.33.$$

Hence, if the procedures cooperate, we can reduce the installed capacity of the operating-theatre in an amount of time per month similar to that necessary to perform around 10 operations. As we have assumed the costs to be linear in the capacity installed, this will generate a proportional cost reduction.

The next step is to assign their corresponding costs to the procedures and the patients. Each surgical procedure will be charged with its variable cost (determined by its rate of patients' arrivals λ_i), plus a fraction of the fixed cost. In our case this fixed cost is $\frac{k}{t_v} = 3k$. The sharing is done by following Equations (2.9) and (2.10), and is given by:

Medical Procedures	$Sh_i(c^f)$	$Sh_i(c)$	$\sigma_j^*(c)$
Knee Replacements (r)	$\frac{1}{24}k$	$12.041k$	$1.0020k$
Cataract Surgery (c)	$\frac{11}{120}k$	$129.091k$	$0.9993k$
Hysterectomies (hy)	$\frac{13}{60}k$	$19.216k$	$1.0103k$
Arthroscopies (a)	$\frac{11}{20}k$	$39.55k$	$1.0128k$
Inguinal Hernias (h)	$\frac{21}{20}k$	$34.05k$	$1.0305k$
Varicose Veins (v)	$\frac{21}{20}k$	$16.05k$	$1.0685k$

The table shows how the part of the fixed costs assigned to each procedure $Sh_i(c^f)$ is increasing in the level of priority of the discipline. The sharing of the total costs $Sh_i(c)$, however, does not respect this ranking, since it is also affected by the variable costs, i.e., the different rates of patients arrivals. In fact, we can see how the fixed costs are only a small fraction of the total cost of the service. Finally, in the last column we show what the part of the total costs corresponding to each patient would be, depending on the medical discipline he belongs to.

2.6.2 Post-operational Costs

We move now to the analysis of the post-operational period. As mentioned in Section 5, we consider the beds as the servers of the system in this part of the process, and we can compute the required number of beds under different scenarios: (S_1) If the procedures share neither the theatre nor the beds; (S_2) If they share the theatre but not share the beds; (S_3) If they do not share the theatre but do share the beds; and (S_4) If they share both theatre and beds.

First of all, it is easy to verify that the condition in Proposition 3 is fulfilled even though the conditions in Corollary 1 are not, as cataract surgery does not fulfill the requirement there. Corollary 2 is fulfilled since the average recovery time of one of the two procedures with a highest priority (the inguinal hernias) is smaller than the average post-operational period for the other medical disciplines. Consequently, there is a saving in the average number of beds when we move from scenario (S_1) to scenario (S_2).

By means of Equation (2.12), we first compute the average number of beds required in scenario (S_1). We also compute the average aggregate number of beds required in scenario (S_2), using (2.13), and with the help of the relative frequencies of arrivals, we assign the corresponding fraction of the total beds to each medical discipline.

As stated in Section 5, however, this does not guarantee that every patient has a bed when he leaves the operating-theatre. Using Erlang's C formula, we can fix an upper bound on the probability of not having a bed available when it is needed. Once this probability is fixed, we can compute for the minimum number of beds required to fulfill it. We will denote this value by b_i.

To avoid an excessive expenditure of resources and a situation in which a large

number of beds are "almost always" vacant, we set the probability that a patient waits to .1. Note that by fixing this upper bound we are, in fact, setting the expected waiting-time of a patient to a very low level.

The following table shows the average number of beds ($\bar{b}_i$) and the number of beds required to guarantee that, with probability .9, no patient has to wait (b_i), in scenarios (S_1) and (S_2):[3]

Medical Procedures	$\bar{b}_i^{(S\_1)}$	$\bar{b}_i^{(S\_2)}$	$b_i^{(S\_1)} = b_i^{(S\_2)}$
Knee Replacements (r)	3.26	3.19	6.26
Cataract Surgery (c)	5.57	5.55	9.43
Hysterectomies (hy)	4.86	4.62	8.21
Arthroscopies (a)	3.42	3.25	6.34
Inguinal Hernias (h)	2.67	2.45	5.20
Varicose Veins (v)	1.5	1.25	3.35

Hence, when the procedures share neither the theatre nor the beds (S_1), the total average number of beds is:

$$\sum_{i\in\{r,c,hy,a,h,v\}} \bar{b}_i^{(S\_1)} = 21.28,$$

and the maximum number of beds is:

$$\sum_{i\in\{r,c,hy,a,h,v\}} b_i^{(S\_1)} = 38.79.$$

One can see how the presence of randomness in the post-operational treatment means that in order to ensure a negligible probability of waiting, the number of beds has to be almost doubled from the reference level (computed in expected terms).

[3]Note that $b_i^{(S\_1)} = b_i^{(S\_2)}$ since we are using Equation (2.11) and, hence, the two stages of the model are treated as independent queuing systems.

We now see how the results differ when the pathologies share the operating-theatre (S_2). The average number of beds is:

$$\sum_{i\in\{r,c,hy,a,h,v\}} \bar{b}_i^{(S\_2)} = 20.33,$$

and the maximum number of beds is:

$$\sum_{i\in\{r,c,hy,a,h,v\}} b_i^{(S\_2)} = 38.79.$$

In this framework, we face the same problem as in scenario S_1. When we want to ensure a low probability of waiting, the required capacity almost doubles, and this means to have lot spare beds to cover the risks.

Moreover, we see how by sharing the operating-theatre we can decrease the average post-operational costs, since it allows us to save one bed. Note that, even if this seems to be a rather small improvement, we are dealing with medical disciplines that have a very short recovery time, and therefore require few beds. Hence, the decrease in the number of beds is around 5%.

The next step is to repeat the analysis for the third and fourth scenarios, that is, when the medical procedures share the beds. We distinguish two situations: one in which each medical discipline has its own operating-theatre (S_3), and another in which there is full co-operation, both in the operating-theatre, and in the recovery period (S_4).

In such scenarios, as beds are shared, we consider the rate of the patients' arrivals as the sum of the expected arrivals from the different procedures, and the recovery time as the expected length of stay in hospital for the different medical disciplines, weighted by the proportion of patients demanding beds in each procedure. The results are summarized in the following table:

$\sum_{i\in\{r,c,hy,a,h,v\}} b_i^{(S\_3)}$	$\sum_{i\in\{r,c,hy,a,h,v\}} b_i^{(S\_3)}$	$\sum_{i\in\{r,c,hy,a,h,v\}} \tilde{b}_i^{(S\_4)}$	$\sum_{i\in\{r,c,hy,a,h,v\}} b_i^{(S\_4)}$
21.28	27.23	20.33	27.23

These results are truly illustrative, and several insights can be highlighted. First, that sharing the operating-theatre is always profitable in terms of the average post-operational costs. When we move from scenario (S_3) to (S_4), we also save one bed.

However, the most interesting comparison is the one between scenarios (S_2) and (S_4) (or analogously between (S_1) and (S_3)), in which we see how crucial it is to share the beds. First, it allows us to decrease the extra capacity required to ensure a low probability of waiting by 50%. Using as a reference scenarios (S_2) and (S_4) for instance, we see that the same waiting probability can be ensured by fixing only 7 extra beds in scenario (S_4), instead of the 18 in scenario (S_2). The same thing occurs in confronting situations (S_1) and (S_3). Moreover, it yields a very important saving in the number of beds that have to be installed to ensure the given probability of waiting. Taking the scenarios in which the medical disciplines share the operating-theatre as a reference, we see how if the medical procedures cooperate in the management of the post-operational period, the need for beds is reduced by nearly 30% (approximately 11 beds).

The reason for this reduction can be explained by the same argument used in Subsection 3.3 for the operating-theatre costs. We are treating the beds as servers, and by allowing the different medical procedures to share the beds, we spread risks optimally among them. Therefore, we set a single extra capacity of beds to account for the potential bad realizations of the random variables, instead of making each medical discipline have its own extra capacity. And this has been shown to generate savings.

However, if we proceed to distribute the costs resulting of the co-operation (S_4),

among the different medical procedures, the cost-sharing game we would face is not an "airport game". Although we can identify which medical disciplines require a greater capacities than others, since we are guaranteeing an almost zero waiting probability for all the patients, the number of beds fixed by the most demanding procedure is not enough to ensure that nobody has to wait.

But we can compute the Shapley value of this cost-sharing game, by simply charging each procedure the average of the marginal contribution to all coalitions containing it. The cost-share of procedure i is computed as an average of the marginal cost (marginal number of beds) inflicted by procedure i on each and every coalition $(T\backslash\{i\})$ of other medical disciplines, and it is given by:

$$Sh_i(b) = \sum_{T \subset N, i \in T} \left[\frac{(N - \#T)!(\#T - 1)!}{N!} \right] (C(T) - C(T\backslash\{i\})).$$

We next present the results:

$Sh_r(b)$	$Sh_c(b)$	$Sh_{hy}(b)$	$Sh_a(b)$	$Sh_h(b)$	$Sh_v(b)$
3.8041	7.3502	6.1996	4.488	3.4734	1.9147

As can be seen, the cost-share assigned to each procedure is increasing in the number of beds that it would require if they do not share the servers. Moreover, if we compare these cost-shares with the ones is Scenario S_2, we see how the savings range between the 22% reduction for cataract surgery, and the 43% savings that varicose veins achieve. In this example ,we observe that the most demanding discipline, in terms of beds required (c), is the one that benefits the least from co-operation and, conversely, varicose veins (whose necessity of beds is the least) enjoys the greatest fraction of the savings from co-operation.

2.7 Conclusions

Surgical waiting-lists have been a persistent and unsatisfactory phenomenon in Public Health Services worldwide, since their very inception. They have always been the subject of a great deal of research.

In this paper, we have modeled the problem of the waiting-lists for surgical treatment, making use of queueing theory. We have considered that both, the arrival of new patients to the waiting-list and the process of treatment have random components. The application of the queueing theory results, therefore, arises naturally.

We have made the simplifying assumptions of considering an exponential distribution of the time between two subsequent arrivals and service times. The other extreme would be to assume arrivals and service times that are always constant. Some authors have suggested that the realistic distribution is often somewhere in between (see, for instance, Worthington (1987)).

We are aware that many hospitals are perpetually in transient conditions because of non-homogeneity in arrivals and sometimes servers rates. Almost all of the existing important results of queueing theory are, however, obtained for equilibrium conditions. Consequently, the application of queueing theory to our model can be useful in pointing out the directions in which to proceed for improving the system. The results obtained should be viewed as approximate indicators of the real performance of the system.

We have concentrated on the costs that operations generate, considering that the higher the resources spent by the hospital, the shorter its resulting waiting-list. The aim of our study was two-fold.

On the one hand, we have studied the effects that the use of a common operating-

theatre by the different medical procedures has on the direct costs of an operation. We have shown that the sharing of the operating-theatre leads to cost reduction.

We, then, studied how such a saving should be allocated to the medical procedures through an optimal tariff. Clearly, we are dealing with a cost-allocation problem. Since the Shapley value is a well-known solution concept with good theoretical and computational properties, we have proposed it as the basis for the computation of the optimal fee per medical procedure.

Furthermore, we have extended our analysis to the patients' recovery time. In order to fulfill the maximum average waiting-time guarantee set by the government, it is not only necessary for the operating-theatre to work properly, but also for there to be an adequate supply of beds to accommodate the patients in hospital. It was relevant, therefore, to analyze the impact of co-operation among medical procedures on the post-operational costs.

We found that the sign of the effect that the sharing of the operating-theatre has on the average post-operational costs depends on the characteristics of the treatments and can not be stated in general terms. We computed two sufficient conditions for making a saving at this second stage of the process as well.

To close the model, we have provided a numerical example to illustrate the main features of our model, on the basis of real data, obtained from a small hospital, concerning the expected number of patients' arrivals and the expected length of their recuperation time. We have applied our theoretical analysis to this particular case and interpreted the results that arise. In particular, we have shown that major savings are obtained when the different procedures cooperate in the managing of beds as well.

We have not addressed an issue which would be of interest for further research. One may think that a server (either an operating-theatre or hospital-beds) that handles all the

types of cases could be more expensive to construct (or to maintain) than a server for only one specific type of patients. However, even if a general purpose server is more expensive, it can be easily shown that there is room for cost reduction provided the difference in the costs is not too large. What would be no longer true is that the cost-sharing game resulting from the operating-theatre can be modeled as an "airport game" and, therefore, we would need other tools to compute the Shapley value of the game.

Finally, it is worth noting that we have concentrated only on the costs derived from non-urgent surgery patients. A more ambitious task would be to build up a model in which both emergency and elective surgery are considered. Dealing with such a model would present important technical difficulties to arrive at any general results. The main reason is that we could no longer assume a "first come, first served" queue discipline, as for each medical procedure emergency patients would be given priority over the others.

Bibliography

[1] Baker, M.J. and associates (1965) "Runway Cost Impact Study", Report presented to the Association of Local Transport Airlines. Jackson, Miss.

[2] 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.

[3] Culyer, A.J. and Cullis, J.G. (1976) "Some economics of Hospital Waiting-lists in the NHS". *Journal of Social Policy* 5, 239-264.

[4] Culyer, J.G. and Jones, P.R. (1985) "National Health Service Waiting List: A Discussion of Competing Explanations and a Policy Proposal". *Journal of Health Economics* 4, 119-135.

[5] Fragnelli, V., García-Jurado, I., Norde, H., Patrone, F. and Tijs, S. (2000) "How to Share Railways Infrastructure Costs?". In: *Game Practice. Contributions from Applied Game Theory.* Patrone, F., García-Jurado, I. and Tijs, S. (eds), 91-101. Kluwer Academic Publishers.

[6] Gross, D. and Harris, C.M. (1997) *Fundamentals of Queueing Theory.* John Wiley & Sons (eds), New York.

[7] Hanning, M. and Winblad Spånberg, U. (2000) "Maximum waiting time a threat to clinical freedom?. Implementation of a Policy to Reduce Waiting Times". *Health Policy* 52, 15-32.

[8] Hillier, F.S. and Lieberman, G.J. (1995) *Introduction to Operations Research* (6th edition). McGraw-Hill, New York.

[9] Iversen, T. (1993) "A Theory of Hospital Waiting Lists". *Journal of Health Economics* 12, 55-71.

[10] Johannesson, M., Johansson, P.O. and Söderqvist, T. (1998) "Time Spent on Waiting Lists for Medical Care: an Insurance Approach". *Journal of Health Economics* 17, 627-644.

[11] Joskow, P.L. (1980) "The Effects of Competition and Regulation on Hospital Bed Supply and the Reservation Quality of the Hospital". *The Bell Journal of Economics* 11, 421-447.

[12] Kleinrock, L. (1975) *Queueing systems.* Volume I: theory, John Wiley & Sons (eds), New York.

[13] Littlechild, S.C. and Owen, G. (1973) "A Simple Expression for the Shapley Value in a Special Case". *Management Science* 20-3, 370-372.

[14] Littlechild, S.C. and Thompson, G.F. (1977) "Aircraft Landing Fees: A Game Theory Approach". *Bell Journal of Economics* 8-1, 186-204.

[15] Moulin, H. and Shenker, S. (1996) "Strategyproof Sharing of Submodular Access Costs: Budget Balance versus Efficiency". Mimeo.

[16] Prabhu, N.V. (1997) *Foundations of Queueing Theory.* Kluwer Academic Publishers, Boston.

[17] Shapley, L.S. (1953) "A Value for n-Person Games". In: *Contributions to the Theory of Games II*, Kuhn, H. and Tucker, A.W. (eds), 307-317, Princeton University Press.

[18] Thompson, G.F. (1971) "Airport Costs and Pricing". Unpublished Ph.D. dissertation, University of Birmingham.

[19] Tijs, S. and Driessen, T. (1986) "Game Theory and Cost Allocation Problems". *Management Science* 32, 1015-1028.

[20] Worthington, D.J. (1987) "Queueing Models for Hospital Waiting Lists". *Journal of the Operational Research Society* 38, 413-422.

[21] Young, P. (1994) "Cost Allocation". In: *Handbook of Game Theory* (vol. II), Aumann, R.J and Hart, S. (eds), 1193-1235. North-Holland.

Chapter 3

Policy Implications of Transferring Patients to Private Practice

3.1 Introduction

Public health services, worldwide, are plagued by over-crowding and lengthy waiting-lists. This unsatisfactory situation has persisted from the very inception of most public health systems and, far from improving, it seems to get more systematic over the years. The general population is particularly sensitive to the congestion within the system as they suffer the direct effects of long waiting-lists for urgently needed operations.

The general discomfort caused by the back-log has been forcing several national health authorities, the Spanish Ministry of Health included, to turn to private hospitals and clinics for assistance in reducing their ever-increasing waiting-lists.[1] The Spanish Health Authority, moreover, in an effort to optimize its health system, not only allows certain patients on its Social Security waiting-lists to be treated at private hospitals, but also uses its own

[1]The British government, in addition to this, has recently decided to allow a significant percentage of its patients awaiting surgery to be operated in France.

operating-theaters outside regular working hours.

At the Spanish regional level, the Catalonian government approved a budget of almost seven million Euros to shorten waiting-lists during 2001. The Valencian Region has been undertaking the policy of transferring patients to private hospitals over the last years. As a consequence, more than 100.000 social security patients were treated at private hospitals from July 1996 to June 2000. Moreover, in this region 4.26 million Euros were spent last year to defray the debt to private clinics that have participated in the "Impact Plan" for reducing surgical waiting-lists. In Galicia 12% of the operations financed by the public sector are performed at private hospitals. This generates an extra cost for the Galician government of 120 million Euros per year.[2]

Such temporary programs, however, cannot solve the problem and can turn out to be extremely costly. Finding the correct balance between cost-containment and improvements in the provision of health care services has, therefore, become a major endeavor in most European economies.

This makes the study of the adequacy and optimality of the policies of distributing patients between the public and private sectors crucial. There is quite a lot of controversy over whether hospital specialists are able to influence and manage these waiting-lists for elective surgery to their own private benefit, which would generate an important negative impact on the public sector budget.

It is common in countries with public health services and waiting-lists, that the doctors who work for government hospitals also have their private practice. In the UK, for instance, most private medical services are provided by physicians whose main commitment is to their public sector duties. A report by the Competition Commission (1994), estimated

[2]This information has been obtained from the journal "La Vanguardia" (19th April 2001), the journal "Información" (4th January 2001) and the Journal "Faro de Vigo" (22nd September 2002).

that 61% of NHS physicians in the UK have significant private practices. In addition to this, and according to Yates (1995), an NHS specialist undertakes, on average, two private operations a week. In the Southern European countries, this phenomenon seems to be even more common.

Furthermore, there is a significant difference between the forms of payment to the doctors in the public and private sectors. In the private practice the fact that the physician charges a fee for his services is widely spread, while in the public sector the physician usually has a fixed salary.

These two features, doctors acting in both private and public sector and different remuneration schemes in both sectors, raise a basic matter. Patient-selection (cream-skimming) by the physicians may appear, i.e., the physicians can have incentives to strategically divert the easiest cases to their private practice.[3]

This behavior, moreover, can hardly be avoided, as the evaluation of the diagnostic information required to assess the severity of a patient can only be performed by a trained physician. The control over the severities of the patients who receive treatment in each sector is, therefore, likely to be out of the monitoring capacity of the Health Authority.[4]

The aim of this paper is to analyze the consequences of transferring patients to private practices, and the circumstances under which the Health Authority should implement it.

In our analysis, it is implicitly assumed that the private sector is operating under capacity. This is also an empirical observation. For instance, Bosanquet (1999) states that

[3]Cream-skimming may also appear in other frameworks. For instance, the editorial of The Economist (1998) addresses the criticism directed towards Health Maintenance Organizations in the US for excluding costly cases.

[4]Arrow (1963) was the first to analyze the health care market, taking the differences of information held by the different agents involved into account. Gaynor (1994) and Propper (1995) provide interesting discussions about this topic.

"at present, there is under-occupation in private hospitals (in the UK), with occupancy rates at 50% or less".

Our starting point is a simple model in which the policy-maker (the Health Authority) contracts a hospital specialist for treating patients with different severities, and reaches agreements with private hospitals to have the remaining patients treated there. The Health Authority agrees to pay a fixed fee per operation performed by the private sector.

Our analysis is appropriate for treatments when the patient's condition is not life-threatening in the absence of treatment. These medical disciplines usually require facilities that both the public and the private sectors possess. Moreover, these non-urgent treatments are precisely the ones included in most plans for diverting patients.

The objective of our work is two-fold. On the one hand, we characterize the physician's behavior when the government undertakes a policy of transferring some of the public health patients to private practice. We show that when the government is not able to monitor the physician's behavior with regard to which severities he treats, a problem of cream-skimming arises. The physician will transfer the least severe cases to the private sector.

On the other hand, we also study how this feature affects the decision of the Health Authority concerning two issues: whether to carry out the policy and, if it is eventually implemented, the proportion of patients that should be transferred to the private sector. We show that the presence of cream-skimming reduces the incentives of the Health Authority to undertake the policy. The reason for this is the increase in the costs borne by the public sector due to the existence of patient-selection. The fact that the physician only transfers the mildest cases to private practice, increases the average severity of those patients who remain in the public sector and the expenditure the Health Authority faces also increases.

We find, moreover, that the relevant measure for evaluating the importance of the

problem of patient-selection is the relative dispersion of the severities of the patients. The higher the dispersion is, the more the physician earns from selecting patients and, at the same time, the greater the impact on the costs borne by the Health Authority is.

We also characterize the distortion that the cream-skimming phenomenon imposes on the characteristics of the policy of transferring patients (when it is eventually implemented). This helps us to establish comparisons with the actual performance of these kinds of measures. In this respect, there is empirical evidence supporting the idea that, when patients on the public waiting-list are transferred to private hospitals, the number of operations that are performed in the public sector in a given period of time (which, in principle, should remain unaltered) decrease slightly. Our model provides rationality to this phenomenon, based on the strategic behavior of the physicians. Since they keep the most severe patients in the public sector, the amount of patients that can be treated there, for a given level of effort, is reduced.

The physician's response to the form of the compensation contract has been widely covered by the literature, which generally compares retrospective with capitation reimbursement methods. The main concern of these works is the effect of the reimbursement rule on either the intensity of health services (see, for instance, Ellis and McGuire (1986, 1990), Selden (1990), Blomqvist (1991), and Rickman and McGuire (1999)), or on the provider selection of who will be treated (see, for instance, Dranove (1997)), or both on the intensity and extent of the treatment (see, for instance, Ma (1994), Ellis (1999)). Only Rickman and McGuire (1999) consider, as we do, the fact that the physician can supply either private health-care to a patient or public.

In our model, intensity of treatment is not considered, as we focus exclusively on the physician's selection of patients. Even if this was the aim of some of the papers mentioned

above, our approach is quite different from theirs. Our analysis is rather positive, as we take the remuneration system as given (and fixed) by the institutional framework and we study the potential strategic behavior of the physician. In our work, cream-skimming does not appear as a consequence of the remuneration system chosen by the Public Authority, but rather due to the different structure of payments in the public and private sectors.

This paper should also be included in the literature that considers a mix of public and private sector services provided by physicians. Iversen (1997) has modelled the impact of public sector waiting-lists on the demand for private care. He concentrates on the patient's decision. He assumes that all patients who are willing to pay for private treatment are served in the private sector. As such, he clearly rules out the possibility of cream-skimming on the part of the doctors.

Patient-selection by the physician is also ruled out in Olivella (2002). He analyzes the incentives of the public health administration (fixing long waiting times for public treatments) to divert costs from the public to the private sector and studies the conditions under which this deviation enhances welfare. In our model, such a behavior does not appear since it is the Health Authority who pays the cost of treating all the social security patients (independently of where they are treated).

Finally, the doctor's strategic behavior plays an important role in Barros and Olivella (1999). However, the different systems of remunerating the physicians in either sector (which is a crucial variable in our model) is not considered in their work. Patient-selection arises mainly from a combination of the rationing policy undertaken by the Health Authority, the criterion of the private physician regarding which severities he is willing to treat and the decision of the patients to leave the queue in the public sector and resort to private treatment (paying a flat fee). In our model, patients are fully insured irrespectively of the sector where

they are treated and no rationing policy is considered. This gives the full power to decide to the physician and, thus, always generates a situation of "full cream-skimming".[5]

The rest of the paper is organized as follows: In the following section we present the model. Section 3 computes the optimal policy in the benchmark scenario. In section 4 we study the behavior of the physician concerning the selection of patients and the response of the Health Authority. In Section 5, we show how our results can be extended to several variants of our model. Finally, Section 6 provides some concluding remarks and the policy implications of our analysis. All of the proofs are in the Appendix.

3.2 The Model

There is a continuum of individuals requiring health care, all of whom demand elective treatment. The size of this population of potential patients is normalized to N. These patients are homogeneous, except for their degree of severity, which is measured by the random variable s. This variable is distributed according to a density function $f(s)$ defined on $(\underline{s}, \overline{s})$ which we assume to be uniform. Let $\Delta s = \overline{s} - \underline{s}$ be the difference between the extremes of the domain for s. A patient with severity s is assumed to obtain a benefit from a treatment defined by Qs, (Q can be, for instance, a monetary value associated to the QALYs).

We consider a situation in which the social pressure on the Health Authority to reduce the excessive congestion in the public health service is severe. To do so, the Health Authority commits to treat all the patients (N) within a given period of time. Note that this construction is equivalent to considering that only a fraction of patients is treated, and that

[5] According to Barros and Olivella (1999), "full cream-skimming" is a situation in which all the mildest patients end up being treated in the private sector.

this amount is exogenously given in our model and reflected by N.

For this purpose, the Health Authority contracts an agent, represented by a specialist, to treat a certain number of patients and reaches agreements with private hospitals to have the remaining patients treated there. We denote by x the number of operations performed in the public sector; hence, $N-x$ patients will be transferred to the private sector. Since we are dealing with public health systems, we assume that patients are fully insured irrespectively of the sector where they are eventually treated.

When treating patients, the Health Authority incurs in two different kinds of costs in the public sector: a transfer T to the physician it contracts and a constant cost of treatment k^{pb} per patient. We take the cost of the public treatment to be linear for the sake of expositional clarity. In Section 5, we provide some insights into the robustness of our results to other more general cost structures, namely, costs increasing in the average severity of the operations, dis-economies of scale or capacity constraints in the public sector.

Moreover, in order to be consistent with real-life observations, we assume that the Health Authority makes a constant payment for each operation sent to the private sector. This payment covers both the fee agreed with the private specialist (w) and the cost of providing treatment to each patient in the private sector (k^{pv}). The private cost of treatment is likely to be linear if the private sector is operating well under capacity.

With this construction, we allow for differences in the cost of providing treatment in both systems. For the sake of clarity, we define $\Delta k = k^{pb} - k^{pv}$ as the difference between the costs of treatment in either sector, disregarding physician's compensations. Although, in principle, we impose no restrictions on the sign of Δk, it may be reasonable to consider the case of $\Delta k > 0$ more likely to occur. This can be sustained by several reasons. First, on the grounds of dis-economies of scale or congestion problems in the public sector, as well as

bureaucratic or administrative inefficiencies. Second, there are other reasons related to the specific nature of the public health sector such as a more stochastic demand (more emergency cases), the training of new physicians or the development of research activities.[6]

We model the physicians' behavior as being that of a single representative agent. As we argued in the Introduction, the fact that the same doctor may work in both private and public practice is a common feature in Europe. We model this by assuming that the doctor who undertakes the operations in the public sector (in the morning, say) also works for a private hospital (in the afternoon).[7] In defining the utility of the physician, therefore, we must not only take his revenues and costs in the public sector into account, but those in his private practice as well.

In order to perform his tasks, the physician has to exert some effort (e^{pb} for the patients he treats in the public sector and e^{pv} for those treated in private). These levels of effort depend on the number of patients he treats in each sector (x and $N-x$, respectively) and on the average severity of the patients ($\widehat{s}^{pb}$ and $\widehat{s}^{pv}$). We define the effort as the product of these two components (amount of patients and average severity). The cost borne by the physician is also affected by a parameter θ, which measures the physician's skills or knowledge to be able to perform his tasks. We consider, however, that all the physicians share the same level of ability, which is common-knowledge among all the agents in the model.[8] As such, the costs of the effort exerted by the physician in each sector are given by:

$$\Psi^{pb} = \Psi(e^{pb},\theta) = \Psi(x\widehat{s}^{pb},\theta)$$

$$\Psi^{pv} = \Psi(e^{pv},\theta) = \Psi((N-x)\,\widehat{s}^{pv},\theta).$$

[6] In favor of this argument, data from Norway indicates that for some types of treatment the price charged by private hospitals is considerably lower than the costs in public hospitals. A further discussion of these cost differences is given in Hoel and Sæther (2000).

[7] The private and public practice of the physician may even be done in the same hospital, under different types of contracts.

[8] In Sub-section 5.2 we will study the robustness of the results to the presence of heterogeneous physicians.

The function $\Psi(\cdot)$ is increasing and convex in the level of effort exerted and decreasing in the physician's ability. Moreover, we assume that $\Psi(0,\theta) = \Psi_{e^i}(0,\theta) = 0$, $\forall i = pb, pv$.

We have chosen a construction with separable cost effort functions for the sake of analytical convenience. It allows us to provide a closed form to the utility of the physician in either sector, what will be very useful for the derivation of the results. However, other more general structures with a single cost function that reflects the cost of effort in the two sectors could also be used. The main features of the model would remain unaltered, provided such a cost function is increasing and convex in the total amount of effort exerted by the physician.

In general, a construction with separable efforts may yield situations in which the physician has an incentive to distribute patients and severities between the two sectors, as a way to decrease his dis-utility by the total effort exerted (cost-induced patient selection). As we will see, this possibility is ruled out in our model by the physician's lack of strategic capacity in the choice of e^{pb}.

We can define the utility function of the physician, as follows:

$$U^s = T + w\,(N - x) - \Psi(x\widehat{s}^{pb}, \theta) - \Psi((N - x)\,\widehat{s}^{pv}, \theta). \tag{3.1}$$

The aim of this work is to study the potential strategic behavior that a physician may have, in his performance as a dual supplier. We, therefore, specifically ignore the possibility of the physician's strategically behaving within either sector, in the sense of exerting little effort (shirking). We consider, then, that the physician will exert the maximum level of effort that he considers compatible with his earnings, either by ethical commitment or because he is fully monitored. Hence, he will treat patients in the public sector as long as his net revenues do not fall below his reservation value (normalized to zero in this model).[9]

[9] In this model the physician treats all the patients he receives in the private sector. As such, there is no room for strategic behavior in the effort he exerts in his private practice.

The Health Authority's surplus derived from the care provided is given by the difference between the social net benefit of the treatment and the social cost that the production of the services generates. Hence, denoting by $\widehat{s}$ the average severity of the potential population of patients, i.e., $\widehat{s} = \frac{\overline{s}+\underline{s}}{2}$, the government's objective function is as follows:

$$H = Q\widehat{s}N - \left[T + k^{pb}x + (k^{pv} + w)(N - x)\right].$$

Or, by re-arranging terms we get:

$$H = (Q\widehat{s} - k^{pv})N - [T + \Delta kx + w(N - x)]. \tag{3.2}$$

Since all the patients eventually receive the treatment they need, maximizing this objective function is equivalent to minimizing the costs derived from undertaking the policy.

Note that we assume that the Health Authority does not take the utility function of the physician into consideration. In other words, the government is not maximizing a social welfare function.

The timing of the game is as follows: At a stage prior to the starting-point of our model, the Health Authority and the private hospital (private physician) bargain over the value of the private fee w that the private physician will receive per operation.[10] At the first stage, the Health Authority contracts a specialist, specifying the salary he will receive (T). At the second stage, the physician takes two simultaneous decisions: On the one hand, he selects the severities that he wants to treat in each sector. On the other hand, he decides on the number of operations he will perform in his public duty, and the remaining patients will be transferred to the private hospital. Finally, the whole population of patients receives treatment and the payoffs are realized.

[10] We show later that, in equilibrium, the bargaining set is not empty, i.e., the maximum wage the Health Authority is willing to pay exceeds the minimum the physician will accept for attending to public patients in his private practice.

We confront two different frameworks. In the first one, it is assumed that the specialist can not select the severities to be treated in either system. Stage 2 is, therefore, only partially active in this initial set-up. In the second scenario we consider that, since the actual severities of the patients can only be known by specialized physicians, the Health Authority cannot monitor the physician in his selection of the patients who are to receive treatment in either sector.

We start by analyzing the optimal policy in the first setting.

3.3 Benchmark Scenario

In this section, we assume that the Health Authority can preclude the physician's selecting the patients he wants to treat in either sector. Hence, patients will be uniformly distributed between the public and the private sectors. We can thus ensure that the average severity of the operations is the same in both sectors, i.e., $\widehat{s}^{pb} = \widehat{s}^{pv} = \widehat{s}$.

In order to guarantee the existence of an interior solution in this framework, we make the following assumption.

Assumption 1 $0 < w - \Delta k < \Psi_x(N\widehat{s}, \theta)$.

Under Assumption 1, the difference between the private fee and the treatment-cost differential has to be positive and bounded above by a certain value. With this assumption we are only requiring that: On the one hand, if the private sector is less costly in terms of the treatment provided, we do not want this difference to be so high that it compensates the fee paid to the private physician. If this was the case, the public sector could purchase all the health services from the private sector instead of providing them, i.e., it would be trivially optimal to send all patients to the private sector. On the other hand, we also require that

the private fee not be so high that the policy of transferring patients is not undertaken, even in this framework where manipulation is not possible.

In order to characterize the solution in this framework, we solve the game by backwards induction. At the second stage, the physician chooses the amount of operations he will perform in the public sector. He will treat patients up to the point at which performing an additional operation would force him to make a loss. Therefore, the number of operations performed in the public sector, x, is such that $\Psi(x\widehat{s},\theta) = T$. We denote it by $x(T)$.

In the first stage, the Health Authority maximizes its objective function. The optimization program that the government faces is as follows:

$$\max_{T} H = (Q\widehat{s} - k^{pv})N - [T + \Delta kx + w(N - x)]$$

$$\text{s.t.} \qquad x = x(T).$$

The following lemma characterizes the optimal sharing of patients between public and private practice and the salary that induces it.[11]

Lemma 1 *In the benchmark scenario, the optimal number of patients treated in the public sector* (x^*) *and the salary the physician receives* (T^*) *are such that:*

$$\Psi_x(x^*\widehat{s},\theta) + k^{pb} = w + k^{pv},$$

with $\Psi(x^*\widehat{s},\theta) = T^*$.

With the above lemma we have computed the optimal policy in the First Best scenario. This will be our reference case for comparison with the results in the next section.

The salary the Health Authority pays to the physician induces him to perform in his public practice the number of operations that equalizes the marginal costs of treating patients in both sectors.

[11] The proof of this lemma is straightforward and therefore we ommit it.

It is straightforward to verify that the optimal level of patients treated in the public sector is increasing in the physician's ability and in the private fee w. We can also easily see the effect of the treatment-cost differential. If $\Delta k > 0$, the Health Authority incurs a smaller cost per operation in the private sector and, hence, is willing to transfer more patients for a given value of w. Moreover, since the fee per operation paid to the physician in the private sector is fixed (independent of the patient's severity), the optimal number of patients treated in the public sector is decreasing in the average severity of the population ($\widehat{s}$).

We proceed now to study the effects of dealing with a physician who can strategically choose the kind of patients to be treated in either sector. This will allow us to analyze the consequences of this potential strategic behavior on the willingness of the Health Authority to undertake the policy.

3.4 Patient Selection

Our concern in this section is to analyze whether the results differ when the physician has the ability to select patients and decide which cases to treat in his public practice and which go to the private sector. In other words, the Health Authority is not able to monitor the physician's choice of the severities of the patients treated in either sector, and this variable cannot be included in the terms of the contract.

As we have argued in the Introduction, this is an issue of great controversy in mixed health-care systems, as the diagnosis process that leads to the assessment of the severity of a given patient can only be performed by a trained physician. This means that the control over the severities of the patients who receive treatment in either sector is probably not possible for the Health Authority. Therefore, as the patients will no longer be randomly distributed

between the public and private sectors, we cannot ensure, in general, that the average severity of the patients treated in both systems is the same.

To characterize the solution in this framework, we proceed to solve again the game by backwards induction. At the second stage, the physician decides on the amount of operations he will perform in the public sector. His optimal number of operations does not depend merely on the salary he receives, since the average severity of the patients he treats $(\widehat{s}^{pb})$ is also a variable of choice. Therefore, x is such that $\Psi\left(x\widehat{s}^{pb},\theta\right)=T$ and we denote it by $x\left(T,\widehat{s}^{pb}\right)$.

The physician also decides which severities he wants to treat in either sector, subject to the restriction that $x\left(T,\widehat{s}^{pb}\right)$ operations have to be performed in his public practice. Therefore, he does not have complete freedom in the choice of $\widehat{s}^{pb}$, as there may be values that are not compatible with the sharing of patients set. The physician will choose the value of $\widehat{s}^{pb}$ (and therefore also of $\widehat{s}^{pv}$) in order to maximize his total revenue. Since he is a dual supplier, he will consider the effects of his strategic behavior concerning his two sources of income. Therefore, the program he faces is:

$$\begin{aligned}\max_{\widehat{s}^{pb}}\ U^{s} &= T+w\left(N-x\right)-\Psi(x\widehat{s}^{pb},\theta)-\Psi((N-x)\,\widehat{s}^{pv},\theta)\\ s.t\qquad x &= x\left(T,\widehat{s}^{pb}\right).\end{aligned}\tag{3.3}$$

In the following proposition we characterize the physician's behavior concerning the selection of patients.

Proposition 1 *For a given sharing of patients between the two sectors (x and $N-x$), the specialist will transfer the least severe cases to the private practice. Formally:*

A patient with severity s will be treated in the public practice if and only if:

$$s \in \left(\bar{s} - \frac{\Delta s x}{N},\ \bar{s} \right).$$

This proposition shows that the physician wants to treat only the mildest cases in the private practice, leaving the most difficult ones for the public sector. This behavior, known in the literature as "cream-skimming", is caused primarily by the difference between the physician's remunerations from the two systems. In the public sector, the physician receives a salary whereas, in the private sector, his earnings are on a fixed fee-for-service basis. Therefore, the more operations the physician performs in his private practice, the higher the earnings he obtains. Furthermore, for a given level of effort exerted, (i.e., for a given cost), the "easier" the operations he performs the more patients he can treat. Note that we are not facing a situation of cost-induced patient-selection. Cream-skimming appears in our model not as an attempt by the physician to incur smaller costs, but rather as a way of increasing his earnings.

At this point the reader may wonder whether the physician's lack of strategic capacity on choosing the number of patients to be treated in the public sector is crucial for the problem of cream-skimming to arise. The answer is no. It can be shown that, under the cost effort structure we have chosen, patient selection by the physician would still appear if he is not restricted to perform the maximum number of operations compatible with his public earnings. With this modification, however, it would not be possible to study the reaction of the health authority to the cream-skimming phenomenon, as the health authority could not use the choice of T as a way to influence physician's decision.

From Proposition 1, we obtain that the outcome of stage 2 consists of a pair $\left(\widehat{s}^{pb}, x\right)$

that simultaneously fulfill the following conditions:

$$\widehat{s}^{pb} = \bar{s} - \frac{\Delta s x}{2N}$$
$$x = x\left(T, \widehat{s}^{pb}\right).$$

We shall now study how this problem of cream-skimming affects the decision of the Health Authority on when to undertake such a policy, and on the amount of patients that should be transferred to the private sector.

In the first stage, when the cost of implementing the policy is being considering, the Health Authority should take the fact that the most severe cases will be treated in the public sector into account.

The maximization program of the Health Authority in this scenario is as follows:

$$\max_{T} \ \left(Q\widehat{s} - k^{pv}\right) N - \left[T + \Delta k x + w\left(N - x\right)\right]$$
$$\text{s.t.} \quad \begin{cases} x = x\left(T, \widehat{s}^{pb}\right) \\ \widehat{s}^{pb} = \bar{s} - \frac{\Delta s x}{2N}. \end{cases} \tag{3.4}$$

The following lemma provides the interior candidate for solution in this framework. Note that when the patient-selection arises, the Health Authority's objective function is much more complicated. In particular, it is no longer true that the function is always concave. The first order necessary condition for optimality is, therefore, not sufficient in general.[12]

Lemma 2 *With patient-selection, in the interior candidate for solution, the number of patients treated in the public sector* (x^m) *and the salary the physician receives* (T^m) *are such that:*

$$\Psi_x\left(x^m\widehat{s}^{pb}, \theta\right) + k^{pb} - \frac{\Delta s}{2N}\Psi_{\widehat{s}^{pb}}\left(x^m\widehat{s}^{pb}, \theta\right) = w + k^{pv},$$

[12]Moreover, we need to impose a regularity condition in the cost of effort function to ensure that the f.o.c is well defined. If the total effect of a change in x on the cost of effort is positive, i.e. $\frac{d\Psi_x\left(x\widehat{s}^{pb}, \theta\right)}{dx} = \Psi_x\left(x\widehat{s}^{pb}, \theta\right) - \frac{\Delta s}{2N}\Psi_{\widehat{s}^{pb}}\left(x\widehat{s}^{pb}, \theta\right) > 0$ for every $x \in [0, N]$, an interior candidate for optimum exists. Under this condition, the proof of the lemma is straightforward and is therefore omitted.

with $\Psi\left(x^m\widehat{s}^{pb},\theta\right)=T^m$ *and* $\widehat{s}^{pb}=\bar{s}-\frac{\Delta s x^m}{2N}$.

With patient-selection, new effects appear in the Health Authority's first order condition which will determine the optimal policy. The cost per operation in the public sector has now increased, as the average severity of the patients treated there is higher. This makes the treatment provided by the public sector more expensive and, hence, leads to a greater transfer of patients to private practice. An additional effect, which goes in the opposite direction, however, appears: $\frac{\Delta s}{2N}\Psi_{\widehat{s}^{pb}}\left(x^m\widehat{s}^{pb},\theta\right)$ reflects how an increase in the number of operations in the public sector has a positive impact on the public costs (through the decrease it induces in the average severity of the patients treated there).

To be able to close the model and perform comparisons with the benchmark case, we need to consider a specific cost of effort function. This will allow us to characterize the effects of the cream-skimming phenomenon on the behavior of the government and on its willingness to undertake the policy.

Hence, hereinafter, the dis-utility of the physician's efforts in either sector is given by:

$$\Psi^{pb}=\frac{1}{2}\left(\frac{x}{\theta}\widehat{s}^{pb}\right)^2 \text{ and } \Psi^{pv}=\frac{1}{2}\left(\frac{N-x}{\theta}\widehat{s}^{pv}\right)^2.$$

Before proceeding to analyze the optimal response from the Health Authority, we need to study the curvature of its objective function. Since the amount of patients that receive public treatment is determined by the salary the physician receives, it is equivalent to the Health Authority's deciding directly on the salary or on the number of patients to be treated. For the sake of notational clarity, let $d=\frac{\Delta s}{\widehat{s}}$ denote the relative dispersion of the patients' severities. The following lemma characterizes the curvature as a function of x.

Lemma 3 *The curvature of the Health Authority's objective function, under patient-selection*

(H^m), is as follows:

1. *If $d \leq 4 - 2\sqrt{3}$, then H^m is always concave.*

2. *If $d > 4 - 2\sqrt{3}$, then:*

   (a) *For any $x \in \left[0, \bar{\beta}(d) N\right]$, H^m is concave at x.*

   (b) *For any $x \in \left(\bar{\beta}(d) N,\ N\right]$, H^m is convex at x.*

   With $\bar{\beta}(d) = \left(\frac{1}{d} + \frac{1}{2}\right) \frac{3-\sqrt{3}}{3}$, $\bar{\beta}'(d) < 0$.

Lemma 3 allows us to study the curvature of the Health Authority's objective function, in terms of the relative dispersion of the patients' severities, measured as the ratio of the difference between the boundary severities ($\bar{s}$ and $\underline{s}$) and the average severity ($\widehat{s}$). This lemma highlights the relevance of the relative dispersion, as it is a measure of how serious the problem of cream-skimming is. The strategic behavior of the physician (in transferring the mildest cases to the private practice) is fostered by the wide range of severities, since his gains in diverting patients are higher. We show that when the relative dispersion of the severities is low, the objective function is still concave in the entire domain. Thus, when the Health Authority does not suffer much from the problem of patient-selection, the program has a unique candidate to optimum. If the severities are sufficiently dispersed, however, the objective function has a convex section, that is bigger the higher the dispersion is. As a consequence of this feature, we may have two candidates to optimum: an interior one and the boundary solution (no transfer of patients to private practice).

The following proposition presents the solution to the government's maximization problem.

Proposition 2 *In the presence of patient-selection, the Health Authority decides, through the choice of the salary, to undertake the policy of transferring patients to private practice if the value of the private fee is below a certain threshold. The higher the relative dispersion of the patients' severities is, the more demanding this condition is. Formally:*

1. T^m *is such that* $x^m < N$ *if* $w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2} G(d)$.

2. $T^m = \frac{1}{2}\left(\frac{N\widehat{s}}{\theta}\right)^2$ *is such that* $x^m = N$, *otherwise.*

Where $G(d) = \begin{cases} 1 - \frac{d}{2} \text{ if } d \leq 4 - 2\sqrt{3} \\ g(d) \text{ if } d \geq 4 - 2\sqrt{3} \end{cases}$ *is a continuous function and* $g(d)$ *is such that* $g'(d) < 0$ *and* $\lim_{d \to \infty} g(d) = \frac{1}{2}$.

This result shows that the presence of "cream-skimming" can lead to a situation in which the Health Authority is no longer willing to transfer patients to the private sector. It is straightforward to verify that, for this particular effort cost function that we have considered here, in the absence of patient-selection, the decision of the Health Authority will be to distribute patients between the two sectors as long as $w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2} = \bar{w}$ (and, by Assumption 1, this condition always holds). If patient-selection by the physician cannot be avoided, this condition is more demanding. For the government to undertake this policy of distribution of patients between sectors, the upper bound of the private fee is now smaller ($\frac{\widehat{s}^2 N}{\theta^2} G(d) < \frac{\widehat{s}^2 N}{\theta^2}$ since $G(d)$ is always less than one).

The reason for this result is that by choosing to pay a salary $T^m = \frac{1}{2}\left(\frac{N\widehat{s}}{\theta}\right)^2$, the Health Authority is completely eliminating the possibility of patient-selection. Inducing the physician (through the salary it pays to him) to treat all the patients in the public sector, there is no chance of avoiding the mildest cases. When the private fee is sufficiently low,

however, the Health Authority decides to suffer the "cream-skimming" problem, in order to bear a lower cost for the patients transferred.

Moreover, the threshold of the private fee from which the Health Authority is willing to carry out the policy, is decreasing in the relative dispersion of the patients' severities. When the relative dispersion of the severities is low, the Health Authority does not suffer the problem of cream-skimming very much (since all the patients have similar levels of severity). In this case, the policy of transferring patients to private practice is undertaken for a wide range of values of w. In particular, when $d \to 0$, the condition for undertaking the policy converges to the one required in Assumption 1. However, as the relative dispersion of the severities increases, the condition necessary for the public authority to reach agreements with private hospitals becomes more demanding. Since the treatment of some patients in the private sector increases the cost per operation in the public sector, the policy is not undertaken unless the private fee is sufficiently low.

We now compare the interior solution in this setting to the optimal one in the benchmark case.

Proposition 3 *When the Health Authority is willing to undertake the policy, the existence of patient-selection implies that:*

i).- If the relative dispersion of the severities is sufficiently low, the Health Authority induces a higher transfer of patients to private practice, provided that the private fee is below a certain value. Otherwise, there is a lower transfer.

ii).- If the relative dispersion of the severities exceeds a critical value, the Health Authority always induces a higher transfer of patients to private practice.

Formally:

1. $x^m < x^*$, *when* $d \geq .7625$ *and when* $d < .7625$ *and* $w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2} \Phi(d)$.

2. $x^m > x^*$, *when* $d < .7625$ *and* $w - \Delta k > \frac{\widehat{s}^2 N}{\theta^2} \Phi(d)$.

With $\Phi(d) = \left[\frac{1}{4d}\left(3d + 6 - \sqrt{d^2 + 4d + 36}\right)\right]$.

This proposition shows that when the policy of transferring patients is undertaken, the consequences of the cream-skimming on the number of patients transferred differ in the relative dispersion of the severities. This result is a consequence of two contrary effects. On the one hand, the marginal cost of treating an extra patient in the public sector is higher with patient-selection, as the average severity of the patients treated is higher. On the other hand, by increasing the number of operations in the public sector we not only save the private fee (w), but also reduce the possibility of cream-skimming and, hence, decrease the expected level of severity that the Health Authority faces. When the relative dispersion is sufficiently high, the negative effect of treating patients in the public sector always dominates and, thus, fewer patients are kept in the public sector, even if we already know that this increases the capacity of the physicians to select patients. In contrast, when the relative dispersion is low, the final result is determined by the value of the private fee. A high level of w implied a relatively low transfer of patients in the benchmark scenario; we show that, in this case, the presence of cream-skimming leads to an even smaller transfer. Conversely, when the first best situation was to transfer a high proportion of the patients (low w) to private practice, the distortion implies to transferring even more. Figure 1 outlines all of these possibilities.

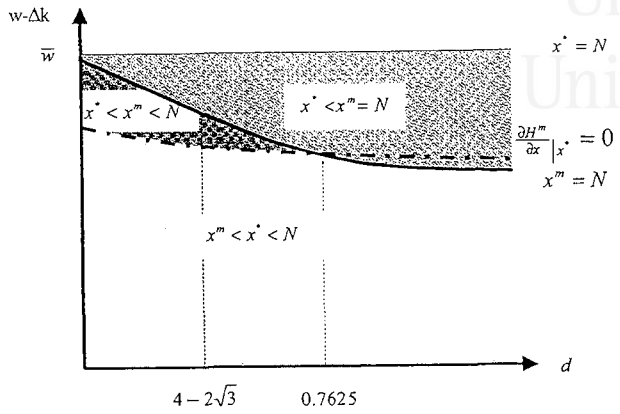

Figure 1: Sharing of patients with and without patient selection.

Proposition 3 can also be interpreted in terms of the salary paid to the physician. In this dimension, however, it is more difficult to obtain the sign of the distortion. The reason for this is that, when the presence of cream-skimming induces a higher transfer of patients to private practice, two effects come into conflict: First, fewer patients are treated in the public sector, which implies lower costs for the physician and a smaller salary, and secondly, the patients who are treated are relatively more severe (hence, they induce a higher salary to cover the physician's effort cost). As a result, the impact of patient-selection on the salary the physician receives is ambiguous in this region. Nevertheless, when the relative dispersion of the severities is low, and the private fee sufficiently high, there is no such ambiguity. In equilibrium, fewer patients are transferred to the private sector. Moreover, due to patient-selection, the patients who are left in the public sector are the most severe ones. In this region, therefore, the existence of cream-skimming induces the Health Authority to pay a higher salary to the physician.

Figure 2 compares the objective functions of the Health Authority and the optimal

sharing of patients under the alternative scenarios we have studied: H^* denotes the objective function in the benchmark case, whereas H^m stands for the one under patient-selection. This illustration is made for the case in which the policy is implemented and the relative dispersion of the severities is sufficiently high ($d > .7625$).

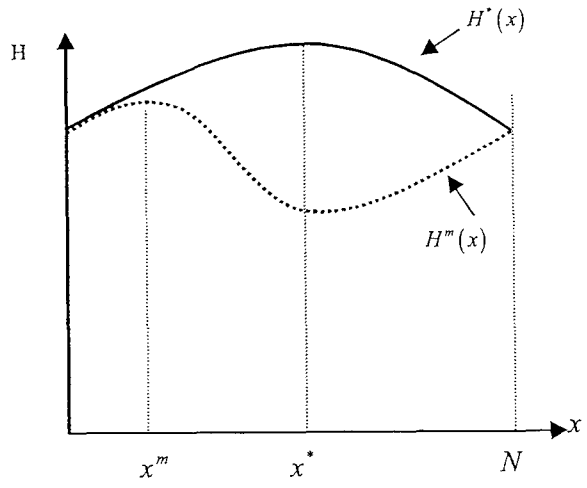

Figure 2: Health Authority's objective functions when $d > .7625$.

We have taken the negotiations between the government and private hospitals, about the value of the fee to be paid per operation performed in the private sector, as given in our model. It is crucial, however, to know whether the equilibrium values we have computed leave room for such a bargaining process. That is, if there are values of w that make the Health Authority willing to undertake the policy $\left(\text{i.e. } w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2} G(d)\right)$ and, at the same time, are acceptable to the physician $\left(w(N-x) \geq \frac{1}{2}\left(\frac{N-x}{\theta}\widehat{s}^{pv}\right)^2\right)$. The result is presented in the following remark:

Remark 1 *In equilibrium, and for $\Delta k \geq 0$, the bargaining set is not empty (for every value of d), i.e., the maximum wage the Health Authority is willing to pay exceeds the minimum that the physician requires to accept public patients in his private practice.*

This remark shows that if the private sector is not more inefficient than the public sector is, we can ensure that there are values of w that are both physician and government compatible, independently of the relative dispersion of the severities. If $\Delta k < 0$, there also exists room for negotiation, provided the relative dispersion of the severities is not too high. Figure 3 illustrates the bargaining set between the Health Authority and the physician for $\Delta k \geq 0$.

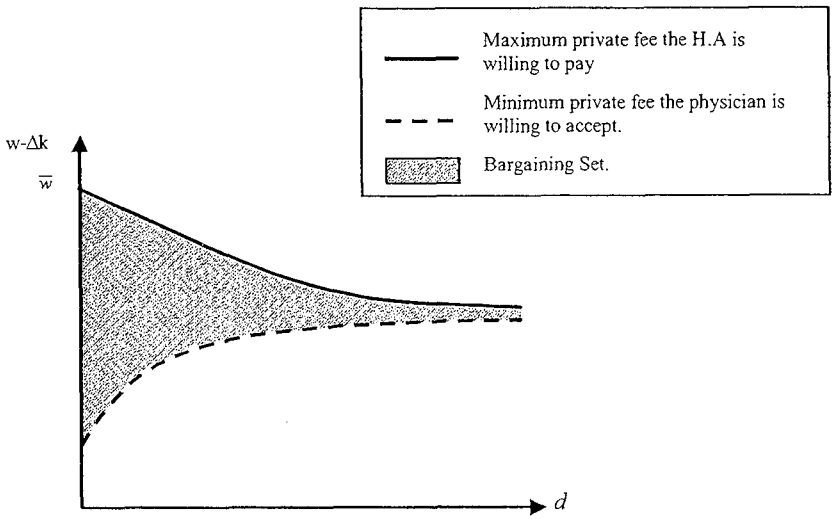

Figure 3: Bargaining set for $\Delta k \geq 0$.

3.5 Comments and Extensions

In this paper we have shown that when physicians are dual providers, a problem of cream-skimming may arise. This strategic behavior makes the government less willing to undertake a policy of transferring some of the public sector's patients to private hospitals. When the policy is undertaken, however, in most of the domain of the variables, more patients are finally treated in private practice than in the absence of cream-skimming.

In this section, we discuss some of the ingredients of our model by proposing alter-

native constructions and providing insights into their impact on the results. A first modification would be to alter the distribution of patients' severities. We have chosen to work with a uniform distribution of patients' severities for analytical convenience, even though a distribution in which the severities are more concentrated around the mean might seem to be more realistic. Despite we do not formally analyze such a possibility in this section, we guess that this would not alter the main features of the model. There would be incentives for patient-selection, and the other results should not be qualitatively altered.

We will devote this section to introduce modifications that affect the structure of the costs of treatment in the public sector, and allow for heterogeneous physicians in the model.

3.5.1 Cost of Public Treatment

Our assumption of a constant marginal cost of treatment in the public sector, was made for the sake of clarity in the presentation. The construction we have chosen allows us to concentrate on the physician's incentives in selecting patients and to fully characterize the Health Authority's response to such behavior.

Nevertheless, other more general structures can be considered for the costs. The alternatives that we have in mind are: (1) Marginal cost of treatment increasing with the average severity of the patients; (2) Marginal costs increasing in the number of treatments provided (dis-economies of scale); and (3) Capacity constraints in the public sector.

In all these set-ups, the incentives of the physician to select patients remain unaltered and, hence, the same problem of cream-skimming arises. Moreover, the crucial measure for assessing the seriousness of the problem continues to be the relative dispersion of the severities. Even if the government's reaction is not qualitatively altered, there are quantitative

differences in the results arrived at under these different cost structures.

The first two alternatives considered (marginal costs increasing with the severity or in the number of operations) alter the curvature of the Health Authority's objective function. Even if the curvature cannot be fully characterized now, the effect of the change in the cost structure on the results is clear. When the marginal costs increase with the severity of the patients, the presence of cream-skimming fosters the non-concavity of the function, as it increases the average severity in the public sector. Consequently, the boundary solution (not undertaking the policy) is more likely to prevail. On the contrary, when the costs are convex in the number of operations, it results in an impulse of the concavity of the program and, hence, in a reduction of the region where the boundary solution is optimal.

To analyze the situation where public costs are affected by capacity constraints, we consider the simplest structure, i.e., costs that are linear in the number of patients, but with the cost parameter increasing when the number of operations exceed a certain threshold. Formally, there exists a level of operations $\bar{x} < N$ such that if $x < \bar{x}$, the costs are $k_1^{pb}x$, while, if $x > \bar{x}$, total costs are $k_1^{pb}\bar{x} + k_2^{pb}(x - \bar{x})$, with $k_2^{pb} > k_1^{pb}$. In this set-up, the analysis of the curvature of the objective function is not different from that of our original model. The only difference is that, under this new cost structure, the boundary solution becomes more costly, since it implies bearing the cost of a higher extra capacity. The condition for the Health Authority to be willing to carry out the policy is, therefore, less demanding here than it was in the original model.

3.5.2 Heterogeneous Physicians

Our analysis assumes that all physicians have the same level of ability and, therefore, this level is observable by the Health Authority (or, what is equivalent, to considering a single

physician).

There are two reasons for choosing such a set-up: First, we wanted to focus on the potential strategic behavior that a physician may have in his performance as a dual supplier. We, therefore, specifically ignored the possibility of the physician to strategically behave within each sector, in the sense of his exerting a low level of effort or shirking off. Dealing with homogeneous physicians allows us to avoid the additional problems derived from the physician's having a double strategic behavior.

Secondly, we wanted to study physicians' incentives to select patients under a fixed contractual structure that is close to the one we observe in many real mixed health care systems. Under these schemes, the earnings of the physicians in either sector are the same, independently of their type.

There is a general consensus in the literature, however, concerning the problems that the different degrees of information of the agents involved in the provision of health care generate (for an overview of principal-agent theory, see Laffont and Tirole (1993)). To be more specific, the providers of medical services generally have more information than the government has, concerning their own skills, i.e., their capacity to treat the patients at a given cost of effort. The important question here, therefore, is whether our results are robust to the existence of heterogeneous physicians. This would introduce a problem of adverse selection in the model, in addition to the one of patient-selection.

Let us now consider two types of physicians: a "high skilled" physician ($\theta = \theta_h$) and a "low skilled" one ($\theta = \theta_l$). Earning a fixed salary is not incentive compatible for the "high skilled" physician, since he can pretend to be "low skilled" and reduce the dis-utility of his effort.

To solve this problem and concentrate on patient-selection, a truthful-revelation

contract should be designed. A contract that ensures the honest revelation by the doctor should include two components: a fixed salary and a bonus. The Health Authority would offer a contract $\{(T_h, B_h), (T_l, B_l)\}$ to the physician, under which he receives a salary T_h, plus a bonus B_h, if he announces θ_h as his level of ability, and analogously for (T_l, B_l). The bonus is included to induce a high skilled physician to sign his contract. In equilibrium, then, it is strictly positive for the high type (informational rent) and zero for the low type.

Under this new sort of contract, however, (and maintaining the same reimbursement scheme in the private sector), the incentives for the physician to select patients remain unaltered. Only the high skilled physician receives a bonus, but this does not depend on his decision regarding the selection of patients. The magnitude of his informational rents is only affected by the decision taken by the low skilled physician. The problem of cream-skimming is, therefore, the same under adverse selection.

The reaction of the Health Authority, however, differs. When the government does not observe the physician's ability, a new effect appears: its incentives to undertake the policy of distributing patients between the two sectors increase, even in the presence of cream-skimming. This effect is due to the fact that the physician's informational advantage allows him to earn informational rents, which can be reduced by transferring more patients to the private sector. Even if sending less patients to private practice alleviates the problem of cream-skimming, it raises the value of the bonus received by the high skilled physician and is, therefore, more costly. As a consequence of this, there is a range of parameter values (in particular private fees) for which the policy of distributing patients between the private and the public sectors is only optimal when the Health Authority suffers from asymmetric information about the physician's ability. The Health Authority is more willing to undertake the policy, as it is a way of relaxing the agency problem suffered and "transferring" the

informational disadvantage to the private sector.

3.5.3 Concluding Remarks and Policy Implications

The motivation behind this study was the enormous congestion that exists in public health services worldwide, which has forced several Health Authorities to devise especial programs to alleviate it. Temporary programs, however, may be extremely costly and, hence, the study of the regularity and adequacy of the measures undertaken becomes a major concern.

We have analyzed a situation in which the Health Authority undertakes a policy to reduce the proportion of patients that remain untreated, transferring a fraction of such cases to private hospitals. We have shown that a problem of cream-skimming arises. Due to the different structure of the physician's remuneration in either system, specialists prefer to treat only the mildest cases in their own private practices. We have, then, shown how this problem makes the Health Authority be more reluctant to implement this policy.

We have also characterized the range of parameters that makes the Health Authority find it more profitable to treat all of its patients in the public sector. Moreover, we found that the crucial variable that measures the importance of the problem the Health Authority faces is the relative dispersion of the patients' severities.

When the policy is undertaken, we study the effects of patient-selection on the amount of patients that are finally transferred to the private sector. We also find that the results are ambiguous. When the relative dispersion of the severities is high enough, the negative effect of treating patients with higher average severities in the public sector always dominates. More patients are, therefore, sent to private hospitals. When we deal with medical disciplines in which the dispersion of the severities is low, the result is determined by the value of the private fee. In this case, the behavior of the Health Authority is always more extreme

than in the absence of cream-skimming.

Our analysis provides some policy recommendations concerning the optimality of this kind of measures. When designing a policy to transfer patients from the public to the private sector, the policy-maker should consider the fact that the difference that exists between the reimbursement systems of the two sectors can create perverse incentives for the physicians. To be more specific, if the reimbursement rule in the public sector is a fixed salary, while in the private practice functions on a fee-for-service basis, as is the case in several countries with mixed health-care systems, a problem of cream-skimming arises. The physicians transfer the mildest cases to the private practice, thus increasing the cost per operation borne by the public sector.

Our results suggest that the decision to undertake the policy or not is also influenced by another important issue, apart from the fee per operation agreed with the private sector. The type of illness and, in particular, how disperse the severities of the patients are is shown to be very important. The wider the range of severities, the more serious the problem of patient-selection becomes and, therefore, the less likely it is that the policy will be undertaken.

Moreover, empirical evidence shows that when the health authority decides to send some patients on the public waiting-list to private hospitals, an undesired effect appears. The number of operations that are performed in the public sector, for instance in a month, slightly decreases in comparison with the ones that were performed before the implementation of the policy. Our model provides an explanation for these empirical findings. Since the patients who remain in the public sector are the most costly ones, for a given level of physician's effort, the amount of patients that can be treated is lower.

Our approach to the policy and its performance has been done mainly from a positive perspective. Given the institutional framework and the existing contractual arrangements in

some real health systems, we have identified a potential problem of patient selection. What is left for further research is a more normative analysis in which possible solutions to the problem can be analyzed. In spite of this, our article provides some hints on potential alternative ways to address the issue. One may think of two possible types of instruments. The first one concerns physicians' reimbursement. Different contractual arrangements would alter physicians' incentives and, hence, they deserve a careful analysis. A second alternative is to consider measures that limit the physician's strategic capacity. Among the possible alternatives we could mention the introduction of either exogenous rules for allocating patients between the two sectors, or legal exclusions that prevent either public hospitals or physicians from providing treatment (outside their regular working hours) to patients that were on their waiting lists.

There are some issues that have not been addressed in this work. First, all of the analysis we have performed is done under the assumption that there is no difference between the quality of the services provided by the public and private sectors. This assumption is reasonable for the kind of treatments we consider (non-urgent elective surgery), but would be clearly inappropriate for modelling other kinds of illnesses that require very expensive or high-tech treatment. To include these medical disciplines in the analysis would require a model with heterogenous quality among the different providers.

Second, one of the characteristics of the model is the fact that the patients are passive through all the process. Given the structure of our model, allowing for active patients would not have any impact on the results. Introducing differences in quality between the two providers and considering unequal waiting times to access treatment in each sector, would certainly give an important role to the patient in the outcome of the policy.

Finally, this analysis would require a first step to study the negotiations between

the Health Authority and the private hospitals concerning the payment for those operations transferred. Endogeneizing this process would be an interesting exercise that would provide the Health Authority with a new regulatory tool to face the cream-skimming problem. Introducing issues such as competition among private providers, however, is outside the scope of this work.

3.6 Appendix

Proof of Proposition 1:

At stage 2, the reduced form of the optimization program the physician faces is as follows:

$$\begin{aligned} \max_{\widehat{s}^{pb}} U^s &= T + w(N-x) - \Psi(x\widehat{s}^{pb},\theta) - \Psi((N-x)\widehat{s}^{pv},\theta) \\ s.t \quad x &= x\left(T,\widehat{s}^{pb}\right). \end{aligned}$$

The derivative is given by:

$$\frac{\partial U^s}{\partial \widehat{s}^{pb}} = -w\frac{dx}{d\widehat{s}^{pb}} - \frac{d\Psi\left(x\widehat{s}^{pb},\theta\right)}{d\widehat{s}^{pb}} - \frac{d\Psi\left((N-x)\widehat{s}^{pv},\theta\right)}{d\widehat{s}^{pb}}. \tag{3.5}$$

By stage 3 we know that $\forall \widehat{s}^{pb}$, $x = x\left(T,\widehat{s}^{pb}\right)$ is such that $\Psi\left(x\widehat{s}^{pb},\theta\right) = T$, which implies that $\Psi\left(x\left(T,\widehat{s}^{pb}\right)\widehat{s}^{pb},\theta\right)$ is constant with respect to $\widehat{s}^{pb}$.

Since $s \sim U(\underline{s},\overline{s})$, $\widehat{s}^{pb}$ and $\widehat{s}^{pv}$ are related by the following equation:

$$\widehat{s}^{pb}\frac{x}{N} + \widehat{s}^{pv}\frac{(N-x)}{N} = \widehat{s}.$$

From the above expression we get:

$$\widehat{s}^{pv}(N-x) = \widehat{s}N - \widehat{s}^{pb}x.$$

Since $\widehat{s}^{pb}x$ is constant with changes in $\widehat{s}^{pb}$, $\widehat{s}^{pv}(N-x)$ is also constant in this variable. Hence:

$$\frac{d\Psi\left(\left(N-x\left(T,\widehat{s}^{pb}\right)\right)\widehat{s}^{pv},\theta\right)}{d\widehat{s}^{pb}} = 0.$$

Therefore, condition (3.5) is reduced to:

$$\frac{\partial U^s}{\partial \widehat{s}^{pb}} = -w\frac{dx}{d\widehat{s}^{pb}},$$

and from here it is straightforward that:

$$\frac{\partial U^s}{\partial \widehat{s}^{pb}} > 0, \ \forall \widehat{s}^{pb}.$$

The physician would like to treat the patients with the highest average severity in the public sector.

We can write $\widehat{s}^{pb}$, using the formula for the conditional expectation. We let C be the subset of severities treated in the public sector (for a given number of operations x). Formally:

$$\widehat{s}^{pb} = E(s|C) = \frac{1}{\Pr(C)} \int_C s \frac{1}{\Delta s} ds.$$

For a given value of $\Pr(C) = \frac{x}{N}$ this expectation is increasing in the location of C in $(\underline{s}, \overline{s})$. Thus, $\widehat{s}^{pb}$ is maximum when C is maximized in the interval $(\underline{s}, \overline{s})$, from which we get that $C = \left(\overline{s} - \frac{\Delta s x}{N}, \overline{s}\right)$. Therefore, the physician chooses to treat the patients with severities in the range $\left(\underline{s}, \overline{s} - \frac{\Delta s x}{N}\right)$ in the private sector and leaves those in the interval $\left(\overline{s} - \frac{\Delta s x}{N}, \overline{s}\right)$ in the public sector. This implies that, in equilibrium, $\widehat{s}^{pb}$ and x are such that:

$$\widehat{s}^{pb} = \overline{s} - \frac{\Delta s x}{2N}$$
$$x = x\left(T, \widehat{s}^{pb}\right).$$

This completes the proof. ■

Proof of Lemma 3:

We study the curvature of the Health Authority's objective function in the restricted domain given by the behavior of the physician at stage 2. From this stage we know that x is such that $\frac{1}{2}\left[\frac{x}{\theta}\widehat{s}^{pb}\right]^2 = T$ and that $\widehat{s}^{pb} = \overline{s} - \frac{\Delta s x}{2N}$. By substituting these two constraints in the

program, we will characterize the curvature of the objective function as a function of x. The optimization program at the first stage is as follows:

$$\max_x H^m = \left(Q\widehat{s} - k^{pv}\right) N - \left[\frac{1}{2}\left[\frac{x}{\theta}\left(\bar{s} - \frac{\Delta s x}{2N}\right)\right]^2 + \Delta k x + w\left(N - x\right)\right].$$

The f.o.c is given by:

$$\left(\frac{-x}{\theta^2}\left[\left(\bar{s} - \frac{\Delta s x}{2N}\right)\right]^2 + \frac{\Delta s x^2}{2N\theta^2}\left[\left(\bar{s} - \frac{\Delta s x}{2N}\right)\right] + w - \Delta k\right) = 0.$$

Or:

$$\frac{x^2\bar{s}}{2N\theta^2}\widehat{s}^{pb} + w - \Delta k = \frac{x}{\theta^2}\left(\widehat{s}^{pb}\right)^2.$$

Computing the second order condition and rearranging terms yields:

$$\frac{\partial^2 H^m}{\partial^2 x} = \frac{1}{\theta^2}\left(-\bar{s}^2 + 3\Delta s\beta\bar{s} - \frac{3}{2}\Delta s^2\beta^2\right), \text{ with } \beta = \frac{x}{N}.$$

This derivative is increasing in β. We compute its roots and we find:

$\beta_1 = \frac{\bar{s}}{\Delta s}\frac{3+\sqrt{3}}{3} > 1$ and $\beta_2 = \frac{\bar{s}}{\Delta s}\frac{3-\sqrt{3}}{3}$. Simple manipulations allow us to re-write β_2 as $\beta_2 = \left(\frac{1}{d} + \frac{1}{2}\right)\frac{3-\sqrt{3}}{3} = \bar{\beta}\left(d\right)$. It is straightforward to see that $\bar{\beta}'\left(d\right) < 0$.

From here, and taking into account that $\beta \leq 1$ since $x \leq N$, it can be shown that:

i) If $\bar{\beta}\left(d\right) \geq 1$, $\frac{\partial^2 H^m}{\partial^2 x} < 0\ \forall x \in [0, N]$, i.e., the objective function is always concave.

ii) If $\bar{\beta}\left(d\right) < 1$, $\frac{\partial^2 H^m}{\partial^2 x} < 0$ for every $\beta \leq \bar{\beta}\left(d\right)$. Thus, we can ensure that for $x \in \left[0,\ \bar{\beta}\left(d\right)N\right]$ the objective function is concave, and for $x \in \left(\bar{\beta}\left(d\right)N,\ N\right]$ it is convex.

In re-writing the conditions for $\bar{\beta}\left(d\right)$ in terms of the relative dispersion of the severities, we find that:

$$\bar{\beta}\left(d\right) \geq 1 \Leftrightarrow d \leq 4 - 2\sqrt{3} \text{ and}$$

$$\bar{\beta}\left(d\right) < 1 \Leftrightarrow d > 4 - 2\sqrt{3}.$$

This completes the proof. ■

Proof of Proposition 2:

From the f.o.c computed in Lemma 3 it is straightforward to verify that $x = 0$ can be never a solution, since:

$$\frac{\partial H^m}{\partial x}_{|x=0} = w - \Delta k > 0.$$

This, together with the other conditions found in Lemma 3, provides a complete characterization of the optimization problem:

1.-If $d \leq 4 - 2\sqrt{3}$, there exists a unique candidate to optimum (x). This solution will be interior, i.e., $x \in (0, N)$, if and only if:

$$\frac{\partial H^m}{\partial x}_{|x=N} = w - \Delta k - \frac{N}{\theta^2}\widehat{s}\left(\widehat{s} - \frac{\Delta s}{2}\right) < 0 \Leftrightarrow w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2}\left(1 - \frac{d}{2}\right).$$

2.-If $d > 4 - 2\sqrt{3}$, there exists, at most, a single interior candidate to optimum (x), such that $x \in \left(0, \bar{\beta}(d) N\right]$, with $\bar{\beta}(d) = \left(\frac{1}{d} + \frac{1}{2}\right)\frac{3-\sqrt{3}}{3}$. The boundary solution $x = N$ is also a potential candidate to optimum, provided that:

$$\frac{\partial H^m}{\partial x}_{|x=N} > 0 \Leftrightarrow w - \Delta k > \frac{\widehat{s}^2 N}{\theta^2}\left(1 - \frac{d}{2}\right)$$

This is a necessary, although not sufficient condition, for $x = N$ to be a solution. Hence, to choose the optimal level of x in this region we need to compare the value function for both candidates. We check the conditions under which it is better to operate on a fraction α of the patients, rather than perform N operations. We find that:

$$\begin{aligned} \text{If } w - \Delta k &< \frac{\widehat{s}^2 N}{\theta^2} g(d), \text{ with } g(d) = \max_{\alpha \in (0,1)} \frac{1}{2}\left[(1+\alpha) - \alpha^2 d\left(\frac{(1-\alpha)}{4}d + 1\right)\right] \\ \exists \alpha^* &\in (0,1) \text{ such that } H^m(x = \alpha^* N) > H^m(x = N). \end{aligned}$$

We can see that $g(d)$is such that $g'(d) < 0$, $\lim_{d\to\infty} g(d) = \frac{1}{2}$.

If $w-\Delta k$ exceeds the above threshold, then we can ensure that the boundary solution is optimal, since for this parameter configuration $\frac{\partial H^m}{\partial x}_{|x=N} > 0$. Therefore, the solution to the principal's problem is:

i) $x^m < N$ if $w - \Delta k < \frac{\widehat{s}^2 N}{\theta_i^2} G(d)$,

where $G(d) = \begin{cases} 1 - \frac{d}{2} \text{ if } d \leq 4 - 2\sqrt{3} \\ g(d) \text{ if } d \geq 4 - 2\sqrt{3} \end{cases}$ is a continuous function and $g(d)$ is such that $g'(d) < 0$ and $\lim_{d\to\infty} g(d) = \frac{1}{2}$.

ii) $x^m = N$, otherwise.

And this completes the proof. ■

Proof of Proposition 3:

From the previous Proposition we know that x^m is interior if:

$$w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2} G(d).$$

The f.o.c of the Health Authority's problem in the presence of patient-selection was given by:

$$\frac{\partial H^m}{\partial x} = \left(\frac{-x}{\theta^2} \left(\overline{s} - \frac{\Delta s x}{2N} \right)^2 + \frac{\Delta s x^2}{2N\theta^2} \left(\overline{s} - \frac{\Delta s x}{2N} \right) + w - \Delta k \right) = 0$$

Considering that $\overline{s} = \widehat{s} + \frac{1}{2}\Delta s$, we can rewrite it in terms of $\widehat{s}$ as:

$$\left(\frac{-x}{\theta^2} \left(\widehat{s} + \left(1 - \frac{x}{N}\right) \frac{\Delta s}{2} \right)^2 + \frac{\Delta s x^2}{2N\theta^2} \left(\widehat{s} + \left(1 - \frac{x}{N}\right) \frac{\Delta s}{2} \right) + w - \Delta k \right) = 0.$$

The f.o.c of the Health Authority's problem under non-manipulability was given by:

$$\frac{\partial H}{\partial x} = \left(\frac{-x\widehat{s}^2}{\theta^2} + w - \Delta k \right) = 0$$

Performing some algebraic manipulations, we find that:

$$\frac{\partial H^m}{\partial x} > \frac{\partial H}{\partial x} \Leftrightarrow \frac{x}{\theta^2} \Delta s \widehat{s} \left[-\left(\frac{x}{N}\right)^2 \frac{d}{2} + \frac{3}{2}\frac{x}{N}\left(1 + \frac{d}{2}\right) - \left(1 + \frac{d}{4}\right) \right] > 0$$

From the inequality above we find that:

$$\begin{aligned} \frac{\partial H^m}{\partial x}_{|x=x^*} &> 0 \Longleftrightarrow \\ w - \Delta k &> \frac{\widehat{s}^2 N}{\theta^2}\left[\frac{1}{4d}\left(3d+6-\sqrt{d^2+4d+36}\right)\right]. \end{aligned}$$

In order to complete the characterization of the solution x^m, we need to check when the inequality above is compatible with the condition guaranteeing that x^m is interior. On combining the above condition with that of Proposition 2, we find that:

1) When $d < .7625$:

If $w - \Delta k < \frac{\widehat{s}^2 N}{\theta^2}\Phi(d)$, then:

$$x^m < x^*.$$

If $w - \Delta k > \frac{\widehat{s}^2 N}{\theta^2}\Phi(d)$, then:

$$x^m > x^*.$$

With $\Phi(d) = \left[\frac{1}{4d}\left(3d+6-\sqrt{d^2+4d+36}\right)\right]$.

2) When $d \geq .7625$:

$$x^m < x^*.$$

And this completes the proof. ■

Bibliography

[1] Arrow, K.J. (1963) "Uncertainty and the Welfare Economics of Medical Care". *American Economic Review* 53, 941-973.

[2] Barros, P.P. and Olivella, P. (1999)."Waiting lists and Patient Selection". D.P. UAB-IAE 499-99.

[3] Blomqvist, Å. (1991) "The Doctor as a Double Agent: Information Asymmetry, Health Insurance, and Medical Care". *Journal of Health Economics* 10, 411-432.

[4] Bosanquet, N. (1999) "A Successful National Health Service". Adam Smith Institute. London.

[5] Competition Commission (1994) "Private Medical Services: a Report on the Agreements and Practices Relating to Charges for the Supply of Private Medical Services by NHS consultants". London: HMSO

[6] Dranove, D. (1987) "Rate-setting by Diagnosis Related Groups and Hospital Specialization". *RAND Journal of Economics* 18-3, 417-427.

[7] Ellis, R.P. (1998) " Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins". *Journal of Health Economics* 17, 537-555.

[8] Ellis, R.P. and McGuire, T.G. (1986) "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply". *Journal of Health Economics* 5, 129-152.

[9] Ellis, R.P. and McGuire, T.G. (1990) "Optimal Payment Systems for Health Services". *Journal of Health Economics* 9, 375-396.

[10] Gaynor, M. (1994) "Issues in the Industrial Organization of the Market for Physician Services". *Journal of Economics and Management Strategy* 3-1, 211-255.

[11] Hoel, M. and Sæther, E.M. (2000) "Private Health Care as a Supplement to a Public Health System with Waiting Time for Treatment". Memorandum from the Department of Economics, no 41/2000. University of Oslo.

[12] 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.

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

[14] Ma, C.-t.A. (1994) "Health Care Payment Systems: Costs and Quality Incentives". *Journal of Economics and Management Strategy* 3-1, 93-112.

[15] Olivella, P. (2002) "Shifting Public-Health-Sector Waiting Lists to the Private Sector". Forthcoming: *European Journal of Political Economy.*

[16] Propper, C. (1995) "Agency and Incentives in the NHS Internal Market". *Social Science and Medicine* 40, 1683-1690.

[17] Rickman, N. and McGuire, A. (1999) "Regulating Providers' Reimbursement in a Mixed Market for Health Care". *Scottish Journal of Political Economy* 46-1, 53-71.

Volver al índice/Tornar a l'índex

[18] Selden, T.P. (1990) "A Model of Capitation". *Journal of Health Economics* 9, 397-410.

[19] Yates, J. (1995) "Private Eye, Heart and Hip: Surgical Consultants, the National Health Service and Private Medicine". London: Churchill Livingstone.

Chapter 4

Should Physicians' Dual Practice Be Limited? An Incentive Approach

4.1 Introduction

It is quite common in countries where there are both public and private health care systems that many doctors work in both sectors at the same time. There are very few studies, however, that analyze the complex relationships that exist between the two sectors and, therefore, the conflicting interests that arise from the doctors' dual activity.

In this article we examine the specific implications that such dual activity has for public health authorities. Our main objective is to analyze the circumstances under which the health authorities benefit from the doctors' dual practice and those under which they lose. Our analyses thus provides a theoretical bench-mark for the evaluation of the optimality

of some existing regulatory frameworks in the field. We will concentrate on two kinds of measures: First, the possibility of offering exclusive contracts (no private practice) to doctors who work in the public health sector. Under such contracts, the doctor agrees to forego his private practice in exchange for a mutually agreed amount of financial compensation. Second, the decision taken by some health authorities of limiting private earnings to public physicians.

The actual situation in some European mixed health care systems is as follows. In Spain, the Law of Professional Incompatibilities that governs the employment of civil servants (Law No. 53/1984), does not prohibit doctors from having private practices. The specific legislation for medical professionals, however, offers those who choose not to do so a fixed monthly bonus in addition to their basic salaries. In the UK, physicians who are employed in the public sector are allowed to operate in the private sector under their NHS contracts. NHS part-time consultants are not limited in their private practice, whereas for full-time consultants their private practice is limited to 10% of their NHS salary. Indeed, most private medical services are provided by physicians whose main commitment is to their NHS duties. A report by the Competition Commission (1994), estimated that 61% of NHS physicians in the UK have significant private work. In addition to this, and according to Yates (1995), a NHS specialist, on average, undertakes two operations a week in the private sector. In France, public hospitals employ both full-time and part-time physicians who can also accept private patients with the restriction that income from private fees is limited to no more than 30% of the physician's total income. Similar arrangements apply in the majority of the European countries which, although characterized as public health care systems, also allow private health care (Johnson, 1995; US Congress, 1995).

The conflicting interests that arise between the doctor's public and private practices can affect both services in many different dimensions. The standard approach in the principal-

agent literature considers that while an agent's external tasks (i.e., private) might provide him with some extra revenue, they also reduces his level of effort in his contracted tasks. In the particular case of doctors, however, we believe that there is another important dimension that has not been addressed by any of the researches of the topic. There are many well-respected doctors whose service in public hospitals have won them the reputation of being "good doctors" and such prestige obviously has a positive influence on their private practice.[1] Whilst they receive a fixed salary in the public service, regardless of the volume of patients they treat, in their private practice their income depends directly on the demand they receive from the patient population. A physician that is perceived by the population as a "good doctor", therefore, will obviously increase his private revenue.

We formalize these ideas in a model with one patient, one specialist, and the health authority. The patient suffers from an illness whose severity is unclear and requires medical attention. In attending to the patient, the doctor is charged with two basic tasks: (a) to diagnose the ailment and its severity, and (b) to provide the required treatment. The doctor's performance is affected by his condition of dual supplier in the following way. If he can cure the patient in a single treatment, it improves his "professional prestige", which then reflects positively on his private practice and thus increases his private income.

The task of the health authority is to draw up the contract that minimizes the social costs. In this regard, there is a general consensus in the literature on the conviction that the doctor usually has access to privileged information compared with the insurers (see Gaynor, 1994). We therefore suppose that neither his diagnostic process nor the treatment he prescribes could be either verifiable or contractible. This general set-up with double moral hazard provides us an appropriate framework to study the role of the physician as

[1] One may think that this prestige effect will be higher in those countries where the physician's requirements to get a job in the public sector are more demanding.

dual supplier. Studying such a role disregarding any of these agency problems would lead to biased results due to the loss of important effects. The incentives included in the doctor's contract will be therefore key factors in determining the sort of behavior he has with regard to his work.

It is within this framework of the doctor's role as a dual supplier that we study the implications that such a situation has for the health authority. It allows us to verify whether it is to the health authority's benefit to either avoid (by offering an exclusive contract to the physician) or limit the doctor's private practice.

On characterizing the doctor's behavior we find that his double role provides certain important strategic effects. On the one hand, his keen interest in curing the patient (and gaining prestige) leads him to over-provide services. In other words, he tends to prescribe stronger (and more expensive) treatments systematically, as a way of ensuring a higher level of successful treatments. He is therefore more likely to divert from the treatment recommended for the diagnosis that was done. This perverse incentive, however, has its positive aspect. If the health authority manages (through the contract) to force the doctor to apply the treatment indicated by the diagnosis, this interest shown in curing the patient generates a beneficial effect. In such a case, the doctor will have an incentive to carry out a very precise diagnosis, thus avoiding, or at least minimizing, the chances of his providing a wrong treatment. In such a case, the doctor's dual practice can, indeed, be positive for the health authority.

We prove that the overall impact of the doctor's dual practice on the health authority's costs, depends basically on the treatment strategy that the health authority decides to follow. If it chooses a plan in which cost-savings prevail over the health losses of the patients, the doctor's dual activity will be negative, causing an increase in the social costs.

On the other hand, if the health authority decides to follow the strategy of sticking to the treatment that the diagnosis suggests, and if the cost associated with a wrong diagnosis is high enough, the effect of the doctor's private practice is positive for the health authority. In such a case, the extra payment that the health authority has to make to induce the doctor to follow the strategy is compensated by the doctor's interest in sticking to a very precise treatment.

We apply the analysis to the study of some actual regulatory policies regarding physicians' dual practice. First, we show that even in cases where the doctor's dual practice is socially detrimental, it is still not optimal for the government to offer him an exclusive contract. The reason for this is that, in designing the incentive contract, the health authority partially mitigates the increase in costs. The saving made from the doctor's exclusive contract is never enough to compensate what has to be paid to him for his losses from not having a private practice.

This result varies if we consider payment structures that are more similar to the ones currently being used by several health authorities to contract specialists. Their systems of remuneration are not usually based on incentives, but rather, on the payment of a flat salary. When we repeat our analysis considering this form of payment, we observe situations in which it is better for the health authority to offer an exclusive contract to ensure that the doctor does not have a private practice. Such results may well explain the very existence of exclusive contracts, which must be considered a "second best" choice.

Secondly, for the case of a regulation based on limiting the physician's private earnings, we find that this measure can be useful to mitigate the physician's tendency to overprovide services. However, we also show that in those cases in which the health authority is highly concerned about the accuracy of the diagnosis the physician performs, this kind of

regulation can be socially detrimental.

The literature has analyzed the physician's response to the form of contract quite extensively. Ellis and McGuire (1986, 1990), Selden (1990), Blomqvist (1991), and Rickman and McGuire (1999), like ourselves, have studied how the different refund rules can lead the doctor to either over-provide or under-provide services. In their analyses, however, none of them has considered the possibility that the doctor might be supplying to two different sectors simultaneously.

The possibility of over-treating a patient, in the sense of providing him with an excessively aggressive treatment appears in Gal-Or (1999), although in a framework completely different from ours. In her model, the physicians have incentives to practice "defensive medicine" in order to avoid paying malpractice costs, as a more aggressive treatment increases the probability of healing the patient.

The sequential modelling of the doctor's decisions (since diagnosis precedes treatment), follows Jelovac (2001) and García-Mariñoso and Jelovac (2002). In these studies optimal contracts are designed respectively for a specialist and a GP in a context with double moral hazard. In these papers, however, the doctor's behavior is not affected by his external activities, as occurs in ours.

An interest has recently arisen, in the literature, in analyzing the implications of the doctor's dual activity. Rickman and McGuire (1999) study the system of optimal refund in the public sector for a doctor who also works in the private sector. Contrary to our analyses, theirs is not carried out in a principal-agent framework and furthermore, they concentrate on the implications of the fact that the doctor can offer both public and private services to the same patient. González (2002) studies the consequences of carrying out a policy of transferring patients from the public to the private sector. She shows that if the doctors are

dual providers, only the less serious patients will be transferred, generating an increase in the cost of implementing the policy.

This article is also related to the analysis of Holmstrom and Milgrom (1991), on the optimality of limiting access to external activities for the employees of a company. Although our model is quite different from theirs, it shares a common characteristic. In both, what determines whether these "outside activities" should be allowed, are not their intrinsic characteristics, but rather their impact on the agent's incentives in their "inside activities".

The rest of the paper is organized as follows: In the following section, we present the model. Section 3 analyzes the physician's behavior. In Section 4 we derive the health authority's optimal contract. Section 5 studies some alternative regulatory frameworks. We analyze the consequences of offering exclusive contracts to physicians and of limiting their private earnings. Section 6 is devoted to provide a rationality for the way in which we have modelled physicians' reputation acquisition throughout the article. Finally, in Section 7, we present our conclusions.

4.2 The Model

There are three agents in this economy: a patient, a physician and the regulator or health authority. They are all risk neutral.

The patient suffers from a specific type of illness. The severity of the illness is measured by a random variable s. We assume that s can only take two values: $\underline{s}$ and $\bar{s}$, which indicate whether the patient suffers from either a low or a high severity of the illness. For the sake of simplicity, we assume that both types of illnesses are equally likely. The patient is perfectly aware that he is ill but does not know just how serious his illness is. He

therefore seeks health care from the medical provider. The patient's utility is defined in terms of his expected health loss.

In our model, the patient can be given two different types of treatment: a "mild treatment" $(\underline{T})$, which can only cure a patient suffering from a low severity, or a "strong treatment" $(\overline{T})$, which can cure both levels of severity. The strong treatment, however, causes a health loss L in the patient when he is suffering from a low level severity. This health loss can be interpreted as the result of an over-treatment.

The physician is supposed to provide the appropriate treatment to cure the patient. To do so, he exerts a level of effort $e \in [0,1]$, in performing the diagnosis, which yields a signal (s^e) about the severity of the patient's condition. The accuracy of this signal depends positively on the level of effort exerted by the doctor in carrying out the diagnosis. We consider a probability of receiving a correct signal defined by the following function:

$$\Pr\left(s^e = \overline{s}|\overline{s}\right) = \Pr\left(s^e = \underline{s}|\underline{s}\right) = \frac{1+e}{2}.$$

In exerting the lowest level of effort $(e = 0)$, the physician does not obtain any further information about the severity of the patient's condition except the common knowledge that either severity is possible with a probability of one-half. By exerting a greater effort, the physician would get a more accurate diagnosis of the severity of the patient. When $e = 1$, the signal is perfectly accurate.

In performing the diagnosis, the physician incurs in some disutility due to the effort he exerts. We denote it by $V(e) = k\frac{e^2}{2}$, where $k > 0$ is a measure of how the physician's marginal cost increases with his level of effort. This effort in diagnosing is not contractible and, as such, he will not be directly reimbursed for it.

Using the information he receives from the diagnosis, the physician decides on which

treatment he will provide to the patient: either the mild treatment or the strong one. If the patient recovers his health with the first treatment prescribed the game ends. Otherwise, the patient receives a new round of treatment. Since we have restricted the analysis to a two-severity illness, the patient is eventually cured after the second treatment. We measure the health loss borne by the patient from receiving a delayed cure for his illness by l.

To summarize, the patient only suffers from a health loss if the appropriate treatment is not provided to him: he either loses L if he is over-treated, or l if he is treated twice. We consider that both L and l are not verifiable for the health authority.

As already mentioned in the Introduction, we are interested in analyzing the implications for the public authority of allowing the physician to offer his services as a private provider. We model this by introducing a parameter $\mu > 0$ that measures the impact of the physician's activity in the public sector on his private revenue. We assume that the prestige of the physician increases (and hence he receives extra earnings of value μ) when he cures the patient with just one treatment in his activity in the public sector. Therefore, μ does not reflect the physician's real private income, but rather, it is a proxy for how such revenue is affected by his activities within the public sector.[2] Although the physician's reputation effect is derived in a very simple form, in Section 6 we will show that our formulation can be interpreted as a reduced form of a more general game in which reputation arises endogenously.

The third agent involved in the model is the health authority. It pays the cost of the treatment provided to the patient, and the payments made to the physician. Concerning the costs of treatment, $\overline{c}$ denotes the cost of the strong treatment, while $\underline{c}$ is the cost of the mild one. We assume that $\overline{c} > \underline{c}$.

The health authority designs the physician's payment contract, which consists of

[2]An alternative construction would be to consider losses in reputation when two treatments are required for the patient to recover from his illness.

three non-negative components: $(\overline{w}, \underline{w}, B)$. $\overline{w}$ is the amount of money that the physician receives if the strong treatment is provided and $\underline{w}$ if he recommends the mild treatment. Furthermore, in the latter case, we consider that the physician receives a bonus B if the patient is cured with just one round of treatment. This bonus can be interpreted as a premium for cost-containment, since the mild treatment is cheaper than the strong one. Moreover, as will be shown later on, B will be an important instrument for offering incentives to the physician.

The health authority designs the contract in such a way that the expected social costs are minimized. Such costs, denoted by C, are the sum of the financial costs (i.e., expected treatment costs and payoffs to the physician) and the patient's disutility (which is measured by his expected health loss).

The timing of the game consists of the following stages. First, the health authority fixes the physician's payment contract, which he can either accept or reject (in which case the game ends). Secondly, the severity of the patient is realized. He seeks health care from the physician, who exerts some level of effort while doing the diagnosis. This provides him with a signal about the patient's severity. Third, after observing the signal, the physician decides on a treatment. If the patient does not recover after the first treatment, the physician provides a new treatment. Once the patient has recovered his health, the game ends.

4.3 The Physician's Behavior

4.3.1 The Physician's Treatment Choice

In our model, the doctor faces a population of patients that can suffer from either a high severity or a low severity illness, with the same probability. Once the doctor has carried

out his effort in the diagnosis, the information that he obtains will allow him to improve his prognosis on the patient's true situation.

The probability of having correctly diagnosed that the severity of the illness is high is as follows:[3]

$$\Pr\left(\overline{s}|s^e=\overline{s}\right)=\frac{\Pr\left(\overline{s}\right)\Pr\left(s^e=\overline{s}|\overline{s}\right)}{\Pr\left(\overline{s}\right)\Pr\left(s^e=\overline{s}|\overline{s}\right)+\Pr\left(\underline{s}\right)\Pr\left(s^e=\overline{s}|\underline{s}\right)}=\frac{1+e}{2}.$$

Similarly, the probability of having correctly diagnosed that the severity of the illness is low is given by:

$$\Pr\left(\underline{s}|s^e=\underline{s}\right)=\frac{1+e}{2}.$$

Hence:

$$\Pr\left(\underline{s}|s^e=\overline{s}\right)=\Pr\left(\overline{s}|s^e=\underline{s}\right)=\frac{1-e}{2}.$$

Once the physician has diagnosed the true severity of the condition, he then decides on the treatment that the patient should receive. Regardless of the severity of the illness, however, the doctor always has the choice between the two alternative treatments: the strong treatment or the mild one.

When $s^e=\underline{s}$, if the doctor prescribes the mild treatment he receives a remuneration $\underline{w}$. Moreover, if the patient's condition is really mild (with a probability of $\Pr\left(\underline{s}|s^e=\underline{s}\right)$), the physician also obtains B and, since the patient is cured with just one treatment, this has a positive impact on the doctor's private practice benefits, measured by μ. If, on the other hand, the patient's condition is severe (with $\Pr\left(\overline{s}|s^e=\underline{s}\right)$) the physician obtains only $\underline{w}$. Finally, if he decides to prescribe the strong treatment, he earns $\overline{w}$ and, furthermore, as the patient is cured, independently of the real severity of his condition, the doctor also receives μ.

[3]Note that since $\Pr(\overline{s})=\Pr(\underline{s})$, then $\Pr\left(\overline{s}|s^e=\overline{s}\right)=\Pr\left(s^e=\overline{s}|\overline{s}\right)$.

Likewise, when $s^e = \overline{s}$, the doctor receives $\overline{w} + \mu$ if he prescribes treatment $\overline{T}$. If he prescribes treatment $\underline{T}$, his remuneration is $\underline{w}$ if the diagnosis is correct. If the diagnosis is wrong (which occurs with a probability of $\Pr(\underline{s}|s^e = \overline{s})$) the doctor earns $\underline{w} + B + \mu$.

In comparing the different payments that the doctor receives from prescribing either of the two treatments, we can conclude that:

If $s^e = \underline{s}$, the doctor will prescribe $\underline{T}$ whenever $e \geq \tilde{e} = \frac{2(\overline{w}-\underline{w})+\mu-B}{B+\mu}$ and $\overline{T}$ otherwise.

If $s^e = \overline{s}$, the doctor prescribes $\overline{T}$ whenever $e \geq \hat{e} = \frac{B-\mu-2(\overline{w}-\underline{w})}{B+\mu}$ and $\underline{T}$ otherwise.

Keeping in mind that the doctor's level of effort is bounded, since $e \in [0,1]$, we can summarize the doctor's decision on the type of treatment to be provided in the following lemma:

Lemma 1 *The treatment that the doctor prescribes is as follows:*

- *If* $2(\overline{w} - \underline{w}) + \mu - B \geq 0$*:*
  - *If* $B < \overline{w} - \underline{w}$*, the doctor always prescribes* $\overline{T}$*, regardless of the signal of the severity of the condition.*
  - *If* $B \geq \overline{w} - \underline{w}$*, when* $e \geq \tilde{e}$ *the doctor prescribes the most appropriate treatment for the severity diagnosed. For* $e \in [0, \tilde{e})$ *the doctor always prescribes* $\overline{T}$ *regardless of the signal.*
- *If* $B - \mu - 2(\overline{w} - \underline{w}) \geq 0$*:*
  - *If* $\mu < \underline{w} - \overline{w}$*, the doctor always prescribes* $\underline{T}$*, regardless of the signal.*
  - *If* $\mu \geq \underline{w} - \overline{w}$*, when* $e \geq \hat{e}$ *the doctor prescribes the most appropriate treatment for the severity diagnosed. For* $e \in [0, \hat{e})$ *the doctor always prescribes* $\underline{T}$*, regardless of the signal.*

If the remuneration plan contracted with the physician is of the first type, he simply decides on either the strong treatment or the most appropriate one (since $\hat{e} < 0$). Moreover, the bonus he receives would have to surpass a certain threshold for him to be encouraged to prescribe the most appropriate treatment. In contrast, when the doctor is employed under the other type of contract, it is generally the strong treatment that is dominated by one of the other two. With either type of contract, the effort that the doctor makes in the previous stage (i.e., the diagnosis) has to exceed a certain threshold, for him to decide to prescribe the most appropriate treatment for the conditions revealed in the diagnosis. When his level of effort is low, the doctor will always prescribe a treatment that is independent of the one the diagnosis recommends but which depends, rather, on the sort of payment structure agreed to in his contract.

The repercussions that the doctor's behavior within the public health service has on his private practice, reflected by μ, affects his decision on what treatment to prescribe.

Remark 1 *The fact that the doctor also works in the private sector means that, in his contracted public health duties, he will always be more inclined to prescribe the strong treatment and less inclined to prescribe the mild one, regardless of the results of the diagnosis.*

It is quite easy to see that $\tilde{e}$ is increasing in μ while $\hat{e}$ is decreasing. In other words, the fact that the doctor's behavior in the public service affects his private income makes him more likely to be in the region where he always decides to prescribe the strong treatment, and less likely to be in the region where he decides to prescribe the mild one. The doctor with a private practice is interested in curing the patient with just one treatment, so that his reputation as "a good doctor" will benefit his private practice. This encourages him to prescribe the strong treatment. This gives us the first insight into the strategic effects of the

physician's work in both sectors: it may encourage a tendency to over-provide services in the public health sector.

In our model, the doctor does not receive the signal of the patient's severity until Stage 3 of the game. Before exerting his effort in the diagnosis, therefore, the doctor anticipates the strategies that he will be able to follow once the signal has been received. In particular, the physician can follow any of the following strategies: (1) Prescribe the most appropriate treatment for the signal received; (2) Always prescribe the strong treatment; (3) Always prescribe the mild treatment.[4]

These three possible strategies not only determine the doctor's expected utility (U) (and the effort he will exert in the diagnosis), but also the expected social costs (C).

The structure of the doctor's expected utility and of the health authority's expected costs, under each of the possible strategies are detailed in the Appendix 1.

The expected costs for the health authority are: $\overline{C}$ if the doctor adopts the strong-treatment strategy, $\underline{C}$ if he chooses the mild-treatment strategy, and C_* if he employs the most-appropriate-treatment strategy.

Likewise, the doctor's utility is: $\overline{U}$ if he adopts the strong-treatment strategy, $\underline{U}$ if he chooses the mild-treatment strategy, and U_* if he employs the most-appropriate-treatment strategy.

4.3.2 The Physician's Diagnosis Decision

In this sub-section, we analyze the doctor's decision on the level of effort he exerts in the diagnosis, for each of the possible treatment strategies he adopts.

[4] In fact, there is a fourth strategy the physician can follow: to prescribe the most inappropriate treatment for the signal received. However, this strategy is always dominated by one of the others, for any level of effort exerted.

If he decides to adopt the most-appropriate-treatment strategy, he will choose the level of effort that maximizes U_*. Furthermore, this level of effort will necessarily overcome the thresholds defined in Lemma 1, for the strategy to be finally adopted.

We define:

$$\varepsilon \equiv \max\{\tilde{e}, \hat{e}\} = \frac{|2(\overline{w} - \underline{w}) + \mu - B|}{B + \mu}.$$

Only for those levels of effort in the diagnosis in which $e \geq \varepsilon$ it could be optimal for the doctor to adopt the strategy of providing the most appropriate treatment for the signal he receives.

The problem the doctor faces is as follows:

$$\begin{aligned} \max_{e} U_* &= \frac{1}{2}\left[\underline{w} + \overline{w} + \mu + \frac{1+e}{2}(B+\mu)\right] - k\frac{e^2}{2} \\ s.t \quad e &\in [\varepsilon, 1]. \end{aligned}$$

The solution to the problem is given by $e_* = \min\left\{\max\left\{\frac{B+\mu}{4k}, \varepsilon\right\}, 1\right\}$.

When the doctor decides to adopt either of the other two treatment strategies, i.e., when he always provides the same treatment, independently of the signal received, (either $\overline{T}$ or $\underline{T}$), the optimal level of effort in the diagnosis is the minimum. A positive level of effort does not have any effect on his payments (these are independent of the signal received and, therefore, of e) and only generates greater costs for him.

The following lemma summarizes all these possibilities.

Lemma 2 *The physician's optimal effort for each of the treatment strategies that he can adopt is as follows:*

- *If the doctor adopts the strong-treatment strategy, then* $\overline{e} = 0$ *and his expected utility is given by:* $\overline{U} = \overline{w} + \mu$.

- *If the doctor adopts the mild-treatment strategy, then* $\underline{e} = 0$ *and his expected utility is given by:* $\underline{U} = \underline{w} + \frac{1}{2}(\mu + B)$.

- *If the doctor adopts the most-appropriate-treatment strategy to the signal, then* $e_* = \min\left\{\max\left\{\frac{B+\mu}{4k}, \varepsilon\right\}, 1\right\}$ *and his expected utility is given by:*

$$U_* = \frac{1}{2}\left[\underline{w} + \overline{w} + \mu + \frac{1+e_*}{2}(B+\mu)\right] - k\frac{e_*^2}{2}.$$

Once the doctor's expected utility has been computed under each strategy, we obtain a series of restrictions that determine when the doctor decides to adopt each of the possible treatment strategies. These restrictions are the ones that the health authority will include later on in his optimization program as incentive constraints. The doctor's decision is summarized in the following lemma.

Lemma 3 *The doctor's behavior concerning the strategy he will adopt is as follows:*

- *To adopt the strong-treatment strategy is preferred by the doctor if:*

$$e_*\left(\frac{B+\mu}{2} - ke_*\right) \leq \overline{w} - \underline{w} + \frac{\mu - B}{2}. \qquad (\overline{IC})$$

- *To adopt the mild-treatment strategy is preferred by the doctor if:*

$$e_*\left(\frac{B+\mu}{2} - ke_*\right) \leq \underline{w} - \overline{w} + \frac{B - \mu}{2}. \qquad (\underline{IC})$$

- *To adopt the most-appropriate-treatment strategy to the signal is preferred by the doctor if:*

$$e_*\left(\frac{B+\mu}{2} - ke_*\right) \geq \left|\overline{w} - \underline{w} + \frac{\mu - B}{2}\right|. \qquad (IC_*)$$

With $e_* = \min\left\{\frac{B+\mu}{4k}, 1\right\}$.

The first result that it is extracted from this lemma is that the only relevant region of effort is the one with $e > \varepsilon$. When $e = \varepsilon$, the doctor is ex-post indifferent to either adopting the optimal treatment strategy or one of the other two that are independent of the signal (see Lemma 1). Therefore, ex-ante, when the doctor includes the cost of his effort in the diagnosis in his decision, the most-appropriate-treatment strategy is trivially dominated by one of the other two.

When the effort under the optimal strategy is interior, its level is increasing in B and does not depend neither on $\overline{w}$ nor on $\underline{w}$. This occurs because B is the only payment that is contingent on healing the patient. In fact, it is a bonus for having diagnosed a low severity correctly, and the accuracy of the diagnosis increases with the effort exerted. Likewise, the effort also increases with μ, since the probability that the patient needs a second treatment (and the doctor does not gain any prestige) is smaller the more accurate the diagnosis is. Finally, k affects the decision on the level of effort negatively.

From the restrictions, it is easy to see that a higher $\overline{w}$ induces the doctor to prescribe the strong treatment systematically, while a greater $\underline{w}$ induces him to always provide the mild one. The effect of the bonus on the doctor's behavior is not so clear. For example, with a high value of B the doctor prefers the strategy of treatment contingent on the signal to giving the strong treatment always. However, if B is sufficiently high, it can result in the doctor's deciding to provide the mild treatment systematically.

The presence of μ has important implications in the choice of the strategy, which we summarize in the following remark.

Remark 2 *The fact that the doctor also works in the private sector implies that, in his public activity:*

- *He has fewer incentives to choose the mild-treatment strategy.*
- *If* $e_* = \min\left\{\frac{B+\mu}{4k}, 1\right\} = \frac{B+\mu}{4k}$, *he has a greater incentive to choose the strong-treatment strategy.*

This remark confirms the prediction made at the end of Subsection 3.1, that the doctor's quest for a good reputation discourages the under-provision of services in the public sector. Also, whenever the diagnosis process is not perfect $\left(e_* = \frac{B+\mu}{4k} < 1\right)$ the doctor will have a higher tendency to over-provide services. If the diagnosis is perfect ($e_* = 1$), the decision to adopt either the most-appropriate-treatment strategy or the strong strategy is independent from μ, since with either the patient is always treated with just one treatment.

From the perspective of the health authority, the fact that the public and the private sectors are related through μ, has opposite effects. In particular, if the health authority wants to induce the doctor to adopt the most-appropriate-treatment strategy, the effect of μ over IC_* is ambiguous. On the one hand, it is less expensive to prevent the doctor from opting for the mild strategy but, on the other hand, the strong strategy becomes more attractive to him. To all this we must add that the effort the physician exerts, once he has opted for the most-appropriate-treatment strategy, is increasing in μ. Therefore, the possible rise in the government's costs to induce a positive effort in diagnosis, can be compensated by the fact that the level of effort finally exerted by the doctor, and thus, the accuracy of the diagnosis, is greater.

4.4 Contract Design

This section studies the optimal contracts that induce the physician to adopt each of the different treatment strategies presented in Section 3. We characterize the optimal

contract under different scenarios and analyze which strategy is preferred by the principal in each case. From the perspective of the health authority, we put special emphasis on studying the repercussions that the doctor's strategic behavior as a result of his dual activity has on the design of the contract and on his treatment strategy choice.

We do the analysis within a framework with limited liability constraints for the doctor. That is to say, we will impose that, under any circumstance, the doctor must receive a certain minimum payment. We denote this value by $M > 0$. Such a restriction, which is quite common in moral hazard models, reflects the existent limitations on the public liabilities that can be imposed on a doctor in the execution of his professional duties. Such limitations arise from the fact that the result of any medical treatment is, to a certain extent, unpredictable.[5]

We study two different scenarios. In the first one, we suppose that there is no agency problem and that the health authority can control the doctor's behavior perfectly, regarding both the effort he exerts in the diagnosis and his choice of treatment strategy. In the second scenario, we give the physician an informational advantage, assuming that neither of the two decisions that he can make are monitored by the health authority.

4.4.1 Contract with Symmetric Information

In this sub-section, we characterize the optimal contract under symmetric information, i.e., considering that both the level of effort that the physician exerts in the diagnosis and the treatment strategy he chooses are verifiable and contractible. Depending on the treatment strategy that the health authority wishes to induce, the payments it offers to the

[5]Moreover, from an analytical point of view, limited liability constraints introduce some kind of risk aversion in the physician's behavior. It avoids situations in which the Principal can implement the first best allocation without incurring in extra costs.

physician and his level of effort will be the ones that minimize the government's expected cost. The health authority has to consider the fact that the doctor's expected utility, under any given strategy, cannot be lower than his reservation utility (PC) and that his liability constraints have to be fulfilled (LLC).

The problem the health authority faces when computing the optimal contract for each of the treatment strategies presented is as follows:

$$\min_{\overline{w},\underline{w},B,e} C$$

$$s.t \quad \begin{cases} U \geq M & (PC) \\ \overline{w} \geq M \\ \underline{w} \geq M & (LLC) \\ B \geq 0 \\ e \in [0,1] \end{cases} \tag{4.1}$$

With $(C,U) \in \left\{\left(\overline{C},\overline{U}\right),\left(\underline{C},\underline{U}\right),\left(C_*,U_*\right)\right\}$ depending on whether the health authority chooses the strong, the mild or the most-appropriate-treatment strategy.

Proposition 1 *The optimal contract* $(\overline{w}^s,\underline{w}^s,B^s)$ *under symmetric information is as follows:*

- *If the strong-treatment strategy is contracted, any contract such that* $\overline{w}^s = M$ *with* $\overline{e}^s = 0$ *is optimal. The associated costs are:* $\overline{C}^s = M + \overline{c} + \frac{L}{2}$.

- *If the mild-treatment strategy is contracted, any contract such that* $\underline{w}^s = M$ *and* $B = 0$ *with* $\underline{e}^s = 0$ *is optimal. The associated costs are:* $\underline{C}^s = M + \underline{c} + \frac{(\overline{c}+l)}{2}$.

Proof. See Appendix 2. ■

Under symmetric information, the level of effort that the health authority demands of the doctor in contracting either the strong-treatment strategy or the mild-treatment strat-

egy, is null, i.e., $\overline{e}^s = \underline{e}^s = 0$. This occurs because in these cases treatment is decided regardless of the patient's diagnosis and, then, no effort in diagnosis is required. This decision by the health authority is consistent with the effort that the doctor would exert in the diagnosis if he had choice on this variable (see Lemma 2). The interaction between the doctor's public and private activities does not affect the social costs. The reason for this is that, although the presence of μ makes less expensive to induce a higher level of effort (since it relaxes the restriction PC), such effect does not really materialize, since, in equilibrium, the health authority will always choose $\overline{e}^s = \underline{e}^s = 0$.

The optimal contract under symmetric information, therefore, is partially undetermined, in the sense that multiple optimal contracts exist. In particular, a fixed salary, i.e. payments such that $\overline{w}^s = \underline{w}^s = M$ and $B^s = 0$ would be optimal in this context.

Before characterizing the optimal contract when the health authority decides to provide the most-appropriate-treatment strategy, we define a new term $\alpha \equiv L + l + \overline{c}$, which will be of use to us in presenting the rest of the results. α reflects the expected increase in the social cost due to an erroneous diagnosis.[6] If this makes that the physician provides the strong treatment to the patient when he needed the mild one, the loss in health associated to the over-treatment is L. If, on the contrary, he provides the mild treatment to a patient that needs the strong one, the loss is double since, to the loss in health (l) it is necessary to add the cost of providing him the strong treatment ($\overline{c}$) in a second round. As we will see in the following proposition, α plays an important role in the determination of the optimal effort in diagnosis.

We restrict the value of the parameters under study by excluding some extreme situations. First, we rule out the cases where the social cost of an erroneous diagnosis is very

[6] It is easy to see that the expected impact of a wrong diagnosis on the social costs is given by $\frac{1-e}{4}\alpha$.

low. Formally, we assume $\alpha \geq k$. Secondly, to avoid situations in which a perfect diagnosis will always be performed, we impose a sufficiently high cost on increasing the accuracy of the diagnosis ($4k \geq \max\{\alpha, \mu\}$).

The following proposition characterizes the optimal contract under symmetric information.

Proposition 2 *Under symmetric information, the optimal contract* $(\overline{w}^s, \underline{w}^s, B^s)$ *when the most-appropriate-treatment strategy is contracted, is as follows:*

- *If* $\mu \leq \frac{k}{2}$, *any contract* $\overline{w}^s \geq M$, $\underline{w}^s \geq M$, $B \geq 0$ *such that* $U_*^s = M$ *is optimal, with:*

$$e_*^s = \begin{cases} \min\left\{\frac{\alpha+\mu}{4k}, 1\right\} & \text{if } \varphi\left(\min\left\{\frac{\alpha+\mu}{4k}, 1\right\}\right) \geq \mu \\ \check{e} & \text{otherwise.} \end{cases}$$

  With $\varphi(e) = \frac{2ke^2}{3+e}$ *and* $\check{e} \in \left(\frac{\alpha+\mu}{4k}, 1\right]$ *such that* $\varphi(\check{e}) = \mu$.

  The associated costs are:

  $C_*^s = M + \frac{1}{2}(\overline{c} + \underline{c}) + \frac{1}{2}\left[\frac{1-e_*^s}{2}\alpha - \mu\left(1 + \frac{1+e_*^s}{2}\right) + k\left(e_*^s\right)^2\right]$.

- *If* $\mu > \frac{k}{2}$, *the optimal contract is such that* $\overline{w}^s = \underline{w}^s = M$, $B = 0$, *with* $e_*^s = 1$. *The associated costs are:* $C_*^s = M + \frac{1}{2}(\overline{c} + \underline{c})$.

Proof. See Appendix 2. ■

There are several insights in Proposition 2 that are worth mentioning. Firstly, if the health authority contracts the most appropriate strategy, the optimal level of effort that demands of the doctor, when μ is relatively low, depends positively on both the costs associated with an erroneous diagnosis (reflected by α), and the value of μ. If the relationship between the doctor's activities in the two different sectors is sufficiently high ($\mu > \frac{k}{2}$), the principal will always contract a perfect diagnosis. The reason for this is that the doctor's level

of effort in the diagnosis affects the government's costs through the participation constraint (PC). If $\mu \leq \frac{k}{2}$, this constraint is active and, therefore, the principal internalizes the higher expense that the higher level of effort implies. On the contrary, if μ exceeds this threshold, any contract that satisfies the limited liability constraints also fulfills (PC), independently of the level of effort required of the doctor. In other words, the doctor is willing to exert a higher level of effort since the cost he bears is compensated for an increase in his private income (through the increase in his prestige as a "good" doctor). This means that a more accurate diagnosis does not incur higher costs for the health authority and, therefore, it chooses $e_*^s = 1$.

Another interesting insight is that, as we saw in the case for the other two treatment strategies, a salary $\overline{w}^s = \underline{w}^s \geq M$ and $B^s = 0$ is an optimal contract in the absence of informational asymmetries.

Finally, we evaluate how the doctor's role as dual provider affects the social costs, and, once again, the value of μ proves to be crucial to the results. If the relationship between the doctor's public and private activities is relatively low, the costs to the health authority (C_*^s) are decreasing in μ. This is because an increase in μ makes the participation constraint of the physician to be fulfilled for lower values of $\underline{w}$ and $\overline{w}$. On the contrary, if the relationship between his public and private activities is high, the physician's participation constraint (PC) is not binding at the optimum $(U_*^s > M)$. As a result, increases in μ would imply extra rents for the doctor (i.e., increases in U_*^s) and would have no effect on social costs.

The following remark summarizes the effects of the physician's dual practice on the expected social costs.

Remark 3 *The fact that the doctor works in both the public and the private sector, in the absence of informational asymmetries, has the following effects:*

- *It does not alter the social costs of adopting treatment strategies that are independent of the signal received in the diagnosis.*

- *When the treatment strategy employed is contingent on the diagnosis, the social costs are decreasing in μ provided μ does not exceed a certain threshold, beyond which they remain unaffected.*

4.4.2 Contract with Asymmetric Information

In this sub-section we consider the case in which neither the level of effort that the doctor exerts in the diagnosis nor the treatment strategy he employs is observable.

The diagnosis of the severity of a patient's ailment and the choice of treatment that he should be given can only be done by a qualified physician. This implies that it may not be possible for the health authority to control the doctor's decisions in such activities. This scenario with asymmetric information will therefore be useful for reflecting the relationship between the physician and the health authority in real-life situations. We only study this double moral hazard situation, without analyzing intermediate scenarios. The reason is that it is precisely the interaction of these two dimensions what allows us to fully characterize the impact of the physician's dual activity.

Under asymmetric information, the problem the health authority faces is similar to the optimization program under symmetric information presented in (4.1). There are two fundamental differences. First, we must include the physician's incentive compatibility constraint with regard to the strategy that the principal wishes to induce (as defined in Lemma 3). That is to say: $\overline{IC}$ if the health authority wants to induce the strong-treatment strategy, $\underline{IC}$ if it wishes to implement the mild-treatment strategy and IC_* if it prefers the most-appropriate-treatment strategy. Secondly, as the doctor's effort in the diagnosis is no

longer contractible, the principal has to take into account that the level of effort that the doctor would exert for each possible treatment strategy contracted is as given in Lemma 2.

The health authority's optimization program is as follows:

$$\min_{\overline{w},\underline{w},B} C$$

$$s.t \begin{cases} U \geq M & (PC) \\ \overline{w} \geq M \\ \underline{w} \geq M & (LLC) \\ B \geq 0 \\ IC \\ e^a = e \end{cases} \tag{4.2}$$

With $(C, U, IC, e) \in \left\{\left(\overline{C}, \overline{U}, \overline{IC}, \overline{e}\right), \left(\underline{C}, \underline{U}, \underline{IC}, \underline{e}\right), (C_*, U_*, IC_*, e_*)\right\}$ depending on whether the health authority chooses the strong, the mild or the most-appropriate-treatment strategy.

The contract that the health authority will offer in each case is presented in the following proposition.

Proposition 3 *The optimal contract* $(\overline{w}^a, \underline{w}^a, B^a)$ *under asymmetric information is as follows:*

- *If the strong-treatment strategy is considered:* $\overline{w}^a = \underline{w}^a = M$ *and* $B^a = 0$, *with* $\overline{e}^a = 0$. *The associated costs are:* $\overline{C}^a = M + \overline{c} + \frac{L}{2}$.

- *If the mild-treatment strategy is considered:* $\overline{w}^a = M$, $\underline{w}^a = M + \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right)$ *and* $B^a = 0$, *with* $\underline{e}^a = 0$. *The associated costs are:*

$$\underline{C}^a = M + \underline{c} + \frac{1}{2}\left[\overline{c} + l + \mu\left(1 + \frac{\mu}{8k}\right)\right].$$

- *If the most appropriate strategy is considered:* $\overline{w}^a = \underline{w}^a = M$ *and:*

  i) If $\mu < 3k$, $B^a = 4\sqrt{k^2 + \mu k} - 4k - \mu$ *with* $e_*^a = \sqrt{1 + \frac{\mu}{k}} - 1 \in (0,1)$.

  ii) If $\mu \geq 3k$, $B^a = k$ *with* $e_*^a = 1$.

  The associated costs are:

$$C_*^a = M + \frac{1}{2}(\overline{c} + \underline{c}) + \frac{1}{2}\left[\frac{1 + e_*^a}{2}B + \frac{1 - e_*^a}{2}\alpha\right].$$

Proof. See Appendix 3. ■

To interpret this proposition, we study the results under each treatment strategy independently. If the government wants to induce the physician to provide the strong treatment systematically, the doctor's interest in building a good reputation for his private practice is, in principle, beneficial to the health authority, as it makes the incentive restriction more easily fulfilled. However, in equilibrium, the limited liability constraints prevent the government from taking advantage of this opportunity to reduce costs. As a result, the optimal contract would be a flat salary, and the social costs would be independent of μ, i.e., $\frac{\partial \overline{C}^a}{\partial \mu} = 0$.

If the principal prefers the mild-treatment strategy, the doctor's interest in curing the patients with just one treatment makes it more expensive to be induced $\left(\frac{\partial \underline{C}^a}{\partial \mu} > 0\right)$. The health authority would have to make an extra payment (increasing in μ) to prevent the doctor from opting for another treatment strategy that would ensure a greater positive impact on his private income (through his gain in prestige).

If, however, the government wishes to implement the appropriate-treatment strategy, the doctor's condition of dual supplier has two opposite effects. On the one hand, his interest in improving his reputation makes him prefer the strong treatment systematically, since it cures all of the patients, which makes it more expensive for the government to induce this treatment strategy. On the other hand, if the doctor finally decides to follow the

most-appropriate-treatment strategy, his interest in curing the patients induces him to exert a higher effort in the diagnosis, which is socially beneficial.[7] The result of this trade-off, is presented in the following corollary.

Corollary 1 *When the health authority induces the doctor to provide the most-appropriate treatment strategy, the physician's dual practice generates:*

- *If* $e_*^a \in (0,1)$:
  - *An increase in the social costs if* $\alpha < \hat{\alpha}(k,\mu)$.
  - *A decrease in the social costs if* $\alpha > \hat{\alpha}(k,\mu)$.

    With $\hat{\alpha}(k,\mu) = k\left(8.\sqrt{1+\frac{\mu}{k}} - \frac{3\mu}{k} - 6\right) \in (2k, 2.3k)\ \forall\mu.$
- *If* $e_*^a = 1$ *the social costs remain unaltered.*

The interpretation of Corollary 1 is quite clear, but two cases have to be distinguished. First, when the diagnosis is not perfect the value of α is crucial, since it reflects the expected social cost of an erroneous diagnosis. When α is relatively small, the health authority does not consider it optimal to induce the doctor to perform a very accurate diagnosis. The effect that really governs this case, therefore, is the higher cost involved in inducing the doctor to choose the most appropriate strategy. This makes the doctor's dual activity socially negative. In contrast, if α exceeds a certain threshold, the doctor's interest in performing a very accurate diagnosis is in line with that of the health authority's interests. As a result, although the government has to make an extra payment to ensure that the doctor chooses the most-appropriate-treatment strategy, it is compensated by the reduction in costs

[7] It can be shown that, despite this incentives to provide a more accurate diagnosis, the level of effort under asymmetric information is always sub-optimal (except when $e^a = 1$), with respect to the one with symmetric information.

derived from the doctor's improved diagnosis. The final effect of the doctor's private practice, therefore, is positive for the health authority.

When the health authority induces the physician to perform a perfect diagnosis, i.e. $e_*^a = 1$, his dual practice has no effect on social costs. The reason is that the health authority can not benefit from the physician's interest in increasing the accuracy of the diagnosis, to gain prestige, as he is already performing a perfect one.

The results of this analysis have their political implications as well. There are certain situations in which it would be in the health authority's interest to prohibit private practice by any doctor who works in the public health sector. This is the case when such dual activity by the doctor generates incentives for him to over-provide medical services. If the government prefers the physician to follow the mild-treatment strategy, or when it opts for a treatment in accordance with the signal received in the diagnosis, and, furthermore, the social costs of an incorrect diagnosis are not very high, the doctor's dual activity is negative from the social point of view. If the health authority has it within its power, it should prohibit the doctor from having a private practice whenever either of the two above-mentioned treatment strategies is chosen.

Labor legislation currently in force in many countries with mixed health care systems does not allow the health authorities to prohibit physicians' dual provision. In the next Section we will study, on the light of the analysis performed, two alternative ways in which this dual practice has been regulated.

4.5 Alternative Regulatory Measures

In the Introduction we presented alternative regulatory frameworks that exist in some countries with mixed health care systems: France, Spain and the U.K. They implement alternative regulations that can be categorized in two branches.

First, the Spanish system, based on exclusive contracts. In Spain, those doctors who work for the public sector are allowed to have their private practices if they wish to. If, however, they decide to forego this privilege, they receive a fixed monthly bonus in return for such exclusive contracting. The pertinent question here, therefore, is whether it is in the interest of the health authority to pay him a bonus to give up his private practice.

Secondly, the French and English systems, where the physicians' dual provision, although permitted, is restricted. In these countries, public physicians' private earnings cannot exceed a certain threshold. Such a threshold is computed in the UK on the basis of physicians' public revenues and in France on the basis of their total income.[8]

In the following sub-sections we will analyze the optimality of these alternative regulations.

4.5.1 Should an Exclusive Contract Be Offered?

In this sub-section we study whether, in an initial stage of the game, it is in the interest of the health authority to offer the physician an exclusive contract with extra economic remuneration for him to forego his private practice. In the case of a contract being offered, the physician can either accept it (and work exclusively in the public sector) or reject it (and be a dual supplier).[9]

[8] It is easy to see that these two posibilities are analitically equivalent.

[9] Note that the analytical difference between the subgames with and without physician's dual practice, is the presence or absence of the parameter μ.

In Sub-section 4.1, we demonstrated that, under symmetric information, the physician's dual activity is never negative from the social point of view. The analysis that follows, therefore, only makes sense when asymmetric information exists between the doctor and the health authority.

The government should only decide to offer an exclusive contract if the savings obtained from the doctor's exclusive attention to the public health sector exceed the extra cost of paying him to give up his private practice.

So far, we have not really considered the doctor's private revenue in itself, but rather the effect that his performance in the public health sector has on it. In this section, however, the value of this revenue is crucial, since it also determines the extent of the sacrifice made by the physician when he is restricted to working exclusively in the public health sector.

We define the value of the doctor's private income, in absence of any relationship between public and private practices, by π. To this value we must add the increase in private income he obtains from his enhanced image and prestige as "a good doctor" (reflected by μ).[10]

We denote the quantity that the health authority offers to the doctor in exchange for his exclusive contract by $R \geq 0$ To analyze the government's behavior, we must determine the costs involved in both scenarios separately: C^E when there is an exclusive contract (equivalent to having $\mu = 0$), and C^{NE} when the doctor works in both sectors. We define:

$$\begin{aligned} C^E &= \min\left\{\overline{C}^a(\mu=0), \underline{C}^a(\mu=0), C^a_*(\mu=0)\right\} \text{ and} \\ C^{NE} &= \min\left\{\overline{C}^a, \underline{C}^a, C^a_*\right\}. \end{aligned}$$

These two values determine the maximum that the government will be willing to pay for an

[10] This sort of modelling implicitly assumes a linear relationship between the increase in the doctor's prestige as a result of his satisfactory performance in the public sector, and the increase in his private income.

exclusive contract ($R_{\max}$). Formally:

$$R_{\max} = \max\left\{0, C^{NE} - C^{E}\right\}.$$

The physician, on the other hand, will be interested in signing an exclusive contract if the remuneration he receives exceeds a certain threshold, $R_{\min}$, defined as:

$$R_{\min} = \pi + U^{NE} - U^{E},$$

where U^{NE} is the doctor's expected utility when he works in both sectors, and U^{E} when he accepts an exclusive contract. $U^{NE} \in \left\{\overline{U}^{a}, \underline{U}^{a}, U_{*}^{a}\right\}$, and $U^{E} \in \{\overline{U}^{a}(\mu = 0), \underline{U}^{a}(\mu = 0), U_{*}^{a}(\mu = 0)\}$, having either value depending on the strategy chosen by the principal in each scenario.

Therefore, there will be room for exclusive contracts if there exist values of $\mu > 0$ and $\pi > 0$ such that $R_{\min} < R_{\max}$.

In the following proposition we analyze this possibility.

Proposition 4 *When the government offers an incentive contract to the doctor, it is never optimal for it to offer him an exclusive contract as well.*

Proof. See Appendix 4. ■

The government should only consider the option of an exclusive contract when the doctor's dual activity implies an increase in costs. It should also be noted that the treatment strategy chosen by the principal can vary, depending on whether the doctor is a dual supplier or not. From the previous analysis we know that there are two cases in which an exclusive contract is never desirable: Either when the government wants the doctor to systematically adopt the strong-treatment strategy, or when both, it prefers a treatment in accordance with the signal of the severity received and the social cost of an incorrect diagnosis is high.

Proposition 4 shows that even in cases where the doctor's dual practice is detrimental to the government, it will not offer exclusive contracts. The reason for this is that, through its incentive contract, the health authority partially mitigates the increase in costs. Therefore, the savings from the fact that the doctor is not a dual supplier are never enough to compensate the doctor for what he loses from not having a private practice.

These results seem to be difficult to reconcile with what actually happens in the real-economy, since in some "mixed" health systems, where doctors are dual suppliers, the health authorities do offer exclusive contracts. It must be remembered, however, that existing remuneration systems are not generally based on incentives.

The following proposition shows how the previous result changes radically if we focus our attention on remuneration systems based on a salary.

Proposition 5 *When the government pays the doctor a fixed salary, i.e.,* $\overline{w} = \underline{w} = M$ *and* $B = 0$, *offering him an exclusive contract can be optimal if* $\overline{c} > 2\underline{c} + l - L$.

Proof. See Appendix 5. ■

Offering a salary to a physician who works in both sectors encourages him to prescribe the stronger treatment systematically, as a means of enhancing his image and reputation. This over-provision of services is not optimal if $\overline{c} > 2\underline{c} + l - L$ (i.e., if the strong treatment is sufficiently more expensive than the mild one). If this condition holds, therefore, there will be values of μ and π for which the fact that the government offers an exclusive contract and the doctor accepts it, is an equilibrium.

Our results provide a rationale for the existence of exclusive contracts, like the ones implemented by the Spanish Health Administration, as a second-best choice. If the health authority can offer incentive contracts there is no reason for any exclusive contracts to be

offered. If, however, the health authority is restricted to payment systems based on a flat salary, the exclusive contracts can be a useful tool for helping to contain expenses within the public health sector.

In addition to this, one may think of other dimensions of the physician's activity with outputs difficult to measure, such as teaching or researching or even treating chronic illnesses. In all these cases, it is likely that the health authority will not be able to offer incentive contracts to physicians (or at least not under the structure we have proposed in this paper) and, hence, exclusive contracts may be a powerful regulatory tool.

4.5.2 Should we Limit Physicians' Private Earnings?

In this sub-section we study the consequences of a regulation that imposes an upper bound on the amount of public physicians' private earnings, on the light of the analysis we have performed. As we already said, this sort of policy is currently in force in countries like France and the U.K.

Firstly, we can compare this regulation with the Spanish one, based on offering exclusive contracts to the physicians. With the Spanish regulation, the public authority has absolutely no power, as it cannot impose the physicians to sign the exclusive contract. In the French-English system, on the contrary, the public authority goes one step beyond. Even if its power is restricted, since it cannot forbid physicians' dual practice, it can limit this dual provision by fixing an upper bound on the physicians' private earnings. From this qualitative difference in the power of the public authority, it is clear that, when the physicians' dual practice is welfare decreasing, its consequences will be less severe under the French-English type regulation.

We move now to a more specific analysis of the French-English system. To do so,

we define by $\Pi_{\max}$ the maximum amount of private earnings allowed by the health authority. In this analysis we consider that $\Pi_{\max} \in (\Pi, \Pi + \mu)$. With this restriction we ensure that, on the one hand, the physician is still concerned by his prestige in the private sector ($\Pi_{\max} > \Pi$), but on the other hand, the regulation is active and imposes a restriction on the amount of private profits the physician can get ($\Pi_{\max} < \Pi + \mu$).

Let us denote by μ' the difference in earnings between a physician with a "high" prestige, and one without it, i.e. $\mu' = \Pi_{\max} - \Pi$. By construction, we have that $\mu' < \mu$. The direct implication of this sort of regulation is, therefore, a decrease in the physicians' incentives to gain prestige as a practitioner, as this has a lower impact on his earnings.

We proceed now to state how this affects the social costs borne by the health authority, depending on the treatment strategy chosen.

Corollary 2 *A regulation that limits physicians' private earnings generates:*

- *If the mild-treatment strategy is chosen, a reduction in the social costs.*
- *If the strong-treatment strategy is chosen, no change in the social costs.*
- *If the most-appropriate-treatment strategy is chosen:*
  - *A reduction in the social costs, for low values of α.*
  - *An increase in the social costs, for high values of α.*

The interpretation of Corollary 2 is clear for the first two cases. When the health authority follows the mild-treatment strategy, physician's dual practice is detrimental as it generates an incentive to over-provide services. In this case, the regulation is beneficial precisely because it reduces such incentives. When the strong-treatment strategy is chosen,

physician's private work has no effect on the social costs and, hence, the regulation has no impact.

However, when the health authority chooses the most-appropriate treatment strategy, the implications of the regulation are not so clear. They depend on whether the social cost of an erroneous diagnosis is high or not.[11] When an incorrect diagnosis of the severity of the disease is socially not very harmful, the double practice of the physician is cost-increasing. The regulation, therefore, helps to mitigate the increase in the costs. On the contrary, when the health authority is concerned about the accuracy of the diagnosis, the physician's interest in gaining prestige is a powerful tool, as it makes the physician increases his effort in the diagnosis. A regulation that limits the practitioner's private earnings, will reduce his incentives to perform a correct diagnosis and will be, therefore, welfare decreasing.

Therefore, our conclusion is that this sort of regulatory policy may be beneficial from a social point of view, although it can generate as a non-desired effect, a reduction on the physicians' incentives to perform an accurate diagnosis.

4.6 Modelling the Acquisition of Reputation

In our model, we have considered a representative physician with a given level of ability that can be measured by k (the marginal disutility of his effort in the diagnosis). The physician provides medical services to a patient in a static framework. In this setting, we have studied the implications of the presence of physician's reputation concerns.

We have assumed that the physician's earnings in the private sector have two components: a fixed part π, which is exogenously given, and a variable part (μ), which is affected by the physician's behavior in his public duties. We have modelled the process of acquisition

[11] The threshold that determines the result is the one obtained in Corollary 1.

of prestige as follows: physician gains prestige (μ) when he provides a treatment to a patient and he cures him in a single round of treatment.

In this section we show that our formulation can be viewed as a reduced form of a more general game. In such a game, the physician increases his ability (it can be interpreted as a process of learning) when he provides the appropriate treatment to the illness. Afterwards, the patient observes the rounds of treatment provided by the physician as an imperfect signal of his learning process. Under this new formulation, gaining or not prestige not only depends on the physician's behavior, but also on the perception that the private patients have about his ability. This more general set-up endogeneizes the physicians' reputation acquisition and provides a rationality for the way in which we have modelled it throughout the article.

Although to model properly the process of learning by the physician it would be interesting to deal with a more dynamic setting, we consider here a simple form for it. If the physician learns after he provides a treatment in the public sector, his value of k is reduced (from k_H to k_L). This "acquired experience" will be used, then, when he offers his services as a private provider.

We are not interested in modelling the provision of medical services in the private sector. Private sector patients, however, are crucial as they determine physician's private earnings. When they require private medical services, they prefer to visit a physician with a high level of ability. We can model this by considering that they will only demand services to a physician, if they consider more likely that he is a physician who has learned. In this case, physician's private revenues will be $\pi+ \mu$, instead of only π.

Under this modelization, it will be in the interest of the private patients to infer whether the physician has learned or not in his public duties. These patients, however, can not observe the physician's behavior. The only thing they can observe is whether the public

patient was cured in a single round of treatment ($1r$) or if he required a second round ($2r$).

Private sector patients will, therefore, use the number of rounds of treatments as an "imperfect signal" of the successful learning of the physician. The signal, although informative, is not perfect because there is a case in which only one round of treatment is provided and, however, the physician prescribes the wrong treatment (this is the case of a low severity patient who is given the strong treatment).

We need to find a structure of beliefs of the private patients that is consistent with the behavior of the agents (physician and patients) in the equilibrium it generates. Let us consider the following structure:

$$q(k_H|1r) = \beta \qquad q(k_L|1r) = 1 - \beta$$

$$q(k_H|2r) = \alpha \qquad q(k_L|2r) = 1 - \alpha,$$

with $\beta \leq \frac{1}{2}$ and $\alpha \geq \frac{1}{2}$. Under these beliefs, patients perceive that after one round of treatment is more likely that the physician has learned, while after two, it is more likely that he has not learned. To show the consistency of the beliefs we proceed as follows.

First, given these perceptions, we study the behavior that the private patients will have. The fact that they interpret one round of treatment as a signal that the physician is high skilled, together with their preferences (they will demand services if they consider more likely that they will be treated by a physician who has learned), generates the following: If they observe one round of treatment in the public sector, they will demand private services from the physician. The physician's private earnings will be: $\pi + \mu$. Analogously, if they observe two rounds of treatment in the public sector, they will not demand private services from the physician, and the physician's private earnings will be: π.

Secondly, given the patients' behavior, we characterize the physician's public performance and the health authority's optimal contract. Note that since the structure of payments

induced by the behavior of the patients coincides with the one we have used in our model, this part of the analysis has already been performed in Sections 3 and 4.

Finally, for this to be consistent in equilibrium, we need to check that the behavior of the physician confirms the beliefs of the patients:

$$q(k_L|1r) = \Pr(\text{right treatment}|1r) =$$

$$= \begin{cases} 1 & \text{If the treatment strategy is the mild one} \\ \frac{1}{2} & \text{If the treatment strategy is the strong one} \\ \frac{\frac{1+\epsilon}{2}}{\frac{3+\epsilon}{4}} \geq \frac{1+\epsilon}{2} > \frac{1}{2} & \text{If the treatment strategy is the appropriate one} \end{cases}$$

Hence, $q(k_L|1r) = 1 - \beta \geq \frac{1}{2} \Rightarrow \beta \leq \frac{1}{2}$.

$$q(k_L|2r) = \Pr(\text{right treatment}|2r) = 0.$$

Hence, $q(k_L|2r) = 1 - \alpha = 0 \Rightarrow \alpha = 1$.

Therefore, the beliefs of the private patients are consistent with the behavior of the physician. This constitutes an equilibrium in which the "reputation effect" arises as the outcome of an imperfect signalling process.

This analysis, therefore, allows us to understand our simple form of modeling physician's prestige, as a reduced form of a more general model in which the "reputation effect" appears endogenously.

4.7 Concluding Remarks

This paper studies, in a moral-hazard framework, the implications that the physician's dual activity has for public health authorities. As mentioned, the fact that many

doctors work both in public and private sector at the same time is common in mixed health care systems.

From the different dimensions in which conflicting interests may arise between the doctor's public and private practices, we have focused on a particular one: The possibility that the physician uses his work in the public sector as a way of improving his professional prestige and, hence, increasing his private revenue. We derived optimal payment contracts for a physician in the public sector and we studied how his incentives are affected, under such contracts, when he is also a private provider.

We have found that the physician's dual practice has conflicting effects. On the one hand, his interest in curing patients and gaining prestige, generates an over-provision of health services. On the other hand, if the health authority is able to control these incentives to over-provide services, then it can benefit from the physician's increased interest in doing a more accurate diagnosis.

Regarding policy recommendations, our analysis suggests that the physician's dual practice can be either welfare improving or reducing, depending on the treatment policy that the health authority wants to implement. If the priority of the health authority is to contain costs, then the doctor's dual activity is negative. If the priority is to minimize patients' health losses, his dual practice affords the objective at a lower cost.

The assumption of equal probability for the two severities of the illness is made for the sake of analytical tractability. In spite of this, some new effects of dealing with other probability structures can be spotted. First, the larger the proportion of patients suffering from a low severity is, the lower the physician's incentives to over-provide services will be. However, a second one may overwhelm this positive effect. Now, the over-provision of services becomes more costly, since it affects to a larger proportion of the population. Therefore, the

final effect over the health authority costs can not be unambiguously determined.

We have imposed no restriction on the relationship between the two types of patients' health losses (l for an unsuccessful treatment and L for an over-treatment). If we wanted to focus on which treatment policy is preferred by the health authority, the relationship between such parameters would be crucial. In any case, our analysis suggests that the physician's dual activity will make the health authority more reluctant to implement the mild-treatment strategy, as it becomes more costly due to the physician's incentives to over-provide services.

In several countries with mixed health care systems the labor legislation in force allows dual activity by the physicians. In spite of this, different measures have been undertaken by the governments to regulate this issue. This work provides a theoretical framework in which the optimality of exclusive contracts and of limits on physicians' private incomes can be addressed. Considering the former, our analysis shows that if the remuneration policy of the health authority is based on a salary, then the exclusive contracts can be useful for cost-containment. However, if we consider incentive payment contracts, it is never optimal to offer an exclusive contract to the physician.

Considering the latter type of regulation, and under an incentive contract, we show that limiting physicians's private income is beneficial from a social point of view, except in those cases in which the health authority is highly concerned about the accuracy of the diagnosis the physician performs. In such a case, this kind of regulation is socially harmful.

4.8 Appendix

Appendix 1. Physician's expected utility and health authority's expected costs under the alternative treatment strategies

1.- Under the strong-treatment strategy:

$$\overline{U} \equiv \overline{w} + \mu - V(e).$$

$$\overline{C} \equiv \overline{w} + \overline{c} + \tfrac{1}{2}L = \overline{U} - \mu + \overline{c} + \tfrac{1}{2}L + V(e).$$

2.- Under the mild-treatment strategy:

$$\underline{U} \equiv \underline{w} + \tfrac{1}{2}(\mu + B) - V(e).$$

$$\underline{C} \equiv \tfrac{1}{2}(\underline{w} + B + \underline{c}) + \tfrac{1}{2}(\underline{w} + \underline{c} + \overline{c}) + \tfrac{1}{2}l = \underline{U} - \tfrac{1}{2}\mu + \underline{c} + \tfrac{1}{2}(\overline{c} + l) + V(e).$$

3.- Under the most-appropriate treatment strategy:

$$U_* \equiv \tfrac{1}{2}\left[\tfrac{1+e}{2}(\underline{w} + \mu + B) + \tfrac{1-e}{2}(\overline{w} + \mu)\right] + \tfrac{1}{2}\left[\tfrac{1+e}{2}(\overline{w} + \mu) + \tfrac{1-e}{2}(\underline{w})\right] - V(e) =$$

$$= \tfrac{1}{2}\left[\underline{w} + \overline{w} + \mu + \tfrac{1+e}{2}(B + \mu)\right] - V(e).$$

$$C_* \equiv \tfrac{1}{2}\left[\tfrac{1+e}{2}(\underline{w} + B + \underline{c}) + \tfrac{1-e}{2}(\overline{w} + \overline{c})\right] + \tfrac{1}{2}\left[\tfrac{1+e}{2}(\overline{w} + \overline{c}) + \tfrac{1-e}{2}(\underline{w} + \underline{c} + \overline{c})\right] +$$

$$+\tfrac{1-e}{2}\left[\tfrac{1}{2}L + \tfrac{1}{2}l\right] = \tfrac{1}{2}\left[\overline{c} + \underline{c} + \overline{w} + \underline{w} + \tfrac{1+e}{2}B + \tfrac{1-e}{2}(\overline{c} + l + L)\right] =$$

$$= U_* - \tfrac{1}{2}\mu\left(1 + \tfrac{1+e}{2}\right) + \tfrac{1}{2}\left[\underline{c} + \overline{c} + \tfrac{1-e}{2}(L + l + \overline{c})\right] + V(e).$$

Appendix 2. Proof of Propositions 1 and 2

Under symmetric information, the optimal payment contract and the optimal level of effort under the different treatment strategies are the solution to the following program:

$$\min_{\overline{w},\underline{w},B,e} C$$

$$s.t \quad \begin{cases} U \geq M & (PC) \\ \overline{w} \geq M \\ \underline{w} \geq M & (LLC) \\ B \geq 0 \\ e \in [0,1] \end{cases}$$

With $(C,U) \in \left\{\left(\overline{C},\overline{U}\right),\left(\underline{C},\underline{U}\right),(C_*,U_*)\right\}$ depending on whether the health authority chooses the strong, mild or the most-appropriate-treatment strategy.

If the health authority contracts both the mild or the strong treatment strategy, since the treatment decision is independent on the physician's diagnosis and inducing effort only increases the health authority costs, the health authority prefers the effort to be minimal, i.e., $\overline{e}^s = \underline{e}^s = 0$.

The (LLC) imply the (PC). Therefore, the health authority chooses the cheapest contract compatible with the (LLC). Then, when the health authority contracts the strong strategy any contract $(\overline{w}^s,\underline{w}^s,B^s)$ with $\overline{w}^s = M$ is optimal. Analogously, when the health authority contracts the mild strategy any contract $(\overline{w}^s,\underline{w}^s,B^s)$ with $\underline{w}^s = M$ and $B = 0$ is optimal.[12]

If the health authority contracts the most-appropriate-treatment strategy, the prob-

[12]Note that under these payment structures, positive levels of effort can be sustained in equilibrium. However, this would not alter the social costs and only would reduce physician's utility.

lem it faces is given by:

$$\min_{\overline{w},\underline{w},B,e} C_* = \frac{1}{2}\left[\overline{c}+\underline{c}+\overline{w}+\underline{w}+\frac{1+e}{2}B+\frac{1-e}{2}\alpha\right]$$

$$\begin{array}{rl} & \frac{1}{2}\left[\underline{w}+\overline{w}+\mu+\frac{1+e}{2}(B+\mu)\right]-k\frac{e^2}{2}\geq M \qquad (PC) \\ & \left.\begin{array}{l}\overline{w}\geq M \\ \underline{w}\geq M \\ B\geq 0\end{array}\right| \quad (LLC) \\ s.t. & \\ & e\in[0,1] \end{array}$$

We will study independently two cases, depending on whether (PC) is binding or not at the optimum.

i) If (PC) is not binding at the optimum, then we can ignore it. Since the health authority's costs are increasing in the payments to the physician, we make the (LLC) binding. This immediately implies that $e_*=1$. One can check that this solution will indeed fulfill that (PC) is not binding (as we have assumed) if and only if $\mu>\frac{k}{2}$.

ii) If (PC) is binding at the optimum, and the limited liability constraints (LLC) are fulfilled, then a necessary condition for the optimal level of effort is that $\varphi(e)=\frac{2ke^2}{3+e}\geq\mu$. $\varphi(e)$ is such that $\varphi'(e)>0$, $\varphi(0)=0$ and $\varphi(1)=\frac{k}{2}$.

Substituting (PC) into the objective function and optimizing with respect to e we find that $e_*=\min\{\frac{\alpha+\mu}{4k},1\}$. Note that e_* will only be the solution provided $\varphi(e_*)\geq\mu$.

When $\varphi(e_*)\geq\mu$, then the solution to the problem is $e_*=\min\{\frac{\alpha+\mu}{4k},1\}$.

When $\varphi(e_*)<\mu$, e_* cannot be the solution. Moreover, since $\varphi(e)\leq\frac{k}{2}\ \forall e$, then if $\mu>\frac{k}{2}$ there does not exist any $e\in[0,1]$ such that $\varphi(e)\geq\mu$. We distinguish two situations:

If $e_*=1$, then $\varphi(e_*)<\mu\Leftrightarrow\mu>\frac{k}{2}$, and we have already shown that in this region there is no solution with (PC) binding.

If $e_* = \frac{\alpha+\mu}{4k}$, and $\mu \leq \frac{k}{2}$ we have: $\varphi\left(\frac{\alpha+\mu}{4k}\right) < \mu$, $\varphi(1) = \frac{k}{2} \geq \mu$ and $\varphi'(e) > 0$. These imply that there exists a value $\check{e} \in \left(\frac{\alpha+\mu}{4k}, 1\right]$ such that $\varphi(\check{e}) = \mu$. This value $\check{e}$ is the optimal level of effort in this region.

Therefore, we have characterized the solution for all the range of parameter values. The optimal contract under symmetric information is presented in Proposition 2.

Appendix 3. Proof of Proposition 3

When the health authority cannot contract either the physician's effort or the treatment strategy, then its optimization program is as follows:

$$\min_{\overline{w},\underline{w},B} C$$

$$s.t \begin{cases} U \geq M & (PC) \\ \left.\begin{array}{l} \overline{w} \geq M \\ \underline{w} \geq M \\ B \geq 0 \end{array}\right| & (LLC) \\ IC \\ e^a = e \end{cases}$$

With $(C, U, IC, e) \in \left\{\left(\overline{C}, \overline{U}, \overline{IC}, \overline{e}\right), \left(\underline{C}, \underline{U}, \underline{IC}, \underline{e}\right), (C_*, U_*, IC_*, e_*)\right\}$ depending on whether the health authority chooses the strong, the mild or the most-appropriate-treatment strategy.

First, note that the participation constraint (PC) is always implied by the limited liability constraints (LLC) when the health authority wants to induce either the strong or the mild treatment strategy. This, in turn, makes that when the most-appropriate treatment strategy is considered the participation constraint is implied by the incentive compatibility constraints.

The physician chooses the level of effort he will exert in diagnosis. When the strategy chosen is either the strong or the mild one, the physician always chooses to exert no effort (i.e., $\underline{e} = \overline{e} = 0$) since it has no effect on his revenue and only implies higher costs.

Hereinafter, we need to study the health authority's problem under the three alternative treatment strategies independently:

i) When the health authority chooses the strong-treatment strategy, its expected costs do not depend on B. Moreover, since B makes the constraint $(\overline{IC})$ tougher, the optimal B is the lowest one such that the constraints are satisfied. Therefore, the liability constraint associated to B is binding at the optimum, i.e., $B = 0$. The same argument applies to $\underline{w}$ and, then, at the optimum $\underline{w} = M$. It is easy to verify that, under $B = 0$ and $\underline{w} = M$, the liability constraint associated to $\overline{w}$ is more demanding that the incentive compatibility constraint $(\overline{IC})$. Since the health authority's costs are increasing in $\overline{w}$, the liability constraint associated to $\overline{w}$ also binds at the optimum.

Therefore, if the strong-treatment strategy is considered the optimal contract is such that:

$$\overline{w}^a = \underline{w}^a = M \text{ and } B^a = 0, \text{ with } \overline{e}^a = 0.$$

The associated costs are: $\overline{C}^a = M + \overline{c} + \frac{L}{2}$.

ii) When the health authority chooses the mild-treatment strategy, an analogous argument to the one used with $\underline{w}$ in i) ensures that $\overline{w} = M$ in the optimum.

We can reduce the health authority's optimization problem as follows:

$$\min_{\underline{w},B} \underline{C} = \underline{w} + \underline{c} + \frac{1}{2}(B + \overline{c} + l)$$

$$s.t \quad \left\{ \begin{array}{l|l} \underline{w} \geq M & \\ B \geq 0 & (LLC) \\ e_* \left(\frac{B+\mu}{2} - k e_* \right) \leq \underline{w} - M + \frac{B-\mu}{2} & (\underline{IC}) \end{array} \right.$$

with $e_* = \min\left\{\frac{B+\mu}{4k}, 1\right\}$.

Since $\underline{C}$ is increasing in $\underline{w}$, $\underline{w} = \max\{M, M'\}$ in the optimum, with $M' = M + e_*\left(\frac{B+\mu}{2} - k e_*\right) - \frac{B-\mu}{2}$.

M' is decreasing in B for all $e_* < 1$ (i.e. $B < 4k - \mu$) and constant if $e_* = 1$. $(B \geq 4k - \mu)$. Moreover, when M' is decreasing in B, its slope is greater than $-\frac{1}{2}$.

When $B = 0$, $M' = M + \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right) > M$. Therefore, the only two candidates to solution (vertexes of the domain) are:

$$B = 0 \text{ and } \underline{w} = M + \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right),$$

$$B > 0 \text{ and } \underline{w} = \max\{M, M + \mu - k\}.$$

In order to choose between these two candidates, it is useful to know that the slope of the level-curves of the objective function is $\frac{dw}{dB} = -\frac{1}{2}$, whereas $\frac{\partial M'}{\partial B} > -\frac{1}{2}$. This directly implies that, if the mild-treatment strategy is considered, the optimal contract is such that:

$$\overline{w}^a = M, \underline{w}^a = M + \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right) \text{ and } B^a = 0, \text{with } \underline{e}^a = 0.$$

The associated costs are:

$$\underline{C}^a = M + \underline{c} + \frac{1}{2}\left[\overline{c} + l + \mu\left(1 + \frac{\mu}{8k}\right)\right].$$

iii) When the health authority chooses the most-appropriate-treatment strategy, the optimal contract is the solution to:

$$\min_{\overline{w},\underline{w},B} C_* = \frac{1}{2}\left[\overline{c} + \underline{c} + \overline{w} + \underline{w} + \frac{1+e_*}{2}B + \frac{1-e_*}{2}\left(\overline{c} + l + L\right)\right]$$

$$s.t \quad \begin{cases} \left.\begin{array}{l} \overline{w} \geq M \\ \underline{w} \geq M \\ B \geq 0 \end{array}\right| \quad (LLC) \\ e_*\left(\frac{B+\mu}{2} - ke_*\right) \geq \left|\overline{w} - \underline{w} + \frac{\mu - B}{2}\right| \quad (IC_*) \\ e_* = \min\left\{\frac{B+\mu}{4k}, 1\right\} \end{cases}$$

Consider first the case where $e_* = 1$ (or equivalently $\frac{B+\mu}{4k} \geq 1$). Then, the health authority's problem can be rewritten as:

$$\min_{\overline{w},\underline{w},B} C_* = \frac{1}{2}\left[\overline{c} + \underline{c} + \overline{w} + \underline{w} + B\right]$$

$$s.t \quad \begin{cases} \left.\begin{array}{l} \overline{w} \geq M \\ \underline{w} \geq M \\ B \geq 0 \end{array}\right| \quad (LLC) \\ \left.\begin{array}{l} \overline{w} \leq \underline{w} + B - k \\ \overline{w} \geq \underline{w} + k - \mu \end{array}\right| \quad (IC_*) \\ \frac{B+\mu}{4k} \geq 1 \end{cases}$$

The problem above is one of linear programming. We want to minimize a function that is increasing in $\overline{w}, \underline{w}$ and B, subject to a series of linear constraints. We find two solutions depending on the value of the parameters:

- If $\mu \leq k$, the optimal contract is such that:

$$\underline{w} = M,\ \overline{w} = M + k - \mu \text{ and } B = 4k - \mu$$

The associated costs are:

$$C_*^0 = M + \frac{1}{2}\left(\overline{c} + \underline{c} + 5k - 2\mu\right)$$

- If $\mu > k$, the optimal contract is such that:

$$\underline{w} = M,\ \overline{w} = M \text{ and } B = \max\left\{4k - \mu, k\right\}.$$

The associated costs are:

$$C_*^1 = M + \frac{1}{2}\left(\overline{c} + \underline{c} + \max\left\{4k - \mu, k\right\}\right)$$

Consider now the case where $e_* = \frac{B+\mu}{4k} < 1$. Then, the health authority's problem can be rewritten as:

$$\min_{\overline{w},\underline{w},B} C_* = \frac{1}{2}\left[\overline{c} + \underline{c} + \overline{w} + \underline{w} + \frac{1}{2}\left(B + \alpha\right) + \frac{B+\mu}{8k}\left(B - \alpha\right)\right]$$

$$s.t \quad \begin{cases} \left.\begin{array}{l} \overline{w} \geq M \\ \underline{w} \geq M \\ B \geq 0 \end{array}\right| & (LLC) \\ \left.\begin{array}{l} \frac{1}{2}\left(B - \mu\right) + \underline{w} - \overline{w} - \frac{(B+\mu)^2}{16k} \leq 0 \\ \overline{w} - \underline{w} - \frac{1}{2}\left(B - \mu\right) - \frac{(B+\mu)^2}{16k} \leq 0 \end{array}\right| & (IC_*) \\ \frac{B+\mu}{4k} < 1 \end{cases}$$

It is easy to see that C_* is convex and increasing in B for all $B \in [0, 4k - \mu]$.

We can re-write the (IC_*) as follows:

$$\frac{(B-\mu)}{2} - \frac{(B+\mu)^2}{16k} \leq \overline{w} - \underline{w} \leq \frac{(B+\mu)^2}{16k} + \frac{(B-\mu)}{2}.$$

Note that both the right-hand-side and the left-hand-side terms of the inequality are increasing in B. This restriction, together with $\overline{w} \geq M$ and $\underline{w} \geq M$, determine the restricted domain of the minimization program.

We analyze how this domain changes with B, taking into account that the optimal level of B is the lowest one such that the constraints are fulfilled. We obtain that the solution to the program has to be on the frontier of the right-hand-side restriction:

$$\overline{w} - \underline{w} = \frac{(B+\mu)^2}{16k} + \frac{(B-\mu)}{2}.$$

Moreover, we find that $\overline{w} = M$, and this implies: $\underline{w} = M - \frac{(B+\mu)^2}{16k} - \frac{(B-\mu)}{2}$. This will only be a feasible value of $\underline{w}$ if it fulfills the initial restriction $\underline{w} \geq M$. It is easy to check that it holds only if $B \in \left[0, \tilde{B}\right]$ with $\tilde{B} = 4\sqrt{k^2 + \mu k} - 4k - \mu > 0$.

Substituting the above values of $\overline{w}$ and $\underline{w}$ in the objective function and minimizing with respect to B, we can see that the objective function is convex in B and attains a global minimum at $B = \alpha$.

It can be shown that $\alpha > \tilde{B}$. This implies, then, that the optimal level of B is $B_* = \tilde{B}$. This can be the solution provided $\tilde{B} < 4k - \mu$, i.e., if $\mu < 3k$. The value of e associated is $e = \sqrt{1 + \frac{\mu}{k}} - 1$.

Summarizing, the unique candidate to solution is a contract such that:

$$\begin{aligned} \overline{w} &= \underline{w} = M \text{ and } B = 4\sqrt{k^2 + \mu k} - 4k - \mu \\ \text{with } e &= \sqrt{1 + \frac{\mu}{k}} - 1 \in (0,1) \text{ if and only if } \mu < 3k. \end{aligned}$$

The associated costs are:

$$C_*^2 = M + \frac{1}{2}(\overline{c} + \underline{c}) + \frac{1}{2}\left(\alpha + \frac{1}{2}\sqrt{1 + \frac{\mu}{k}}\left[4\sqrt{k^2 + \mu k} - 4k - \mu - \alpha\right]\right)$$

If $\mu \leq k$, comparing C_*^0 with C_*^2 we find that the solution with $e_* < 1$ always dominates. If $k \leq \mu < 3k$, comparing C_*^1 with C_*^2 we find that the solution with $e_* < 1$ always dominates.

Finally, if $\mu \geq 3k$ the solution is $e_* = 1$ with $\overline{w} = \underline{w} = M$ and $B = k$.

Therefore, the optimal contract under asymmetric information if the most-appropriate treatment strategy is considered, is as described in Proposition 3.

Appendix 4. Proof of Proposition 4

Several cases have to be studied independently:

i) If $C^E = \overline{C}^a(\mu = 0)$, with $C^{NE} = \min\left\{\overline{C}^a, \underline{C}^a, C^a_*\right\}$.

This is the case in which the health authority wants to induce, when the physician signs an exclusive contract, the strong-treatment strategy. Since $\overline{C}^a(\mu = 0) = \overline{C}^a$, then $C^{NE} \leq \overline{C}^a(\mu = 0)$. Therefore, it is never optimal for the health authority to offer an exclusive contract to the physician.

ii) If $C^E = \underline{C}^a(\mu = 0)$ and $C^{NE} = \underline{C}^a$.

The maximum the health authority is willing to pay for an exclusive contract is:

$$R_{\max} = \max\left\{0, \underline{C}^a - \underline{C}^a(\mu = 0)\right\} = \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right).$$

The physician will sign the exclusive contract if he receives at least:

$$R_{\min} = \pi + \underline{U}^a - \underline{U}^a(\mu = 0) = \pi + \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right).$$

Both conditions are compatible (i.e., $R_{\max} \geq R_{\min}$) only if $\pi \leq 0$, which is a contradiction. Therefore, in this case, an exclusive contract is never offered.

iii) If $C^E = \underline{C}^a(\mu = 0)$ and $C^{NE} = \overline{C}^a$.

Proceeding analogously, we find that an exclusive contract is offered if:

$$R_{\max} \geq R_{\min} \Leftrightarrow \pi + \mu \leq \frac{\overline{c}}{2} - \underline{c} + \frac{L - l}{2}.$$

Since $C^{NE} = \overline{C}^a$, this necessarily implies that $\overline{C}^a < \underline{C}^a$, and this is true if:

$$\frac{\overline{c}}{2} - \underline{c} + \frac{L - l}{2} < \frac{1}{2}\mu\left(1 + \frac{\mu}{8k}\right).$$

Both conditions are simultaneously fulfilled only in the extreme case when $\pi \to 0$, and provided $\mu > 8k$, which contradicts our assumption $\mu < 4k$.

Therefore, $R_{\max}$ do not exceed $R_{\min}$ and an exclusive contract is never offered in this case.

iv) If $C^E = \underline{C}^a(\mu = 0)$ and $C^{NE} = C^a_*$.

Applying the same argument than in the cases above, we can check that an exclusive contract will be offered to the physician (i.e. $R_{\max} \geq R_{\min}$) if:

$$\frac{1}{4}(\overline{c} + L - l) - \frac{1}{2}\underline{c} - \frac{\alpha}{4}\left(\sqrt{1+\frac{\mu}{k}} - 1\right) - \frac{\mu}{2}\left(1 + \frac{\sqrt{1+\frac{\mu}{k}}}{2}\right) + \frac{k}{2}\left(\sqrt{1+\frac{\mu}{k}} - 1\right)^2 \geq \pi.$$

Since $C^{NE} = C^a_*$, this necessarily implies that $C^a_* < \underline{C}^a$, and this is true if:

$$\frac{\alpha}{4} + \frac{\mu}{2}\left(1 + \frac{\mu}{8k}\right) - \frac{1}{2}\left[\alpha + \frac{1}{2}\sqrt{1+\frac{\mu}{k}}\left(4\sqrt{k^2+\mu k} - 4k - \mu - \alpha\right)\right] \geq \frac{1}{4}(\overline{c} + L - l) - \frac{1}{2}\underline{c}.$$

Both conditions are compatible if:

$$\frac{\mu}{4k}\left(\frac{\mu}{4k} - \sqrt{1+\frac{\mu}{k}}\right) + \frac{1}{2}\left(\sqrt{1+\frac{\mu}{k}} - 1\right)^2 - \sqrt{1+\frac{\mu}{k}}\left(\sqrt{1+\frac{\mu}{k}} - 1 - \frac{\mu}{4k}\right) - \frac{\pi}{k} \geq 0.$$

This condition never holds, for any $k > 0$, $\pi > 0$ and $\mu \in (0, 4k)$. An exclusive contract, therefore, will never be offered.

Appendix 5. Proof of Proposition 5

Under asymmetric information, if the health authority pays the doctor a fixed salary such that $\overline{w} = \underline{w} = M$ and $B = 0$, the doctor will prescribe the strong treatment systematically. Then:

i) If the health authority wants to induce the strong-treatment strategy, an exclusive contract will never be offered.

ii) If the health authority wants to induce the mild-treatment strategy, it will be willing to offer an exclusive contract provided that:

$$R_{\max} \geq R_{\min} \Leftrightarrow \pi + \mu \leq \frac{\overline{c}}{2} - \underline{c} + \frac{(l - L)}{2}.$$

The right term of the inequality is always positive, since we are in the region in which the health authority chooses the mild-treatment strategy, i.e., in the region in which $\overline{c}+L > 2\underline{c}+l$.

Then, provided $\overline{c} + L > 2\underline{c} + l$ there exist values of $\mu > 0$ and $\pi > 0$ for which the health authority is interested in offering an exclusive contract that is acceptable by the physician.

Bibliography

[1] Blomqvist, Å. (1991) "The Doctor as a Double Agent: Information Asymmetry, Health Insurance, and Medical Care". *Journal of Health Economics* 10, 411-432.

[2] Competition Commission (1994) "Private Medical Services: a Report on the Agreements and Practices Relating to Charges for the Supply of Private Medical Services by NHS consultants". London: HMSO

[3] Ellis, R.P. and McGuire, T.G. (1986) "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply". *Journal of Health Economics* 5, 129-152.

[4] Ellis, R.P. and McGuire, T.G. (1990) "Optimal Payment Systems for Health Services". *Journal of Health Economics* 9, 375-396.

[5] Gal-Or, E. (1999) "Optimal Reimbursement and Malpractice Sharing Rules in Health Care Markets". *Journal of Regulatory Economics* 16, 237-265.

[6] Garcia-Mariñoso, B. and Jelovac, I. (2002) "GPs' Payment Contracts and their Referral Practice". Manuscript UAB.

[7] Gaynor, M. (1994) "Issues in the Industrial Organization of the Market for Physician Services". *Journal of Economics and Management Strategy* 3-1, 211-255.

[8] González, P. (2002) "Policy Implications of Transferring Patients to Private Practice". IVIE Working Papers WP-AD2002-12.

[9] Holmstrom, B. and Milgrom, P. (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design". *The Journal of Law, Economics and Organization* 7, 24-52.

[10] Jelovac, I. (2001) "Physicians' Payment Contracts, Treatment Decisions and Diagnosis Accuracy". *Health Economics* 10, 9-25.

[11] Johnson, N. (1995) *Private Markets in Health and Welfare: An International Perspective.* Berg Publishers Ltd., Oxford.

[12] Rickman, N. and McGuire, A. (1999) "Regulating Providers' Reimbursement in a Mixed Market for Health Care". *Scottish Journal of Political Economy* 46-1, 53-71.

[13] US Congress (1995) "Hospital Financing in Seven Different Countries". Office of Technology Assessment. Background Paper OTA-BP-H-148. Washington DC, US Government Printing Office.

[14] Yates, J. (1995) "Private Eye, Heart and Hip: Surgical Consultants, the National Health Service and Private Medicine". London: Churchill Livingstone.