Waiting time distributions and national targets for elective surgery in UK: theoretical modelling and duration analysis

Sofia Dimakou

City University

Department of Economics

A Dissertation submitted for

the degree of Doctor of Philosophy

at City University

July 2013

Contents

List of Tables

List of Figures

Acknowledgments

There is a number of people I would like to thank both for their contribution to my research but also for their personal support. First of all, I would like to thank my supervisor Mireia Jofre-Bonne for her constant encouragement and advice. I appreciate deeply her comments, suggestions and help. I am also indebted to David Parkin and Nancy Devlin with whom we started together this research project. It has been a privilege to be under their guidance and companionship. Many thanks to John Appleby for all the interesting conversations we had on HES data and the useful advise he provided. Financial support from City University is gratefully acknowledged.

I am also grateful to all the friends and colleagues at the Department of Economics and back home for sharing fruitful discussions and bearing my anxieties and frustrations in stressful times, especially to Professor Glycopantis for the motivation he has always inspired and Zia Sandique and Federica Maiorano for their continuous encouragement and true friendship. A big thanks to Henrique Basso for his intuition, insight and valuable help.

Lastly, I am deeply grateful to my parents, George and Angeliki, and to my siblings, Ourania and Michael, for their endless encouragement, motivation, help and continuous love throughout this journey.

Declaration

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without any further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgment.

Abstract

Waiting times for elective surgery constitute a key performance indicator for the NHS. The principal policy response has been to introduce maximum waiting time targets against which performance is measured and rewarded. The aim of this thesis is to shed light on the mechanism of patients' admittance for elective surgery in UK by examining the whole distribution of their waiting times from an empirical and theoretical perspective.

In Chapter 2, we empirically investigate the effect of government targets on the distribution of patients' waiting times by applying duration analysis techniques to waiting time data from 2001/02 and 2002/03 for three specialties: general surgery, trauma & orthopaedics and ophthalmology. In Chapter 3, we examine further the variation in the way hospitals and surgeons manage their waiting lists by exposing detailed patterns regarding the shape of the survival and hazard curves of patients' waits. We use an expanded dataset (1997/98 to 2005/06) both in a cross-sectional and across time framework controlling for factors such as size, type and performance rating for hospitals and activity level for doctors. We also address the issue of the evolution of waiting time distributions over time.

Chapter 4 provides a theoretical supply model on how a hospital manages its stock of patients given its objective function and the constraints it is faced with. We derive the optimal waiting time distribution and identify important factors that could explain the differences between the observed empirical patterns.

Abbreviations

AFT	Accelerated failure time
CABG	Coronary artery bypass grafting
cdf	Cumulative distribution function
CF	Cumulative function
CT	Computerised tomography
GP	General practitioner
HES	Hospital episodes statistics
KM	Kaplan-Meier
MRI	Magnetic resonance imaging
NHS	National Health Service
OECD	Organisation for Economic Co-operation and Development
pdf	Probability density function
PF	Probability function
PH	Proportional hazard
PTCA	Percutaneous transluminal coronary angioplasty
SUS	Secondary Uses Services
UK	United Kingdom

CHAPTER 1

Introduction

A common characteristic of all health care systems is that people often have to wait in person to see a physician or to receive treatment. They devote their personal time to forming a queue -time that could be spent elsewhere- which will grow until the cost in time due to waiting equals the value of the good or service received by the marginal individual (Cullis, Jones and Propper, 2000). However, in some health care systems, demanders of non-emergency hospital care are allocated by specialised physicians to explicit waiting lists. Queuing in such circumstances does not impose a cost in wasted time to the occupants of the list as they queue in absentia (Lindsay and Feigenbaum, 1984). These waiting lists constitute a feature of tax-financed systems, where coverage is universal and consumers face zero price at the point of demand of health care.

Various reasons have been suggested to account for the existence of waiting lists. Firstly, they serve as a means of prioritisation of patients on behalf of physicians, usually based on the clinical urgency of medical conditions. Secondly, they facilitate the scheduling of available resources by using theater ses-

sions at most efficient and flexible way, for example, by combining long complex procedures with quicker routine ones (Appleby *et al.*, 2005b). Moreover, waiting time might correspond to a time period in which some clinical conditions might improve and thus operations prevented. Hence, their existence enables clinical purposes based on the argument that every surgical procedure comes with some kind of risk (Mullen, 1992). Consultants might have various incentives to keep their lists, such as encouraging implicitly or explicitly their patients to seek private care when the former work both at the National Health Service (NHS) and for the private sector (Hamblin, Harrison and Boyle, 1998). It is evident that patients who prefer to be treated by particular physicians would have a preference to follow them in private sector instead of remaining in the lengthy waiting lists of NHS (Yates, 1987).

Most importantly, waiting lists serve as rationing devices in health care systems where price is zero at the point of demand (Barzel, 1974). The number of patients waiting at any point in time is being determined by the rate at which people leave the list, by being admitted for surgery, self-deferring; being removed due to clinical reasons or dying, relative to the rate at which people join the list, by the decision by a consultant to admit them. Various determinants can influence the entry to or the exit from the waiting list. At the same time, Gravelle, Smith and Xavier (2003a) have demonstrated that waiting times and waiting list sizes act as signals that have an impact upon both supply and demand.

Long waiting lists and extensive waiting times are observed for elective surgery -that is, routine, non-emergency clinical procedures- such as hip or knee replacement, cataract surgery, general surgery, cholecystectomy, prostatectomy, vaginal hysterectomy, varicose vein surgery, inguinal hernia repair, coronary artery bypass grafting (CABG) and percutaneous translunimal coronary angioplasty (PTCA). On the other hand, emergency care is not rationed

at all and is directly supplied by the hospital accident and emergency departments. Moreover, even within the group of different types of elective surgery it has been reported that the waiting times for less urgent procedures such as hip replacement and cataract surgery are systematically higher than the waiting times for more urgent procedures such as CABG and PTCA (Siciliani and Hurst, 2003; MacCormick and Parry, 2003).

The existence of long waiting lists and high waiting times has been an issue in many countries. In the United Kingdom (UK) it has persisted since the launch of NHS in 1948 and remains as one of the most significant concerns of the English health care system. It is also present in the Netherlands (Brouwer *et al.*, 2003) and in Sweden (Hanning, 1996), where long waiting lists have for many years remained a serious quality problem on the health policy agenda. One study attempting to compare the extent of the problem among 20 countries of the Organisation for Economic Co-operation and Development (OECD) and investigating the possible causes for the variation in waiting times revealed that countries with the highest waiting times were the UK and Finland followed by Denmark, Norway, Australia and Spain (Siciliani and Hurst, 2003). As a result, shortening waiting lists and reducing waiting times represent a significant health policy concern leading governmental bodies to set targets and develop other strategies for the amelioration of the problem.

Focusing on the UK, in 1948, NHS inherited a waiting list of around 500,000 patients but ever since and up to the late 1990s, the number of patients waiting has been growing rapidly. More importantly, the actual waits that patients faced until treatment have been also quite extensive. The highest peak was reached in 1998 with 1.3 million patients awaiting admission for hospital treatment half of whom had to wait at least 6 months and 6.5% of whom at least a year. In addition, more than 450,000 people had to wait more than 3 months

for an outpatient appointment[1].

Long waiting lists and the time spent on these have been a persistent source of health policy and political concern in the UK. The existence, complexity and persistence of the problem has been stimulating the interest of patients, physicians of primary and secondary care, managers of trusts, politicians and policy makers. The public's main concern lies in the speed with which the queue moves rather than the number of people waiting in front of them. The more they have to wait the more anxious they become while their health status could gradually deteriorate affecting both their personal and social life. General practitioners and specialised consultants act according to their own perception of best clinical practice prioritising patients according to medical urgency; yet, at the same time, waiting lists do reflect upon their professional prestige. Managerial staff aim at increasing the performance of the organisations they work for, focusing on efficiency criteria and at the same time trying to abide by national standards. Policy makers develop measures and initiatives to serve the basic principles embodied in the NHS; universal coverage, free access to health care, high quality of services delivered, efficiency and equity. Thus, considerable resources and effort have been directed to reducing waiting lists and waiting times in UK as they undoubtedly represent a key indicator of NHS's overall performance.

The purpose of this thesis is to shed light on the mechanism of admitting patients for elective surgery in the NHS by examining the distribution of their waiting times. We depart from other studies that focus either on mean waiting times or proportions of patients waiting more than a predetermined period (e.g. 6 months) by analysing the whole spectrum of patients' waits, that is the entire waiting time distribution. Moving one step further, an endeavour to identify the factors that influence these waiting time distributions is made. In this respect, the effect of the national target regime on elective waits plays a significant role.

[1]Historical time series available on http://www.performance.doh.gov.uk/waitingtimes/index.htm.

The analysis is motivated by the extensive concerns on long lists and waiting times in the NHS since its creation. In fact, understanding the factors leading to particular waiting time distributions could enable physicians, health care managers and policy makers handle them in a more efficient way. In consequence, patients' dissatisfaction of the health care system's tardiness could be weakened.

The aim of this thesis is three-fold. In the first part, we empirically investigate the effect of maximum targets on waiting times for elective surgery in the NHS. Compared to the existing literature on the effect of targets on NHS performance, the new element in this work is the application of duration analysis, which enables the study of the whole waiting time distribution. The use of this approach is motivated by the fact that waiting times reflect a spell, with a well-defined start (decision to join a list) and finish point (hospital admission for treatment). Based on the insights provided by the first part, we further study the level of variation of patients' waiting time distributions by highlighting their management by suppliers. In the second part, the focus on a much more in-depth analysis of the diversity of waiting distributions succeeds in exposing hospital tactics and physicians' behaviours with respect to their patients' waits. This is also an empirical work that exploits the advantages of the methodology of duration analysis, yet in a less-aggregated and across time aspect that results in depicting distinct shapes and patterns of the survival and hazard functions of waiting times. However, the need to understand and provide a profound explanation of the patterns of distributions observed is imperative. This is where the third part of the study comes; to accomplish this, we move on developing a theoretical supply model of how a hospital manages its stock of patients for elective surgery. In particular, it is a utility maximisation problem that derives the optimal behaviour of the hospital with respect to the waiting time distribution of the patients it treats, given the constraints it faces.

In that respect, this thesis contributes to the strand of literature that engages with theoretical models of waiting times.

The introductory chapter is organised as follows. The next section describes the elective health care system in UK. Then the various policy initiatives to cut down the long waits are presented. In Section 1.3, we depict the existing literature on waiting times, both theoretical and empirical. The last section demonstrates the rationale and aims of the current study and provides a presentation of the next chapters.

1.1 The main characteristics of the system of elective surgery in UK

The British NHS is a central government agency and constitutes the principal provider of health care in the UK. It is publicly funded and patients are not charged directly for the services they receive. Almost all citizens are registered within a NHS general practitioner (GP) who acts as a gatekeeper between primary and secondary health care (Gérvas *et al.*, 1994; Glied, 2000; Forrest, 2003). If patients experience any ill-health symptoms, they schedule an appointment with their family GP who, after performing a series of diagnostic tests and if it is found to be necessary, refers them to a hospital specialist. The specialist, if necessary, orders further tests and after evaluating the patients' condition, decides whether or not they need to be admitted as inpatients. If this is the case, patients, depending on how severe their condition is, are either admitted immediately to receive treatment or are allocated to a waiting list for future elective surgery.

It is worth emphasising the different levels of waiting elapsed between the first clinical symptoms one faces and his/her admission for a surgical procedure (Hamblin, Harrison and Boyle, 1998). As illustrated in Figure 1.1, the first

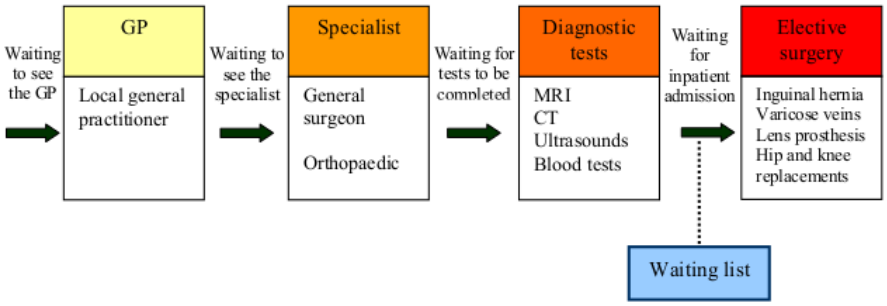


Figure 1.1: Levels of waiting times within the NHS.

level incorporates the wait to see a GP, which is the first representative of the health care sector the patient gets in contact with. If the patient's medical status requires further investigation he is referred to a specialist; therefore he has to wait a further time period for a first appointment and subsequent evaluation by the doctor. This corresponds to the second level of waiting, which is surely longer than the first one. It is now the specialist's turn to perform diagnostic tests and appropriate exams in order to deliver the correct diagnosis; these might consist of blood tests, ultrasounds, x-rays, computed tomographies (CTs) and Magnetic Resonance Imaging (MRIs). Thus, another level of wait within the system is the one that includes the time for all relevant tests to be completed. Finally, the last step in this 'waiting ladder' comprises of the waiting time between joining the waiting list of a surgeon and admitted to hospital for the actual operation. This thesis aims at investigating this last wait that will be referred to as 'waiting time for elective surgery'. Although there are different levels of waiting throughout the whole system, they are definitely not independent to each other; waiting lists management affects the outpatient follow-ups of operated patients as well as the outpatient inflow[2].

Historically, until April 1991, the GPs had no incentives to restrain refer-

[2]For an analytical review of the levels of waiting for an elective surgery see Appleby *et al.* (2005a), page 21 and Harrison and Appleby (2005), page 7.

rals for elective procedures as the cost of those was covered by geographically defined entities, the Health Authorities. In 1991, as part of NHS reforms[3], GPs were given the option of becoming fundholders; they could hold a fixed budget to cover the cost of a range of elective procedures for their patients. 'District Health Authorities' represented the second type of purchaser in the internal NHS market and were responsible for buying health care services for the population of specific geographical areas. Under the same reforms, hospitals were renamed as NHS trusts and, although remaining public sector bodies, had to compete with each other in order to negotiate contracts with the two purchaser authorities mentioned above.

This reform created a direct incentive to contain referrals by GPs as they were allowed to retain any budgetary surplus and use it to improve their patients' care in other ways. However, there is some controversy with regards to the net impact of the above mentioned policy shift. Le Grand, Mays and Mulligan (1998) suggested that such restraint did not occur, while others showed that fundholders had fewer patients admitted (Gravelle, Dusheiko and Sutton, 2002). Moreover, Croxson, Propper and Perkins (2001) suggested that fundholding gaming took place during the period of the 1991 reforms of NHS. They argued that fundholders had an incentive to increase their referrals in the year before they became fundholders and reduced them just after.

Fundholding was abolished in 1999 and was replaced by new organisations, the Primary Care Groups (PCGs), which were later to become separate legal entities as Primary Care Trusts (PCTs). All practices had to join their local PCTs, which were responsible for the health care of their population by holding a budget to cover all types of NHS expenditure[4]. Lastly, a new type of

[3]For information on the NHS internal market see Appleby *et al.* (1990), Le Grand (1991); Propper (1995); Le Grand, Mays and Mulligan (1998); Propper, Wilson and Söderlund (1998); Propper and Söderlund (1998); Propper, Burgess and Green (2004) and Propper, Burgess and Gossage (2007a).

[4]Department of Health (1997a,b).

NHS trust, called foundation trust, was introduced in 2004 allowing for greater financial and operational autonomy on behalf of the hospital. These trusts remained within the NHS's performance inspection system and were part of the government's goal to de-centralise the public services[5].

1.2 Policy initiatives to tackle waiting lists and waiting times for elective surgery in UK

Long waiting lists and excessive waiting times have stimulated the interest of policy makers since the launch of NHS. A number of different initiatives have been developed through the years to deal with this persistent problem. These policy measures can be directed to influence the demand or supply of elective care and can be of regulatory, financial, managerial or informational flavor.

Some examples of such policies over the years have been the following: increases in funding, increases or better use of existing capacity (e.g. number of doctors, nurses, beds), increases of operating and technical support (e.g. activities of the National Access Patient Team, Waiting List Action Team, Task Forces), set up of treatment centres and enhancement of day surgeries, introduction of specialty programmes (e.g. targeting specialties with very long waits such as ophthalmology and orthopaedics), greater involvement of the private sector, increases in booking admissions, guidance and management on prioritisation of patients and referrals, switching activity away from hospitals (e.g. support of GPs with special interests and expansion of nursing roles in community settings), increases of the initiatives of patients and staff (e.g. 'patient choice' and 'payment by results'), initiatives to reduce the number of people on the lists, monitoring of waiting lists and publication of routine statistics and

[5]More information can be found at http://webarchive.nationalarchives.gov.uk/+/dh.gov.uk/en/healthcare/secondarycare/nhsfoundationtrust/index.htm.

performance measures either on the system overall (star-rating system) or on specific fields (national targets for outpatients and inpatients)[6].

Before reviewing in detail the use of performance measures and especially the introduction of targets on inpatient waits, we will briefly pinpoint a couple of issues. Interestingly, informing patients about the time they are going to spend on lists is important as it eliminates the distress and the anxiety they feel due to the uncertainty concerning the time they will receive treatment (Propper, 1990). The main role of the National Booked Admissions Programme, which was firstly launched in 1998 as a pilot and was expanded later, was to give patients the opportunity to be made aware of the date of their admission[7]. Although, previously, most patients in need for an elective procedure were added to a waiting list. However, the use of the booking system revealed management difficulties for hospitals that had to deal with variation in emergency demand.

An additional initiative that strengthened the patients' rights, gave them the opportunity to be able to choose another hospital if they faced long waiting times. Initially, this was set up as a strategy to reduce the number of people waiting more than 6 months for surgery, yet later on, the view that patients should select hospitals and not hospitals patients was widely established. At the beginning, choice was introduced on a pilot basis for heart patients but from December 2005 all patients requiring elective surgery were offered the choice between 4-5 hospitals at the point where their GP decides to refer them to a specialist[8,9]. At the same time, a new system called 'Payment by Results'

[6]More information on the policy initiatives can be found in Harrison and Appleby (2005). The authors further divide the policies adopted by the government into three categories: phase 1 (1997-2000), phase 2 (2000-2004) and phase 3 (2005-2008 and beyond).

[7]Comparisons of different admission methods will be discussed in Chapter 2.

[8]Building on the best: Choice, responsiveness and equity in the NHS - Department of Health (2003), Choose and book: Patient's choice of hospital and booked appointment - Department of Health (2004), Choice at six months. Good practice - Department of Health (2005).

[9]More information on the evaluation of the London Patient Choice Project was published by Burge *et al.* (2005).

was introduced to allow for the fact that money had to follow the patient to the chosen hospital. This programme aimed at linking a hospital's income to the amount of work it performed, while at the same time it created incentives for the hospital to improve performance in order to attract more patients and reduce costs so as to survive financially.

Performance measures

Performance measures have been commonly used in the public sector in order to improve accountability, increase productivity and reduce consumers dissatisfaction of the services provided. This process involves a series of actions starting with the development of well-defined and observable performance indicators followed by their publication and subsequent monitoring (Bird *et al.*, 2005). Performance management could comprise of various types of measures that could be less or more explicitly introduced. Yet, they are strongly linked to either monetary or non-monetary rewards. Besides the field of education, the health care service constitutes one of the commonest public sectors applying aggressive targets in their fight against under-performance.

Some basic information regarding the performance of hospitals in England and Wales has been available since the early 1980s. The 'Health Services Management Centre' of the University of Birmingham was the first to use performance indicators based on routinely collected data. Following this attempt, the government initiated a programme within NHS (it came as a series of grey books) by publishing 123 performance indicators for the local health authorities (Smith, 1990). Since 1999 there has been a large increase in the number of published data on performance indicators of health care providers, with the introduction of waiting times targets among the most important ones.

Besides the beneficial effect of performance indicators (publication of data, systematic monitoring of agents, increased productivity, increase of the amount

of effort towards targeted tasks, decrease of patient disutility and dissatisfaction), a few problems have appeared in their use in the public sector; it is evident that targets can produce perverse effects as well. Among the most important is misinterpreting them due to complexity and inconsistency as well as having the incentive to distort behaviour by the employment of manipulation, gaming and inappropriate responses. Another significant worry is that well monitored aspects could receive a higher priority increasing the peril to shift attention away from unmonitored, less concrete and poorly measured fields such as quality of services (Propper and Wilson, 2003).

According to Smith (1993), performance measurement could encourage various undesirable outcomes such tunnel vision (shift away from unmonitored but important tasks so as to concentrate to targeted features only), sub-optimisation (managers pursuit their personal objectives instead of the ones set up centrically), myopia (avoidance to consider long term attributes and concentration on short term issues), convergence (organisations respond to performance management by avoiding extreme performance behaviour resulting in a convergence of all responses), ossification (behaviours are characterised by a strong attachment to conventional patterns denying any innovative methodology), gaming (distort behaviour so as to demonstrate achievement of the performance standards) and misrepresentation (commitment of fraud).

These problems tend to appear quite frequently in multi-product public organisations that have multiple principals and face multiple incentives. For example, take under consideration the different number of actors that are involved in such an organisation: the users of the service, the payers, the providers, the managers of the providers and the politicians of various levels of the government. It not difficult to comprehend the extensive range of goals and incentives these agents have[10].

[10]For a more detailed discussion of these incentives look at Burgess and Ratto (2003) and

A characteristic example of such an organisation is the English health care sector. The introduction of explicit national performance targets took place with the publication of the NHS Plan in 2000[11]. The proposed strategy consisted of maximum waiting times for elective surgery that health care providers had to meet, with rewards and penalties for successful and unsuccessful performance. Rewards included greater autonomy for hospitals that performed well, promotions of managers and positive advertising of the institution due to the publication of waiting times. Penalties comprised of threats, demotions and dismissals of managers in hospitals that performed poorly and 'shaming' of the trust through the release of its poor performance. In particular, nobody had to wait more than 18 months by the end of March 2000, 15 months by March 2002, 12 months by March 2003, 9 months by March 2004 and 6 months by December 2005. Similar targets that focused on specific fields were introduced for outpatients, cancer patients, patients attending 'Accident & Emergency' and patients waiting to book an appointment with their GP[12].

Furthermore, an annual 'star rating' system consisting both of a small number of 'key targets' and a series of indicators in a 'balanced scorecard' was applied for acute hospitals from 2001 to 2005[13]. Scores ranged from 0 (unsuccessful trusts) to 3 (successful trusts) stars based on the trusts' performance according to the criteria mentioned above. It is worth mentioning that six out of the nine key targets were set up for waiting times with the other three being financial balance, hospital cleanliness and improvement of the working lives of staff.

In 2004, the government introduced a new maximum waiting time target

Besley and Ghatak (2003).

[11]The NHS plan: A plan for investment, a plan for reform, Department of Health (2000).

[12]A list of all the different waiting list targets announced since 1997 can be found at Appleby and Coote (2002), page 26.

[13]Department of Health (2001, 2002), Commission for Health Improvement(2003), Healthcare Commission (2004, 2005).

of 18 weeks from the initial referral of a GP to a specialist until admission for surgery. From the patient's perspective, this new waiting time target was quite desirable as it attempted to reveal hidden delays not previously measured that indeed lengthened their actual waiting for treatment. Indeed, the government moved into setting a target that takes into account the complete patients' journey within the NHS; in other words it considers the total time patients wait from GP referral to surgery, including any delays in waiting for diagnostic tests and receiving the results. This time period is now known as 'Referral To Treatment' (RTT)[14].

After describing the NHS system, its evolution and the various policy initiatives to tackle waiting times, we turn to explore the theoretical and empirical literature on healthcare provision.

1.3 Related Literature

1.3.1 Theoretical models

The existing literature on waiting times has been extensive[15]. Economic models of waiting lists and waiting times have predominantly been developed in the context of demand and supply models of health care markets. We first review a series of basic demand models and then move towards the presentation of frameworks that focus on the supply side of health care market, which is also the focus of this thesis.

Lindsay and Feigenbaum (1984) develop a demand model of waiting list queues in order to demonstrate how the healthcare market reaches an equilibrium and how changes in the waiting time, the service rate and other variables affect this equilibrium. They investigate the decisions of the marginal patient in

[14]Information on RTT can be found at http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/ReferraltoTreatmentstatistics/DH_089757.

[15]See Cullis *et al.* (2000) for a survey.

a steady state in which the numbers of people joining the list are equal to those leaving the list. They consider two main assumptions; first, individual demand is assumed to be unpredictable. This assumption provides a rationale for the existence of queues and for the waiting lists to be clearing the market. Second, delay in the receipt of the good lowers the value to the demander. Assuming that the position in the queue is linked to the date of delivery, the value and the number of people joining the list will be influenced by the length of the list. Hence, it is the diminishing value placed on obtaining such a good that results in demand meeting supply. Their main theoretical finding suggests that increased service rate (capacity) is not necessarily resulting in reduced waiting lists. The authors do not model the supply of elective care but just assume that it is affected by a vector of unknown factors and the waiting time.

Extending on the framework of Lindsay and Feigenbaum (1984), a stream of theoretical contributions has developed on similar lines in the literature. These include Cullis and Jones (1986), Gravelle, Dusheiko and Sutton (2002), Gravelle, Smith and Xavier (2003a,b) and Martin and Smith (2003). Some authors take under consideration the patient's choice between the public and the private sector (Goddard, Malek and Tavakoli, 1995; Martin and Smith, 1999) and find that long waiting times in the public sector may encourage patients to seek private surgery. Besley, Hall and Preston (1998, 1999) show that longer waiting lists in the NHS are associated with greater demand and purchase of private health insurance or private surgery out-of pocket.

In the majority of the articles mentioned above, waiting time is modelled as a static phenomenon where perfect clearance in the health care market is assumed. Thus, waiting times are known and deterministic. Among the first to introduce a dynamic element in the analysis is Worthington (1987, 1991) who applies queuing theory to hospitals waiting lists. A further step towards dynamic modelling is achieved by Gravelle, Smith and Xavier (2003a), who

employ the methodology of dynamic optimisation in discrete time. They model the utility function of the hospital manager to include not only current but lagged waiting time indicators. They do not consider an equilibrium between demand and supply, yet, for comparative-statics purposes, the authors assume that providers care only about the effect of current actions on current utility.

Furthermore, one might be interested on the path of reaching a equilibrium and more importantly on what happens in disequilibrium conditions. The formulation of a comprehensive dynamic model is achieved by Siciliani (2006) who uses optimal control theory to model hospital incentives within a continuous time dynamic framework[16].

With respect to the stand of literature that investigates the supply side of elective care one of the first basic models was proposed by Iversen (1993). Inversen focuses on the elective system of the National Health Service in Norway and after constructing the production function of health services he models the long-run equilibrium of a non-cooperative game between the hospital and its sponsor, the government. He discovers that under this type of game, where the hospital decides first and the government follows, excessive waiting times exist under a stackelberg equilibrium. On the other hand, in a Nash equilibrium where the two players act simultaneously, no excessive waits appear. In another work, Iversen (1997) investigates the effect of the private sector on the waiting times of NHS patients. The results suggest that concerns about minimising the costs of public health care can also lead to maintaining longer waiting lists, which could induce a shift of patients to the private sector.

Farnworth (2003) builds on Iversen's work and develops a theoretical model of how interactions among hospitals that charge different prices for health services can determine and affect the equilibrium (average) waiting time. His

[16]This work is actually modeling the supply of elective services. Based on how expected waiting time is perceived, the optimal path towards the (steady state) equilibrium is depicted. The results are also extended to incorporate stochastic demand.

findings suggest that, under specific circumstances, an increase in the price charged to one hospital can lower the waiting time for all.

Another three contributions that examine whether waiting times are welfare improving under specific behaviours or decisions of patients and doctors are those of Olivella (2002), Hoel and Saether (2003) and Barros and Olivella (2005). The first shows that the movement of patients from the public to the private sector due to long waits on the former is welfare improving when the public sector operates with over capacity and the private sector with under capacity. The analysis is conducted within a framework in which the list is prioritised by the level of severity of patient's health status. The second work reaches similar conclusions within a different structure. They also analyse welfare considerations when there is a concern for equity and examine the optimal tax/subsidy for private health care.

The work of Barros and Olivella (2005) focuses more on the strategic behaviour of doctors in selecting the milder cases to be treated in the private sector (cream skimming). This study relates to the literature on prioritisation of a list within the same specialty[17] and addresses the rationing of public treatment, that is, only cases that meet particular criteria are admitted. Given the admissions requirement (rationing policy) and the co-existence, in the same doctor, of private and public practice, there is scope for patient selection. The find that full cream skimming[18] takes place only with intermediate rationing policies and partial cream skimming with very stringent (or lax) rationing policies. Additional studies on cream skimming of low severity patients are those of Ellis (1998) and González (2005). Moreover, Xavier (2003), Siciliani (2005) and Brekke *et al.* (2008) examine further the effect of competition on waiting times. Unlike the other studies on competition, in which it is assumed that the

[17]The work on prioritisation is mainly concerned with across specialties differences. See Cullis *et al.* (2000) for more details.

[18]That is, all milder cases are treated in the private sector.

hospitals are local monopolists, the above papers model competition within a Hotelling or a duopoly framework.

The last chapter of this thesis develops a model that also focuses on the supply side of healthcare provision. It does not explicitly consider competition with other hospitals, nor the strategic interaction with the government. The aim of our model is the derivation of the whole waiting time distribution of a given list, and the exploration of relevant supply side factors and their effect on the optimal (steady state) admittance pattern. In that sense, the work of Dixon and Siciliani (2009) is similar. Motivated by the drive to compare and link the two main sources of NHS data on individual waiting times, they construct two measures (i) the waiting time distribution of patients on a list (at a census date) and (ii) the distribution of waiting times of patients treated. They show under different conditions what those two measures capture and how they are linked[19]. In Chapter 4 we develop the waiting time distribution of measure (ii) within a supply model.

1.3.2 Empirical models

In the UK, economic models of supply and demand have also been tested empirically (Lindsay and Feigenbaum, 1984; Martin and Smith, 1999, 2003; Gravelle *et al.*, 2003a; Martin *et al.*, 2007). These studies use cross-sectional or panel data to investigate the responsiveness of demand for or supply of health services to waiting times.

In particular, Martin and Smith (1999) use HES data for 1990/1991 and find a small elasticity of demand with respect to waiting time (-0.20). They further conclude that increased resources may reduce waiting times without greatly stimulating utilization. Yet, they use aggregate data of a small area

[19]However, they do not consider a model.

(ward) level in England and their model equations are identified by exclusion restrictions (Windmeijer *et al.*, 2005). Gravelle *et al.* (2002) propose a model of the admission process for cataract surgery for a three year period in an English Health Authority and conclude that admission rates are negatively related to waiting times and distance to hospital (elasticity was equal to -0.25). Martin and Smith (2003) use small area panel data to estimate supply and demand functions for seven specialties and all specialties combined, finding that the demand elasticity varies between specialties, but is always small. Gravelle, Smith and Xavier (2003a) develop a dynamic demand and supply model, again tested using panel data, and find that supply increased and demand decreased (estimated elasticities for two different models are -0.30 and -0.21) in response to measures of the previous period waiting time. The authors again use aggregate hospital utilization data at health authority level for 24 quarters during the years 1987-1993.

Siciliani, Stanciole and Jacobs (2009) use data for 137 hospitals during the period 1998-2002 and estimate the elasticity of hospital's costs with respect to waiting times (in both cross-sectional and panel data). The results indicate (whenever significant) inelastic costs and a U-shaped relationship between costs and waiting times.

Empirical evidence of the effect of waiting time targets in the UK will be analysed in detail in Chapter 2. Indicatively, it is mainly conducted in a before and after framework, as for example in Harrison and Appleby (2005), Hauck and Street (2007), Propper *et al.* (2007b, 2008a). Propper *et al.* (2010) introduce a measure of 'target pressure' and examine the kernel densities of waits before and after the policy initiative.

The use of duration analysis in the study of waiting times literature is scarce and sporadic. MacCormick and Parry (2003) investigate the differences in waiting times for four diagnoses of elective general surgery using a sample of

918 patients in a tertiary level hospital in New Zealand. The authors use the non-parametric techniques of duration analysis to compare the waiting times of patients for the four diagnoses. Their findings show that different waiting time thresholds exist for different diagnoses. The authors, therefore, are in favour of the application of distinct waiting times that will reflect and correspond to the natural history of each disease.

Levy *et al.* (2005) examine whether extra funding for CABG operations in British Columbia, Canada (1991-2000) had an effect on the time patients spent on waiting lists. In particular, the authors compare the waiting time until surgery for equal proportions of patients in synthetic cohorts before and after the funding became available. Their data set consists of 9,231 records whose waiting times are treated as duration times. The authors find that time to CABG shortens after the introduction of the supplementary funding –possibly due to hospital capacity utilization– and at the same time this effect is not uniform across different priority groups.

Furthermore, Sobolev *et al.* (2005) conduct a survival analysis of a cohort of 1,928 patients waiting to be admitted in the Division of Vascular Surgery at Queens University in Kingston, Canada for four different surgical procedures due to vascular disease. Duration analysis reveals different patterns of waiting time distributions and variations among different procedures are compared. Mainly, they observe shorter times for more urgent groups, although in some comparisons less urgent patients had a significant chance of admission to more urgent cases. Patient-related delays in scheduling operations, availability of hospital resources, anticipated length of stay in the IC and cancellations of booked surgeries may account for this phenomenon. The concern that waiting times for elective surgery might be determined not only by how many patients are on the list and the urgency of their condition but also by waiting list management has been reported previously by the same authors (Sobolev *et al.*, 2000, 2001).

The studies mentioned above attempt to address and investigate different questions using duration analysis. Although the adopted design is not the same among them, they clearly show that duration analysis can reveal various patterns of waiting times distributions. Some of the limitations of these studies involve small sample size, observation of not statistically significant results and in some studies exclusion of censored observations. Our work employs the technique of duration analysis[20] for elective surgery in the UK. We deviate from the above mentioned studies in terms of both coverage and scope. Firstly, we provide a wide and systematic analysis of UK waiting time data both across time and at different degrees of disaggregation (hospitals, operative procedures, physicians). Secondly, we use a much larger sample. Most importantly, we utilise duration analysis, and its unique insights, to study the effect of waiting time targets.

1.4 Rationale and aims of the study

As mentioned previously, this thesis is related to the strand of literature that examines the elective health care system in UK. Interest in the elective waiting list mechanism is motivated by its complexity and the view that waiting for treatment represents a key indicator for NHS performance. The core strategy in England since the publication of the NHS Plan has been to use maximum waiting times targets every hospital should abide to.

The main purpose of this thesis is to investigate the elective admission process by taking under consideration the waiting times of all patients on the list and not just measures of central tendency such as mean or median waiting time. In consequence, we look into the whole waiting time distribution of patients, empirically, with the application of duration analysis techniques and

[20] Our empirical methodology is discussed in detail in Chapter 2.

theoretically, with the development of a supply model that derives endogenously the entire optimal waiting time distribution of patients.

The second chapter of the thesis explores the impact of government targets on the distribution of waiting times of elective patients in the NHS[21]. How does the probability of admission for any given patient vary during the time that they wait and how is it affected by the targets? What incentives do targets create to hospital managers? Do hospitals change their behaviour as a result of the targets?

In order to address these issues we estimate the survival and hazard functions of patients' waits at the level of specialty and operation of all NHS hospitals in UK. We start examining the waiting times distribution at such an aggregated level due to the universal nature of the policy initiative; maximum targets were addressed to all elective patients of every medical specialty in all trusts all over UK. We also explore differences by type of admission (waiting list, booked, planned surgeries) and by a small set of seven trusts geographically spread across England. In order to identify changes of responses to targets we utilise data for two years, 2001/2002 where the maximum target was set for 15 months and 2002/2003 where the target was trimmed down to 12 months. We focus on three elective specialties; general surgery, trauma and orthopaedics and ophthalmology.

The third chapter of the thesis aims at examining further the variation in the way hospitals and physicians manage their waiting lists. The rationale of this piece of work lies in exposing detailed *patterns* regarding the shape of survival and hazard curves of patients' waits. It is an informative piece of investigation as we learn more about the hospitals' and doctors' behaviour on waiting list

[21]The second chapter of the thesis is derived from Dimakou *et al.* (2009). Regarding the division of responsibilities among the co-authors, I was responsible for all the analyses undertaken while Professors David Parkin and Nancy Devlin were the project supervisors. John Appleby provided useful information on HES data.

administration and patient admittance for treatment. To achieve this goal, we expand the initial empirical analysis by exploring wait data for a greater time period (financial years from 1997 to 2005) for an expanded set of hospitals at less aggregated levels. We further examine the waiting time distributions of elective patients by specific consultants.

In particular, we apply the methodology of duration analysis either in a cross-sectional or an across time framework controlling for factors such as size, type and performance rating for hospitals and activity level for doctors. We also address the issue of the evolution of waiting time distributions over time.

In the fourth chapter of the thesis, we build on our empirical findings and develop a supply model of how a hospital manages its waiting lists given its objective function and the constraints it is faced with. This chapter focuses at matching important empirical patterns of the waiting times distributions of hospitals' patients and identifying possible factors that could explain the observed distributions. Two are the distinct features of our theoretical framework (i) the dynamic element of the model and (ii) the derivation of the entire optimal waiting time distribution of patients treated at the steady state. On that basis, we also construct and compare the corresponding survival and hazard functions. We solve the model numerically and perform several comparative statics exercises.

As a whole, this thesis explores the mechanism of patients' admittance for elective surgery in UK and their associated waiting time distribution from both an empirical and a theoretical perspective. The former reveals both the great level of variability in the shape of the survival and hazard curves of patients' waiting times and their evolution over time. The latter contributes to the theoretical literature on waiting times models by deriving endogenously the waiting time distribution and provides valuable insights on the potential factors that may explain the distinct empirical behavioural patterns.

CHAPTER 2

The impact of government targets on waiting times for elective surgery in the NHS

2.1 Introduction

Long waiting lists and extensive waiting times for elective surgery have been a persistent source of health policy and political concern in the UK and other OECD countries (Siciliani and Hurst, 2003). Waiting lists function, in part, as a non-price rationing device to reconcile the differences between supply and demand that arise when coverage is universal and those demanding -patients or their agents- face zero price at the point of demand (Cullis, Jones and Propper, 2000). Waiting times and waiting list sizes act as signals that have an impact upon both the supply of and demand for health care (Gravelle, Smith and Xavier, 2003a).

Policies intended to reduce waiting lists may impact on either supply or de-

mand or both. Supply-side responses include extra funding for elective surgery, tackling supply bottlenecks, provider monitoring and management of waiting lists. Demand management includes promulgating guidelines for appropriate referral and explicit methods for prioritising patients. Historically, UK NHS policies on waiting tended to reflect a view that waiting lists were a backlog of untreated patients, which could be addressed by short-term increases in activity (Hamblin, Harrison and Boyle, 1998).

More recently, the emphasis of policy has shifted from waiting lists to waiting times, on the grounds that patients and policy makers are more concerned about the speed with which the queue moves -and thus the time spent on the list- rather than the number of people waiting.

While current NHS policy combines a number of the supply- and demand-side strategies noted above, the main strategy in England since the publication of the NHS Plan in 2000[1] has been to use waiting times targets (Harrison and Appleby, 2005). These stipulate maximum waiting times for elective surgery that health care providers should meet, with rewards and penalties for successful and unsuccessful performance. The inpatient waiting time target has been progressively reduced from 18 months in March 2000 to 15 months in March 2002, 12 months in March 2003, 9 months in March 2004 and 6 months in December 2005 (Appleby *et al.*, 2005b). The most important feature of the targets lies in their national and universal character; they apply to all elective patients of every medical specialty in all trusts all over UK. Furthermore, in 2004[2], the government presented a new target for the end of 2008, which incorporated a maximum wait of 18 weeks from initial referral of a patient by a general practitioner for an outpatient consultation and any subsequent treatment, including a hospital stay if needed. It is apparent that in recent years more attention

[1]The NHS plan: A plan for investment, a plan for reform, Department of Health (2000).

[2]The NHS Improvement Plan: Putting people at the heart of public services, Department of Health (2004).

is drawn on the complete patients journey within the NHS, that is the total waiting time of patients from GP referral to surgery.

Since 2000, there has been substantial progress in reducing waiting lists, overall waiting times and the number of patients waiting over 6 months, as Figure 2.1 demonstrates.

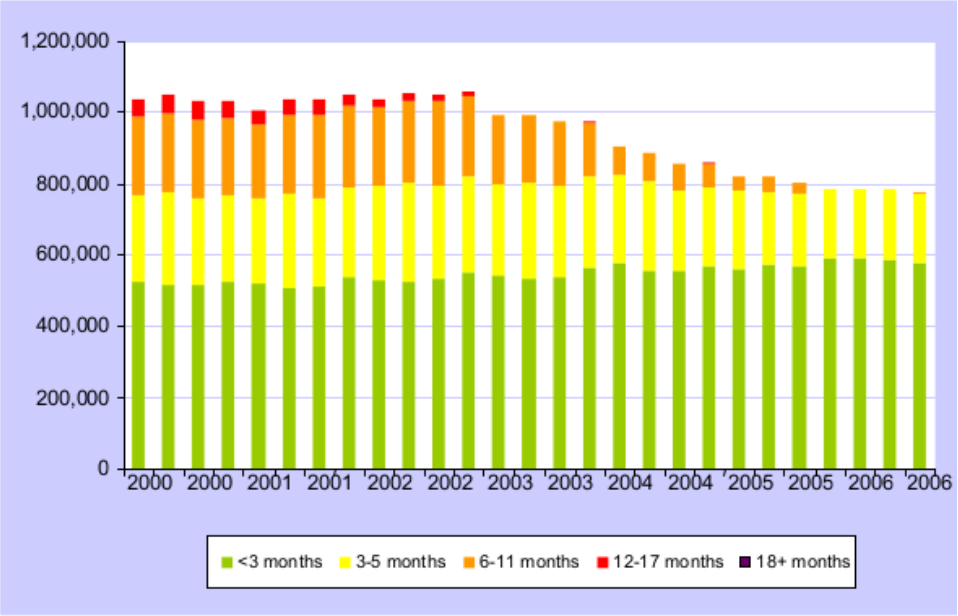


Figure 2.1: Provided-based inpatient waiting lists and times in the English NHS, Quarterly, 2000-2006. Source: King's Fund, London.

However, seven years after the introduction of the NHS plan and despite the considerable resources and effort that have been directed to reducing waiting lists and waiting times, the number waiting for inpatient treatment is around 750,000, with just over 1,000 still waiting 6 or more months, and just over a quarter still waiting over 3 months. A further one million are waiting for an outpatient appointment; and around one million are waiting for various diagnostic tests to be carried out, although the extent of the overlap between these latter figures is impossible to determine[3].

[3]http://www.performance.doh.gov.uk/waitingtimes/index.htm

There are a number of concerns relating to the use and impact of these targets, especially, as they become even more stringent. Principal among these is the extent to which targets distort clinical priorities, by changing the order, and thus speed, with which patients are treated. The National Audit Office reports that 20% of consultants surveyed in three specialties stated they had changed the order they had prioritised patients in order to meet government targets[4].

In one respect, this changed order of priorities might be considered not as an unwanted or unanticipated side effect, but rather as an intended outcome. Although there is evidence to suggest that the length of a patient's wait may have influenced clinical decisions to admit even before the introduction of targets (Harrison and Appleby, 2005), presumably the targets reflect an explicit view that whatever clinical or social factors previously determined priority for treatment did not place sufficient weight on time waited, in particular, maximum time waited. However, if providers are meeting targets by substituting less urgent cases, with less ability to benefit, for more urgent cases, with higher ability to benefit, then this would be a potential cause for concern on both economic and ethical grounds.

The challenge in analysing the number, importance and effect of these changes in admission decisions arising from the targets is that the admission criteria without targets are neither clearly specified nor consistent. Individual clinicians assess patients' conditions according to their own personal judgements of clinical urgency. There are neither gold standard admission criteria nor any systematic scoring system in widespread use in the UK to aid between-patient prioritisation. Cullis and Jones (1976) advocated such an approach over 25 years ago, and there are examples of such systems from other countries

[4]Inpatient and Outpatient waiting in the NHS. Report by the Comptroller and Auditor General., National Audit Office (2001).

(Siciliani and Hurst, 2003; Hadorn and Holmes, 1997).

More fundamentally, neither the way in which providers meet the targets nor what differentiates successful from unsuccessful hospital trusts with respect to the targets are clear. For example, targets may be met mainly by increasing surgical throughput -reducing waiting times for all patients- or by substituting higher wait for lower wait patients, or a mix of both. The targets create incentives which might be expected to affect both manager and clinician behaviour.

Evidence on the effects of the targets tends to focus on average waiting times or the total number of those waiting for specified periods; the effect on the distribution of waiting times is less well understood. Harrison and Appleby (2005) compared waiting times distributions for orthopaedic surgery before and after the introduction of targets, with differences in the distributions used to identify changes in admission patterns. The results suggested that "...any reordering of cases had less to do with substituting very short wait (presumed urgent) cases with longer wait (presumed less urgent) cases but rather that the latter displaced some (less urgent) 'filler cases' -that is, those with short operating times which could be used to make best use of available theatre time". However, these results rely on relatively crude before-and-after comparisons of waiting times distributions, so the conclusion remains somewhat speculative. The same view is also supported by Propper *et al.* (2008a) who state that shorter waiting times might have been achieved by targeting less needy patients but actually it was not evident that it happened.

Only a few other papers in the existing literature of waiting times examine the effect of targets on the elimination of long waits in the NHS and the subsequent improvement of hospitals performance. As mentioned in chapter 1, Smith (1990, 1993), Propper and Wilson (2003) and Bird *et al.* (2005) review the effects of performance management in health care and in the public sector. They all agree that such indicators can be manipulated by public organisations

(e.g. hospitals) so as to falsely declare accordance to targets. Yet, they do not provide any theoretical or empirical model to support their views.

Another attempt to reveal the effects of the different policy initiative of waiting time targets in England is made by Alvarez-Rosete *et al.* (2005) and Bevan and Hood (2006a,b). They all conclude that the target regime trimmed down the waiting times of elective patients[5,6]. However, the former study only employs descriptive statistics in a simple before-and-after analysis of a very small number of indicators, where both the quality and comparability of the data between the countries tested is questionable, while in the latter, statistical analysis of waiting times is only limited to plain illustrations of frequencies of patients waiting specific periods of time, with no consideration at all on wait distributions or more sophisticated modelling. Further, the presence and mechanism of gaming is not straightforwardly demonstrated.

Thus, the segment of research that followed (Hauck and Street, 2007; Propper *et al.*, 2007b, 2008a,b, 2010; Besley *et al.*, 2008, 2009) utilise a mixture of econometric models to estimate the effect of targets on performance in the health sector[7]. They all take advantage of the different timing and nature of

[5]Alvarez-Rosete *et al.* use performance data for England, Scotland, Wales and Northern Ireland in an attempt to compare a variety of indicators (health indicators such as mortality ratios and life expectancies, NHS expenditure per capita, availability of hospital beds, staff numbers, activity measures and waiting times) for years 1996/1997 (before the devolution) and 2002/2003 (after the devolution).

[6]The two studies by Bevan and Hood examine whether the star-rating system for NHS trusts improved performance in England. In the first, the authors, after presenting the percentages of patients waiting longer than 6 and 12 months for treatment for 1999-2005, conclude that the stricter policy of '*naming and shaming*' in England created lots of pressure to reduce waiting times for elective surgery. As a result, it ameliorated performance but at the same time gaming was evident. In the second, the authors, after providing assumptions regarding the way the government sets priorities and measures performance (idea of synecdoche) and regarding the gaming (defined as hitting the target and missing the point), they conclude that targets have operated as '*management and terror*' in order to achieve wait reductions drawing parallels with the Soviet regime.

[7]Hauck and Street analyse routine data collected over a six year period in three English and one Welsh hospital close to the English-Welsh border employing the techniques of OLS and PROBIT estimation. Propper *et al.* (2007b, 2008a) use the difference-in-difference methodology to compare performance in reducing waiting times in England and Scotland while Besley, Bevan and Burchardi (2008, 2009) undertake the same analysis for England and Wales.

the policy measures of elective treatment (natural experiment) between England and Scotland/Wales and similarly report greater reductions of extensive waiting times of elective patients in England compared to the other two.

Interestingly, Propper *et al.* (2010) introduce a measure of 'target pressure' defined as the ratio of the number of patients at the end of the previous quarter whose waiting times will exceed the end of the quarter target if left untreated divided by the total number of patients waiting at the end of the previous quarter. By illustrating the kernel densities of waits before and after the policy initiative, they suggest that the distribution of waits in England was slightly pulled leftwards at the right tail, especially for 2003/2004. However, these comparisons are not very informative. The authors conclude that targets achieved their objective and led to a significant fall of inpatient waiting times, however, not due to the gaming activities that had been advocated by some. Furthermore, among their findings of the effects of the target regime are the increase of elective admissions, the unchanged order in which patients were treated, the same percentage of urgent cases and the absence of any impairment of quality indicators. Although some evidence of waiting list manipulation was evident -an increase of suspensions and removals- it did not cause any alterations to overall patient outcomes.

The aim of this chapter is to empirically identify the impact of government targets on the distribution of waiting times in the NHS. We depart from the above-mentioned studies by exposing the whole spectrum of patients' waits and its possible alteration due to the imposition of universal targets. The relevant literature does not pay much implicit attention to the examination and evolution of waiting time distributions; due to the fact that previous studies rely on relatively crude before-and-after comparisons of average waiting times or wait distributions, the effect of targets on the distribution of waiting times remains to be investigated. Therefore, there is scope for further research so as

to enlighten the waiting list system and its response to policy initiatives.

The focus on a more in-depth analysis of the waiting time distributions is motivated by the following considerations. Firstly, the same average waiting time might correspond to totally different waiting time distributions. Although averages of waiting times might remain constant over years, a whole series of distinct hospitals' behaviours with respect of admissions rates might have taken place. Secondly, it would be helpful to acquire as much information as possible regarding the ways hospitals and physicians behave in the process of decision-making -decision to join a list, decision to admit for treatment- of patients' flow into the system.

To achieve this, we employ the techniques of duration analysis that are proven a valuable tool in dealing with variables that represent time periods; on top of this, they reveal the overall distribution of such spells. We use administrative data on three specialties -general surgery, trauma and orthopaedics and ophthalmology- for all the NHS trusts in England.

Specifically, we address the following questions: (i) How does the probability of admission for any given patient vary during the time that they wait? (ii) How is the probability of admission for any given waiting time affected by the targets? (iii) Can variations in waiting times be explained by clinical, patient, or provider-level characteristics? (iv) What implications may be drawn from our results with respect to providers managerial responses to the targets?

The structure of the remainder of the chapter is as follows. The next section outlines the features of the methodology employed and describes the data. In Section 2.3, the main findings of the empirical analysis are discussed. These include empirical estimation of the waiting time distributions at different levels of disaggregation and in relation to the introduced waiting time targets, as well as covariate analysis determining the importance of patients' and clinical characteristics on the waiting times. Section 2.4 provides concluding remarks.

2.2 Methodology: Duration analysis of waiting times data

Duration analysis is the usual label applied in econometrics to a set of techniques known in biomedical sciences as survival analysis, in other social sciences as time-to-event analysis and in engineering as failure-time analysis or reliability analysis. It consists of parametric and non-parametric methods for estimating survival and hazard functions, explained below, which permits comparison of the duration of states of interest for different groups and estimation of the impact of explanatory variables on duration (Cox and Oakes, 1984; Collett, 2003)[8].

In Economics and other disciplines interest is drawn to variables that come in the form of a specified time period or duration, which is the time elapsed until a certain event occurs[9]. A well defined origin (date of entering the state of interest) and exit (date of leaving the state of interest) define the spell length or spell duration. In Economics, these data are referred to as *duration data*.

In the context of this study, we are interested in the duration of waiting, which is a state initiated by referral to join a list and terminated by admission for treatment. The duration of waiting is treated as a continuous variable in our empirical analysis. The following graph depicts the process of creating duration spells of various lengths (waiting times) for all individuals admitted for surgery within 2002/2003.

[8]For more information on survival analysis techniques see Miller, Gong and Muñoz (1981); Allison (1984); Kalbfleisch and Prentice (1980); Fleming and Harrington (1991); Elandt and Johnson (1999) and Hosmer, Lemeshow and May (2011).

[9]Examples of applications from different disciplines that use duration analysis techniques are shown in Table 2.8 at the end of this Chapter.

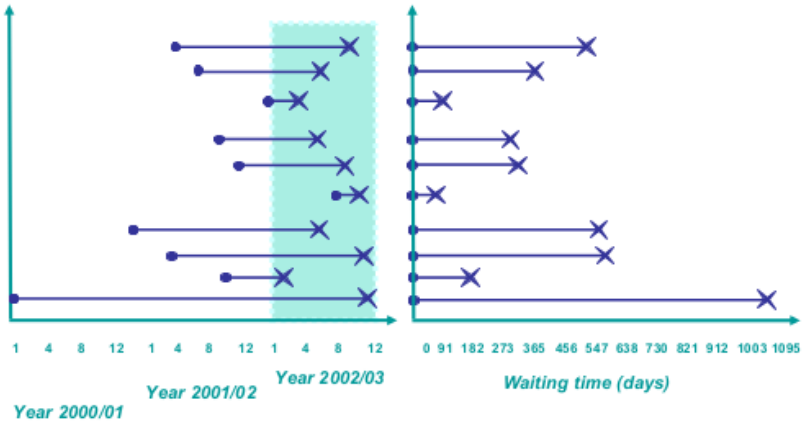


Figure 2.2: Wait spells of patients admitted for surgery during 2002/2003.

Duration analysis offers several advantages in analysing waiting time data. First, as mentioned before, it looks at the distribution of waiting times, which generates deeper insights than methods that focus on average waiting times. Secondly, it takes account of the fact that waiting times are not usually normally distributed. Thirdly, it allows for censored observations, where either the date of referral or of admittance is not known. However, in this study the data set does not include censored observations; this will be explained in the next section of the chapter.

2.2.1 Survival and Hazard functions

Two key concepts in duration analysis are survival functions and hazard functions. Let t be the length of a completed spell. It is the realisation of a continuous non-negative random variable T with a probability density function (pdf), $f(t)$, and a cumulative distribution function (cdf), $F(t)$. In the survival analysis literature, $f(t)$ is also known as the "failure probability function" and $F(t)$ as the "failure function".

The failure function (cdf) represents the probability that duration time is

less or equal than some value t.

$$F(t) = P(T \leq t), \quad t \geq 0. \tag{2.1}$$

This implies that the survival function, $S(t)$, which is the complement of $F(t)$, is equal to:

$$S(t) = P(T > t) = 1 - F(t), \quad t \geq 0. \tag{2.2}$$

It represents the probability that an individual survives from the time origin to some time beyond t. The mean survival time is the area under the survival curve:

$$E(t) = \int_0^\infty S(t)dt.$$

Both $F(t)$ and $S(t)$ are probabilities and therefore inherit the properties of probabilities. In particular, the survival function lies between zero and one; it is equal to one at the start of the spell ($t = 0$) and is then decreasing as t increases. The first order derivative is negative, while the second can be either positive or negative:

$$0 \leq S(t) \leq 1, \quad S(0) = 1, \quad S(\infty) = 0, \quad \frac{\partial S}{\partial t} \leq 0, \quad \frac{\partial^2 S}{\partial t^2} >< 0.$$

The pdf is the limiting value of the probability of failing to survive within the interval t and $(t + \Delta t)$ as Δt tends to zero:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}. \tag{2.3}$$

This density function does not summarise probabilities; it may be greater than one in value but it is always non-negative:

$$f(t) \geq 0.$$

Easily seen from the following two formulae, the pdf is the slope of the cdf and the cdf is the area under the curve of the pdf:

$$f(t) = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t} \quad \text{and} \quad F(t) = \int_0^t f(t)dt.$$

The hazard function, $h(t)$, is widely used to express the risk or hazard of failure at some time t and is obtained from the probability that an individual fails to survive in the interval $(t, t + \Delta t)$ conditional on survival up to that time.

$$h(t) = P(t \leq T < t + \Delta t | T \geq t).$$

This conditional probability can be expressed as a probability per unit time by dividing by the time interval Δt, to give a rate. The hazard function $h(t)$ is the limiting value of this quantity as Δt tends to zero. So,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{2.4}$$

which is equal to $\frac{f(t)}{S(t)} = \left|\frac{S'(t)}{S(t)}\right|$.

The function $h(t)$ is also known as the hazard rate, the instantaneous death rate, the intensity rate, the force of mortality or the inverse Mills ratio. In the context of our application, $h(t)$ is the instantaneous rate of admission.

It should be stated that the hazard rate is not a probability and therefore does not have the properties of probabilities. Like the $f(t)$, the only restriction implied by its specification is that $h(t) \geq 0$.

Moreover, the integrated or cumulative hazard function, $H(t)$, shows the expected number of failures that have occurred by time t:

$$H(t) = \int_0^t h(t)dt.$$

It follows that

$$h(t) = -\frac{\partial}{\partial t}\big[\log S(t)\big], \quad S(t) = \exp\big[-H(t)\big] \quad \text{and} \quad H(t) = -\log S(t). \quad (2.5)$$

It is important to note that whatever functional form is chosen one can derive all the other functions from it. In practice, it is advantageous to model the hazard function and derive the other functions from that.

2.2.2 Specific functional forms of the survival and hazard functions

The hazard rate $h(t)$ is particularly useful in duration data analysis as it represents the risk of failure at some time t. Sometimes, there is available information as to how this rate could change over time and specific distributions of $h(t)$ are adopted. Yet, more often, there is no obvious shape for it and any distribution over non-negative values could be a possible candidate.

The commonest functional forms that have been used in the literature comprise of the exponential, weibull, log-normal, log-logistic, generalised gamma and gompertz. These distributions have been described by several authors using different parameterisation; we mainly follow the specification given by Collett (2003). The various functional forms for the survival and hazard functions of the above distributions are summarised on Table 2.1.

2.2.3 Estimation of the survival and hazard functions

The survival function can be estimated by various methods among which the most important are the life table estimator and the Kaplan-Meier estimator (KM). The life table is one of the oldest techniques to present lifetime data by illustrating the survival experience of a cohort of individuals who are grouped

Table 2.1: Functional forms for the survival and hazard functions

	Hazard function	Survival function	Parameters
Exponential	λ	$e^{-\lambda t}$	λ
Weibull	$\lambda\gamma t^{\gamma-1}$	$e^{-\lambda t^{\gamma}}$	λ, γ
Gompertz	$\lambda e^{\theta t}$	$\exp\left(\frac{\lambda}{\theta}(1-e^{\theta t})\right)$	θ, λ
Lognormal	$\frac{\frac{1}{\sigma\sqrt{2\pi}}t^{-1}\exp\left(-\frac{(\log t-\mu)^2}{2\sigma^2}\right)}{1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)}$	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	μ, σ
Log-logistic	$\frac{e^{\theta}kt^{k-1}}{1+e^{\theta}t^{k}}$	$(1+e^{\theta}t^{k})^{-1}$	θ, k
Generalised gamma	$\frac{\frac{\theta\lambda^{\rho\theta}t^{\rho\theta-1}\exp\left(-(\lambda t)^{\theta}\right)}{\Gamma(\rho)}}{1-\Gamma_{(\lambda t)^{\theta}}(\rho)}$	$1-\Gamma_{(\lambda t)^{\theta}}(\rho)$	λ, θ, ρ

Source: Collet (2003)

in well defined time intervals. However, our interest lies on the non-parametric estimation of the survival function using the KM or product limit estimator.

To begin with, suppose we have a single sample of ungrouped complete duration times. Their survival function can be estimated by the 'empirical survival function' given below:

$$\hat{S}(t) = \frac{\text{number of observations with duration spells} \geq t}{\text{number of observations in the data set}}.$$

We assume that the estimated $\hat{S}(t)$ is constant between two adjacent failures, hence when we plot it against t a step-function is created. Similarly, the 'empirical distribution function' $\hat{F}(t)$ is given by $\hat{F}(t) = 1 - \hat{S}(t)$.

To incorporate censored observations in the analysis some modification is needed. Kaplan and Meier (1958) provided such extension by the introduction of the "product limit or Kaplan-Meier" estimate. It is defined as follows: Suppose that we have information on n individuals and that there are k distinct times $t_1 < t_2 < ... < t_k$ at which failures occur. Some of these individuals might be right-censored (denoted as c). Thus, we suppose that there are k failures times amongst the individuals where $k \leq n$. If we rank them in ascending order, the j-th is given by $t(j)$, for $j = 0, 1, 2, ...r$. If d_j represents the number

of individuals observed to fail at time t_j, n_j the number of individuals at risk at time t_j, and δ an infinitesimal time interval, then the probability that an individual fails within the interval $(t_j - \delta, t_j)$ is given by d_j/n_j. The estimated probability of survival is given by $(n_j - d_j)/n_j$. Given the way these intervals were constructed, the probability of surviving from t_j to $(t_{j+1} - \delta)$ is one and the joint probability of surviving from $(t_j - \delta)$ to t_j and from t_j to $(t_{j+1} - \delta)$ can be estimated by $(n_j - d_j)/n_j$. As δ tends to infinity, this becomes the estimate for the probability of surviving between t_j and t_{j+1}. The KM estimate for the survivor function is obtained by the following formula:

$$\hat{S}(t) = \prod_{t=0}^{j} \frac{n_j - d_j}{n_j}.$$

This represents the product of one minus the number of failures divided by the number of individuals at risk or the product of one minus the "exit rate" at each of the duration times. From the estimated $S(t)$ one can derive the estimates for $F(t)$, $f(t)$, $H(t)$ and $h(t)$.

The standard error of the KM estimate, that is the square root of the estimated variance of the estimate, is given by the Greenwoods formula:

$$se\{\hat{S}(t)\} \approx \hat{S}(t) \left\{ \sum_{t=0}^{t} \frac{d_t}{n_t(n_t - d_t)} \right\}^{\frac{1}{2}}.$$

At the tails of the $\hat{S}(t)$ distribution, the estimate of the variance using the above formula can underestimate the true variance. Alternative formulas are used to avoid such problem (Peto *et al.*, 1977).

The KM estimate is formed as a product of a series of estimated probabilities and is the limiting value of the life table estimate as the number of intervals tends to infinity and their width tends to zero. For the above reason, it is also known as the product-limit estimator of the survival function.

An alternative way to estimate the survival function especially for smaller samples is the Nelson-Allen estimate. For more details, visit Collett (2003).

Estimation of the hazard function can be also achieved by the life-table and the KM estimates. In practice, estimates of the hazard function obtained by the KM estimator $\left\{\hat{h}(t) = \frac{d_j}{n_j(t_{j+1}-t_j)}\right\}$ tend to be irregular. However, there are a number of ways for smoothing the hazard function, such as the kernel smoothed estimate, enabling the identification of any possible patterns.

2.2.4 Comparison of the survival functions of two or more groups of duration data

The simplest way of comparing the survival times obtained from two groups of individuals is to graphically plot the corresponding estimates of the two survival functions on the same axes. Moreover, hypothesis testing enables us to assess whether a possible difference between the survival curves of the two groups is real or a result of chance variation. Two of the commonest non-parametric procedures to evaluate group differences are the log-rank test and the Wilcoxon-Breslow-Gehan test. Below we summarise the different methods and formulas for testing the equality of survival functions across groups.

Definitions-Hypotheses-Test statistics

Let $t_1 < t_2 < ... < t_k$ represent the ordered failure times, d_j be the number of individuals observed to fail at time t_j, n_j is the number of individuals at risk just before t_j, and d_{ij} and n_{ij} denote the same things for group i, where $i = 1, 2, ..., r$.

The *null* hypothesis that there is no difference in the survivor experience of the individuals in the r groups is tested against the *alternative* that at least

one of these is different from the others.

To access the above hypotheses we evaluate the difference between the observed number of individuals in group i that fail at time t_j and the expected number of failures given the null hypothesis is true.

The expected number of failures group i at time t_j is $e_{ij} = n_{ij}d_{ij}/n_j$.

By summing the differences between the observed and expected failures over the total number of failure times, in the i groups, we get a test statistic of:

$$U = \sum_{j=1}^{k} W(t_j)(d_{1j} - e_{1j}, ..., d_{rj} - e_{rj}),$$

where $W(t_j)$ is a positive weight function that takes the value of zero when n_{ij} is zero.

Placing different weights at each failure time (t_j) gives rise to different tests:

Test	$W(t_j)$
Log-rank	1
Wilcoxon-Breslow-Gehan	n_i
Tarone-Ware	$\sqrt{n_i}$
Peto-Peto-Prentice	$\tilde{S}(t_j)$
Fleming-Harrigton	$\hat{S}(t_{i-1})^p\left[1 - \hat{S}(t_{i-1})^q\right]$

Note that $\tilde{S}(t_j)$ is the estimated KM survival function value for the combined sample at failure time t_i and $\tilde{S}(t_j) = \prod_{\ell:t\leq t_j}\left(1 - \frac{d_\ell}{n_\ell+1}\right)$.

2.2.5 Duration analysis of waiting times

In our application, a survival function shows the probability of a person remaining -or surviving- on the waiting list beyond a given time. In other words, the survival function shows the percentage of people admitted to hospital from the waiting list and the variations in this proportion as waiting time increases. It

also provides an estimate of the average waiting time as the integral of the survival function, though for data that are not censored this is merely an alternative calculation to a straightforward mean. An advantage of survival functions is the ability to observe patterns of waiting list behaviour over time; the same average waiting time might be generated by very different distributions of waiting time, reflecting different ways of managing lists. We estimate the survival functions using the non-parametric KM estimator described above. Survival functions can be compared between different groups, defined by, for example, illness, treatment, doctor or patient characteristics, and the log-rank test can be used for statistical testing of differences between them. The impact of variables that affect waiting time patterns can also be analysed.

The hazard function shows the rate at which patients leave the waiting list at a given time. Specifically it represents the probability that an individual is admitted for surgery conditional on having waited in a list up to that time. For example, if the hazard function is constant -the instantaneous rate remains the same at all times- this generates an exponential survival function of the form $S(t) = e^{-(\lambda)t}$, where λ is the hazard rate. The advantage of examining hazard functions is that it may reveal patterns of waiting list behaviour that would not otherwise be apparent. For example, if management effort in clearing waiting lists varies over time, the probability of a patient being admitted at a particular time will vary -as the length of time that they have had to wait increases, the probability of being admitted may rise, fall or remain constant.

Regression analysis can also be employed on duration data. Parametric estimation models for durations have two flavours, which depend on assumptions about the hazard rate. Proportional hazard (PH) models assume that there is a baseline hazard function that depends on time but not on other variables that affect duration, and is therefore common to all individuals. These other variables, which are usually assumed to be time-invariant, essentially scale the haz-

ard function for each individual. A valuable technique in estimating PH models is the semi-parametric Cox regression, which does not require any assumption about the hazard rate, simply the impact on it of the other variables. Accelerated failure time (AFT) models allow scaling to vary over time. Although these are therefore more flexible, they are entirely reliant on assumptions about the underlying hazard function; there is no equivalent of Cox regression.

2.3 Waiting times data

2.3.1 The structure of Hospital Episode Statistics

In UK there are two official waiting time datasets. The first comprises the Hospital Episode Statistics (HES) database which provides the waiting times of patients treated in a given financial year[10]. Each year's data consists of all admissions within that year and records both the date the patient was placed on the waiting list and the date he was admitted to the hospital to receive treatment. Thus, the measure of waiting time is constructed by the difference of the two dates. The second source of data stems from the Department of Health waiting list returns[11], a cross-sectional measure which contains the time patients spend on the list at a particular census date. The above measures of waiting times are fundamentally different. For more details see Dixon (2004) and Dixon and Siciliani (2009). Due to the fact that the first offers complete waiting times of treated patients -even thought the spells are generated retrospectively- while the second constitutes a snapshot of incomplete waiting times currently on the list, it serves best the scope of our research to work with HES.

Data were provided by the HES database of the Department of Health. This

[10] http://www.hesonline.nhs.uk

[11] http://www.performance.doh.gov.uk/waitingtimes/index.htm and http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/HospitalWaitingTimesandListStatistics/index.htm

service is run by the 'Health and Social Care Information Centre'[12] through 'Northgate Information Solutions'. HES data cover episodes of care for NHS-funded admitted patients that were treated in NHS trusts, primary care trusts or in the independent sector. It constitutes a record-level database of hospital admissions in all NHS trusts in England since 1989, with more than 12 million inpatient records per year, and outpatient records since 2003, comprising more than 40 million records per year. Although private hospitals are not included, private patients treated in NHS hospitals are. Information on personal, administrative, geographical and clinical characteristics is submitted on an annual base by NHS Trusts to the 'NHS Wide Clearing Service' and from December 2006 to the 'Secondary Uses Services' (SUS). The latter, after making the data available to the commissioners, include them in their database. At a particular and pre-arranged date, SUS send an extract from their database to HES. After this point HES data is fixed as opposed to SUS data that keep changing. The next step in the processing cycle comprises the validation and cleaning of the extract by HES. Data quality reports and checks are completed at various stages in the above cycle.

As stated in the previous paragraph, before making information available, HES data quality team is responsible for cleaning all incoming data[13]. HES performs checks and makes corrections where errors are identified. This cleaning system consists of four stages; provider mapping, automatic cleaning, manual cleaning and derivation. They also generate quality reports and derive new variables that might be of interest for providers and HES users. Yet, the fact that data are gathered from a large number of trusts with each of them having a different administrative system can create concerns about the quality of data as a whole. There might be well-organised trusts that provide high quality data,

[12]It was previously known as the 'NHS Information Centre'.

[13]http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=376

but at the same time there might be trusts with poor information systems. Thus, HES sets special data quality indicators to improve the quality of the data collected. It is worth mentioning that the quality of the data is inspected at various stages of the process; at SUS and both during the HES administration of it and even after its publication.

Furthermore, data security and patient confidentiality issues are quite critical so as to limit the possibility of identifying particular individuals -either patients or consultants- through sensitive information. That is the reason why HES provides tables with data in an aggregated form while a 'Data Access Advisory Group' is responsible for dealing with requests for sensitive data. The original HES protocol, which was updated in 2009, and the HES user guide supply guidelines for the handling of data aiming at protecting the privacy of individuals and meeting all the required security standards[14].

More specifically, the data are collected by financial year and among others include information on specialty, diagnosis, operation, healthcare resource group, admitting hospital, type of admission, waiting times, length of stay and patient characteristics such as age, sex, ethnicity and residence. Through the years HES has been updating the items it provides and has been enriched with new variables such as codes of GP practice, pseudonymised consultant codes and socioeconomic domains.

2.3.2 Exploratory data analysis

HES data for 2001/2002 and 2002/2003 covering the English NHS are analysed. Due to the fact that we have the waiting times for all the admissions recorded in each year the data are complete with respect to date of admission; hence they are not right-censored. This analysis focuses on elective care; it excludes emergency and maternity cases, which are not counted as part of waiting lists.

[14]http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=331

Data are included on all three principal routes to admission for elective surgery: waiting lists, where there is no exact date of admission; booked admissions, where there is an exact date for admission; and planned, where there is an exact date of admission for a course of treatment over time or a second operation.

As shown in Table 2.2, waiting times for elective surgery are asymmetrical; for both years, they are positively skewed with long right tails and have an average of at least 130 days for all methods of admission (waiting lists, booked and planned) and 155 days if admission takes place by waiting lists. Another feature of interest is the huge standard deviations observed for all the cases, revealing the extent of the variability of waits.

Table 2.2: Descriptive statistics of the variable *waiting time*.

	2001/2002		2002/2003	
	elective	*waiting list*	*elective*	*waiting list*
Number of observations	1639007	1052279	1709155	1042757
Mean	130.29	155.67	134.39	161.43
Standard deviation	157.93	155.73	159.78	154.17
Median	70	101	75	112
Minimum	1	1	1	1
Maximum	4236	3691	7577	5102
Skewness	2.94	1.92	3.26	2.09
Kurtosis	21.95	11.32	29.58	16.74

Figure 2.3 illustrates the kernel densities of the patients' waiting times admitted from a waiting list. The kernel density estimation is a non-parametric way to estimate the probability density function (pdf) of a random variable. It is a data smoothing process in which the smoothing parameter (the bandwidth) has a strong influence on the derived estimate. The kernel densities of Figure 2.3 illustrate that overall waiting times distributions are skewed to the right with the bulk of the distribution lying way below 500 days of wait. However, as the rest of the chapter (and Chapter 3) will show, duration analysis is a much more informative estimation technique.

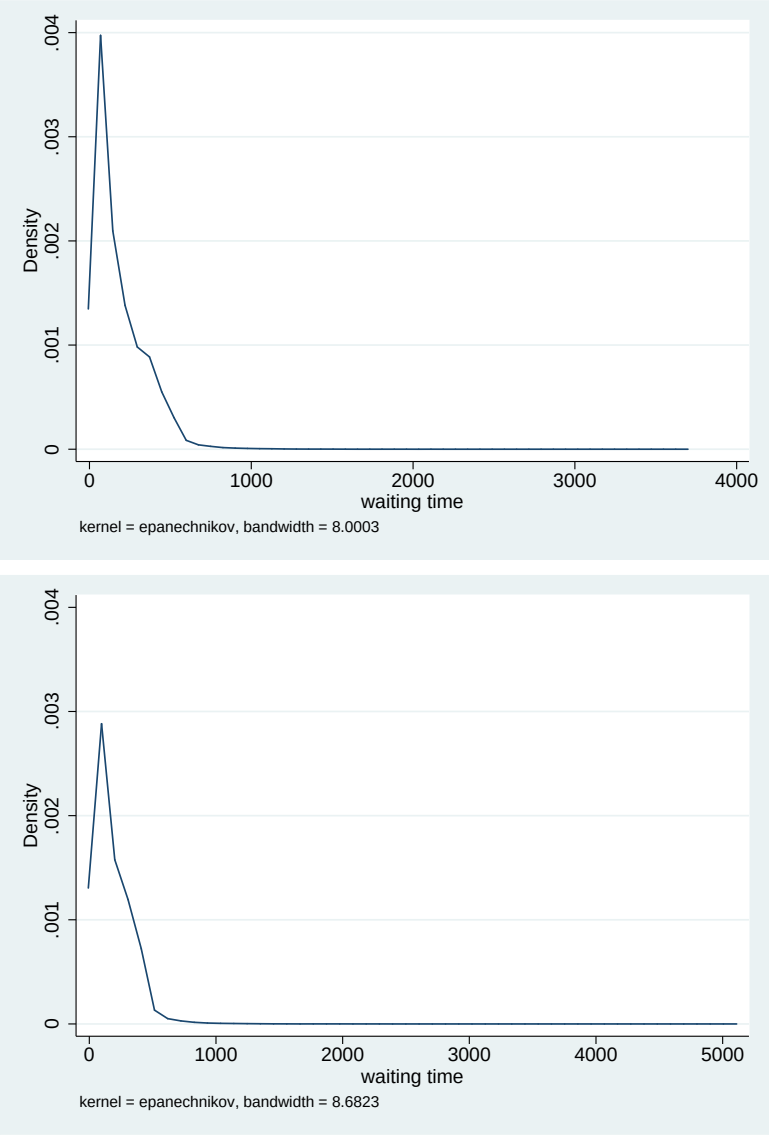


Figure 2.3: Kernel densities of waiting times, 2001/2002 *(top)* and 2002/2003 *(bottom)*.

The probability of remaining on the waiting list past a certain point in time (survival function) may be more interesting than expected waiting time, especially for policymaking. In particular, with regards to the impact of waiting time targets, which is a major focus of this chapter, the kernel density can only provide information on whether the bulk of the distribution lies around the set

targeted waiting time. However, the hazard function can lend more insight on the 'failure' mechanism, that is, the response of the admission mechanism to the set target. The hazard rate allows to approximate the probability of exiting the list within an incremental interval, conditional on having 'survived' up to that point. It thus approximates the conditional probability of leaving the list given the amount of time on test, rather than the unconditional probability (pdf), and as such is more meaningful.

As we will see, we can utilise the hazard rate to essentially establish whether there is increased probability (or not) of patients being treated 'around' the interval of the waiting time target. The duration analysis technique can help us with the question: given that the patients' duration is approaching the waiting time target, what is the probability of exiting the list? The unconditional probability (pdf) does not take into account the time that has elapsed and the pre-announced waiting time target.

We obtained HES data on each episode, including specialty, diagnosis, operation, admitting hospital and type of admission, and on the characteristics of the patient whose episode it was, including age, sex, ethnicity, and residence. The data were anonymous with respect to patients.

We evaluate data from three specialties: general surgery; trauma and orthopaedics and ophthalmology. These were chosen because together they constitute more than 50% of the patients waiting for elective treatment. Initial analysis reveals some patients who appear to have waited an implausibly long time -some greater than ten years- which is most likely the result of coding problems; consequently, the 0.1% of patients whose waits appeared to be longer than three years were excluded.

We analyse the data at three levels. The incentives associated with waiting times targets apply to the hospital, so it is appropriate to analyse overall hospital waits. However, given the possibility of systematic differences in waiting list

management between specialties within any hospital, we examine the waiting times separately for the three specialties. Moreover, it is possible that waiting lists are managed differently for particular operations, so we also focus on the four most frequently performed procedures within each specialty, as described in Table 2.3.

Table 2.3: The four most common procedures in each of the three surgical specialties.

Surgical specialties and their common procedures	**Percentage**
General surgery	
Excision of gall bladder (total cholecystectomy)	26% of all general surgery
Ligation of varicose veins of leg	
Excision of lesion of skin	
Primary repair of inguinal hernia	
Trauma & Orthopaedics	
Release of entrapment of peripheral nerve at wrist	28% of all orthopaedic surgery
Total prosthetic replacement of hip joint using cement	
Total prosthetic replacement of knee joint using cement	
Endoscopic operations on semilunar cartilage	
Ophthalmology	
Extirpation of lesion of eyelid	71% of all ophthalmologic surgery;
Incision of capsule of lens	lens prosthesis accounts for 62%
Prosthesis of lens	
Cauterisation of lesion of retina	

Source: Hospital Episode Statistics

2.4 Results

We start off by analysing the managing of waiting times according to clinical characteristics; that is specialty, operative procedure and admission source. We then explore the overall waiting times of a selected set of NHS hospitals for both years. In all levels of analysis, the survival and hazard functions are estimated non-parametrically and the long-rank test for significance of differences is performed. Also, note that the HES data are very rich and a fine level of disaggregation of the data is possible. This generates a very large set of possible analyses; here only a fraction of the results generated is reported. Key differences or similarities between the data that were selected for presentation, and those not shown, are identified.

2.4.1 Estimation of survival and hazard functions

Variation in the probability of admission by specialty and operative procedure

Figure 2.4 presents survival curves for all patients admitted[15] in each of the three specialties during 2001/2002. In effect, these show the proportion of patients who remain on the waiting lists at each point in time. At time 0, all patients are on the list and the curve falls as they leave the list by being admitted. The graphs have been truncated at two years, as this helps to display more clearly the patterns for the very large majority of patients that have waiting times below that level.

The shortest waiting times are for general surgery and the longest are for trauma and orthopaedics, but examining the shapes of the curves shows a more complex picture than that. Until around three months the rate at which patients are admitted from a list is similar for ophthalmology and trauma and

[15]In Figures 2.4, 2.5 and 2.6 admission method is by waiting lists.

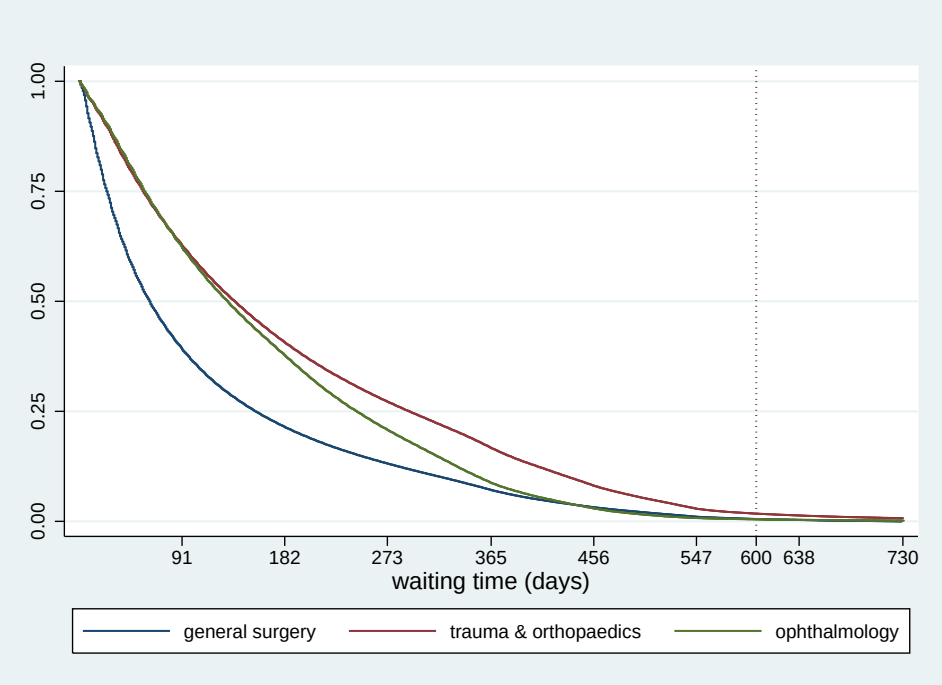


Figure 2.4: Kaplan–Meier survival curves for three specialities, 2001/2002.

orthopaedics, both being very much slower than for general surgery. After that the admittance rate becomes relatively faster for ophthalmology until, by 15 months, survival on the ophthalmology waiting list is similar to that of general surgery. The corresponding survival curves for 2002/2003 (not shown) are very similar. The log-rank test for equality of the survival functions demonstrates statistically significant differences between waiting times for the three specialties for both years, suggesting systematic differences in waiting times and admission patterns between surgical specialties.

Figure 2.5 shows the estimated hazard functions for 2001/2002 and 2002/2003, with the national waiting list targets for those years represented by the bold dashed lines. The hazard functions are all characterised by 'peaks'; there is initially an increasing probability of admission as waiting time increases, this reaches a maximum, after which there is a decreasing probability. The exact nature of this effect is different in each specialty.

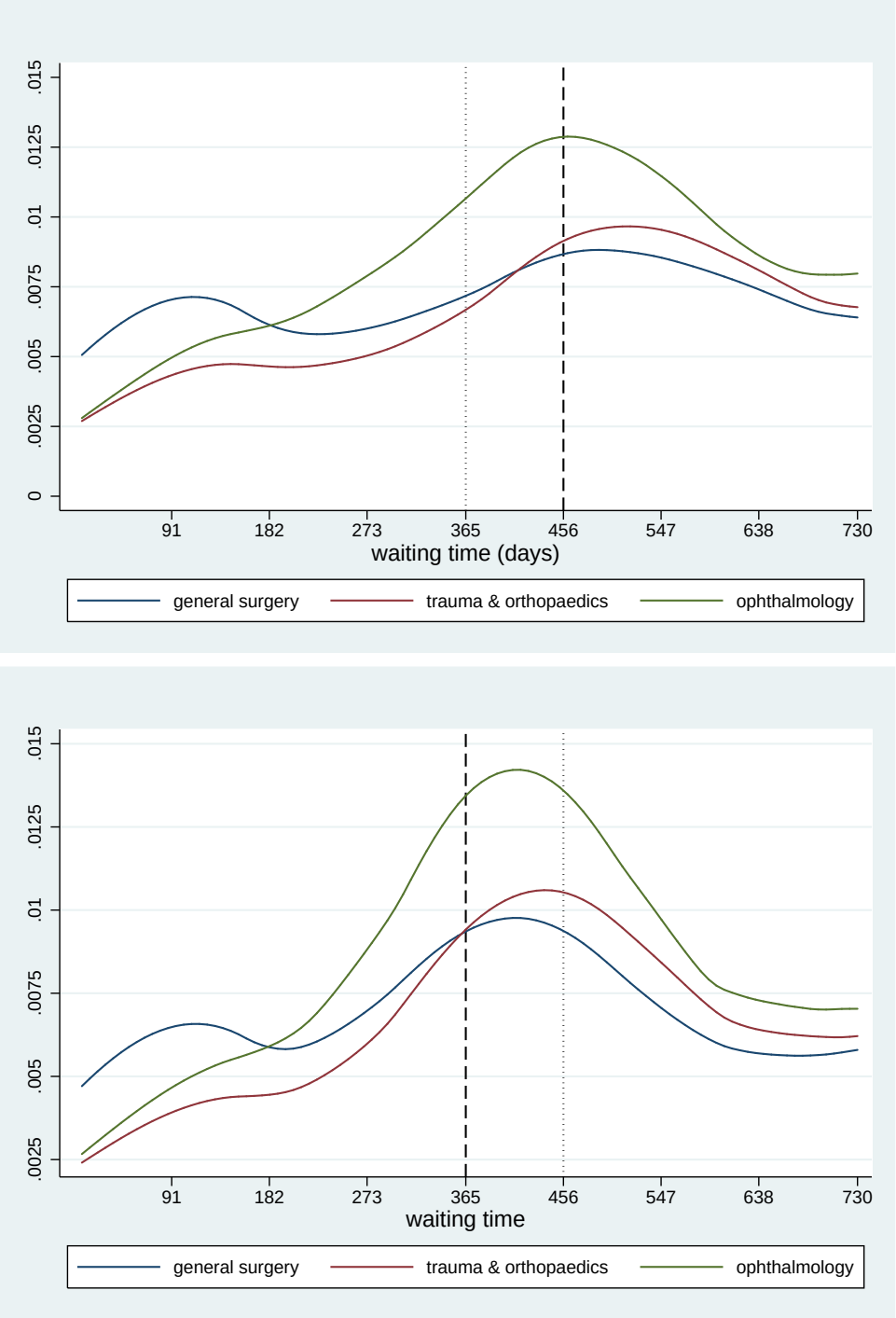


Figure 2.5: Hazard curves for three specialities, 2001/2002 *(top)* and 2002/2003 *(bottom)*.

In 2001/2002, for general surgery, increased waiting list activity is observed by peaks in the curve at 4 and 15 months; for ophthalmology, at exactly 15 months; and for orthopaedics between 15 and 18 months. The waiting list target for that year was 15 months.

The waiting time target for 2002/2003 was 12 months. The hazard curves for that year show, in every case, that the peaks occurred earlier than in 2001/2002. For general surgery, the main peak in probability of admission in 2002/2003 reduced to 12 months; for orthopaedics between 12 and 15 months; and for ophthalmology to around 14 months.

We now turn to examine the management of the lists in terms of the most common procedures in each specialty. The top graph in Figure 2.6 shows the survival curves for the four most common procedures within general surgery for 2001/2002 and the bottom graph shows the corresponding hazard curves. The shortest waiting times were for excision of skin lesions, followed by inguinal hernia, gall bladder excision and varicose vein ligation. The log-rank test for equality of the survivor function showed that these differences are statistically significant.

Each operation has a peak close to the waiting time target of 15 months, though excision of skin lesions has an additional earlier peak at 3 months. For 2002/2003 (not shown), the peaks for every procedure again occur earlier, between 12 and 15 months, coincident with the lower target time. However the earlier peak for excision of skin lesions at 3 months is unchanged. Earlier peaks are observed in other occasions (see Chapter 3). The most prominent possible explanation for this phenomenon is prioritisation in the admission process, also verified in the theoretical model of Chapter 4. Given that hospitals undertake some sort of prioritisation of cases based on clinical urgency, we would expect an early peak due to more urgent cases being scheduled for earlier periods. In addition, another plausible explanation could be some sort of informal/internal

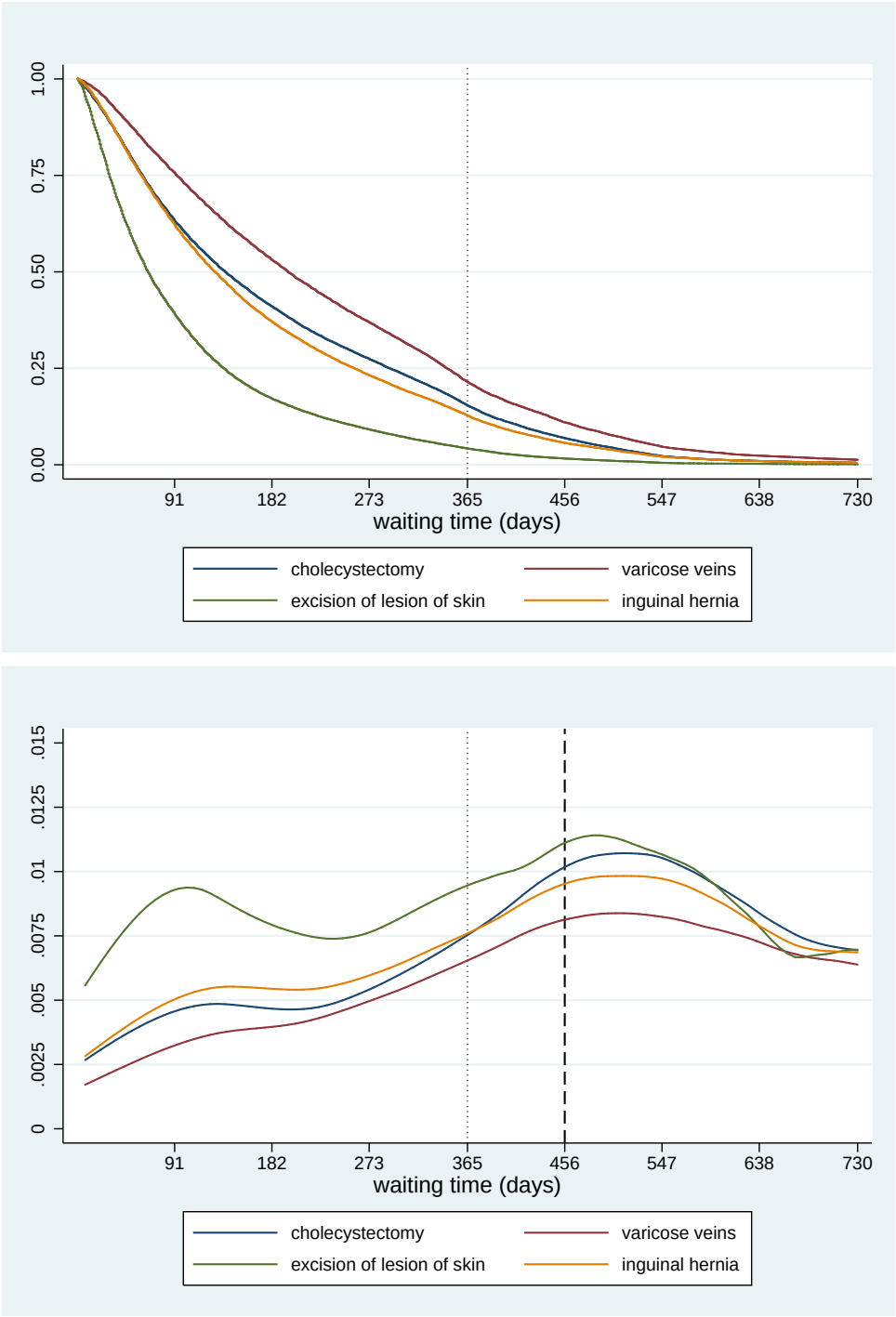


Figure 2.6: Survival *(top)* and hazard *(bottom)* curves for the four most common general surgical procedures, 2001/2002.

management of the list based on targets set by the hospitals at distinct points, say 3 months, 6 months etc, before the pre-announced universal and formal waiting time targets. Thus, earlier peaks could also represent clinical preference for a particular date of admissions, unrelated to the set target.

These results are broadly replicated in analyses of procedures in the other two specialties. In each case there are statistically significant differences in waiting times between the most common procedures within each specialty, and peaks in the hazard functions are at around 15 months for 2001/2002 and between 12 and 15 months for 2002/2003 -both coinciding with the prevailing waiting time target for those years.

Variation in the probability of admission by type of admission

While patients can be admitted to hospital from a variety of sources, the three principal routes are waiting lists, booked admissions, and planned admissions. Waiting times targets apply only to patients on waiting lists. Figure 2.7 shows the survival curves for 2001/2002 by admission method and Figure 2.8 shows the hazard curves for both 2001/2002 and 2002/2003.

Booked admission patients have the lowest waiting times and waiting list patients the longest; the differences between the three routes are statistically significant. The survival curves show that the proportion of patients waiting for planned admissions reaches that of waiting list patients at around 15 months. Booked admissions have two hazard curve peaks, a larger one at 3 months for both years and a smaller one at 15 months in 2001/2002 and at 12 months for 2002/2003.

The hazard rates for planned admissions have very small variations over waiting times, but the very slight peaks that are observable are similar to those for booked admissions. However, waiting list admissions have a more notable

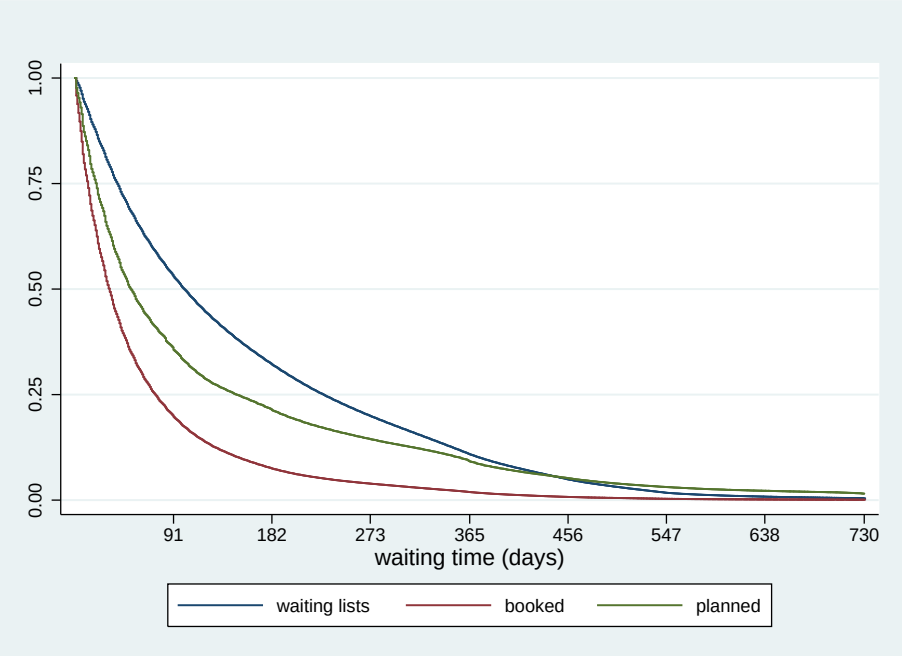


Figure 2.7: Survival curves by type of admission, 2001/2002.

peak which reduces from around 15 months in 2001/2002 to between 12-15 months in 2002/2003.

These results also point to the direction of a change in the list management due to the implementation of the target, since the 'move' of peaks is more apparent in the case of waiting lists, rather than planned or booked admissions. A somewhat contrary observation is that we would not expect any peaks at the target times for planned admissions because these are not subject to waiting times targets; however, there are in fact peaks (albeit slight). One plausible explanation is that part of the managerial response to targets may include reclassifying patients between waiting list and planned admissions.

Variation in the probability of admission by hospital

We further analyse the data for a selection of seven trusts to highlight the variations between different hospitals. These were selected to give a reason-

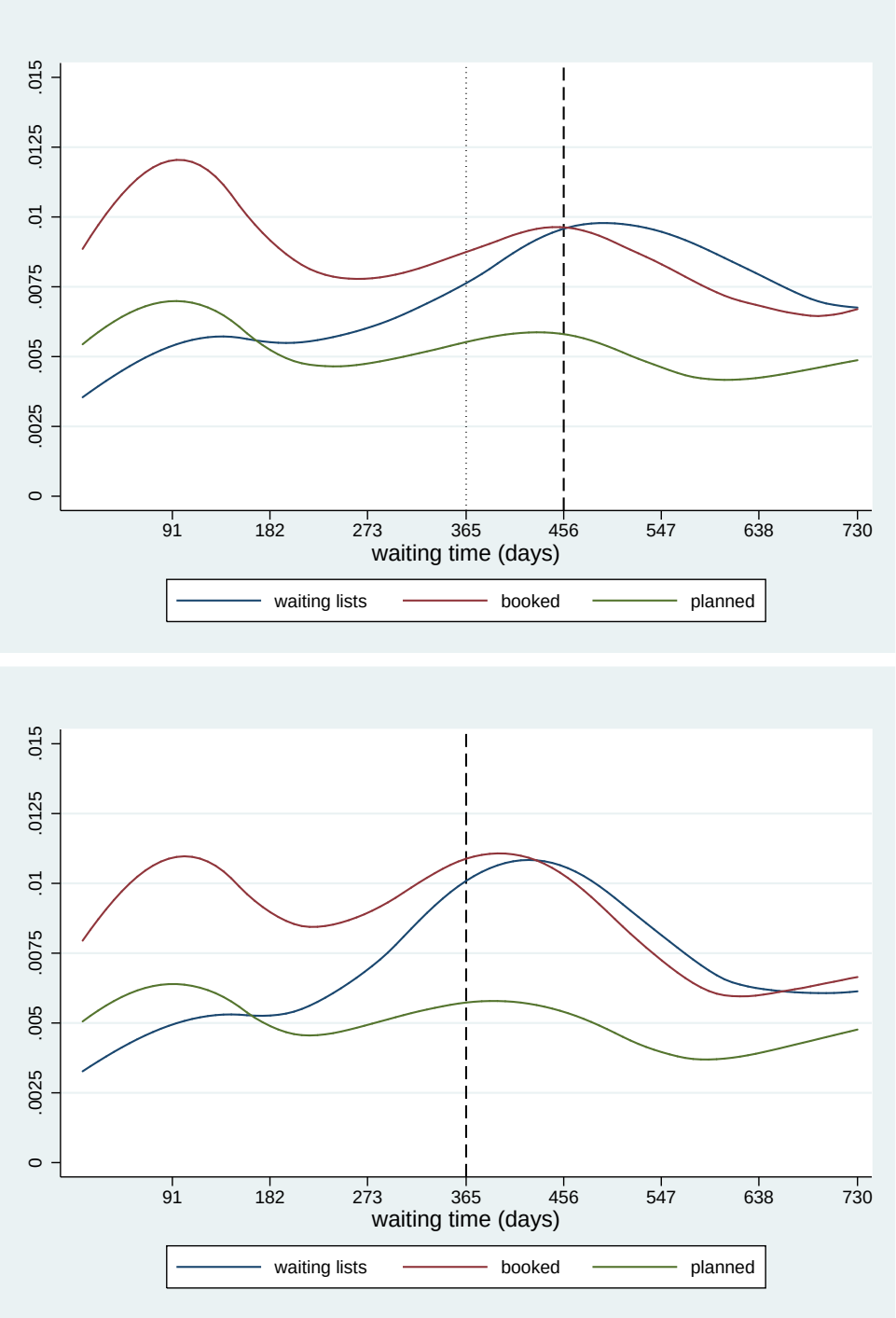


Figure 2.8: Hazard rates by type of admission, 2001/2002 and 2002/2003.

able geographic spread across England, and Table 2.4 shows the NHS Regional Offices within which they are located.

Table 2.4: Seven trusts and their equivalent NHS Regional Offices.

Trust	**Regional Office**
Manchester University Hospital	North West
Nottingham University Hospital	Trent
Southampton Hospital	South East
Newcastle Upon Tyne Hospitals	Northern and Yorkshire
Birmingham Heartlands and Solihull	West Midlands
Royal Free Hospital	North Central London
Guys and St.Thomas Hospitals	South East London

Figure 2.9 shows the hazard curves for each trust for each year. The patterns and peaks in admission probabilities at each point in time differ greatly between providers. Some do not have notable peaks; for these providers, the probability of admission did not vary much over waiting time. In particular, the hazard curve for the Royal Free Hospital is almost horizontal in both years. All of the peaks, for those providers that have them, occur at lower waiting times in 2002/2003 compared to 2001/2002, although the extent of this differs. For example Newcastle Upon Tyne Hospital has a main peak of between 6 and 12 months for both years, with only a small difference between them. Other providers changes were much larger, for example, the main peak for Birmingham Heartlands & Solihull reduced from a little more than 15 months in 2001/2002 to 12 months in 2002/2003, an almost exact mirroring of the change in targets.

Much variation is observed at a hospital level. Indeed the implementation of the targets in the UK identifies cases, to a more or lesser extent, that appear to exhibit increased instantaneous probabilities of patients admission ('peaks' in the hazard curves) around the time of the target. This disaggregated analysis is extended in Chapter 3, in an attempt to identify patterns of variation across hospitals and even physicians, as well as across time.

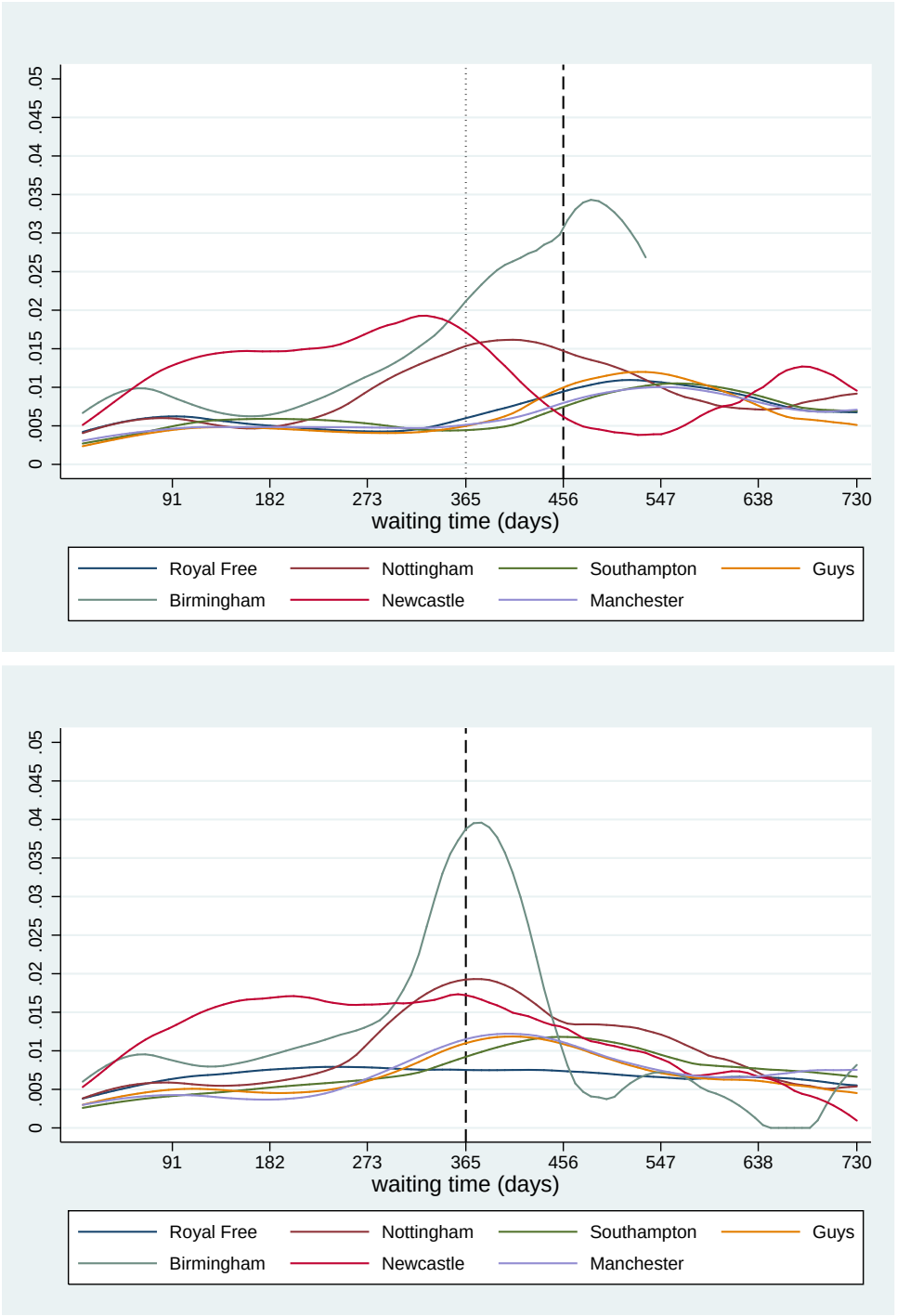


Figure 2.9: Hazard rates for seven NHS Hospital Trusts, 2001/2002 *(top)* and 2002/2003 *(bottom)*.

2.4.2 Duration analysis with covariates

In this section, we explore the adjustments of the survival and hazard rate functions for the effects of covariates representing patient and clinical characteristics that may impact on waiting times, using parametric PH and AFT models under different distributional assumptions and Cox regression.

For the AFT models, the dependent variable is waiting time until admission; for the PH models it is the hazard rate. For both models, the independent variables are Age, (calculated as age minus 50, the average for the sample), and a series of dummy variables representing sex, admission category, main specialty, patient classification and ethnicity (see Table 2.5). Together, these define a reference group of people that are male, 50 years old, NHS patients, admitted in general surgery, admitted to inpatients and white. The dummy variables are therefore Female; Private patient; Orthopaedics; Ophthalmology; Day case; Black; Indian; and Other ethnicity. We analysed data for both years, but because the results were very similar we report analyses only of the 2002/2003 data.

Table 2.6 shows the results from four alternative AFT models. It should be noted that the very large sample size means that the statistical significance of each coefficient is a poor guide to its practical significance[16]. The only coefficients that are not statistically significant are those for Black, in all specifications, and for Female in the AFT Log-normal specification. All models are statistically significant with high log-likelihood values; the sign and magnitude of the estimated coefficients is similar regardless of the particular assumptions made about the distribution. We are unable to distinguish statistically between the different models according to goodness of fit criteria, and indeed they produce very similar results.

[16]See Johnson (1999), and references therein, for an analysis on the relation between sample size, significance and power of tests.

Table 2.5: Description of variables

Name	**Description of variable**
Age at the start of episode	It is the age in years calculated from the date an episode starts and the date of birth. Recode of patients younger than 1 year old.
Sex	It defines the sex of patients: • male • female
Administrative category on admission	It is an administrative measure on category of admission. It includes: • NHS patients • private patients
Main specialty	It is the specialty under which the consultant is contracted by the hospital. We are going to investigate the following specialties: • general surgery • trauma and orthopaedics • ophthalmology
Patient classification	Patients are classified as • ordinary admissions • day cases
Ethnic group	It specifies the ethnic group of patients. Although from April 2001 new codes have been created, the old ones are still accepted. For our analysis they have been recoded as: • white • black • indian • any other ethnic group
Date of decision to admit	It is the date a consultant decided to include a patient in the waiting list for a surgical procedure. The patient might be admitted immediately or some time later.
Date of admission	It is the date the patient was admitted to the hospital.
Waiting time	It is the period in days from the date of decision to admit to the date of admission.

Because of this, we will restrict our discussion to the AFT-Exponential model. The antilog of the constant term is the average waiting time for the reference group as defined above, i.e., $e^{4.91} = 136.5$ days. Although Age has a significant coefficient, this translates into an increase of less than one days waiting time for a one-year increase in age, other things being equal. The changes in waiting times for Female or Indian are also less than one day. Such differences

are obviously of no account -but other findings are of more interest. Other things being equal, private patients wait on average 99 fewer days; orthopaedic and ophthalmology patients wait on average 74 and 71 more days; day case patients wait on average 30 fewer days; ethnic groups other than Black, White and Indian wait on average 33 fewer days.

Table 2.7 shows the results from the PH models. All covariates are, again, statistically significant, apart from Black. The results from the parametric models are consistent statistically, functionally and quantitatively with those from the Cox regression.

Since there is very little to choose between the different models, the results presented above concerning the impact of the independent variables may be taken as representative. Some variables have no real impact on waiting times, such as age and sex; however, some -such as whether the patient is NHS or private- have an impact on waiting times that is significant in both statistical and practical terms.

Table 2.6: Accelerated failure time models

T	Exponential		Weibull		Log-normal		Log-logistic	
	Coefficient	$P > \|z\|$	Coefficient	$P > \|z\|$	Coefficient	$P > \|z\|$	Coefficient	$P > \|z\|$
Age	.0005929	0.000	.0006069	0.000	.000834	0.000	.0005195	0.000
Female	.0076326	0.001	.0079359	0.001	.0026647	0.362	-.0064518	0.024
Private Patient	-1.293627	0.000	-1.288302	0.000	-1.416566	0.000	-1.471763	0.000
Orthopaedics	.4333493	0.000	.4277083	0.000	.7011317	0.000	.748772	0.000
Ophthalmology	.4204603	0.000	.4141631	0.000	.6748901	0.000	.7456852	0.000
Day case	-.2501083	0.000	-.2502762	0.000	-.1594466	0.000	-.2244831	0.000
Black	-.0047612	0.633	-.0047061	0.629	-.0128596	0.291	-.0161637	0.176
Indian	-.0441387	0.000	-.0442778	0.000	-.0204272	0.026	-.0373673	0.000
Other ethnicity	-.2797464	0.000	-.2782713	0.000	-.3077501	0.000	-.3418351	0.000
Cons	4.916261	0.000	4.929258	0.000	4.16032	0.000	4.262388	0.000
ln $\_p$			.0237477	0.000				
P			1.024032					
$1/p$			.976532					
ln $\_sig$					.2005121	0.000		
Sigma					1.222028			
ln $\_gam$							-.3788122	0.000
Gamma							.6846742	
Log likelihood	-1104677.4		-1104356.9		-1150796.2		-1147369	

Table 2.7: Proportional hazard models

T	Exponential		Weibull		Gompertz		Cox	
	Hazard ratio	$P > \|z\|$	Hazard ratio	$P > \|z\|$	Hazard ratio	$P > \|z\|$	Hazard ratio	$P > \|z\|$
Age	.9994073	0.000	.9993787	0.000	.9993014	0.000	.9992734	0.000
Female	.9923965	0.001	.9919063	0.001	.9911281	0.000	.9875895	0.001
Private Patient	3.645988	0.000	3.74066	0.000	3.751843	0.000	3.594825	0.000
Orthopaedics	.648334	0.000	.6453342	0.000	.6463513	0.000	.6437025	0.000
Ophthalmology	.6567444	0.000	.6543478	0.000	.6590856	0.000	.6672428	0.000
Day case	1.284164	0.000	1.292128	0.000	1.297126	0.000	1.319649	0.000
Black	1.004773	0.633	1.004831	0.629	1.005844	0.559	.9981193	0.850
Indian	1.045127	0.000	1.046385	0.000	1.048196	0.026	1.044269	0.000
Other ethnicity	1.322794	0.000	1.329707	0.000	1.328031	0.000	1.339732	0.000
$\ln\_p$			.0237477	0.000				
P			1.024032					
$1/p$			.976532					
Gamma					.0003357	0.000		
Log likelihood	-1104677.4		-1104356.9		-1103564.3			

2.5 Concluding remarks

Statistics on average waiting times, and on performance against targets, show that the NHS has made considerable progress in improving its performance in this respect since 2000. The analysis conducted in this chapter confirms that conclusion. By examining not only the central tendency but also the nature of the underlying distribution of waiting times, we offer new insights into the way waiting targets are being managed, as well as revealing some important issues.

First, the hazard functions by specialty, by procedure and to a lesser extent by hospital, show that the period of time at which the probability of admission peaks coincides with the prevailing waiting times target; and that the introduction of shorter targets coincides with a reduction in the waiting time at which this peak occurs. Is this causal or a coincidence? It cannot be claimed that the conditions that prevailed in the two periods were identical. In addition to changed waiting times targets, there were other measures, including increased spending on elective surgery and the greater use of private sector hospitals to overcome supply bottlenecks in the public sector. However, although these measures may have had an influence on the ability to reduce waiting times, the coincidence of the timing that we have observed is highly suggestive that the targets have influenced the way in which waiting times have been reduced[17].

Specifically, one interpretation of the observed peaks is that management and surgeon efforts have been directed to avoiding breaching the institutional targets, since the rewards and penalties focus on the number of patients treated before or after the target, and that there is therefore an increasing probability of admission as the target approaches. However, once a target has been breached, the extent to which it is breached is less important, so the probability of admis-

[17] As shown in the following chapter, similar patterns in admission rates are observed for the coming years while the targets get stricter.

sion falls and priorities are directed elsewhere. If increased resources for elective surgery were the sole explanation for reduced waiting times, one might instead expect an equal effort devoted to the admission and treatment of everyone irrespective of whether their admission is before, at or beyond the target.

Further evidence suggesting the dominance of targets as the driver of change comes from the observation that, both at the level of procedures and of hospitals, the peaks in the probability of admission that are evident at or around the target wait changed over the two years, while other observable peaks, such as those at 3 months for all general surgical procedures and for booked and planned admissions, remained unchanged. One could speculate that these peaks represent some clinical preferences for a date of admission or perhaps an approximate division between more and less urgent cases, but since these clinical thresholds are below the targets they are not affected. Thus, prioritisation of the list with urgent cases treated quicker can explain the earlier peaks observed. With regards to later -after the target- peaks, as in the case of of Newcastle Upon Tyne Hospital (Figure 2.9), this can be explained again by some clinical preference for a 'maximum' acceptable waiting time. In addition, if penalties from breaching the target are proportional to the 'length' of breach, then one could justify an increased admissions rate at a later point.

Another striking finding from the analyses is the wide variation in waiting time distributions and implied admission tactics by hospital, specialty and by procedure. We suggest that the differences between trusts reflect the level and type of activity employed by them at any given time. Some hospitals exhibit great effort to tackle excessive waits; some manage only the longest waiters; while others appear to tackle the whole spectrum of the waiting times distribution. This is consistent with the findings of qualitative research (Appleby *et al.*, 2005a). One possible interpretation is that, notwithstanding the significant differences that are evident between entire hospitals in their pattern of

admissions, and the observable response in the probability of admission to a change in waiting times targets at all levels, the managerial influence that this implies is more generally overwhelmed by the decision processes of individual clinical teams, each pursuing quite different priorities and admissions criteria. Targets may be met equally, but the ways in which they are met are quite different.

An alternative interpretation is that the differences in the probability of admission between specialty or procedure are determined by inherent differences in patient characteristics, such as the severity of their condition, operation type, clinical difficulty in performing it, and whether the surgery can be performed as a day-case. However, in the absence of data enabling comparisons of the health status of patients (Appleby and Devlin, 2004) and the lack of any objective means of differentiating between surgical procedures on these grounds, for example judging the excision of skin lesions to be clinically more important than varicose veins, such explanations would be highly speculative.

Can variations between individuals' waiting times be explained by clinical, patient or provider-level characteristics? These results suggest that they can, although more analyses are needed to answer this properly. From an equity point of view, it is useful to know that characteristics such as age and sex do not affect waiting times in any important way, and that we can be confident about those findings because of the very large sample size. Some findings suggest that more investigation is required, for example the difference in waiting times for 'Other' ethnic groups. However, some large differences are of immediate interest.

Of particular interest is the very large difference between NHS and private patients -which in this context is private patients using NHS accommodation or services- suggesting that private patients have a considerable advantage in access compared to NHS patients, even though the two groups use exactly the

same facilities and services. Private work undertaken by the NHS is at the discretion of the responsible NHS body, usually the hospital. The main reason for carrying out such work is to generate income -around 300 million in 2001/2002- and to recruit and retain consultants in areas where their access to private facilities, and hence private income, is limited. The rules and guidance governing the treatment of private patients in NHS facilities are complex. However, a recent code of conduct[18] re-emphasises a key point of previous guidance; private work in NHS hospitals should not 'interfere with the organisation's obligations to NHS patients'. Our finding that patients treated privately by the NHS had significantly shorter waits than other, NHS, patients suggests, but does not prove, that private patients can buy priority of access over NHS patients.

Also, as suggested, it is possible that trusts may have to some extent met their targets by adjustments such as reclassifying patients included on waiting lists as planned cases and reclassifying day-cases as outpatients[19].

The application of duration analysis to waiting times data offers an important means of improving the understanding of waiting times management and of gauging the behavioural responses to policy measures. Research at the next chapter uses these techniques to analyse differences between the waiting times distributions of trusts who differ by size and type and those who are successful and unsuccessful in meeting the waiting times targets in each period, as well as examining the extent to which differences in the probability of admissions are evident at still greater levels of disaggregation, such as by individual consultant.

[18] A Code of Conduct for Private Practice: Guidance for NHS Medical Staff. Department of Health, London (2003).

[19] Inappropriate Adjustments to NHS Waiting Lists. National Audit Office (2001).

Table 2.8: Examples of applications from different disciplines that use duration analysis techniques

Studies	*Subject*
Medicine and biology	
Gehan, 1965	Comparison of times until remission between two groups of leukaemic patients (control group and group treated with 6-mercaptopurine).
Feigl and Zelen, 1965	Analysis of mortality rates of two groups of leukaemia patients alongside with measurement of their white blood count, which is a continuous explanatory variable.
Pike, 1966	Times elapsed between induced carcinogenesis and mortality from vaginal cancer in rats experiments.
Crowley and Hu, 1977 Turnbull et al., 1974 Gail, 1972	Investigation of the effect of heart transplant to the survival of patients of the Stanford heart transplant programme.
Dobbs, 1980 Murray et al, 1993	Survivorship of total hip-replacements.
Muenchow, 1986	Comparison of the times till the arrival of any flying insect on male and female flowers of the species Clematis ligusticifolia.
Engineering	
Nelson and Hahn, 1972	Analysis of the time needed for motorettes to fail operating by increasing the temperature.
Shooman, 1983; Musa et al, 1987	Software reliability.
Bullough et al, 1999	Reliability analysis for structural integrity assessment of a UK nuclear plant.
Jardine et al, 2007	Application of the weibull proportional hazards model to aircraft and marine engine failure data.
Peiravi, 2010	Reliability of air to air missile fuse electronics.
Economics and Sociology	
Lancaster, 1979,1990 Nichell, 1979 Butler et al, 1985	Duration analysis of unemployment spells.
Tuma et al, 1977	The impact of income maintenance on marital dissolution and remarriage rates.
Kennan, 1985	Analysis of the duration in days of 62 contract strikes that commenced within 1968-1976 in US manufacturing.
Forster and Jones, 2001	Exploration of the role of tobacco taxes in starting and quitting smoking using duration analysis.
Chung et all, 1991	Examination of the length of time until an inmate is arrested after being released from prison.
Dolton and von der Klaauw, 1995	Modelling the exit of teachers from their profession.

CHAPTER 3

Variability of waiting time distributions by hospitals and doctors

3.1 Introduction

The analysis undertaken in the previous chapter has established that targets have played an important role in the cut of long waiting times. We conclude that, first, the period of time at which increasing probability of admissions takes place -illustrated by peaks in hazard curves- coincides with the prevailing waiting times target. It is also evident that the introduction of stricter waiting time targets leads to leftward shifts of the new peaks. Second, the patterns of admissions for elective surgery by a waiting list -illustrated by different survival curves- vary between different specialties, operative procedures and providers. Initially, the analysis concentrated on the differences between specialties as a whole, the second part revealed waiting times variations by type of operation and the last part explored differences between selected hospitals across England.

The aim of this chapter is to examine further the variation in the way hospitals and physicians manage their waiting lists. We pose questions in two directions: the first focuses on the behaviour of the hospitals. How do trusts admit patients for elective surgery? Are the waiting time distributions of patients of distinct hospitals exhibiting similarities? What are the commonest patterns of wait distributions? How have waiting time distributions evolved through years? The second direction focuses on the behaviour of physicians. Is there variation in the way doctors manage their waiting lists? How do doctors with similar activity levels regulate the flow of their patients? What is the behaviour of doctors with respect to elective admission across time?

This chapter is also motivated by the small area variation literature that emphasizes large differences in the utilization rates of medical services between geographical regions[1]. Even when controlling for age and sex, a great proportion of variation in the use of hospitals remains (e.g. number of admissions, length of stay, average expenditure per patient and input of physician effort). Possible explanations proposed by this literature include the different way medical teams practice medicine[2], differences in the incidence and prevalence of the disease, socioeconomic or ethnic characteristics of the population and different supply of health care resources. Here we focus on revealing variation in the waiting time distribution of patients and hence the admission rates of different hospitals and physicians.

A more detailed investigation of hospitals' behaviour and the identification of distinct admissions patterns sets the basis of Chapter 4. In Chapter 4 we develop a supply-side theoretical model for the optimal admissions behaviour of hospitals and the derived waiting time distribution. Based on the disaggre-

[1]Indicatively, Wennberg and Gittelsohn (1973) Folland and Stano (1990), Cohen *et al.* (1992), Eibich and Ziebarth (2013).

[2]For example, different surgical styles, different beliefs among doctors regarding the efficacy of procedures and more defensive medicine against a more aggressive view.

gated and detailed empirical observations of Chapter 3, we can replicate these distinct patterns in our theoretical model, and most importantly, relate them to particular supply-side factors.

In addressing these questions, there is an additional aim, to assess whether or not duration analysis should be more widely applied to waiting time data and thus represent a useful tool for systematic exploration of the hospitals' waiting time distributions and their evolution through time. A simple average waiting time cannot reveal the ways a hospital employs to change its behaviour in order to face the problem of long lists and waiting times. In other words, we also aim at evaluating the potential role of duration analysis to capture the changes in hospitals' tactics, that are expressed by changes of the waiting time distributions of their patients.

The present study also implements the techniques of duration analysis to HES data, yet, for an expanded time period that ranges from 1997 to 2005. Focusing on hospitals, the investigation involves comparisons by overall waits, by specialty and operative procedure (thus moving towards a less aggregated analysis of the waiting time distributions). We begin with a more detailed examination of the waiting time distributions of two of the seven trusts introduced in the previous chapter (part I). This study further attempts to compare the wait distributions of hospitals controlling for attributes such as size, type and performance rating (part II). The final part of the analysis on hospitals is devoted to the evolution of their waiting time distributions over time (part III). Focusing on physicians with similar activity levels, again, the examination consists of comparisons by specialty and operative procedure.

Specifically, the analysis comprises of the estimation of survival and hazard curves performed either in a cross-sectional (comparisons of a set of hospitals/doctors at specific years) or an across time framework (waiting times distributions of one hospital/doctor through years). The selection of this ap-

proach is supported by two arguments; the first emerges from the technical advantages of both the long-rank test, that succeeds in revealing statistically significant differences between distinct survival curves and the non-parametric estimation of hazard functions that exposes differences in the probability of patients' admissions. Since there are no known mathematical functional forms to fit the empirical admission patterns as waiting time increases, application of non-parametric analysis is advantageous. The second reflects the nature of the available HES data that do not include variables on hospitals characteristics (e.g. information on capacity, hospital workload and finances). The study by Propper *et al.* (2010) uses data from various sources[3], yet there is a risk of under- or over-estimation of results due to the lack of proper matching between data from different sources. This is the reason why we do not perform any further regression analysis.

The rationale of this chapter of the thesis lies in exposing detailed patterns regarding the shape of survival and hazard curves of patients' waits. It is an informative piece of investigation as we learn more about the hospitals' and doctors' behaviour on waiting list administration and patient admittance for treatment. Using trusts and consultants as units of analysis allows us to observe in a more disaggregated level how health care providers with different characteristics perform within the NHS. On top of that, it is among their own interest to show performance excellence and achieve NHS goals by supplying high quality services when needed. No one can argue against the imperative role that 'time until treatment' plays for patients, institutions, doctors and the state itself.

The main finding of this chapter is a significant variation in the waiting time distributions of patients among hospitals and even more when comparing

[3]HES database of England, the Scottish morbidity record for Scotland, census data for England, public expenditure statistical analyses data, ONS population trends, data on workforce and finances of English hospitals from the Department of Health.

doctors. However, one can rightly ask which are the factors responsible for producing specific distributional patterns. What is the role of hospital budget, capacity levels, cost of treatment and national targets in managing waiting lists? Which attributes could account for changes in the wait distributions? Development of a theoretical supply model that attempts to interpret the mentioned differences of the distribution of patients' waiting times is presented in Chapter 4.

The structure of the rest of the chapter is as follows. A brief presentation of the data set is presented in the next section, followed by a graphical illustration of the results and relevant concluding remarks.

3.2 Data

We further obtained HES data on the specialties of general surgery, trauma and orthopaedics and ophthalmology for 9 years (1997/1998 to 2005/2006)[4]. A graphical presentation of yearly admissions by specialty and type of admission is presented in Figures 3.1 and 3.2, while Figure 3.3 demonstrates the kernel densities of elective waits (admission method: waiting list) by year.

The majority of the procedures are general surgeries followed by orthopaedic ones while ophthalmologic procedures have the smaller percentages. Yet, there is a steady increase on admission numbers of all specialties through years. Moreover, there is evidence of a decline of patients admitted via waiting lists with a simultaneous raise of booked and planned admissions.

Although data were truncated at 730 days, estimation of the kernel densities reveals the before-mentioned point that waiting times are positively skewed.

[4]The expressions '1997/1998' and '1997' will be used interchangeably throughout the thesis. They both refer to the financial year starting in April 1997 and finishing in the end of March 1998. The same stands for all the years of the analysis.

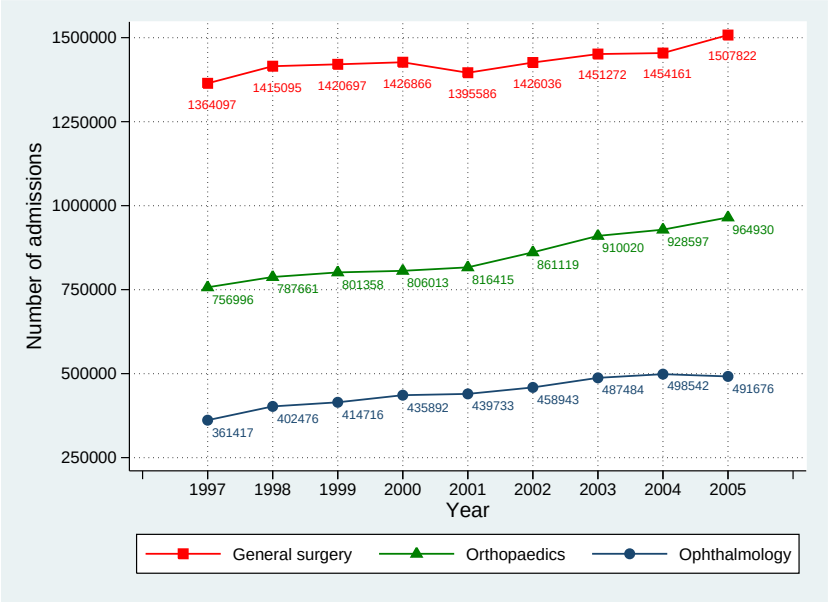


Figure 3.1: Number of yearly admissions by specialty.

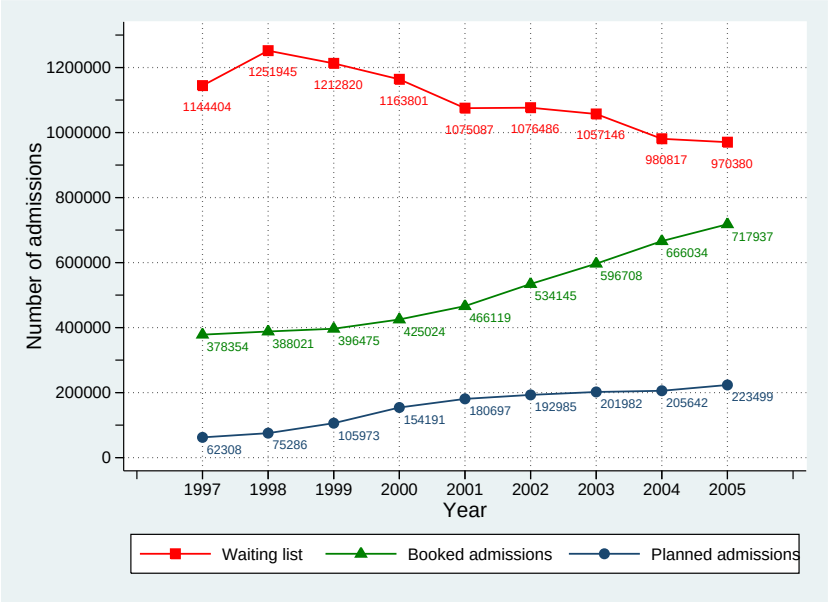


Figure 3.2: Number of yearly admissions by type of admission.

The first part of the distribution represents patients whose waits are quite short. It is clear that there is an increase of the number of patients with such waits admitted for surgery. The second part of the distribution illustrates the

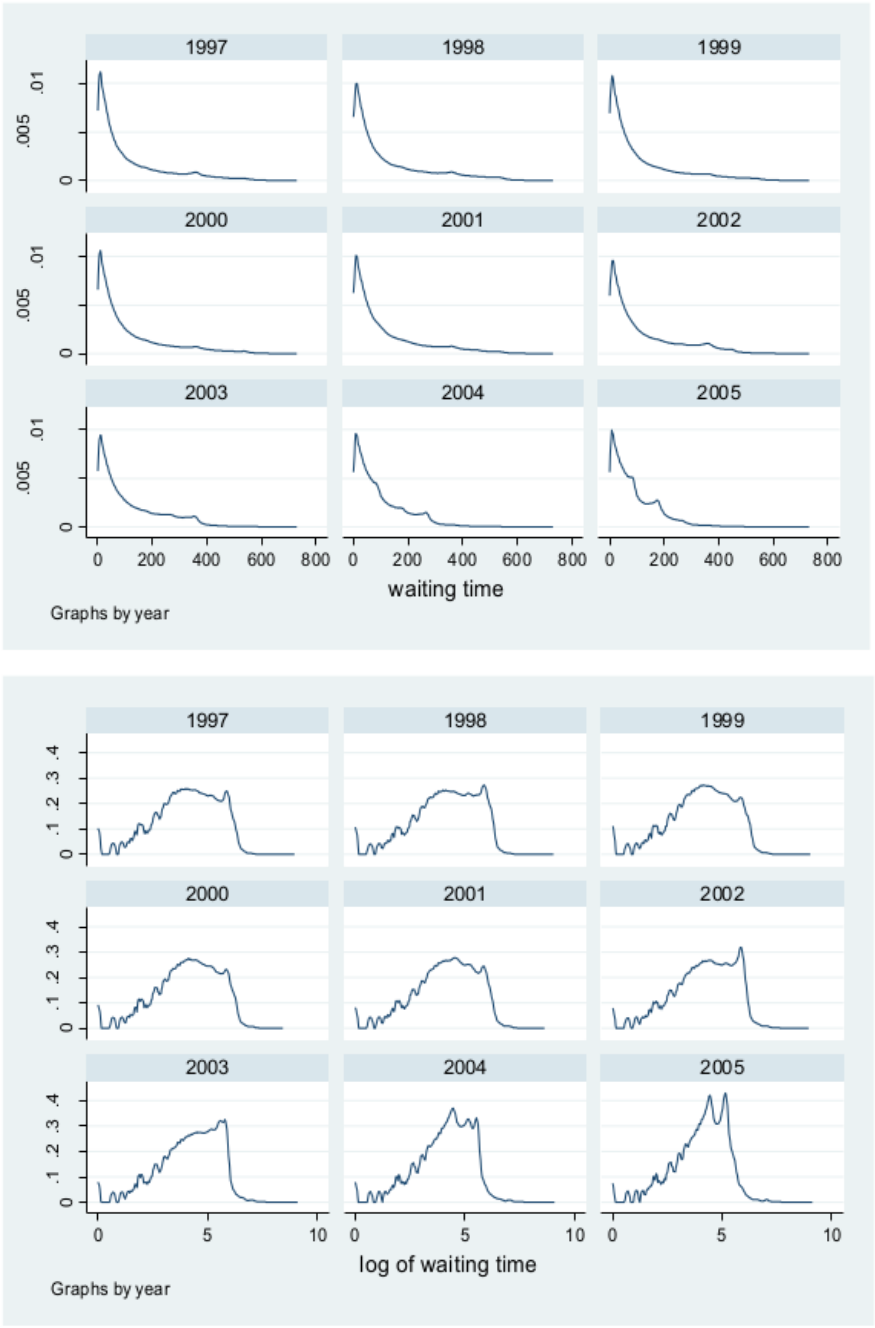


Figure 3.3: Kernel densities of waiting times by year, waiting time *(top)* and log of waiting time *(bottom)*.

gradual decrease of the number of patients who have waited longer periods. Long right tails reveal that some patients wait far too long to be treated. It is obvious that for years 2003, 2004 and 2005 the second part of the distribution changes. Moreover, the graphs of the log of waiting times gradually shift leftwards at the right tail suggesting a cut of the long waiters. However, although it is obvious that there are changes on waiting time distributions over years, yet this illustration cannot easily expose them. A more informative presentation that uses duration analysis techniques is provided in the next section of the chapter.

Since the aim of the analysis is to examine the level of variability among hospitals, data are further classified according to size and type of NHS trust. Taking under consideration information on NHS trust clustering by the Department of Health[5] we classify hospitals by size (large acute, medium acute, small acute) and by type[6] (acute, specialist, teaching). After excluding trusts with missing data over the 9 year period, a set of 52 hospitals remains for further examination. Moreover, based on the performance star-rating system adopted by NHS ranging from 2000 to 2005, we develop an additional grouping of hospitals consisting of good and bad performers.

Clusters reflect the largest part of hospital's activity; in acute trusts, a large proportion of their expenditure covers acute activity while specialist trusts offer specialist acute services (e.g. orthopaedic and children's hospitals). We define as excellent performers those trusts that acquired a 3 star rating for all five years the performance star rating system was in action and as bad performers those that had the worst evaluations (zeros and 1 star ratings). In particular, our set of hospitals consists of: seven large acute trusts, thirteen medium acute hospitals, twelve small acute hospitals, twelve teaching hospitals (five in London

[5]http://ratings2004.healthcarecommission.org.uk/cluster.asp. We further matched this information with a list of trust clusters we obtained from the Department of Health.

[6]The HES variable on the type of provider has not been helpful.

and seven outside London) and eight specialty hospitals (four orthopaedic, three children's and one specialist for reconstructive surgery and rehabilitation).

In this chapter, we are focusing on consultant teams[7] that exhibit high activity levels and particularly on those that performed more than 1000 surgeries per year for at least one year of our 9-year dataset.

3.3 Results

Our results comprise of two subsections. The first part examines the waiting time distributions of patients either among different hospitals in a cross-sectional framework and of the same hospital over time. The second part focuses on the investigation of waiting time distributions by physicians.

3.3.1 Behaviour of trusts

This subsection consists of a much less aggregated analysis for two of the seven trusts examined in Chapter 2, the cases of Birmingham Heartlands & Solihull and Royal Free Hampstead (part I), an illustration of distinct patterns of wait distributions through a selection of survival and hazard curves of specific trusts grouped by size, type or performance rating (part II). Lastly, an example of evolution of waiting time distributions over time is presented using data for Hammersmith hospital (part III).

Part I: Less aggregated analysis by hospital

The case of Birmingham Heartlands & Solihull

Figure 3.4 shows the survival and hazard curves of overall waiting times for elective surgery for Birmingham Heartlands & Solihull. For 2001/2002, at time 0, all patients are on the list, at around 2 months (57 days) 50% of them have

[7]The HES variable is described as 'Pseudonymised consultant team code'.

been moved from the list to be treated and at around 4.5 months (143 days) the proportion increases to 75%. The same pattern characterises the waiting time distribution of elective patients for 2002/2003. Hazard curves reveal notable peaks at a little more than 15 months for 2001/2002 and 12 months for 2002/2003. It is obvious that this trust attempts to adjust to national targets by changing the probability of moving patients from its lists for admission. What is not clear, though, is whether it maintains the same behaviour for all specialties and different operative procedures.

Analysis at the level of specialty and type of operation illustrates variability in trust responses (Figures 3.5 and 3.6). Figure 3.5 demonstrates the estimated survival and hazard functions by different specialties for years 2001/2002 and 2002/2003. In particular, patients waiting for general surgery tend to wait shorter periods than patients scheduled for orthopaedic or ophthalmologic procedures. According to 2001/2002 survival curve, 50% of patients is moved off the general surgery list at around 1 month (34 days), the orthopaedic list at 4 months (111 days) and the ophthalmologic list at 6.5 months (194 days). At around 8 months the survival curves of the last two specialties intersect. Similar results are observed for the subsequent year, yet the difference between orthopaedic and ophthalmologic specialties diminishes.

In addition, the probability of admission does not remain constant and exhibits different patterns for each specialty. For general surgery, increased waiting list activity is observed as peaks in the curve for people waiting 3, 8 and 11 months for 2001/2002 and 2, 6, and 12 months for 2002/2003; for orthopaedic surgeries, between 12 to 15 months for 2001/2002 and at 12 months for 2002/2003; for ophthalmology the greater peak is located at around 15 months for 2001/2002 and a little less than 12 months for 2002/2003. An interesting finding in the bottom right graph of Figure 3.5 is the presence of a late peak of orthopaedics at 547 days -and an increasing hazard at 730 days-

that probably represent the hospital's efforts to clear the last occupants of the list. This particular peak of orthopaedics molds the overall survival curve of Birmingham Heartlands & Solihull for 2002/2003[8]. Two insights can be drawn from the specialty level analysis. Firstly, it is clear that the trust does not adopt the same behaviour in managing different surgical lists; general surgery waiting lists follow a very different pattern than the other two specialties. Secondly, it shows the effects of waiting times targets; the probability of admission for those whose wait approaches a target increases and falls when their wait exceeds the target.

Figure 3.6 shows the estimated survival and hazard functions by different selective operative procedures. The shorter waiting times are for cholecystectomy, inguinal hernia and varicose vein procedures and the longest for lens prosthesis and hip replacement. Furthermore, it is worth emphasising that the three general surgical waiting lists do not consist of patients waiting more that 1 year, while lens prosthesis and hip replacement do. The hazard curve for 2001/2002 reveals the following patterns: for cholecystectomy, peaks are at 2 and 6 months, for inguinal hernia at 2 and 8 months, for varicose vein at 2 and 7 months, for lens prosthesis at almost 15 months and for hip replacement at between 12 and 15 months. The hazard curve for 2002/2003 demonstrates that as the targets become tougher, the peaks change towards the lower waiting times. A peak at exactly 12 months, which is the target of that year, is observed for hip replacements and a little less than 12 months characterises lens prosthesis.

[8]See Figure 3.4.

The case of Royal Free Hampstead

Birmingham Heartlands & Solihull exhibits great management activity to tackle excessive waiting times for its patients. Quite different behavioural responses are observed by the Royal Free Hampstead trust in London. Analysis of the overall waiting times of the trust reveals that Royal Free patients have to wait a little longer for treatment, relative to the other hospital, as the rate of admission changes slower (Figure 3.7). For 2001/2002 there is a small peak at around 3 months and another one at around 17 months while for the following year the hazard curve remains almost constant after a small increase at the beginning. Therefore, after about four months, the probability of admission for elective surgery is independent of the time patients spent on waiting lists.

According to official returns of waiting lists for elective surgery[9], Birmingham Heartlands & Solihull had achieved the national waiting time targets for 2001/2002, while Royal Free had not. One limitation of the official waiting list statistics of that year was that broad time intervals of waiting were used (eg. patients waiting for 12-17 months). Thus, we cannot calculate exactly the number of people waiting more than 15 months. However, none of their patients had to wait more than 12 months at the end of 2003, thus both hospitals achieved the waiting time target for the following year. Conversely, based on NHS performance ratings, both trusts had no inpatients waiting longer than the standards for both years tested[10].

Yet, less aggregated analysis at the level of the three specialties reveals greater variation in the shape of survival and hazard curves (Figure 3.8). General surgeries have the quickest admissions compared to the other two specialties. However, ophthalmologic procedures are managed quicker for patients having waited above 456 days in 2001/2002 and above 547 days in 2002/2003.

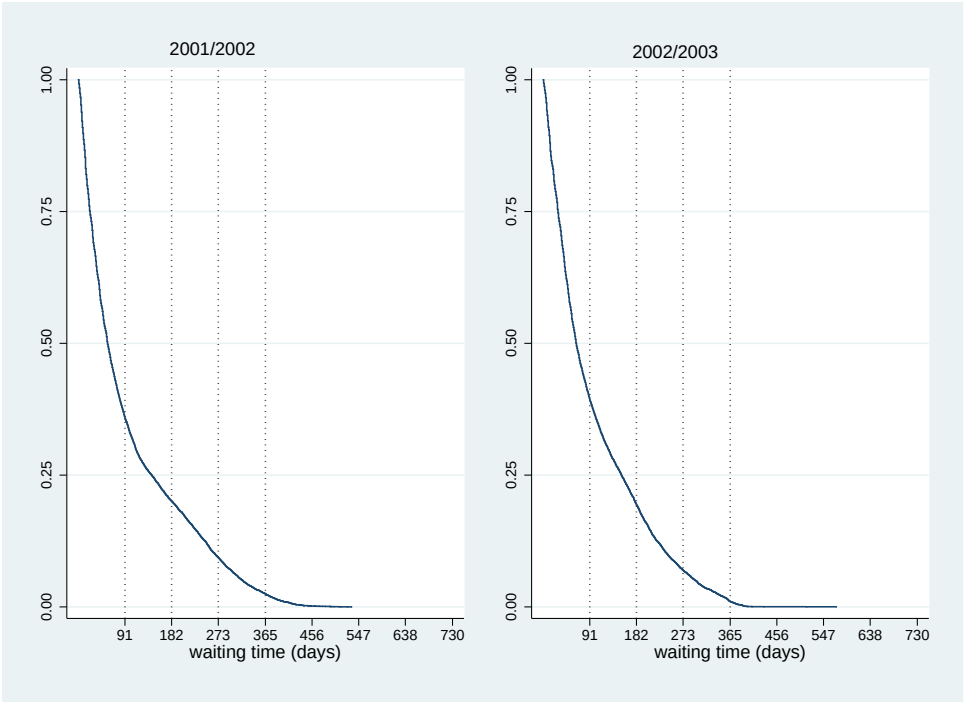
[9] http://www.performance.doh.gov.uk/waitingtimes/index.htm

[10] Department of Health, NHS Performance Ratings 2001/2002, 2002/2003.

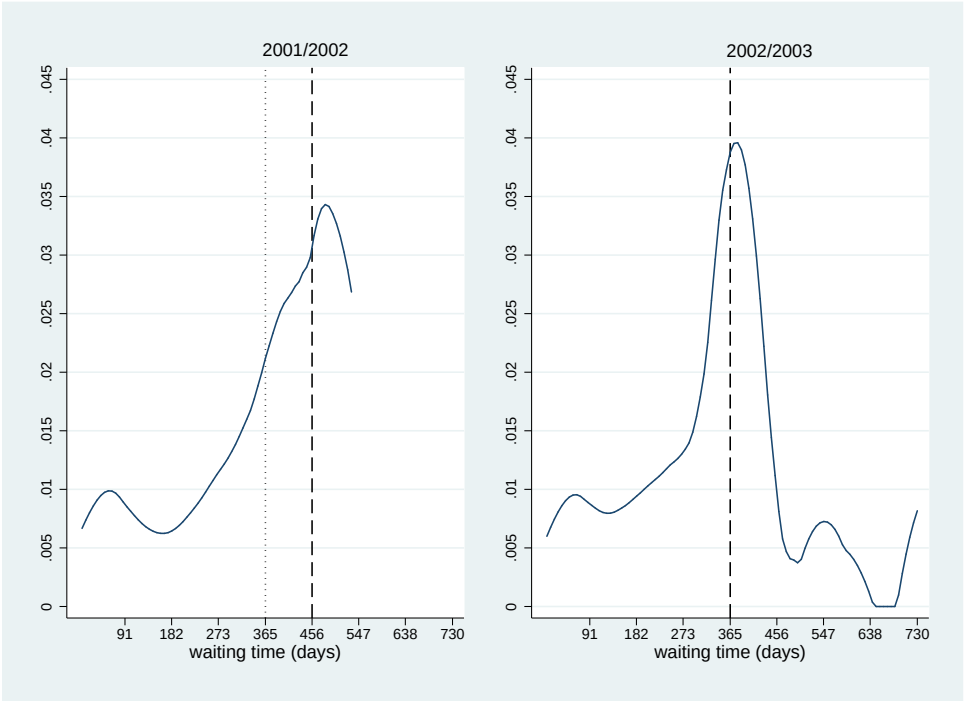
Interestingly, comparing the two survival curves of ophthalmology, we find that there is an important decrease of long waiters in the second year and a simultaneous increase of patients waiting less than 182 days. Similar to Birmingham Heartlands & Solihull the survival curves of ophthalmology and orthopaedics intersect.

In 2001/2002, the probability of admission for general surgeries spikes at 3 and a little after 15 months while a year later the first peak is of much higher intensity and the second peak (noticeably milder) moves forward. The other specialties follow different patterns with the presence of higher intensity peaks at around 456 days for ophthalmologic and 547 days for orthopaedics that do change significantly the next year.

At the less aggregated level of waiting lists for different operations, the image is much clearer (Figure 3.9). An seen for Birmingham Heartlands & Solihull, the three types of general surgery exhibit quicker admission rates, however, for 2001/2002, patients waiting more than 273 days for lens prosthesis are admitted faster than varicose veins and cholecystectomy patients. For 2001/2002, varicose vein procedures have constant hazard rates, cholecystectomy is characterized by a peak at 15 months while lens prosthesis exhibits a high intensity peak at the same period. On the other hand, the hazard curve of hip replacements is monotonically increasing with maximum probability of admissions at 638 days. In the case of inguinal hernia, we observe an early peak at around 3 months and after remaining constant for some time it starts increasing again. For 2002/2003, the hazard curves follow completely different patterns.

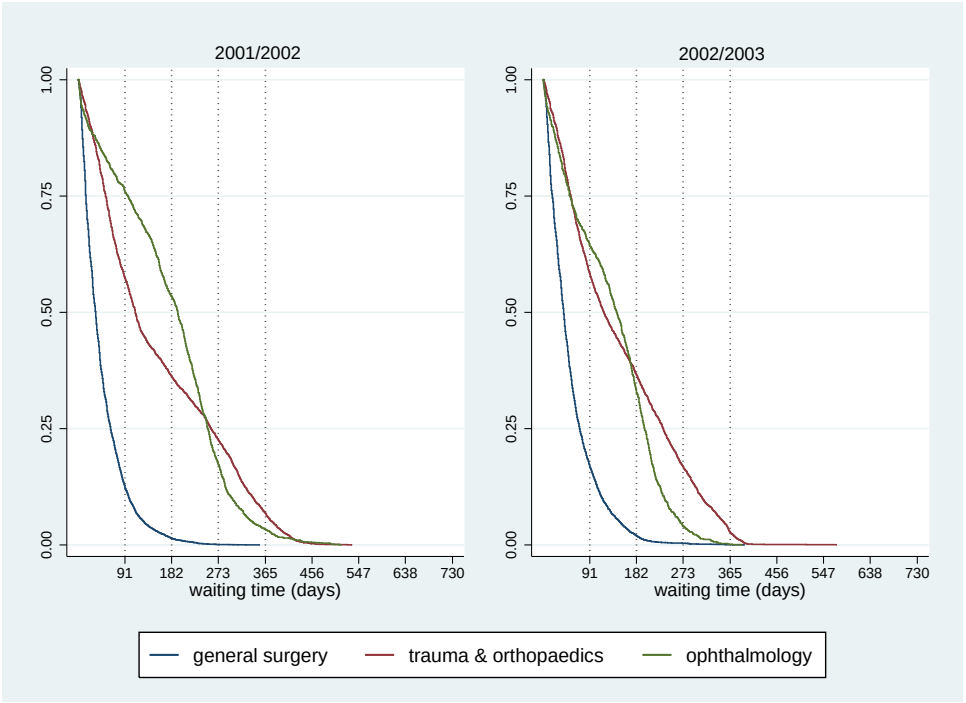


(a) Survival curves

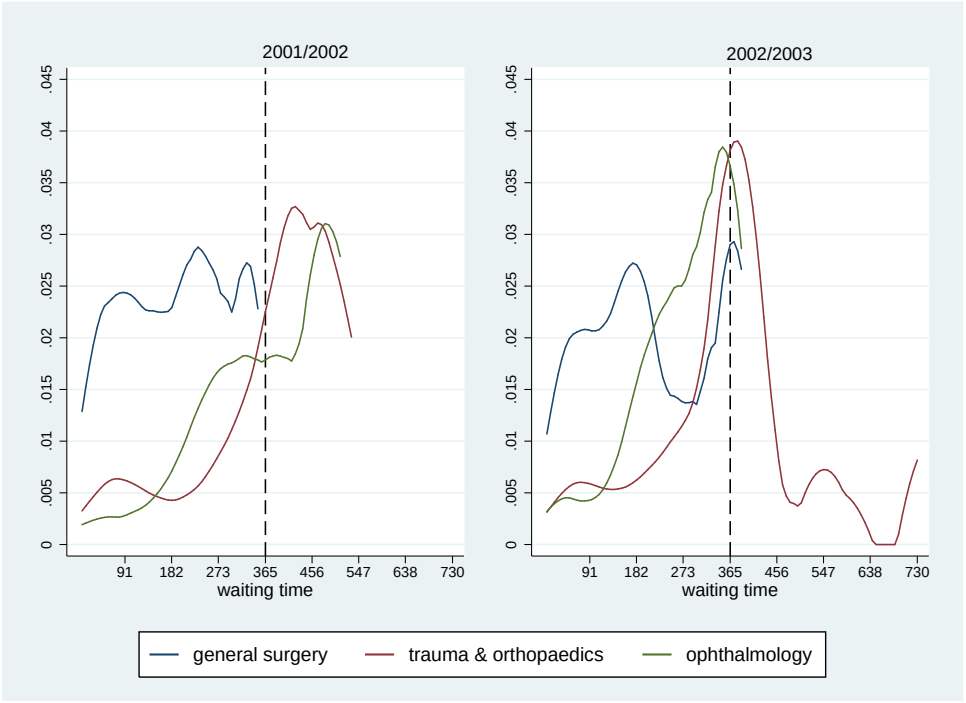


(b) Hazard rates

Figure 3.4: Overall waiting times of Birmingham Heartlands & Solihull for years 2001/2002 and 2002/2003.

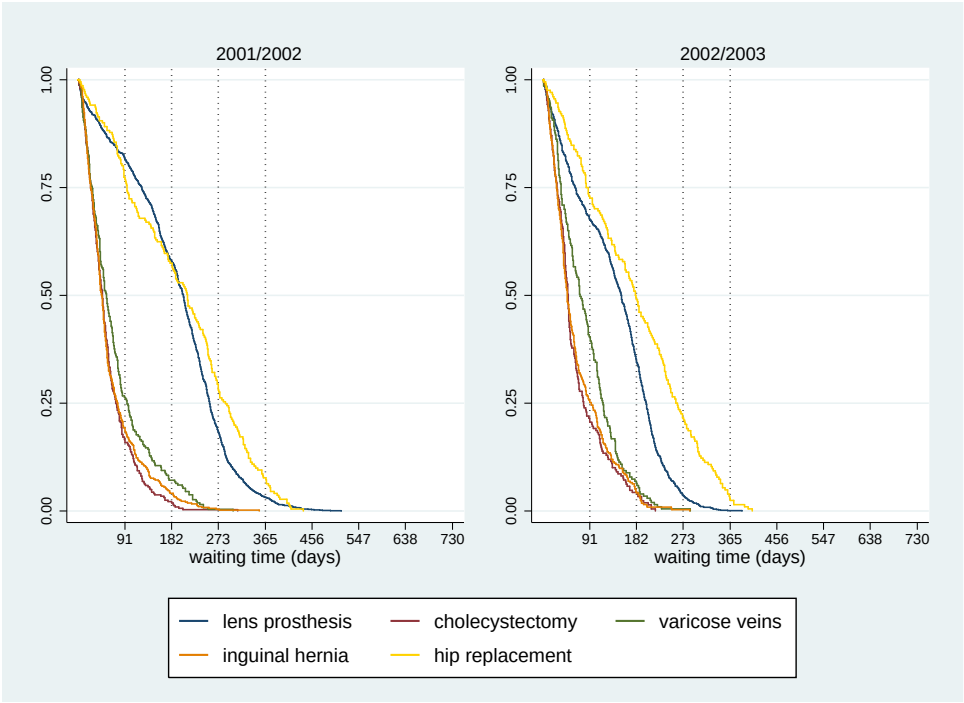


(a) Survival curves

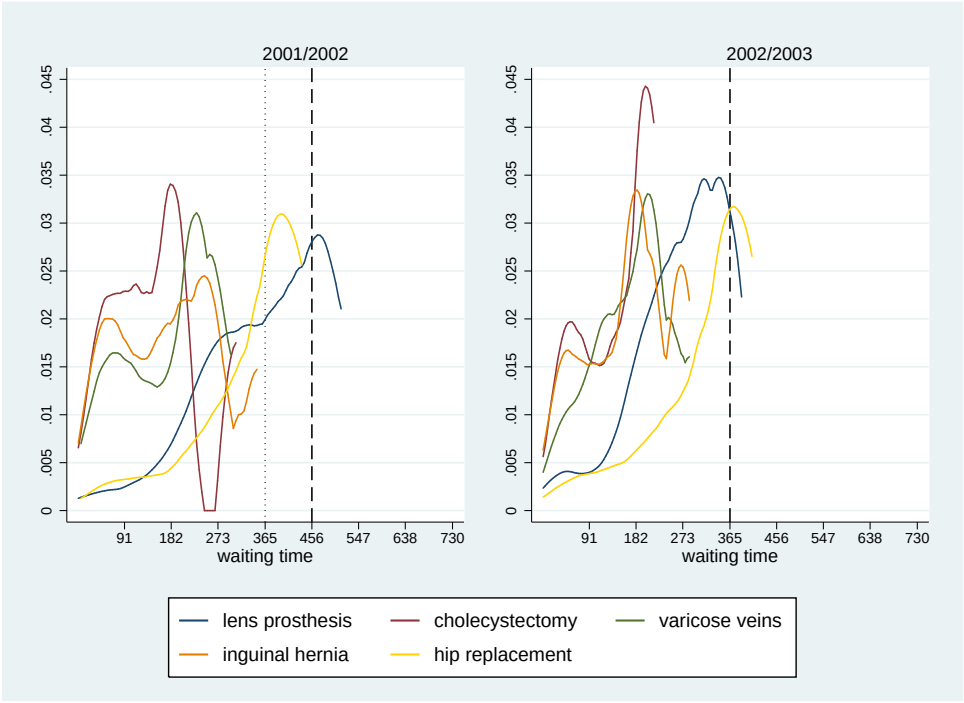


(b) Hazard rates

Figure 3.5: Waiting times by specialty of Birmingham Heartlands & Solihull for years 2001/2002 and 2002/2003.

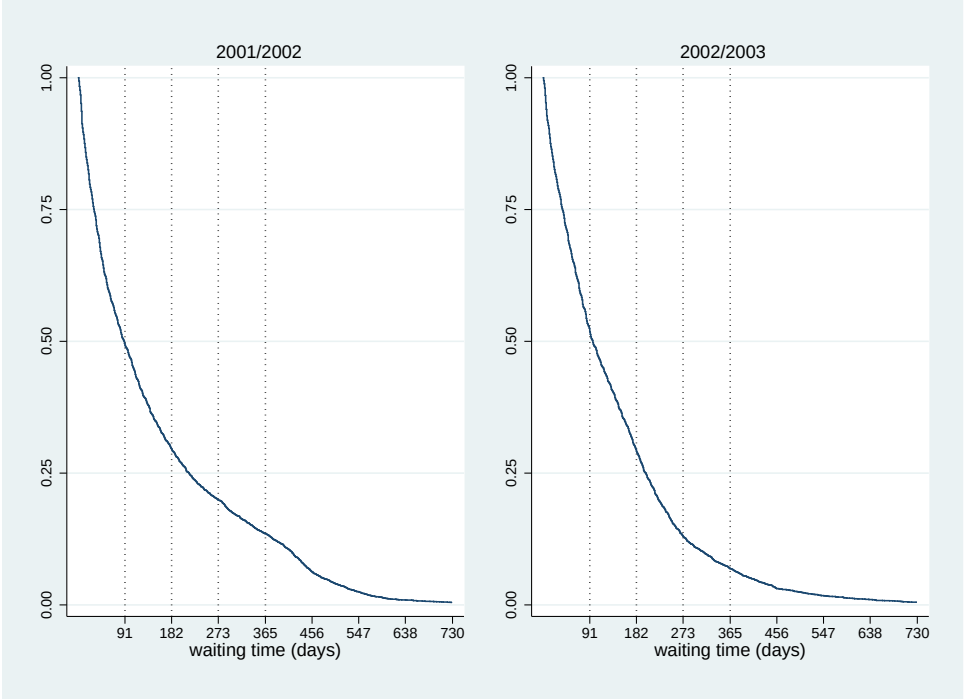


(a) Survival curves

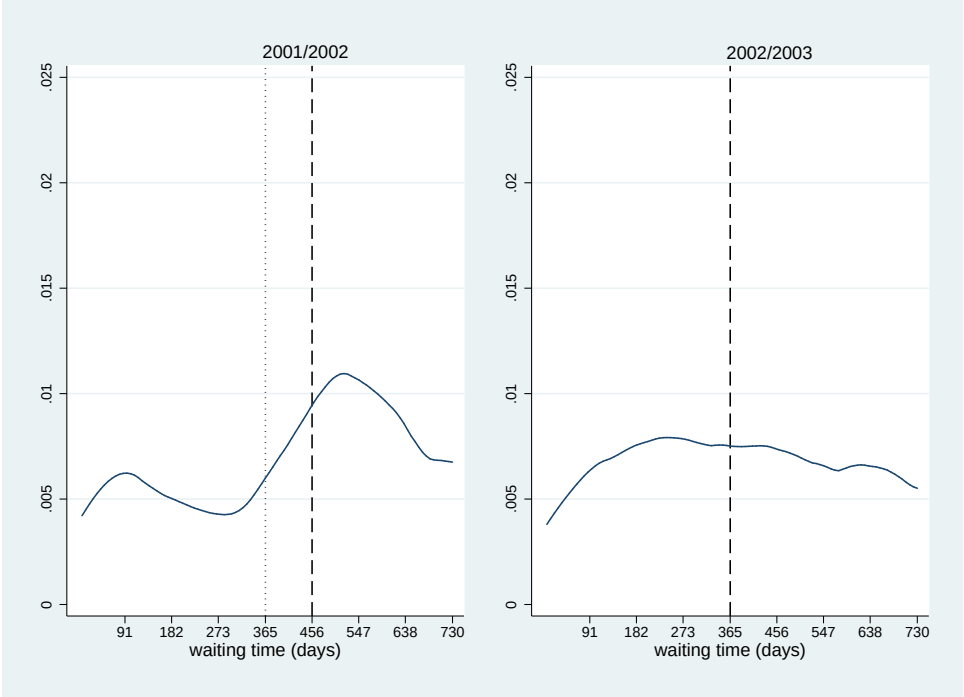


(b) Hazard rates

Figure 3.6: Waiting times by operation of Birmingham Heartlands & Solihull for years 2001/2002 and 2002/2003.

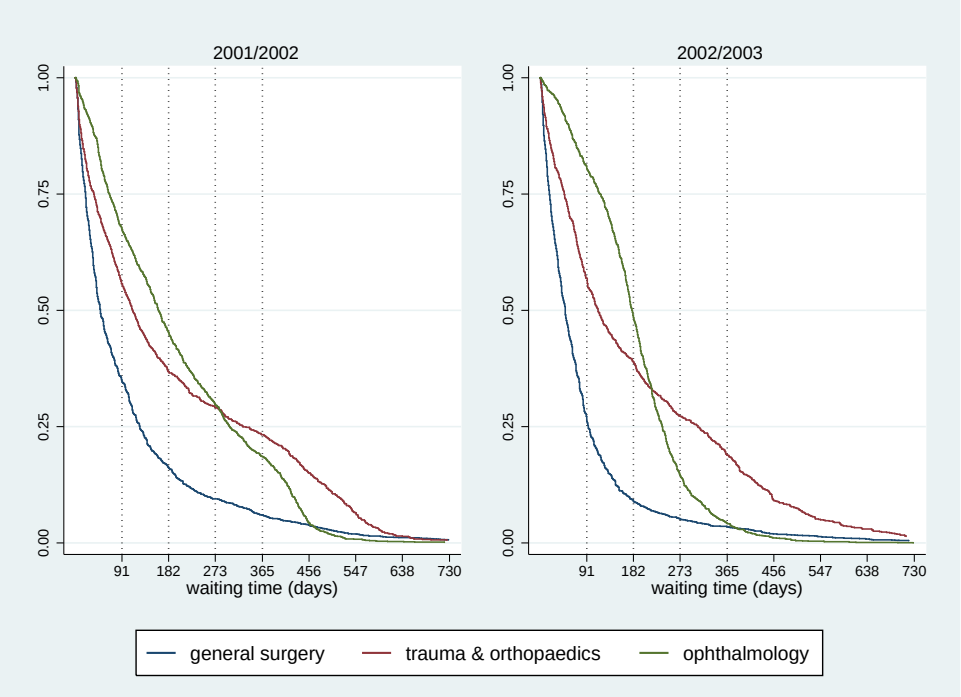


(a) Survival curves

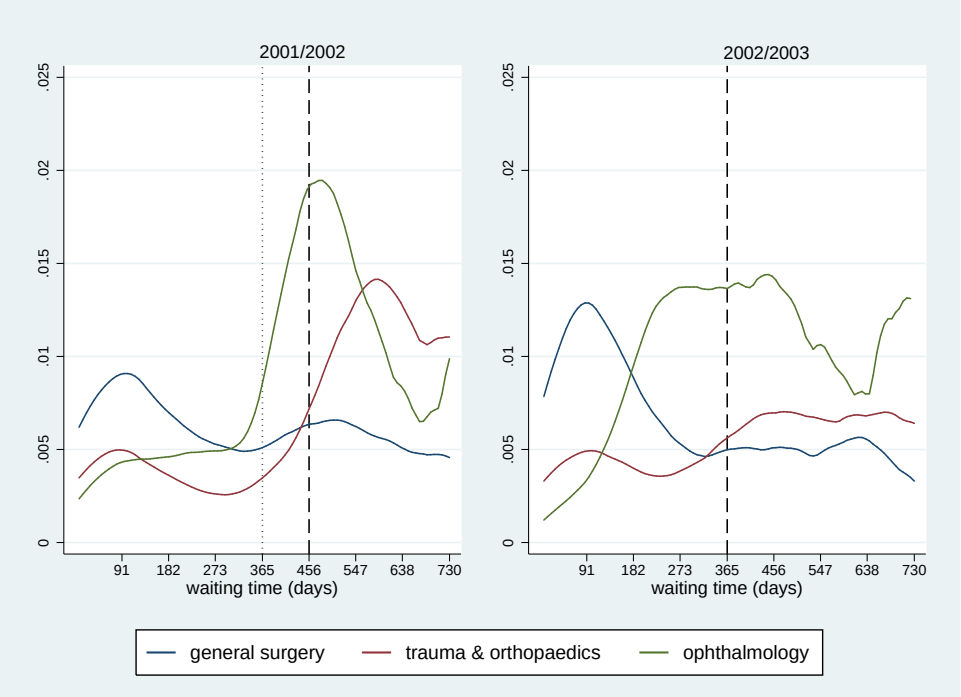


(b) Hazard rates

Figure 3.7: Overall waiting times of Royal Free Hampstead for years 2001/2002 and 2002/2003.

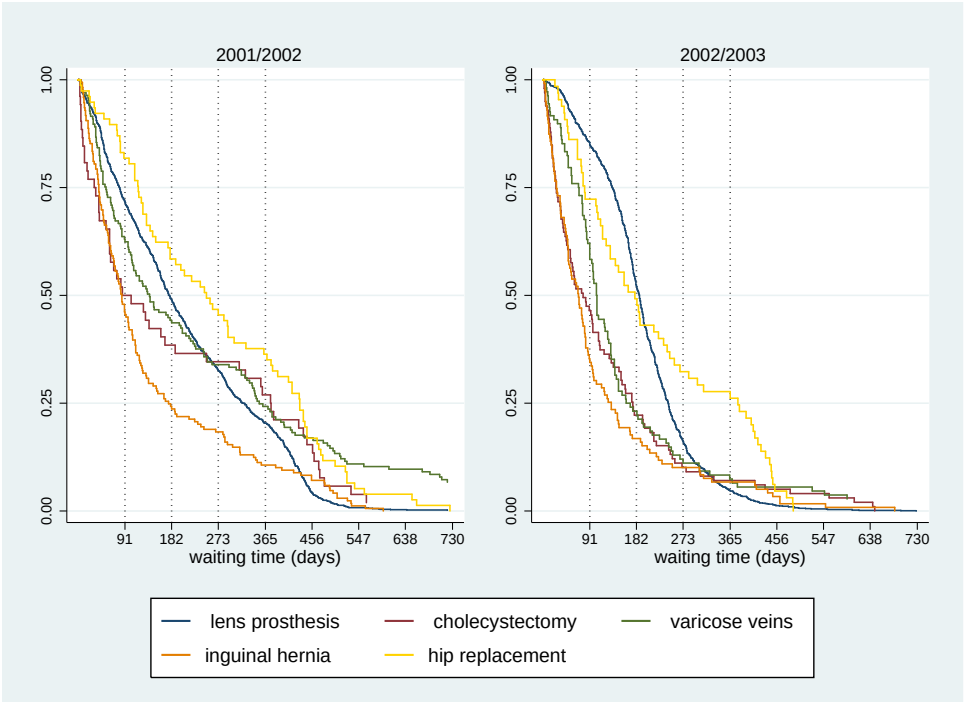


(a) Survival curves

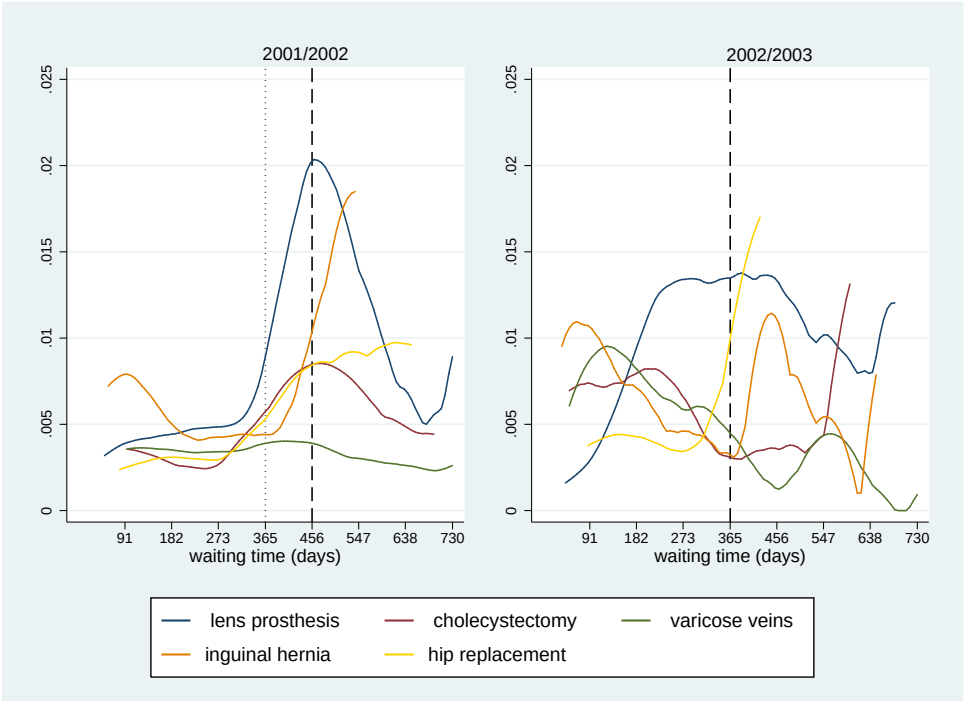


(b) Hazard rates

Figure 3.8: Waiting times by specialty of Royal Free Hampstead for years 2001/2002 and 2002/2003.



(a) Survival curves



(b) Hazard rates

Figure 3.9: Waiting times by operation of Royal Free Hampstead for years 2001/2002 and 2002/2003.

Part II: Comparison across different types of hospitals

The next step in our decomposition comprises of comparisons of different groups of hospitals and aims at revealing patterns in their survival and hazard curves. Figures 3.10 and 3.11 illustrate the survival curves of seven large acute trusts for 1999/2000 and 2000/2001 respectively.

As mentioned already, a survival function is represented as a monotonically decreasing function that declines from 1 to 0 as waiting time increases. Hence, all survival curves initiate at 1 as all patients are initially on the list and as they are admitted for treatment the curves gradually decrease until they reach 0 when the list clears. However, distinct patters on the wait distributions are evident. For the financial year 1999/2000, as the survival curves of Bradford and Berkshire & Battle are far away from the origin they exhibit the worse admission rates. On the other hand, the rest of the curves are closer to the origin with the one of Wirral showing quicker admission rates.

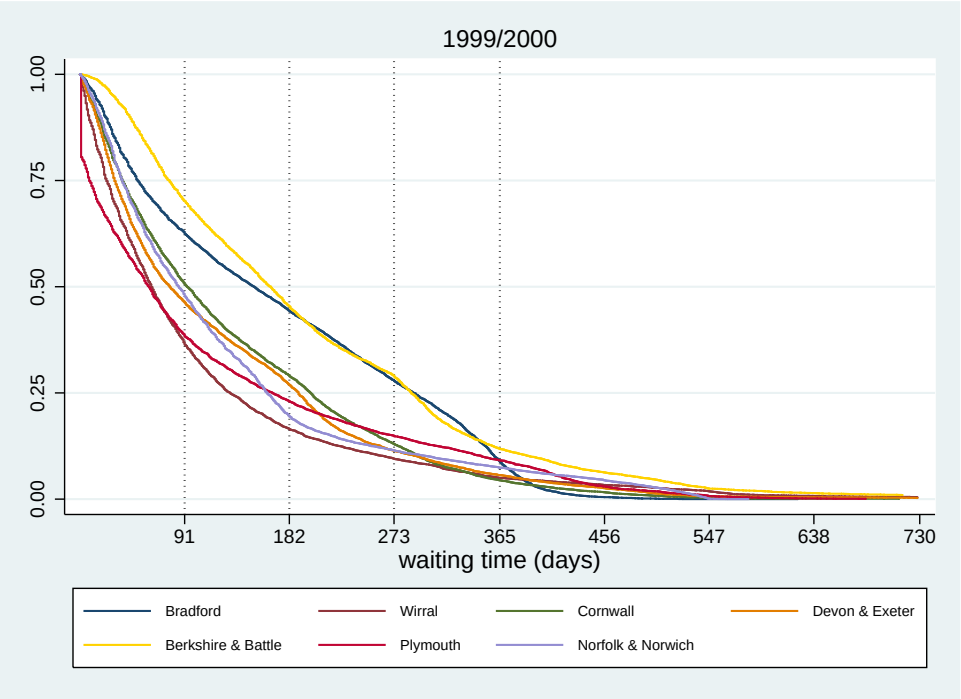


Figure 3.10: KM curves for large acute hospitals for 1999/2000.

The log-rank test for equality of the survivor functions demonstrates statistically significant differences between waiting times for the seven trusts.

Additional information can be derived from a closer observation of the curves and their various shapes. Let us first discuss the most important aspects of the observed variations. Using the terminology of Weon and Je (2011, 2012), survival curves vary in terms of 'shape' and 'scale'. Scale refers to changes in the position of the curve (closer or further away from the origin), while shape refers to changes in the slope of the survival curve. Clearly, the slope of the survival curve is changing as duration increases; however, there are plenty of instances where the change in the slope is more abrupt. The sharpest change in slope is evident with a change in curvature (sign of the second derivative). However, we also observe cases where, without a change in the sign, the magnitude of the second derivative changes drastically at particular points in time. Note that these prominent changes in the shape of the survival function are captured as spikes in the hazard curves.

Noticeable observations in our example (Figure 3.10) are: (i) There is a steep fall (up until around 20 days) of the survival curve for Plymouth trust which implies augmented removal of short waiters off the list. After this striking fall the curve decreases slower until it reaches 0. (ii) The majority of the survival curves decrease monotonically without any change in curvature (they remain convex, e.g. Wirral Hospital). (iii) Taking a look at the admission rates of Berkshire & Battle a different pattern emerges; the curve is first concave and for patients waiting more than 273 days it becomes convex. Finally, due to the fact that survival curves exhibit various shapes many of them intercept with others in one or more points.

The tactics of hospitals are different for 2000/2001 (see Figure 3.11). Plymouth behaves much worse relative to the previous year, especially with regards to the waiting times of patients until 365 days. Comparing its admission rates

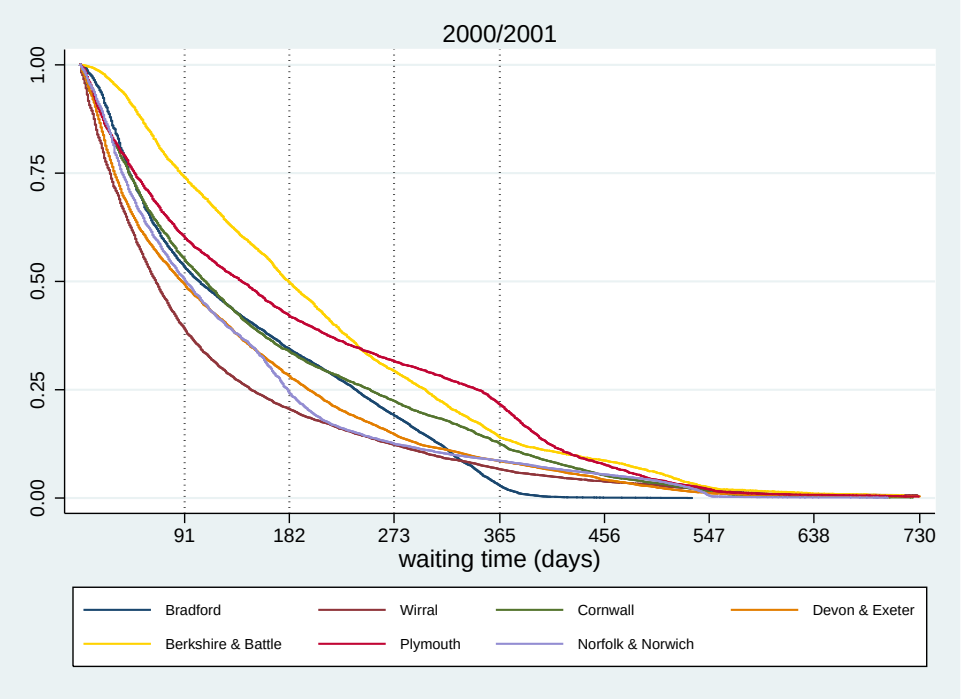


Figure 3.11: KM curves for large acute hospitals for 2000/2001.

with the ones of Berkshire & Battle we infer that although it admits quicker patients waiting up to 250 days it delays substantially the treatment of long waiters. It is worth mentioning that its survival curve changes curvature more than once. Furthermore, the managing of the list could be identical in specific parts of the distributions as in the cases of Bradford and Cornwall (their curves coincide up until 200 days) or with Wirral and Norfolk & Norwitch (they have similar admittances in middle of the distributions).

Figure 3.12 illustrates the hazard rates for the seven large acute trusts for 1999/2000 and 2000/2001. In the first year, two trusts exhibit high-intensity peaks, that is increasing probabilities of admissions, for patients waiting 547 (Norfolk & Norwitch) and 456 days (Bradford). There are lower-intensity wider peaks such as the ones by Plymouth and Devon & Exeter. Additionally, there are trusts with almost constant hazard rates (Berkshire & Battle and Wirral).

Figure 3.12: Hazard curves for large acute hospitals, 1999/2000 *(top)* and 2000/2001 *(bottom)*.

In 2000/2001, some peaks shift leftwards (Bradford and Plymouth), some shift rightwards (Cornwall) while others remain unchanged (Devon & Exeter and Wirral).

To sum up, it is evident that although all trusts are exhibiting high surgical activity levels (large acute trusts) they develop distinct waiting time distributions that change over years. This is also the case when we perform comparisons among medium acute, small acute, specialist, teaching and good/bad performance trusts. Selected diagrams from each category aim at exposing the great level of variation between their waiting time distributions. Log-rank tests for equality of the survival functions was performed for all comparisons demonstrating statistically significant differences between all relevant waiting times.

Figures 3.13 and 3.14 depict the waiting time distributions among thirteen medium acute trusts. Although hospitals exhibit similar activity levels they manage quite differently their waiting lists. In the first figure, for the year 1998/1999, there is one trust that admits 75% of its patients after having waited 3 months or less (Poole) while in other trusts (Royal Surrey County, Worthing & Southlands, Newham) the same percentage is achieved for patients whose waits are up to around 400 days. The rest of the hospitals are characterised by admission rates between the two extremes. An additional pattern not observed previously belongs to Walsall, whose curve falls linearly almost until 0, leaving a huge right tail representing the clearance of the last occupants of its list. In 2004/2005, all survival curves have shifted leftwards towards the origin and are much more concentrated than before, implying reduced waiting times for all waiters. Comparison of two hospitals that show similar behaviour at the beginning (Walsall and Bromley) shows different tactics later on with the former focusing on the long waiters and the latter on handling well the short waiters.

In line with the survival curves, the hazard rates among medium acute trusts exhibit great variability that ranges, in the first year, from distinctive peaks to

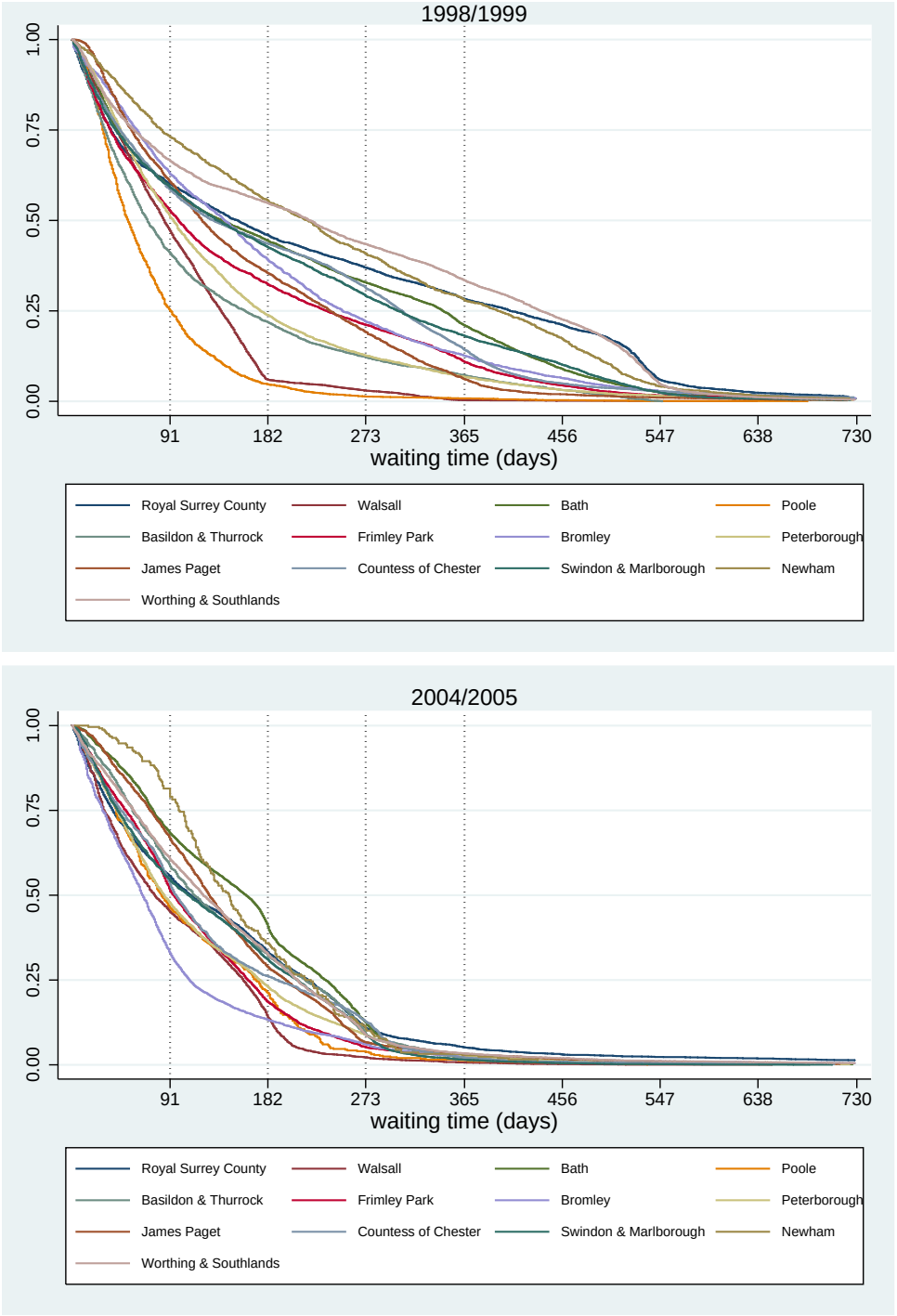


Figure 3.13: KM curves for medium acute hospitals, 1998/1999 *(top)* and 2004/2005 *(bottom)*.

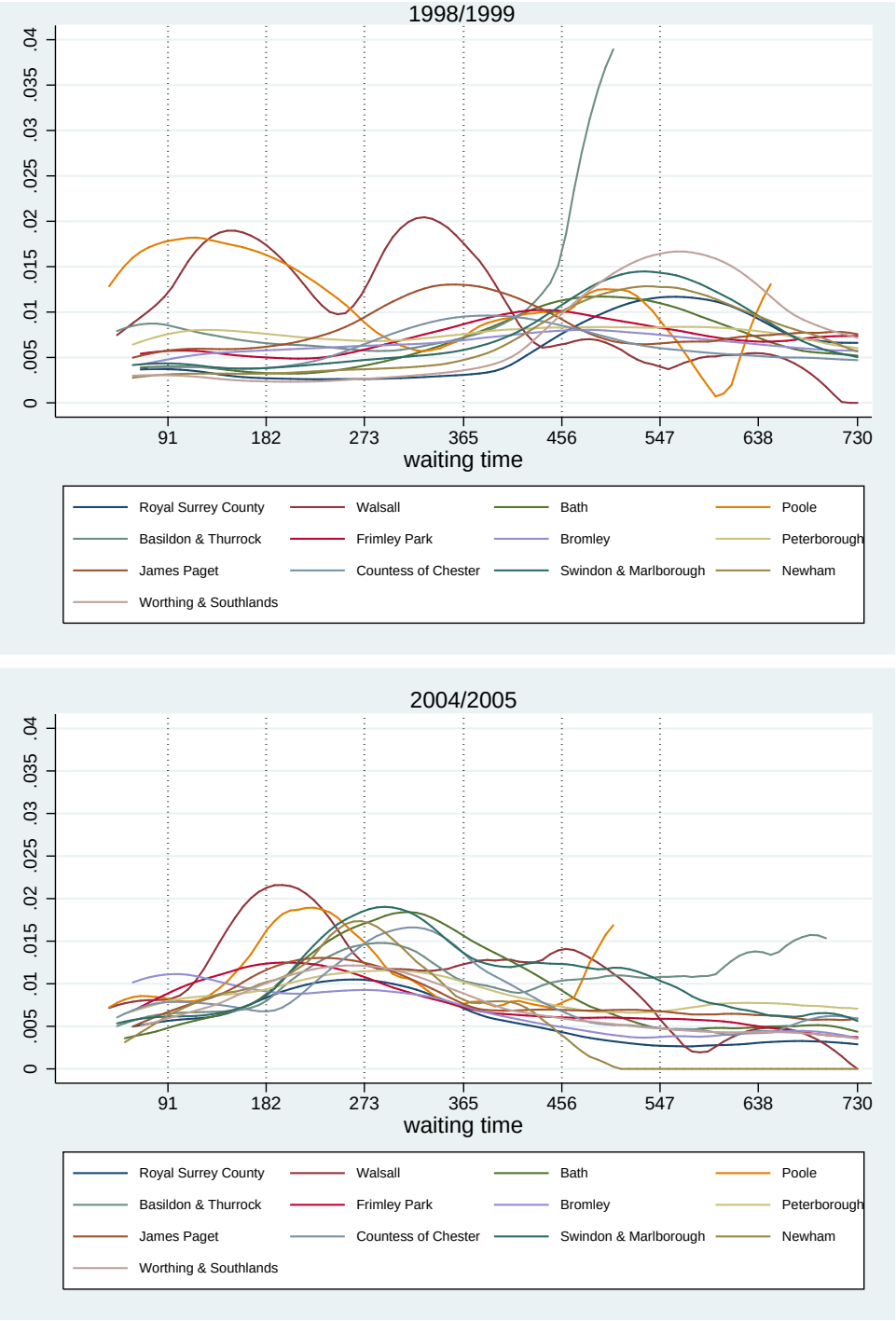


Figure 3.14: Hazard curves for medium acute hospitals, 1998/1999 *(top)* and 2004/2005 *(bottom)*.

multi-peak and quasi constant hazards. Walsall hospital exhibits two peaks, the first one corresponding to the visible change in the slope of the survival curve at about 180 days. Poole has a quite wide and smooth peak with high admittance rates spanning from about 91 to 182 days (consistent with a smooth and close to the origin survival function). The hazard of Basildon & Thurrock ends up at a steep increasing rate, although Bromsley appears to have a quite constant instantaneous admittance rate.

Further, in year 1998/1999 the majority of peaks are located around 547 days while 6 years later increasing probability of admissions is observed for people with shorter waiting times. It is clear that peaks have moved leftwards; one plausible interpretation of which could be attributed to the hospitals' efforts to meet national targets. As a result the majority of the peaks are now situated around 365 days, with some of course before and some after this point.

Figure 3.15 demonstrates the survival and hazard curves of twelve small acute hospitals for 2005/2006. Due to a smaller overall number of admissions, there are survival curves with visible steps, as is the case for East Somerset, South Warwickshire and Royal West Sussex in the first graph. Considerably quick admissions characterise East Somerset until the first three months of wait; only about 10% of patients wait for more than three months, but they take long to be treated. On the other hand, Royal West Sussex performs the worst; 50% of elective patients will have to wait more than 6 months in order to be treated. Patients are initially taken off the list at a decreasing rate and later on at an increasing one. The rest of the hospitals, clustered between those two, exhibit similar behaviour.

As regards to the equivalent hazard curves, the hazard curve for East Somerset starts off and remains much higher for until about three months. The very increased instantaneous probability of admissions observed is attributed to the fact that about 90% of the list is admitted for surgery within three months. It then decreases until it becomes zero between 182 and 273 days of wait; no patients are admitted during this period (this is also reflected in the constant survival). As the survival steps down, the hazard exhibits another peak of smaller intensity, and this process is repeated until the list is cleared. Many trusts have increased probability of admissions at around 6 months, which is the target of that year, yet there are others with peaks at extended waiting times (e.g. at 9 months or even more than a year).

We now turn to specialist hospitals. Analysis undertaken among four orthopaedic hospitals reveals differences on the scale of their survival curves (Figure 3.16). The Royal Orthopaedic Hospital admits quicker its patients, followed by the Royal National Orthopaedic, Nuffield Orthopaedic and finally by Robert Jones & Agnes Hunt Orthopaedic. The last two curves intersect more than once for patients that wait more than a year. The corresponding survival times for

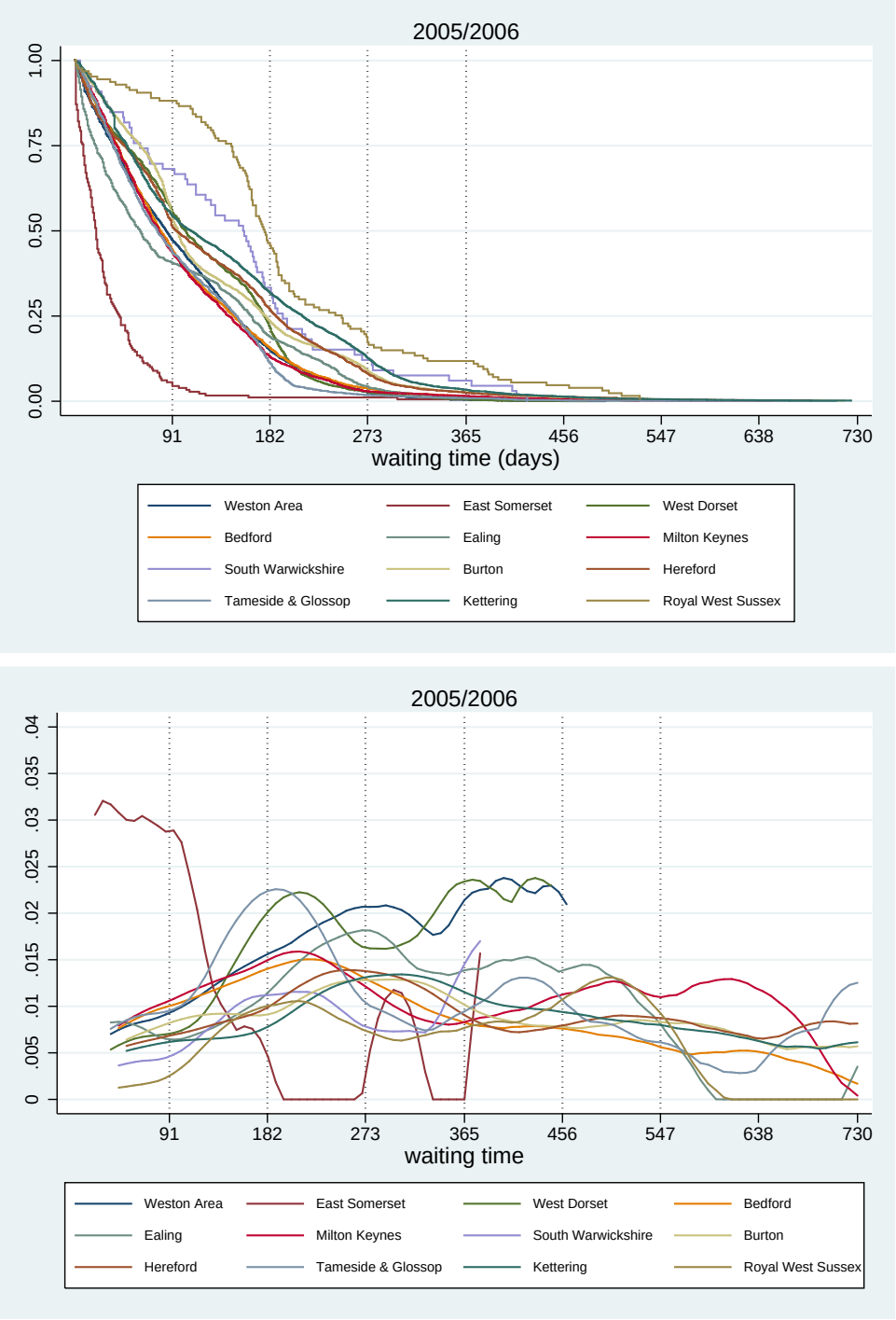


Figure 3.15: Survival and hazard curves for small acute hospitals for 2005/2006.

the 25% , 50% and 75% admission rates of the four hospitals are illustrated in Table 3.1.

Table 3.1: Survival times for four orthopaedic hospitals.

		Proportion of patients still on the list		
Hospital	**Patients**	**75%**	**50%**	**25%**
The Royal National Orthopaedic	4149	34	97	223
Nuffield Orthopaedic	5420	55	139	303
Robert Jones & Agnes Hunt Orthopaedic	4154	93	187	336
Royal Orthopaedic Hospital	8792	32	76	147

A slightly different pattern appears when we control for the type of surgery (total hip replacements only): (i) Besides an intersection at 91 days, the first two curves (yellow and blue) are quite similar to their corresponding overall waiting time graph (ii) This is not the case for the other two hospitals. Robert Jones & Agnes Hunt Orthopaedic admits quicker patients whose waits range from 91 to 400 days compared to Nuffield Orthopaedic. These findings suggest that more disaggregated analysis -by operative procedure- reveals patters that depart from the overall waiting time, and additionally variation still persists even when we control for the same treatment procedure (in terms of type and size of resources required).

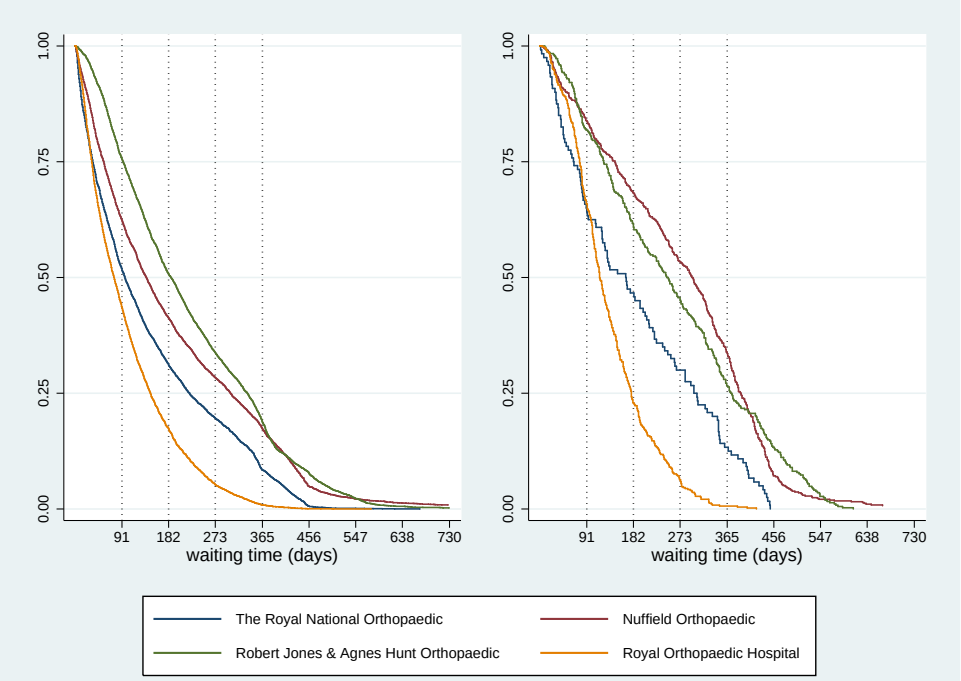
Increasing probabilities of admissions expressed by peaks in the hazard curves are as follows: According to the order of their survival curves, the Royal Orthopaedic Hospital has a two-mode peak between 365 and 456 days, followed by the peak of the Royal National Orthopaedic at 456 days, the wider peak of low intensity by Nuffield Orthopaedic again at 456 days and lastly the one by Robert Jones & Agnes Hunt Orthopaedic at 547 days. As for the hip replacements, Royal National and Robert Jones & Agnes Hunt Orthopaedics have monotonically increasing hazard rates that diverge from the overall admission pattern while the other two have quite similar curves to it.

Figures 3.17 and 3.18 demonstrate the waiting time distributions of a set of teaching hospitals in London for years 2002/2003 and 2005/2006. For the first year, the admission rates by St George's are the worst of all as its curve is far away from the origin. It is worth mentioning the different tactics by Hammersmith and Chelsea & Westminster hospitals with the former handling quicker the short waiters (<200 days) while delaying admission to long waiters compared with the latter that does the opposite. In 2005/2006, all curves are more concentrated and have shifted leftwards showing better waiting list administration. Interestingly enough, Hammersmith decreases significantly the admission rates for the short waiters, or in other words, the 75% of the patients with the shorter waits as illustrated by the convexity of its survival curve for this segment. Among all, Hamstead performs the best and Chelsea & Westminster the worst.

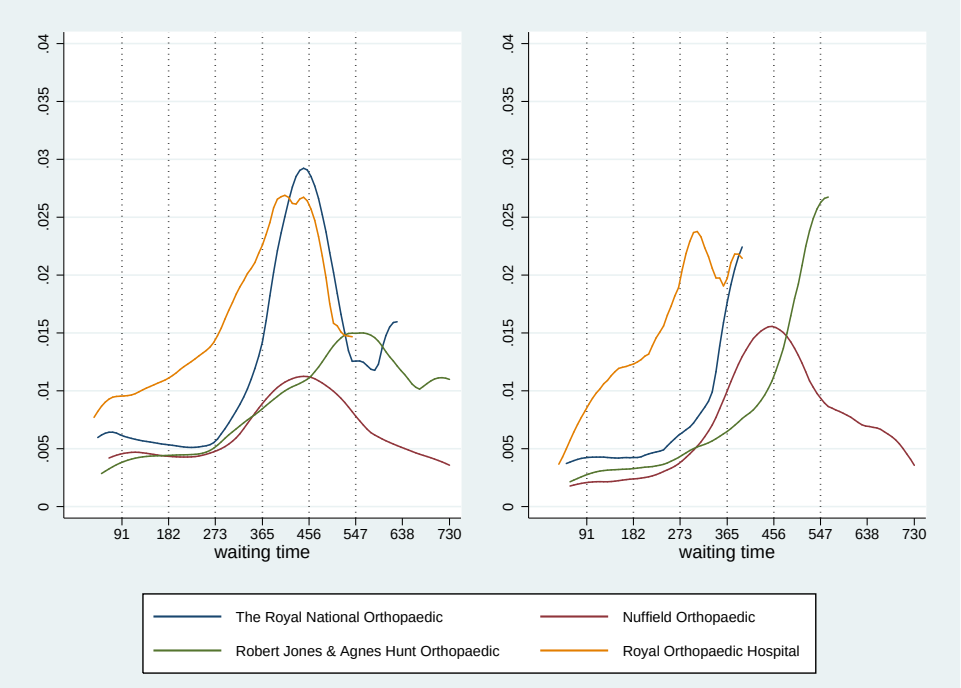
In 2002/2003, with the exception of Hammersmith that exhibits a high intensity peak between 365 and 456 days, the rest of the hospitals have low intensity wider peaks and Hamstead a constant hazard rate. Three years later, all trusts exhibit earlier peaks, a remark supporting the argument of increased elective activity to catch the national targets. However, we also observe delayed peaks that probably account for efforts to treat patients that have remained in the lists for an unexpectedly long time.

Figure 3.19 depicts the 2002/2003 survival and hazard curves of eight hospitals that had different performance ratings between financial years 2000 to 2004. In particular, Basildon & Thurrock, Devon & Exeter, Countess of Chester, Sunderland and Queen Victoria have scored excellent (3 stars) in the performance rating for NHS trusts during the period 2000/2001 to 2004/2005. On the other hand, Weston area, Bristol and Bath have showed bad performance grades. Although the patterns of survival rates appear to vary substantially by hospital, results are not as expected with some of the good performers having slow

admission rates (e.g. Countess of Chester) and bad ones having curves close to the origin (e.g. Bristol). Besides these two exceptions, the rest of the trusts behave as expected. Peaks on the hazard curves are taking place for various waiting times suggesting a great amount of variability in hospital's decision to admit its patients.



(a) Survival curves



(b) Hazard rates

Figure 3.16: Overall waiting times *(1st column)* and hip replacements *(2st column)* in four orthopaedic hospitals for 2002/2003.

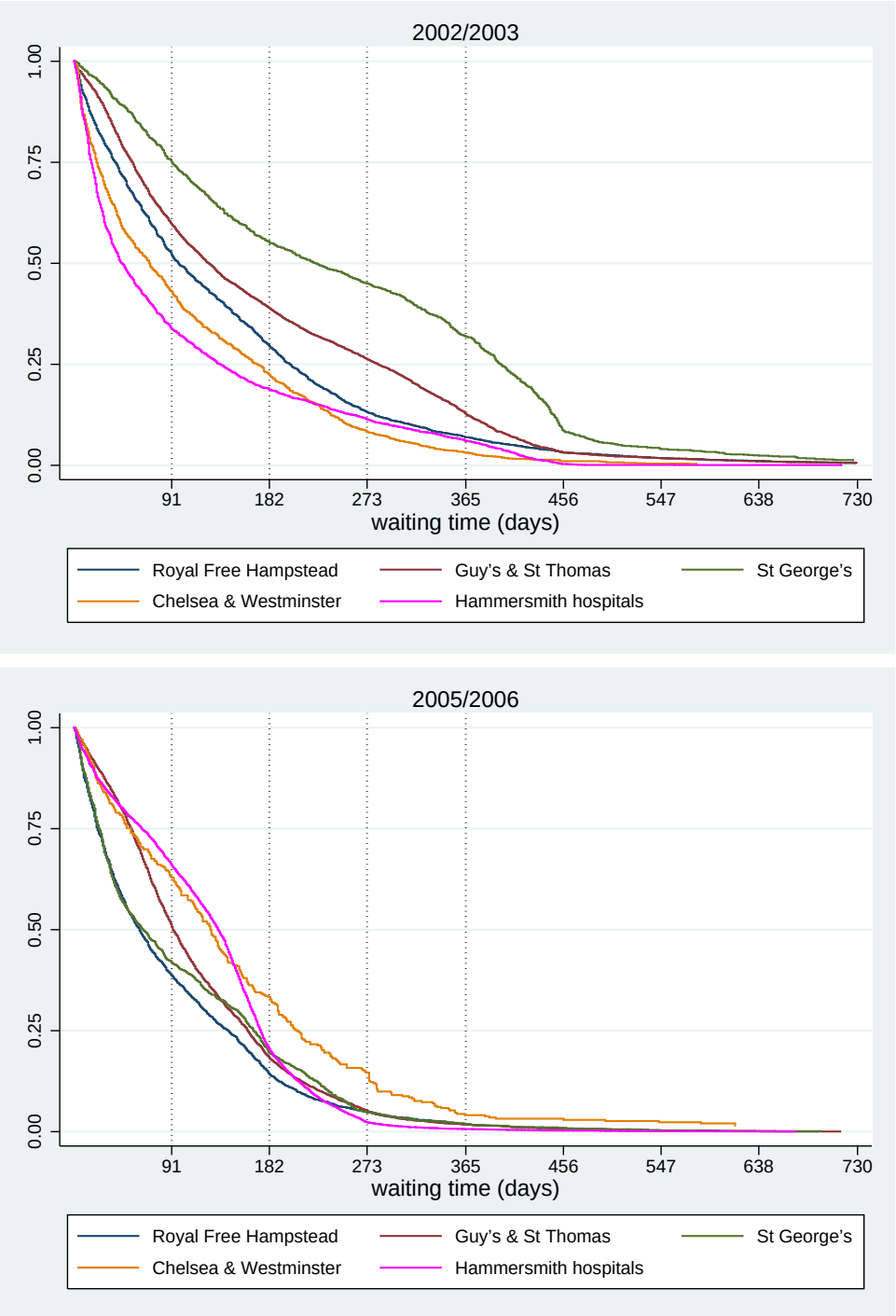


Figure 3.17: Survival curves for teaching hospitals in London, 2002/2003 *(top)* and 2005/2006 *(bottom)*.

Figure 3.18: Hazard curves for teaching hospitals in London, 2002/2003 *(top)* and 2005/2006 *(bottom)*.

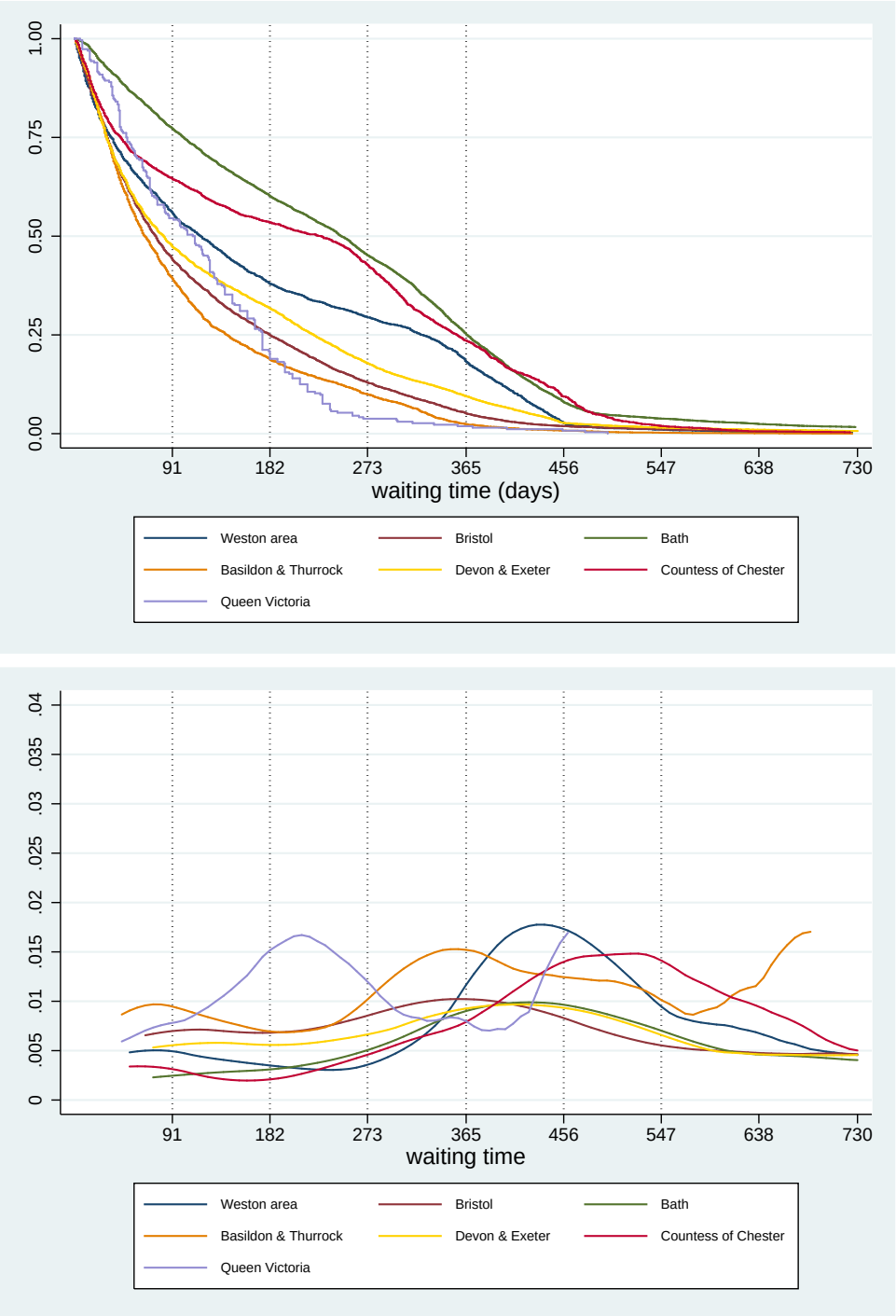


Figure 3.19: Survival and hazard curves for good and bad performers for 2002/2003.

Part III: Comparisons over time

In an attempt to understand how admission rates and increasing probabilities of admission evolve over years, the next step in our analysis explores the waiting time distribution of hospitals over time. Of all the analyses performed we have selected to present the example of Hammersmith hospital as it shares common characteristics with many of the other trusts. Figures 3.20 and 3.21 demonstrate the survival and hazard curves of Hammersmith from March 1997 to April 2006 while Figure 3.22 summarises distinctive patterns of the admission rates of nine trusts between 1997 and 2005.

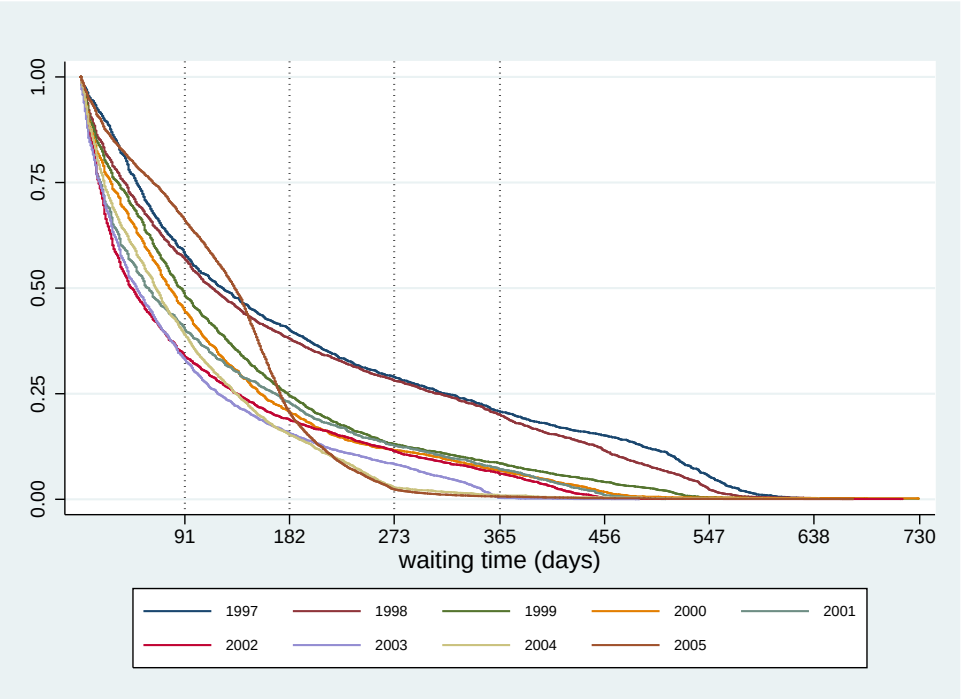


Figure 3.20: Evolution of survival curves of Hammersmith from 1997 to 2005.

The trends in the survival curves of Hammersmith hospital differ markedly between the nine year period. Admission rates for the first two years are quite slow but they do ameliorate gradually as time passes. One can observe that the curves shift leftwards year by year, mainly in a parallel manner, implying a proportional decrease of the waiting time of patients (scale change in the

survival curves). It is also clear that much effort is devoted to reduce extremely long waits; while 20% of patients had to wait more than a year in 1997 and 1998, less than 10% wait more than a year during the following four years and none is waiting such a period in 2003, 2004 and 2005. Of great importance is the fact that for the last two years Hammersmith increased the waiting times of people waiting less than 6 months compared to the equivalent waits of previous years. Graphically, this is represented by a rightwards pivot of the first part (0-182 days) of the waiting times distribution that also led to a change in the curvature of the survival curves. Even though the hospital cuts further long waits, at the same time, it delays the admission of short waiters. In essence, as years pass, Hammersmith hospital clears its stock of patients quicker, however, after some point in time there is a trade-off between short waiters and long waiters.

Comparison of the survival curves for 1997 and 2005 of another eight trusts besides Hammersmith hospital (Figure 3.22) reveals different behaviour by hospitals across time. Two distinct patterns are evident after the nine year period: (i) there are hospitals that manage to reduce the waiting times of the subset of elective patients that experiences long waits while at the same time increase the waiting times of short waiters and (ii) there are hospitals that achieve to decrease the waiting time of all their patients on lists.

Graphically, the first outcome is generated by an intersection between the two survival curves and the second by a leftward shift of the 2005 survival curve. Clearly, the level of substitution of high wait for low wait patients in the first case varies in terms of magnitude; it can be of the same (the created parts before and after the intersection of the two lines have equal area as is is the case for Hamstead and Nuffield hospitals) or different size (the two areas are unequal as we observe for South Manchester, Southampton and Hammersmith hospitals). Under the last scenario, the reduction of long waiters might be greater compared

to the increase of short waiters (Southampton and Hammersmith) or vice versa (South Manchester). Even when hospitals increase the admission rates of all their patients (e.g. the cases of Great Ormond Street, Norfolk, St George's and Bradford hospitals), the change in scale of the new survival curve differs; the shift could be parallel or not, quite big or smaller.

In Figure 3.21, increasing probabilities of admission represented by peaks are present for every year. In fact, what characterises the evolution of the hazard distribution over time is a leftward shift of these peaks; in 1997 the peak is at around 600 days, in 1998 moves to exactly 547 days, the next year shifts slightly leftwards resulting at 456 days in 2000 and 2001 (while of greater intensity), in 2002 it progresses towards 365 days that actually reaches in 2003 and finally in 2004 and 2005 the peaks are located between 365 and 182 days after two additional leftward shifts.

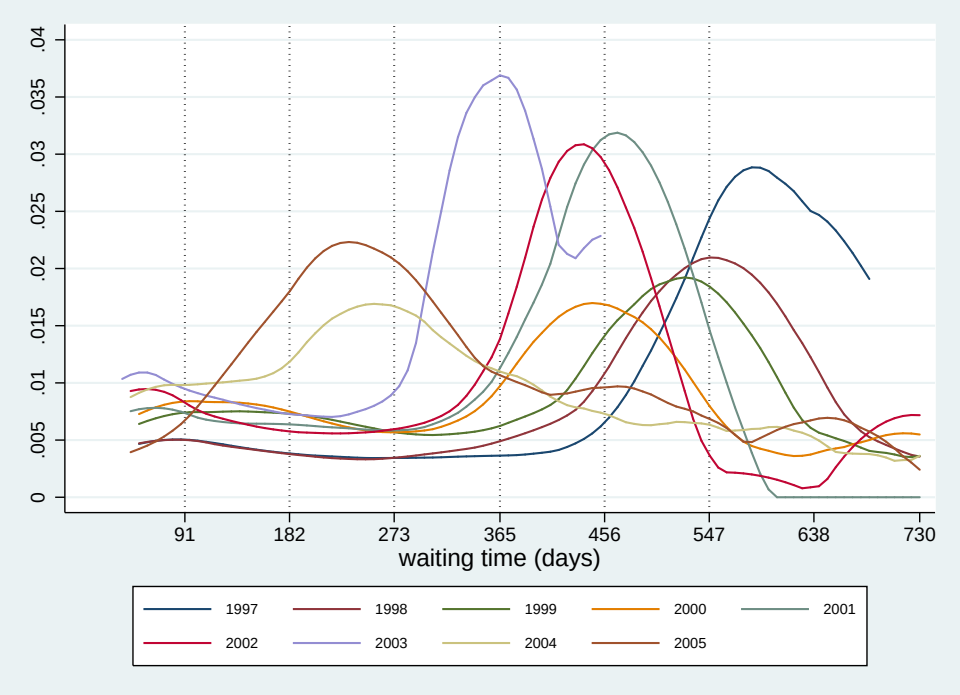


Figure 3.21: Evolution of hazard curves of Hammersmith from 1997 to 2005.

This behaviour is consistent with the general way of conduct of many trusts

that increase elective activity as targets approach and decrease it after the targets. Sometimes peaks coincide with the corresponding target while in other points in time they are very close to them. For instance, in the case of Hammersmith hospital, in 2001, the peak coincides with the national target of 15 months announced for that year while in 2005 it is clearly moving to catch up the target of 6 months.

In conclusion, the hospitals are trying to catch up with the new tougher targets by increasing their efforts in admitting patients with shorter waiting times. In other words, in order not to breach the increasingly stricter targets their hazard rates tend to shift leftwards in accordance with the target movement. It seems that after an action (new target issued by the government), a reaction follows (shift of the peaks in patients waiting times distributions).

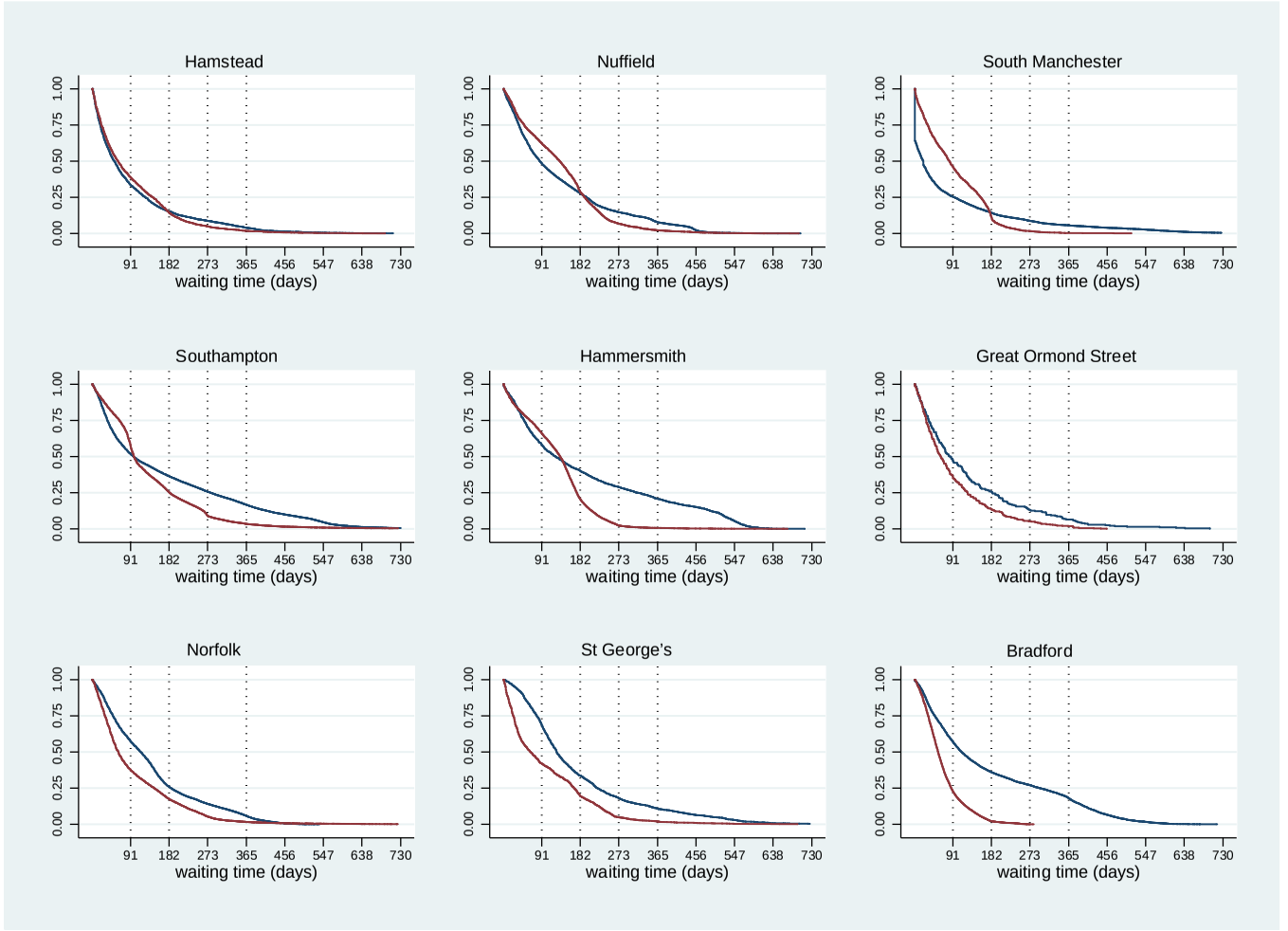


Figure 3.22: Patterns of survival curves in 1997 *(blue line)* and 2005 *(red line)*.

3.3.2 Behaviour of physicians

Focusing on how physicians manage their waiting lists we find evidence of great variation in the waiting time distributions of their patients. Figures 3.23, 3.24, 3.25 and 3.26 demonstrate the survival and hazard rates for a set of high activity general surgeons either by overall waiting list or by operative procedure, for particular financial years. Analysis of the evolution of waiting time distributions over time is presented for two physicians that exhibit distinct behavioral patterns (Figures 3.28 and 3.29).

Our results are as follows: (i) Elevated level of variability is evident in the survival curves of general surgeons in 2004. Plasticity in the survival curves ranges from smoothly decreasing curves to curves characterised by substantial change of curvature as waiting time increases. In accordance with this finding are results from a more disaggregated analysis performed in 2000 by a single operation (primary repair of inguinal hernia).

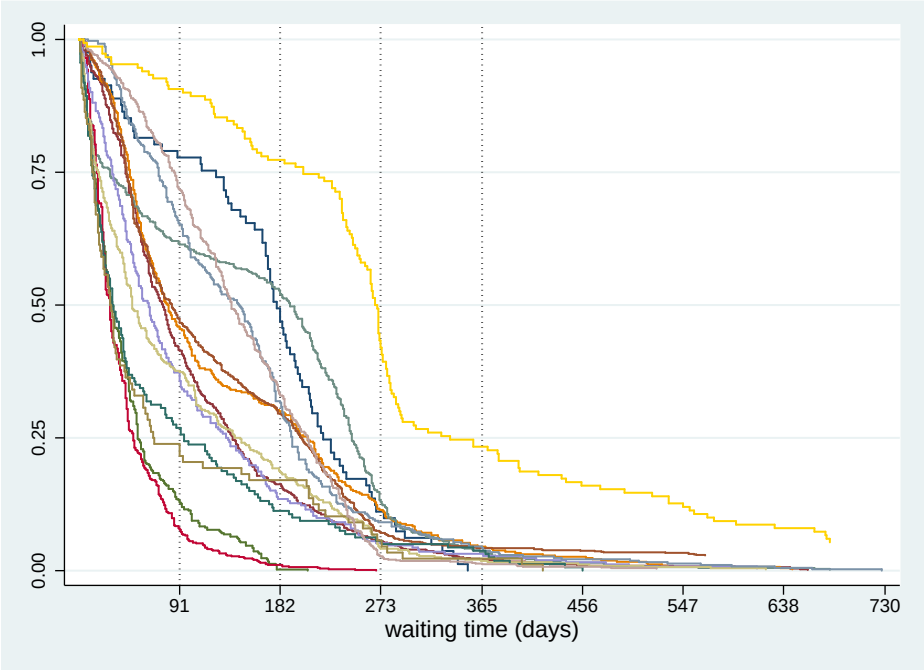


Figure 3.23: Survival curves of high activity general surgeons in 2004/2005.

As expected, the survival curves are stepwise because the number of patients is much smaller relative to the aggregate waiting time distribution (for all procedures). Even though we control for high activity surgeons and type of procedure, variation is extensive. As a consequence, there are doctors that admit the majority of their patients within 3 months while others delay considerably their admittance.

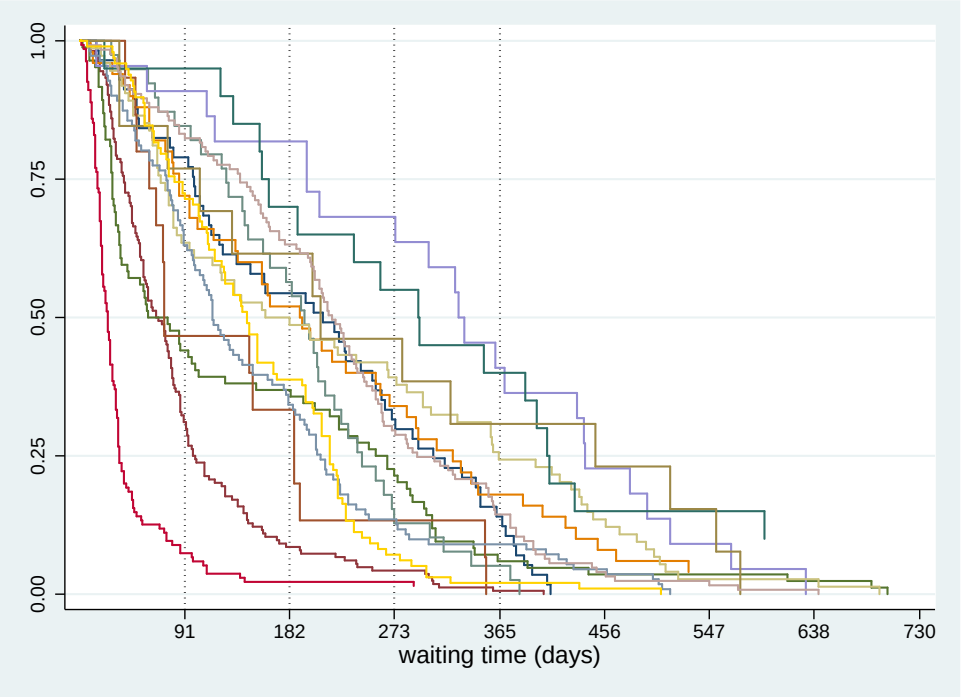


Figure 3.24: Survival curves for primary repair of inguinal hernia in 2000/2001.

(ii) The presence of great variation in the shape of hazard curves is another finding of this analysis. In Figure 3.25, increasing probabilities of admission are observed for patients having waited various periods. These peaks can be quite wide or steep, of high or low intensity and additionally there are doctors that admit patients in the same rate and have almost constant hazard curves.

(iii) Furthermore, we can point out the behaviour of two doctors (green and red lines) that show early peaks of great magnitude for patients waiting less than

3 months. This pattern that involves the presence of peaks at the beginning of the hazard function is related to their survival curves. The equivalent survival curves of the two doctors in Figure 3.23 are the first two that start decreasing steeply from 1. On the other hand, in the same figure, the doctor's survival curve represented by yellow, that is characterised by a concave part, implying considerably delays for patients' treatment, corresponds to the last curve to start rising in the hazard rates graph. Finally, monotonically increasing hazard rates are also present.

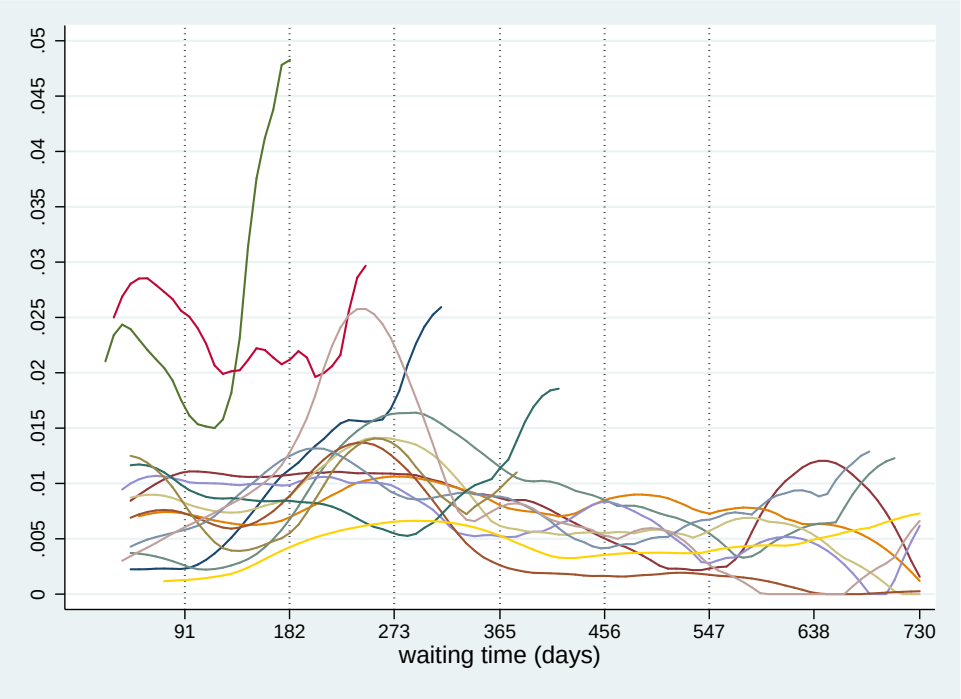


Figure 3.25: Hazard curves of high activity general surgeons in 2004/2005.

Similar patterns are observed when we control for the type of operation. Figure 3.27 illustrates a summary of five distinct patterns of the hazard rates with the equivalent survival curves. Of great interest is the red hazard line that starts with an early strong peak, then falls, has a couple of peaks at around 3 months, decreases again until it reaches 0, has a final peak and becomes 0 once

again. This corresponds to the survival curve located closer to the origin that exhibits a steep fall and represents a doctor that removes quickly its patients from the list.

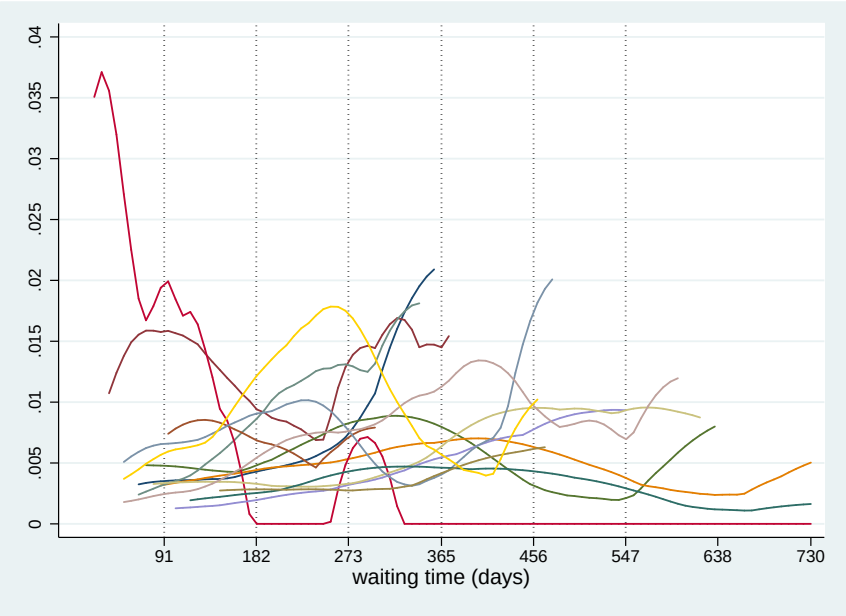


Figure 3.26: Hazard curves for primary repair of inguinal hernia in 2000/2001.

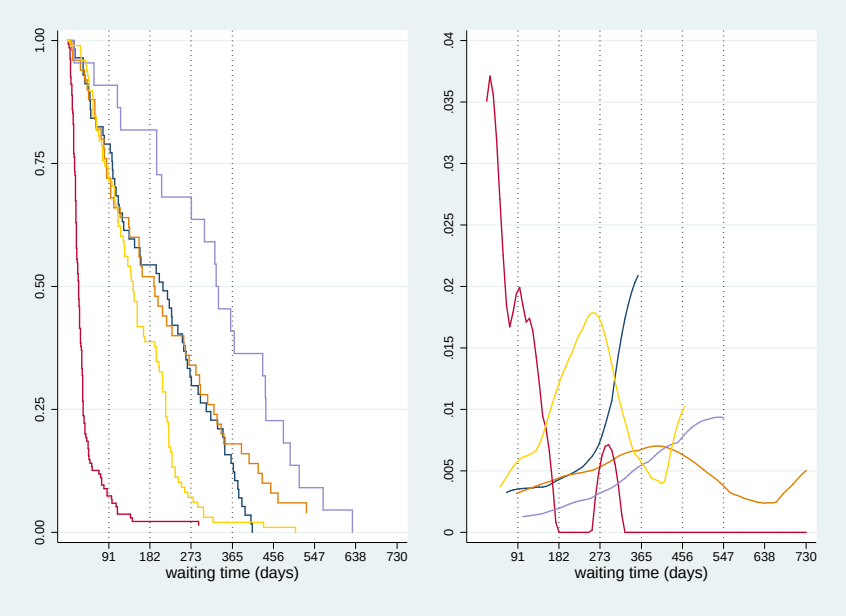


Figure 3.27: Five patterns of survival and hazard curves for doctors performing primary repair of inguinal hernia in 2000/2001.

Over time analysis of the waiting time distributions by physicians comprises the final part of this chapter. We conclude that doctors manage their waiting lists differently across time although they all face similar pressure regarding compliance to national targets.

In particular, variability of the survival and hazard curves of patients' waits appears in a greater level compared to the analysis based on hospitals. We observe reductions of the waiting times of all patients (parallel shifts of survival curves towards the origin), trade-offs of short waiters for long waiters (parts of the waiting time distributions shift leftwards and other parts rightwards) but at the same time increases in the waiting times of all or subsets of patients. Although doctors' behaviour does not follow the pattern of gradual smooth changes in patients' waiting time distributions as demonstrated for Hammersmith hospital, there is evidence for substantial reductions of long waiters.

Notwithstanding we performed various analyses, we decided to present the waiting time distributions of two doctors showing different admission rates with the first having more concentrated survival curves over time than the second. In accordance with this, the hazard curves of the second exhibit greater variation regarding the peaks. In general, the shape of the curves can be monotonically increasing, exhibiting one or more peaks earlier or later as waiting time increases.

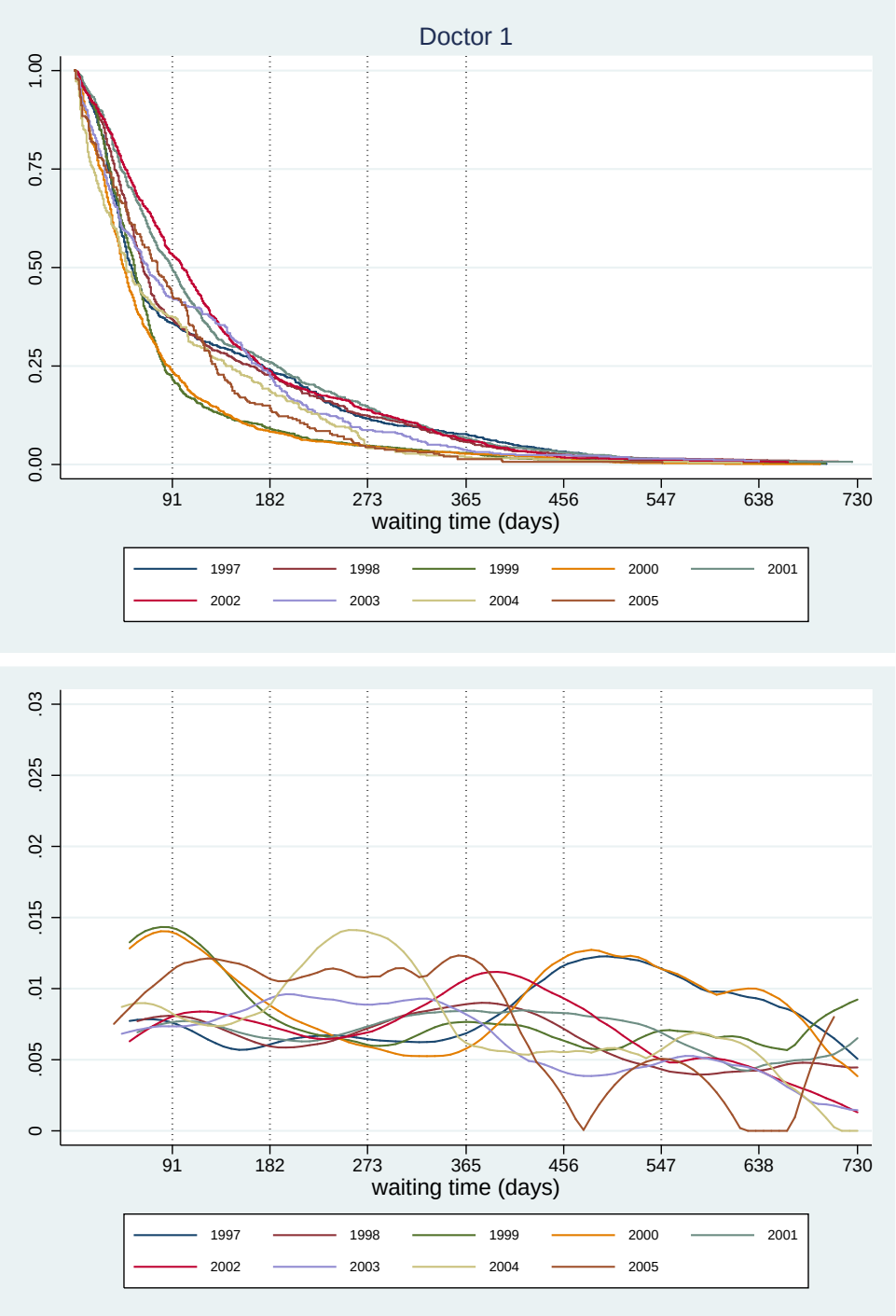


Figure 3.28: Evolution of survival and hazard curves by physician - doc1.

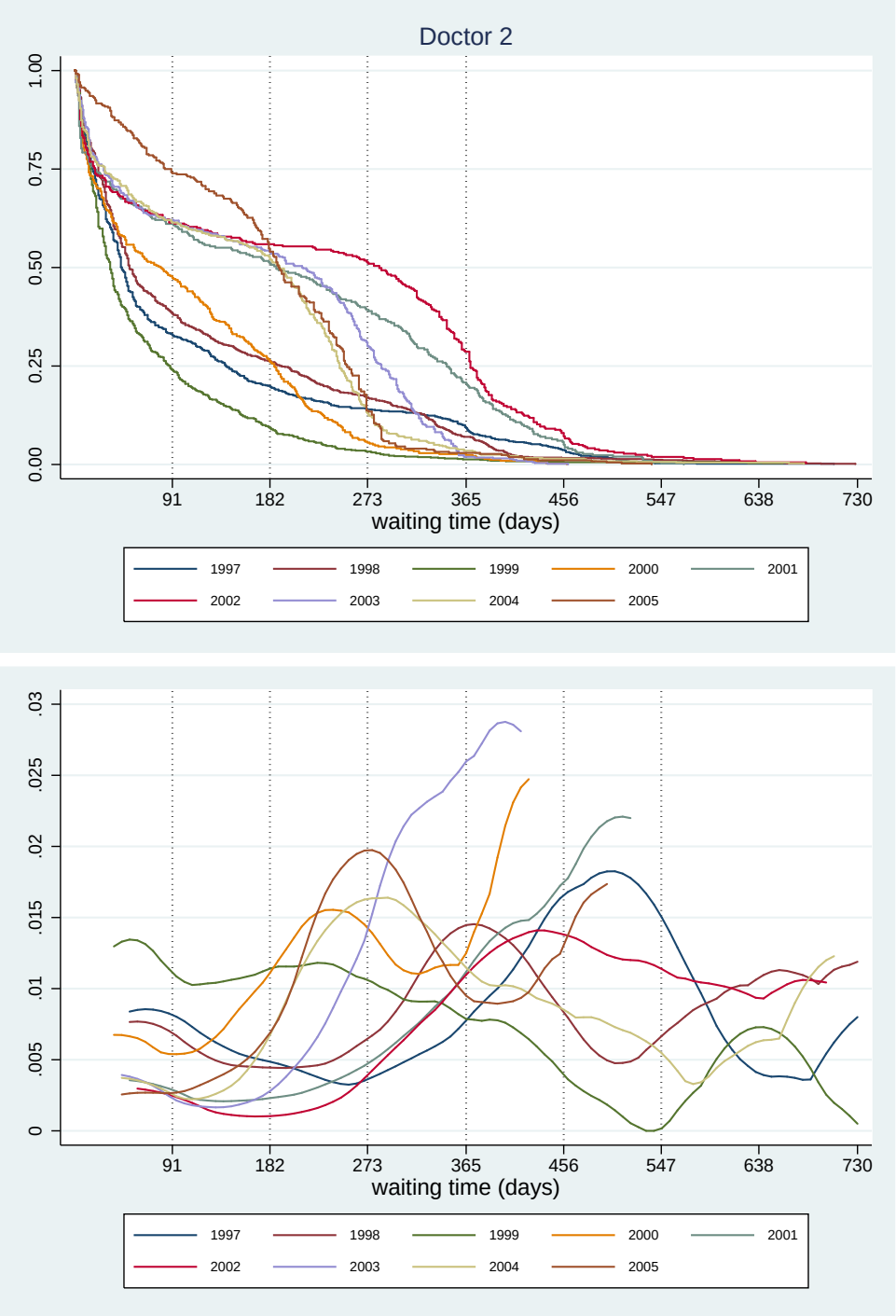


Figure 3.29: Evolution of survival and hazard curves by physician - doc2.

3.4 Concluding remarks

This chapter investigates the way hospitals and consultants manage their waiting lists by estimating the survival and hazard functions of the waiting times of their patients. We find evidence of great variation in waiting time distributions and implied admission tactics both by hospital and by physician. In addition, the more disaggregated the analysis the greater the observed variation. On top of that, greater variability is present in the doctors' survival and hazard curves compared with those of hospitals.

The patterns of survival rates appear to vary substantially by hospital and by doctor. In particular, there is significant variation in both the shape and scale of their survival curves. Our findings suggest that some curves, while retaining the same curvature, exhibit abrupt changes in the magnitude of the slope, while others alter from convex to concave or vice versa. These distinct differences in the slope (magnitude or sign) of the survival functions, correspond to spikes in the hazard curves. Additionally, there are survival curves that move closer to the origin and others that shift rightwards, in a parallel way or not. The different shapes reflect differences in the second order derivative of the removal rate of individuals from the list and variation in the shifts implies that hospitals admit all patients with a slower or quicker rate compared to others.

Consequently, trends in hazard curves also differ markedly between various sets of hospitals and consultants. We observe trusts/doctors with notable peaks of high intensity, others with very short wider peaks and finally some with constant hazard rates expressed as straight lines. Furthermore, it is worth mentioning two additional issues: the number of peaks varies and they can be located earlier or later in the waiting time distribution. Lastly, we report the appearance of cases in which hazards are illustrated as monotonically increasing probabilities of admission. These results indicate differences in the management

of the lists, in the decision process and admissions criteria, even when we control for particular characteristics of the list (type of hospital, operation, and even by physician). At hospital level we compare waiting time distributions controlling for size (large, medium, small acute), type (acute, specialist, teaching) and performance rating and indeed confirm sufficient heterogeneity. A promising path for future research would entail a more systematic empirical analysis of the heterogeneity in hospitals' behaviour. This can be inferred by a regression analysis, similar to the one conducted in Chapter 2, in which the focus would now be on supply side factors (size of hospital with number of beds or number of admissions as proxies and type of hospital in dummy variable format, hospital's budget, performance rating etc).

Focusing on comparisons across time, the effect of targets on waiting time distributions is evident; lots of peaks are situated exactly at the wait period that coincides with the set target for the specific financial year. However, first, there might be additional peaks along the distribution representing increased admission activity for different groups of patients with respect to their waiting times and second, there are cases in which peaks do not coincide with corresponding targets. One plausible explanation of this finding could be that trusts' and doctors' decisions are influenced by several factors besides meeting targets to ameliorate performance, such as severity of patients' medical status, hospital budget and available resources.

In addressing the question of how waiting time distributions by hospitals have evolved over time the answer would include the following points. First and foremost, much effort has been directed in reducing extremely long waits. The new element in the study is that we find that some hospitals admit all their patients quicker while others cut long waiters by delaying the admission of short waiters whose waits are by far below the corresponding target. In essence, as years pass, hospitals clear the stock of their patients quicker, how-

ever, after some point in time there is a trade-off between short waiters and long waiters. Regarding the evolution of the hazard curves over time, hospitals tend to increase their efforts in admitting patients with even shorter waiting times in order to catch up with the stricter targets. In other words, in order not to breach the increasingly tougher targets their hazard rates tend to shift leftwards in accordance with the target movement. Although this way of conduct is also observed by doctors, analysis at that level exhibits much more variation.

Furthermore, all the analyses in this chapter support the usefulness of duration analysis techniques and their wider application in the hospitals setting. Illustration of waiting time distributions of patients by KM and hazard curves can be proven an informative tool of auto-evaluation by every hospital. They could have the opportunity to assess how well they manage their waiting lists and detect consultants that could improve their performance. However, one could possibly argue that one limitation is that this methodology can be used in a retrospective manner as it handles completed spells. This caveat can be easily overcome as these techniques can allow for the utilisation of censored observations.

To conclude, this chapter brought in the surface various patterns of waiting time distributions as both survival and hazard curves differ even when we investigate the admitting behaviour of one health care provider. Interpreting these patterns constitutes the aim of the following chapter that develops a theoretical model on how the hospital manages its waiting list, focusing on the supply side of healthcare provision. Although we previously emphasised the significant role of targets, we are now attempting to examine additional attributes that could explain specific distributional patterns. One of the main questions we seek to explore is: How changes in the objective and cost function of the hospital, level of capacity and budget, demand for elective health care, severity of patients' medical status and national targets affect the distribution of patients' waiting time?

CHAPTER 4

A theoretical model of waiting times for elective surgery

4.1 Introduction

This chapter presents a theoretical model of how a public hospital manages its stock of patients for elective surgery. It aims at identifying the optimal waiting time distribution of a hospital given its objective function and the constraints it faces. Patients that have been referred to clinical physicians by general practitioners and are eligible for surgery join the NHS waiting lists of the former. The patients clinical pathway comes to an end when they are admitted to hospital to receive treatment, after having waited for a period of time. The focus of the model is on the selection mechanism that the hospital employs when taking patients out of its list. It is primarily a supply side model that illustrates the trade-offs in the decisions the hospital makes while managing its waiting lists, given its overall attitude towards admissions and the constraints it is faced with.

The closest analyses to this paper come from Iversen (1993), Olivella (2002) and Siciliani (2006). Although they all develop supply models of health care, they have quite distinct rationales and raise different issues. Yet, none of them obtain the optimal waiting time distribution of patients, which is the main contribution of this paper. Iversen (1993) models the long-run equilibrium of a non-cooperative game between the hospital and the government in an attempt to understand the mechanism of waiting in the National Health Service in Norway. According to Inversen, the hospital derives utility when the number of admissions increases and when the waiting time is reduced and maximises its objective function subject to its resources that are set by the governments decision for budget allocation. The government's utility function incorporates a part that reflects its willingness to pay function net of the costs, since it is financing the budget of the hospital. He concludes that excessive average waiting times exist only under a Stackelberg equilibrium[1].

Olivella (2002) analyses the consequences of prioritisation in waiting lists by introducing different severity levels of patients' health status. In particular, he assumes that the actual waiting time of a patient is a function of the average waiting time and the severity of the patient's health status. Waiting time is increasing with increases of average waits and can either increase or decrease with severity, for different severity levels. He maximises a utilitarian social welfare function by minimising the social costs of health care. Three separate situations are examined in which: (i) the private sector does not exist (ii) the private sector sets the fees monopolistically and (iii) the public sector regulates the private fees[2].

Siciliani (2006) formulates a dynamic model using optimal control theory to

[1]Farnworth (2003) builds on Iversen's framework and develops a theoretical model of how interactions among hospitals that charge different prices for health services can determine and affect the equilibrium expected waiting time. He discovers that, under specific circumstances, an increase in the price charged to one hospital can lower the waiting time for all.

[2]Iversen (1997) studies only the latter case under the assumption of zero total profits.

examine hospitals' incentives within a continuous-time dynamic framework. He analyses not only the steady state, but also possible optimal paths towards the steady state. He also considers the effects of exogenous shocks on both demand and supply of healthcare. The utility function of the hospital is influenced by the number of treatments supplied, the current size of the waiting list and the expected waiting time. Two different specifications for the formulation of waiting time are assumed. In the first case, expectations are myopic and the waiting time is perceived to be proportional to the waiting list. In the second case, specialists estimate future supply on the basis of current supply, hence waiting time is proxied by clearance time. The first model results in an increase of supply and waiting list towards the steady-state value given that the initial waiting list size is lower than the steady-state value. The author also derives the optimal supply path to reach waiting list targets which depends on the time horizon. The second reaches similar conclusions given the responsiveness of demand is low.

The last two papers allow for the interaction between the public and the private sector while the first one does not. In Iversen (1993) and Siciliani (2006), the emphasis is on the hospital decisions on optimal expected waiting time. However, in Olivella (2002), the focal point lies in the influence of waiting times on patients' welfare raising at the same time some prioritisation issues[3]. An important aspect of our work is that the emphasis lies on waiting time distributions, while the existing literature on waiting times utilises average waiting times only. Dixon and Siciliani (2009) is an exception. Their paper describes and maps the distribution of patients already treated (HES data) with the distribution of patients waiting on the list (waiting list returns). At the steady state, a comparison between the two distributions is performed, under different

[3]A study by Barros and Olivella (2005) develops a supply model of waiting lists for public hospitals when private doctors are able to select the patients they treat. The authors conclude that doctors do not necessarily treat the milder cases.

assumptions on the hazard function, however, the waiting time distributions are not derived within a model.

Our main contribution is the derivation of the whole waiting time distribution in the maximisation problem. The hospital's optimisation problem is solved at the steady state and obtains the optimal number of patients admitted for surgery for every single waiting time. That means that we derive the optimal steady state probability density function of waiting times. On that basis, we also explore and compare the corresponding survival and hazard functions. Within this context, we focus on identifying important determinants of the hospital's admissions patterns, their implications and their connection with the empirical findings of the previous two chapters.

This chapter is organised as follows. The next section provides the description of the model, the waiting list and distribution and the hospital's maximisation problem. Sections 4.3, 4.4 and 4.5 present the numerical solution and several comparative statics exercises. We identify, thus, the impact on the hospital's managing of the waiting list of several supply side factors. The influence of the introduction of universal waiting time targets is also analysed. The last section concludes.

4.2 Model

The model has two main elements: a set of patients that are currently waiting to be treated and the hospital that supplies healthcare. The government will be introduced through the policy measures it sets up in order to reduce waits and monitor the hospital's performance (waiting time targets). As already stated, the focus of the model is on the supply side and in particular on the optimal behaviour of the hospital while managing its waiting list.

Our model determines the waiting time of each patient treated, allowing us

to examine the optimal behaviour of the hospital when admitting patients, who have been waiting different periods of time for treatment and pinpoint a set of important determinants. The two main players of the model, patients currently waiting to be treated, and the hospital and analysed below.

4.2.1 Patients

There are currently L_t patients in the waiting list. Patients in the waiting list are characterised by the severity of their disease, denoted by $s = 1, 2, ...p$, where s is increasing in severity, and the time they have been in the list, their waiting time[4], or equivalently, their duration, $d = 1, 2, ...q$. Thus, d denotes the period elapsed between joining the waiting list of a specialist and admittance for surgery at the hospital. The minimum possible waiting time is one period ($d = 1$) and the maximum time a patient can wait is q. That is, patients do not wait infinitely so that a maximum finite duration exists. Thus, $i(s, d) \in L_t$ is one of the patients, at time t, that has been in the stock for duration d with severity level s. Denote k_t the overall number of patients from the list that are being treated at any period t, hence $k_t \subseteq L_t$. The overall number of patients treated consists of subsets of patients classified according to their duration and severity levels, $k_{d,s,t}$; thus $k_{2,1,t}$ shows the number of patients treated at t with duration two and level of severity one.

We do not explicitly consider the decay in health from waiting. A patients' severity does not alter with time, and hence with their waiting. In other words, even if waiting time does cause deterioration to the health status, it does not make the patient's condition to move to the higher severity level and consequently, s is not a function of d.

The inflow of patients in the list, and equivalently, the demand for elective

[4]The terms 'waiting time' and 'duration' will be used interchangeably throughout the chapter.

health care at the beginning of time t is given by:

$$x_t = f(E_{t-1}(d), Z_t)$$

where $E_{t-1}(d)$ denotes the expected or perceived waiting time based on available information at $t-1$, and Z_t is a vector of demand factors. These could include socio-economic conditions and morbidity rates. For example, it is obvious that increased morbidity indexes would lead to a rise in the demand of health services. Given that the focus of the model is on the supply side, we treat Z_t, which can be viewed as potential demand for health care, as exogenous and fixed. The inflow of patients is decreasing in expected duration; the higher the expected waiting time at the beginning of t, the lower the demand for public health care. As standard in the literature[5], waiting times act as rationing devises in order to equilibrate demand and supply, similar to what prices do. Thus, extensive expected waiting times can reduce demand of elective surgeries, by discouraging GPs from making referrals and specialists from adding patients to their lists or encouraging patients to seek private insurance.

We set the inflow to be:

$$x_t = Z_t - \theta E_{t-1}(d) \tag{4.1}$$

where Z_t is the exogenous level of potential demand for health care and θ denotes the sensitivity of the inflow of patients to expected duration. Although the negative relation between expected duration and inflow is not explicitly modeled here, it can depend on the option of private health care provision. The expectation formation on waiting time will be formally determined in Section 4.2.3.

[5]See for instance Cullis *et al.* (2000), pages 1215-16 and 1229, Goddard *et al.* (1995), Iversen (1997), Besley *et al.* (1999), Martin and Smith (1999), Gravelle *et al.* (2002), Siciliani and Hurst (2005) and Siciliani (2006). Section 1.3.2 (pages 32 - 33) in Chapter 1 review empirical findings on the elasticity of demand to waiting time.

4.2.2 Hospital

The utility of the hospital

The hospital's utility for health care provision at any point in time t is given by

$$U_t = g(k_t) = \sum_d \sum_s g(k_{d,s,t}). \tag{4.2}$$

The hospital derives utility from treating patients of distinct severity levels at different durations. $g(k_{d,s,t})$ denotes the hospital's (monetary or non-monetary) gain from treating k patients of severity s and duration d. Recall that in our framework the waiting time (d) is not a choice variable, but it is endogenously determined. The hospital chooses optimally the number of patients of each severity and duration at time t, and this choice determines the average waiting time implicitly.

The properties of the utility function are summarised in the following three assumptions:

Assumption 1 *For a given number of patients treated of the same severity level (i.e. fixed k and s), the higher the waiting time, the lower the hospital's utility.*

That is, the hospital gains more utility from treating today 100 severe patients that have waited for two months rather than having them waiting for four months.

$$g(k_{d_1,s,t}) > g(k_{d_2,s,t}) \quad \text{where } k \text{ and } s \text{ are unchanged and } d_2 > d_1$$

This implies that the hospital will prefer to treat as many people as possible faster; and in the relevant literature[6] it is equivalent to the assumption that the more a patient waits the higher are his/hers waiting costs, and thus the

[6] See Iversen (1993) and Siciliani (2006).

lower the hospital's utility. This assumption reflects the hospital's consideration of the well-being of the patient (some form of altruism), as well as explicit benefits provided to hospitals by the health care system when times and lists are managed appropriately.

Assumption 2 *For a given number of patients treated of the same duration (i.e. fixed k and d), the higher the severity, the higher the hospital's utility.*

That is, treating 100 severe cases that are waiting for two months gives more utility than treating 100 mild cases that are waiting for two months.

$$g(k_{d,s_1,t}) < g(k_{d,s_2,t}) \quad \text{where } k \text{ and } d \text{ are the same and } s_2 > s_1$$

Differentiating patients according to the severity of their disease allows for some form of prioritisation, which is reflected in the hospital's utility. Similar to the model here, Olivella (2002) also allows for different severity levels in a continuous manner. The positive utility from more severe cases stems not only from an assumed altruistic hospital character but in essence from the role of the hospital per se; to treat the ones that most need it. At this point we should mention the fact that this view constitutes one of the NHS core principles; good healthcare available to all with specific goals to meet the needs of everyone, to be free at the point of delivery and to be based on clinical need, not ability to pay. Thus, some prioritisation is allowed based on the gravity of the medical condition of the patients waiting.

Alternatively, assumptions 1 and 2 can be regarded as the benefits of patients from treatment according to their needs (severity of disease and quick admission). This formulation, in which the benefits of the patients enter into the utility function of the provider, is also present in Ellis (1998) and Ellis and McGuire (1986).

Assumption 3 *For the same d and s, $g(k_{d,s,t})$ is concave in $k_{d,s,t} \in [0, k]$ and exhibits a turning point.*

Thus, up until some point, increasing the number of patients treated (of the same duration and severity) increases the hospitals' utility. However, from that threshold level of activity and onwards, the hospital's utility declines as more patients are treated. The same is assumed for example in Siciliani (2006) for average waiting time. Assuming otherwise, (i.e. monotonically increasing utility) could create problems with the management of the list; the hospital would use up all of its budget and capacity to treat as many patients as possible within the same period, at the detriment of the remaining patients, whose treatment would be postponed for much later (due to the lack of adequate resources). We discuss this feature further while presenting the solution to the model, and also show the implications of relaxing assumption 3 in Section 4.3.

The cost of the hospital

With respect to the cost of health care provision, we assume that the hospital is capacity constrained and has a budget allocated for elective surgeries given by B_t. The hospital's cost from providing health care can be decomposed into two separable parts:

$$C_t = c(k_t; \bar{k}) + \sum_d \sum_s c_t(k_{d,s,t}).$$

The first part is the hospital's scale cost and is denoted by $c(k_t; \bar{k})$ while the second is the hospital's duration and severity specific cost and is denoted by $c_t(k_{d,s,t})$. The former is a function of the overall number of treated patients (k_t) in relation to the limit number of patients, $\bar{k}$, the hospital can treat given its capacity. When $Z_t > \bar{k}$, the hospital cannot treat all the patients that demand

elective healthcare at t and thus a waiting list and waiting times emerge[7]. In addition, whenever optimal $k_t > \bar{k}$, the hospital operates above its capacity. The second part of the cost function is sensitive to the patients' waiting times and severity levels. Assumption 4 relates to hospital's scale cost and Assumptions 5 and 6 to hospital's patient-specific cost.

Assumption 4 *Once the capacity limit of the hospital is reached the scale cost, $c(k_t; \bar{k})$, is increasing in k_t.*

Thus, even though the hospital would prefer to treat all its patients immediately and unless there is a capacity expansion (increase of inputs e.g. more operating theaters, more clinic space, more beds, more doctors), treating more patients at t irrespective of their severity or their durations becomes increasingly costly. This formulation is analysed in more detail in Section 4.3 (pages 160 – 161).

Assumption 5 *For the same severity and a given number of treated patients, treating quicker is more costly.*

The duration-specific cost assumes that for the same severity and number of treatments, cost is deceasing in duration. Thus, treating today 100 patients of a severity level that have just entered in the list is more costly than treating today 100 patients of the same severity that have been on the list for two months. In other words, the quicker a patient is treated, the higher the cost for the hospital and similarly the slower a patient is treated, the lower the cost for the hospital. We therefore assume that the hospital's cost due to waiting is monotonically decreasing in duration.

[7]When $Z_t < \bar{k}$, the hospital can treat all the patients demanding healthcare will idle capacity ($k_t < \bar{k}$), provided that its budget is sufficient. In this case, all patients are treated at t, no waiting list is formed and duration is one for all L_t.

The relation between average waiting time and the cost of the hospital has been analysed by a number of contributions. Originated by Iversen (1993) such a relationship is not necessarily monotonic, some empirical evidence of which is provided by Siciliani *et al.* (2009). Inversen argues that for low waiting times, an increase in duration reduces the provider's cost. However, there might be a point over which higher levels of duration increase costs. This increase might be driven by higher administrative and medical resources required to manage a long waiting list.

In our model, we take under consideration Inversen's first point by allowing the hospital's cost to decline in duration. This further implies that it is hard for the hospital to treat patients quickly or equivalently some waiting allows the hospital to better time manage the list and its resources. Otherwise, a list would never be created. The increased costs due to long waits will be introduced in our model distinctly from Iversen (1993). Long waits will be monitored and 'punished' with the implementation of an extra non-smooth cost when we include the implementation of waiting time targets by the government. The introduction of such targets, which corresponds to a fundamental health care policy change will be analysed separately in Section 4.4.2.

Assumption 6 *For the same waiting time and a given number of treated patients, treating more severe cases is more costly.*

The severity-specific cost assumes that for the same duration and number of treatments the cost is increasing in severity. This could be so in terms of both medical materials and number of personnel and/or hours of work, since a more severe case might require a more complex treatment and a prolonged length of stay at the hospital after surgery. The same assumption is supported in both Ellis's (1998), Olivella's (2002) and Barros and Olivella's (2005) frameworks

that incorporate patients of different severity levels.

Before setting the maximisation problem of the hospital, the following subsection analyses the waiting list, the distribution of waiting times and the steady state framework under which the hospital operates.

4.2.3 Distribution of waiting time and the Steady State

We start by describing the evolution of patients' treatment at a certain hospital at time t. The variables of interest are the inflow, the stock and the outflow of patients of severity s at calendar time t. Denote $x_{s,t}$ the inflow, or new arrivals, to the list of severity s at the beginning of t. $k_{d,s,t}$ is the amount of patients of severity s that were treated at the end of t and that had waited d periods to be treated. Note that the minimum possible duration is one period, thus, patients that enter the list at the beginning of a period and exit at the end of the same period are said to have duration one[8]. Lastly, $\Psi_{d,s,t-1}$ represents the stock of patients of severity s that are waiting for d periods at time t and are yet to be treated.

[8]In other words we assume that entry to the list takes place only at the beginning of a period and admission for treatment at the end of a period. So, no one can wait for less than one period ($d = 1$).

The hospital's list at period t

At any point in time t the entire stock of people waiting of severity s is

$$
\begin{aligned}
L_{s,t} = \quad & x_{s,t} \quad \text{(new inflow at } t\text{)} \\
& + (x_{s,t-1} - k_{1,s,t-1}) \quad \text{(untreated patients from } t-1\text{, } \Psi_{2,s,t-1}\text{)} \\
& + (x_{s,t-2} - k_{1,s,t-2} - k_{2,s,t-1}) \quad \text{(untreated patients from } t-2\text{, } \Psi_{3,s,t-1}\text{)} \\
& + (x_{s,t-3} - k_{1,s,t-3} - k_{2,s,t-2} - k_{3,s,t-1}) \qquad (= \Psi_{4,s,t-1}) \\
& + (x_{s,t-4} - k_{1,s,t-4} - k_{2,s,t-3} - k_{3,s,t-2} - k_{4,s,t-1}) \qquad (= \Psi_{5,s,t-1}) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.3) \\
& \qquad\qquad . \\
& \qquad\qquad . \\
& \qquad\qquad . \\
& + (x_{s,t-(q-1)} - k_{1,s,t-(q-1)} - k_{2,s,t-(q-2)} - ... - k_{(q-1),s,t-1}) \qquad (= \Psi_{q,s,t-1}) \\
& + (x_{s,t-q} - k_{1,s,t-q} - k_{2,s,t-(q-1)} - ... - k_{q,s,t-1}) = 0
\end{aligned}
$$

The list, therefore, includes the number of patients that entered at t ($x_{s,t}$) and the number of untreated patients from the previous periods ($t-1, t-2, ..., t-(q-1)$). In particular, the number of untreated patients from $t-1$ consists of the inflow of patients at $t-1$ minus the ones that received treatment at the end of $t-1$ ($x_{s,t-1} - k_{1,s,t-1}$). We set this equal to $\Psi_{2,s,t-1}$ which shows the number of patients on the current list, $L_{s,t}$, that have waited for one period and are currently (at time t) waiting for a second period. Similarly, the number of untreated patients from $t-2$, that is, the number of patients currently on the list that are waiting for at least three periods, is given by the inflow of patients at $t-2$ minus the ones that received treatment at $t-2$ and the ones that received treatment at $t-1$ ($x_{s,t-2} - k_{1,s,t-2} - k_{2,s,t-1} = \Psi_{3,s,t-1}$). The same stands for

the untreated patients from the previous periods $(t-3, t-4, ..., t-(q-1))$. Given that q is the maximum duration any patient can wait, the 'bottom' of the list is given by the untreated patients at t that joined the list in time $t-(q-1)$. These patients, $\Psi_{q,s,t-1}$, are waiting for q periods and will be treated at t with probability one. The last line of (4.3), which is by construction equal to zero, shows that the untreated patients from $t-q$ were 'cleared' in period $t-1$ after having waited for the maximum duration.

The waiting list at t can also be expressed as,

$$L_{s,t} = x_{s,t} + \Psi_{2,s,t-1} + \Psi_{3,s,t-1} + \Psi_{4,s,t-1} + ...\Psi_{q,s,t-1} = x_{s,t} + \sum_{d=2}^{q} \Psi_{d,s,t-1}$$

For notation simplicity let $\Psi_{1,s,t-1} = x_{s,t}$, then we can write $L_{s,t} = \sum_{d=1}^{q} \Psi_{d,s,t-1}$.

Treating patients at period t

Faced with a waiting list of size $L_{s,t}$, the hospital moves towards getting patients off it to be admitted for surgical treatment. The admittance process is as follows. Firstly, the hospital will treat $k_{1,s,t}$ patients from the new arrivals; $k_{1,s,t}$ denotes the number of patients treated at t with waiting time one ($d = 1$). In addition, the hospital treats $k_{2,s,t}$ from the untreated stock of $t-1$ ($x_{s,t-1} - k_{1,s,t-1} = \Psi_{2,s,t-1}$). Thus, $k_{2,s,t} (\leq \Psi_{2,s,t-1})$ is the number of patients treated at t with a waiting time of two periods. In the same way, $k_{3,s,t}$ is the number of patients the hospital treats at t from the untreated stock of $t-2$ that have waited for three periods, and so on. Schematically, for each duration

we have:

$$
\begin{aligned}
d = 1: &\quad x_{s,t} - k_{1,s,t} = \Psi_{2,s,t} \\
d = 2: &\quad x_{s,t-1} - k_{1,s,t-1} - k_{2,s,t} = \Psi_{3,s,t} \\
d = 3: &\quad x_{s,t-2} - k_{1,s,t-2} - k_{2,s,t-1} - k_{3,s,t} = \Psi_{4,s,t} \\
d = 4: &\quad x_{s,t-3} - k_{1,s,t-3} - k_{2,s,t-2} - k_{3,s,t-1} - k_{4,s,t} = \Psi_{5,s,t} \qquad (4.4) \\
&\qquad\qquad . \\
&\qquad\qquad . \\
&\qquad\qquad . \\
d = q-1: &\quad x_{s,t-(q-2)} - k_{1,s,t-(q-2)} - ... - k_{(q-2),s,t-1} - k_{(q-1),s,t} = \Psi_{q,s,t} \\
d = q: &\quad x_{s,t-(q-1)} - k_{1,s,t-(q-1)} - ... - k_{(q-1),s,t-1} - k_{q,s,t} = 0
\end{aligned}
$$

The total number of patients of severity s that the hospital chooses to treat in period t is given by the following summation:

$$k_{s,t} = \sum_{d=1}^{q} k_{d,s,t}$$

Note that the patients treated at t having waited for q periods, $k_{q,s,t}$, should be exactly equal to the patients that entered at $t-(q-1)$ and were still untreated by $t-1$, $\Psi_{q,s,t-1}$. This is shown in the last row of (4.4).

At the end of period t, the untreated stock of patients that will transfer to next period's $(t+1)$ waiting list will be the sum of $\Psi_{2,s,t}, \Psi_{3,s,t}, ..., \Psi_{q,s,t}$. Apart from this set of patients the $t+1$ waiting list will also include the new arrivals $(x_{s,t+1})$. In the same fashion as $L_{s,t}$ we have:

$$L_{s,t+1} = x_{s,t+1} + \Psi_{2,s,t} + \Psi_{3,s,t} + ...\Psi_{q,s,t} = x_{s,t+1} + \sum_{d=2}^{q} \Psi_{d,s,t} = \sum_{d=1}^{q} \Psi_{d,s,t}$$

where as before we set $x_{s,t+1} = \Psi_{1,s,t}$.

The same evolution of treatments applies for both low and high severity conditions. Aggregating for the severity levels at each duration d and noting that

$$k_{d,t} = \sum_{s=1}^{2} k_{d,s,t} \quad \text{and} \quad \Psi_{d,t} = \sum_{s=1}^{2} \Psi_{d,s,t}$$

the overall next period list, including all levels of severity, would be

$$L_{t+1} = x_{t+1} + \Psi_{2,t} + \Psi_{3,t} + ...\Psi_{q,t} = \sum_{d=1}^{q} \Psi_t. \tag{4.5}$$

Finally aggregating for both the severity and duration levels the total number of patients treated at time t for all severity levels and durations is

$$k_t = \sum_{d} \sum_{s} k_{d,s,t}.$$

The distribution of waiting time

Since we have constructed the hospital's waiting list and have described the way the hospital treats patients, we can now present the waiting time distribution. In our theoretical model waiting time is modelled as a discrete variable. The distribution of waiting time depicts the whole spectrum of the relative frequency of patients having waited distinct periods of time until treatment at t. This is the probability function (PF) of waiting time, denoted as $f(d)$. Given the fact that waiting time is discrete, PF is equal to:

$$f(d) = P(D = d).$$

Although the PF shows the probability of a patient having waited for d periods until treatment, the cumulative function (CF) corresponds to the probability of having waited d periods or less:

$$F(d) = P(D \leq d).$$

In the waiting time literature, another two representations of the waiting time distribution are important; the survival function and the hazard function. The former is defined as the complement of CF, that is $S(d) = 1 - CF$, and hence it represents the probability that an individual is waiting more than d periods:

$$S(d) = P(D > d).$$

The hazard function, usually denoted as $h(d)$, depicts the probability that a patient is removed from the list to be treated with waiting time d conditional on having waited on the list up to that duration. Thus,

$$h(d) = P(D = d|D \geq d).$$

The following table describes the PF, the CF, the survival function and the hazard function of waiting time[9].

Table 4.1: Waiting Time Distribution

d	*f(d)* $P(D = d)$	*F(d)* $P(D \leq d)$	Survival Function $P(D > d)$	Hazard Function $P(D = d\|D \geq d)$
0	0	0	1	0
1	$\frac{k_{1,t}}{k_t}$	$\frac{k_{1,t}}{k_t}$	$1 - \frac{k_{1,t}}{k_t} = \frac{\sum_{d=2}^{q} k_{d,t}}{k_t}$	$\frac{k_{1,t}}{k_t}$
2	$\frac{k_{2,t}}{k_t}$	$\frac{k_{1,t}+k_{2,t}}{k_t}$	$1 - \frac{k_{1,t}+k_{2,t}}{k_t} = \frac{\sum_{d=3}^{q} k_{d,t}}{k_t}$	$\frac{k_{2,t}}{\sum_{d=2}^{q} k_{d,t}}$
3	$\frac{k_{3,t}}{k_t}$	$\frac{k_{1,t}+k_{2,t}+k_{3,t}}{k_t}$	$1 - \frac{k_{1,t}+k_{2,t}+k_{3,t}}{k_t} = \frac{\sum_{d=4}^{q} k_{d,t}}{k_t}$	$\frac{k_{3,t}}{\sum_{d=3}^{q} k_{d,t}}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$q-1$	$\frac{k_{q-1,t}}{k_t}$	$\frac{\sum_{d=1}^{q-1} k_{d,t}}{k_t}$	$\frac{k_{q,t}}{k_t}$	$\frac{k_{(q-1),t}}{k_{(q-1),t}+k_{q,t}}$
q	$\frac{k_{q,t}}{k_t}$	1	0	1

[9]Appendix A analyses the continuous counterparts of the waiting time distribution, which were employed in the duration analysis of Chapters 2 and 3.

The average waiting time of patients

The mean waiting time of patients treated at time t is calculated as

$$\bar{d}_t = \sum_{d=1}^{q} d f_t(d) = \sum_{d=1}^{q} d \frac{k_{d,t}}{k_t} = 1 \times \frac{k_{1,t}}{k_t} + 2 \times \frac{k_{2,t}}{k_t} + ... + q \times \frac{k_{q,t}}{k_t}.$$

Note that an important distinction of this work is the derivation of the whole waiting time distribution. This implies that we do not need to relay on any assumption about the evolution of the waiting list and hence the evolution of average waiting time. However, we need to determine how expected waiting time will be formulated at the beginning of time t (i.e. before the actual durations are realised), since the demand for health care (the inflow of patients) takes place at the beginning of time t. Two different expectations formations can be assumed. If individuals are myopic, then expectation formation will be based on all available information up until the end of period $t-1$, thus

$$E_{t-1}^{ME}(d) = \sum_{d=1}^{q} d f_{t-1}(d) = \sum_{d=1}^{q} d \frac{k_{d,t-1}}{k_{t-1}} = 1 \times \frac{k_{1,t-1}}{k_{t-1}} + 2 \times \frac{k_{2,t-1}}{k_{t-1}} + ... + q \times \frac{k_{q,t-1}}{k_{t-1}}$$

where superscript ME stands for myopic expectations. Expectations are formed backwards, thus individuals, while deciding whether to enter in the list assume that the hospital will behave in the future, the same way it behaved in the past. Perceived average waiting time at the beginning of t is assumed to be equal to last period's realised average duration. If, on the other hand, individuals are forward looking, then expectation formation will be rationally set given the available information up until the end of $t-1$,

$$E_{t-1}^{RE}(d) = E_{t-1}\left(\sum_{d=1}^{q} d \frac{k_{d,t+d-1}}{x_t}\right) = E_{t-1}\left(1 \times \frac{k_{1,t}}{x_t} + 2 \times \frac{k_{2,t+1}}{x_t} + ... + q \times \frac{k_{q,t+(q-1)}}{x_t}\right)$$

where superscript RE stands for rational expectations. Here individuals form expectations on how the hospital will behave in the future and thus on how the hospital will treat the new cohort of patients that will enter at t. This is rational formation, since it takes into account that the future service rates of

the hospital need not be the same as the past ones.

The steady state

It becomes clear that our system is dynamic, since current flows affect future stocks and flows. As will be seen in the next section, the way the hospital chooses to treat patients today will have an impact on the future waiting list and consequently on the future decisions of the hospital. For instance, as shown in equation (4.5), the $t+1$ list depends upon outflow decisions at t, which in turn are influenced by previous decisions.

Within this dynamic framework we focus on the steady state behaviour of the hospital, and thus on the steady state waiting time distribution. At the steady state (i) inflow and outflow of patients at any given point are equal[10], that is, $x_t = k_t$ and (ii) the optimal number of patients treated from each duration and severity ($k_{d,s,t}$) may depend on the waiting time elapsed between decision to join a list and actual treatment, but is independent of calendar time t. Consequently, both the inflow and outflow of patients are also time-invariant.

Under this condition, the steady state waiting list and the steady state waiting time distribution are described in Table 4.2. Note that at the steady state, the survival function shows the proportion of untreated patients for each waiting time, that is, $\frac{\Psi_d}{k} = \frac{\Psi_d}{x}$. The steady state expected duration is derived as

$$E(d) = \sum_{d=1}^{q} df(d) = \sum_{d=1}^{q} d\,\frac{k_d}{k} = 1 \times \frac{k_1}{k} + 2 \times \frac{k_2}{k} + ... + q \times \frac{k_q}{k}, \qquad (4.6)$$

and, since it is time-invariant and $k = x$ it is equivalent for both the myopic and the rational execrations formation.[11]

[10] If for all t, $x_t > k_t$ the waiting distribution would be explosive.

[11] The waiting time distribution can also be viewed as the number of periods required in order to treat current inflow, x_t. This is how Dixon and Siciliani (2009) construct the distribution and it is identical to ours at the steady state.

Table 4.2: Waiting Time Distribution at the Steady State

Duration	List	PF	Survival Function
0	0	0	1
1	x	$\frac{k_1}{k}$	$1-\frac{k_1}{k}=\frac{\Psi_2}{k}$
2	$x-k_1$	$\frac{k_2}{k}$	$1-\frac{k_1+k_2}{k}=\frac{\Psi_3}{k}$
3	$x-\sum_{d=1}^{2}k_d$	$\frac{k_3}{k}$	$1-\frac{\sum_{d=1}^{3}k_d}{k}=\frac{\Psi_4}{k}$
.	.	.	.
.	.	.	.
.	.	.	.
q	$x-\sum_{d=1}^{q-1}k_d$	$\frac{k_q}{k}$	0

The main differences between our theoretical waiting time distributions and the empirical ones in Chapters 2 and 3, stem from the treatment of time and the steady state condition. In contrast to the empirical distributions, here, time is discrete, both in terms of the passage of time and the waiting time (duration). In other words, entry and exit from the list takes place only at distinct points in time and the time of wait (duration) is also discrete. Consequently, the probability, survival and hazard functions here denote the discrete counterparts of the continuous empirical ones developed in the previous chapters. In addition, our theoretical model is solved under the steady state condition, in which inflow and outflow are equated; an assumption that cannot be made when dealing with empirical observations.

The main implication of those differences is on the shape of the hazard functions. In the theoretical model, since we are at the steady state, and thus the list clears, the theoretical hazards always reach one. On the contrary this is never observed with the Kaplan-Meier hazard estimates of the HES dataset. Moreover, the discrete hazard function can be interpreted as a conditional prob-

ability, while the continuous hazard function cannot.

4.2.4 Hospital's maximisation problem

The hospital maximises its utility function, $g(k_{d,s,t})$, for the whole spectrum of $k_{d,s}$ at time t subject to its constraints,

$$\begin{aligned} \max_{\{k_{d,s,t}\}_{d,s}} \quad & E_0 \sum_{t=0}^{\infty} \sum_{d=1}^{q} \sum_{s=1}^{2} g(k_{d,s,t}) \\ \text{Subject to} \quad & \sum_{d} \sum_{s} c(k_{d,s,t}) + c(k_t) \leq B_t \\ & 0 \leq k_{d,s,t} \leq \Psi_{d,s,t-1} \\ & x_t = Z_t - \theta E_{t-1}(d) \\ & d \leq q \end{aligned}$$

The first constraint corresponds to the budget constraint of the hospital[12]. The second constraint states that the amount of patients of duration d and severity s treated at time t ($k_{d,s,t}$) must be between zero and the number of untreated patients in the list for that duration and severity. In other words, the number of people selected for treatment at time t cannot exceed the corresponding number of people waiting. Third, the hospital takes the evolution of patients inflow into account, and lastly we impose that the maximum waiting time is q (set to 36 in the numerical simulations).

At the steady state the number of entries to the list is equal to the number of patients treated at any point in time ($x_t = k_t$) and the optimal $k_{d,s,t}$ are time-invariant. Consequently, the hospital's maximisation problem becomes:

[12]Here, unlike in Ellis and McGuire (1986), the budget allocated to the hospital is exogenously given. In the numerical solution (Section 4.3) the budget value is tied to the treatment cost relative to the hospital's capacity, representing some sort of a cost-based reimbursement system.

$$
\begin{aligned}
\max_{\{k_{d,s}\}_{d,s}} \quad & \sum_{d=1}^{q}\sum_{s=1}^{2} g(k_{d,s}) \\
\text{Subject to} \quad & \sum_{d}\sum_{s} c(k_{d,s,}) + c(k) \leq B \\
& 0 \leq k_{d,s} \leq \Psi_{d,s} \\
& k = Z - \theta E(d) \\
& d \leq q
\end{aligned}
$$

Recall that $k = \sum_d \sum_s k_{d,s}$, the steady state expected duration is defined in equation (4.6) as $E(d) = \sum_d d\frac{k_d}{k}$ and $\Psi_{d,s} = k_s - \sum_{h=1}^{d-1} k_{h,s}$. In addition, note that at the steady state the restrictions that $k_{d,s} \leq \Psi_{d,s}$ are satisfied as long as $k_{d,s}$ are non-negative[13]. Thus, at the steady state the Lagrange function reads:

$$
\begin{aligned}
\max_{\{k_{d,s}\}_{d,s}} \mathfrak{L} = & \sum_{d}\sum_{s} g(k_{d,s}) + \lambda\left(B - \sum_{d}\sum_{s} c(k_{d,s}) - c(k)\right) \\
& + \sum_{d}\sum_{s} v_{d,s}k_{d,s} + +\mu\left(Z - \theta E(d) - k\right)
\end{aligned}
\tag{4.7}
$$

where λ is the lagrangian multiplier of the hospital budget constraint, $v_{d,s}$ is the lagrange multiplier of the Kuhn-Tucker constraint $k_{d,s} \geq 0$, and μ is the multiplier for the condition that ensures that the steady state inflow and outflow are equal.

Solving the hospital's problem gives rise to $2(d\times s)+2$ Karush–Kuhn–Tucker (KKT) conditions. For each $k_{h,m}$ where $h = 1, 2, ...q$ and $m = 1, 2,$

[13] At the steady state,

$$k_{d,s} \leq \Psi_{s,d} \Leftrightarrow k_{d,s} \leq k_s - \sum_{h=1}^{d-1} k_{h,s} \Leftrightarrow k_s - \sum_{h=1}^{d} k_{h,s} \geq 0 \Leftrightarrow \sum_{h=1}^{q} k_{h,s} - \sum_{h=1}^{d} k_{h,s} \geq 0 \Leftrightarrow \sum_{h=d+1}^{q} k_{h,s} \geq 0$$

which holds given that $k_d, s \geq 0$.

$$\frac{\partial \mathfrak{L}}{\partial k_{h,m}} = \frac{\partial \sum_d \sum_s g(k_{d,s})}{\partial k_{h,m}} - \lambda \left(\frac{\partial \sum_d \sum_s c(k_{d,s})}{\partial k_{h,m}} + \frac{\partial c(k)}{\partial k_{h,m}} \right) + v_{h,m}$$
$$- \mu \left(\theta \frac{\partial E(d)}{\partial k_{h,m}} + \frac{\partial \sum_d \sum_s k}{\partial k_{h,m}} \right) = 0$$
$$\frac{\partial \mathfrak{L}}{\partial v_{h,m}} = k_{h,m} \geq 0,\ v_{h,m} \geq 0 \quad \text{and} \quad v_{h,m} k_{h,m} = 0$$
$$\frac{\partial \mathfrak{L}}{\partial \lambda} = B - \sum_d \sum_s c(k_{d,s}) - c(k) \geq 0,\ \lambda \geq 0 \quad \text{and} \quad \lambda \frac{\partial \mathfrak{L}}{\partial \lambda} = 0$$
$$\frac{\partial \mathfrak{L}}{\partial \mu} = Z - \theta E(d) - k = 0$$

Given that we do not allow for interaction terms in both the hospital's utility ($\sum_d \sum_s g(k_{d,s})$) and the treatment-specific cost ($\sum_d \sum_s c(k_d, s)$) functions, the derivative of the Lagrange function with respect to $k_{h,m}$ simplifies to:

$$\frac{\partial \mathfrak{L}}{\partial k_{h,m}} = \frac{\partial g(k_{h,m})}{\partial k_{h,m}} - \lambda \left(\frac{\partial c(k_{h,m})}{\partial k_{h,m}} + \frac{\partial c(k)}{\partial k_{h,m}} \right) + v_{h,m} - \mu \left(\theta \frac{\partial E(d)}{\partial k_{h,m}} + 1 \right) = 0.$$

From this we can derive the optimal number of patients of each severity level treated after having waited d durations as a function of all the structural parameters (denoted $\mathfrak{z}$) of the model, $\forall \{d, s\} \quad k^*_{d,s} = k^*_{d,s}(\mathfrak{z})$. Although the first order conditions can be derived analytically, given the number of KKT conditions, the maximisation problem is solved numerically in Matlad after the parameter values are inserted (employing the *fmincon* command).

4.3 Numerical Solution

The solution to the hospital's problem and the corresponding waiting time distribution will be obtained numerically under different functional forms and structural parameters of the hospital. Apart from the restrictions implied in Assumptions 1-6, generally accepted by the literature, empirical information on

hospital's gain from treatments and their cost structures as function of waiting times is limited. We start by assuming a set of functional forms for the key elements of the model and then perform a series of comparative statics for alternative forms. The key empirical data that we use to discuss the appropriateness of each model specification are the waiting time distributions, described in detail in the previous chapters.

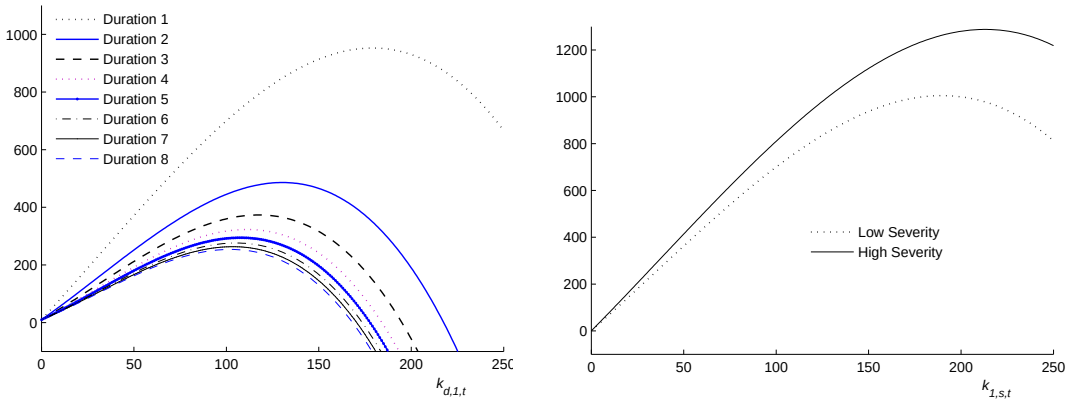
The utility of the hospital, $U_t = \sum_d \sum_s g(k_{d,s,t})$ is a function of $(d \times s)$ variables. The main specification for $g(k_{d,s,t})$ is assumed to be a polynomial of third order,

$$g(k_{d,s}) = a_{d,s}k_{d,s}^3 + b_{d,s}k_{d,s}^2 + c_{d,s}k_{d,s} + e,$$

where $a_{d,s} < 0$, $b_{d,s} > 0$, $c_{d,s} > 0$ are functions of duration and severity and $e \geq 0$ is a constant. Time t is suppressed for simplicity.

This specification fulfills Assumptions 1-3 laid out in Section 4.2.2. The cubic function ensures that $g(k_{d,s})$ is concave in $k_{d,s} \in [0, k]$ for a given set of d, s, which implies that after some threshold level of activity the hospital's utility declines as the number of patients treated from the same severity and duration increases. Moreover, the fact that $a_{d,s}, b_{d,s}, c_{d,s}$ depend on duration and severity allows for a differentiation of the hospital's utility with regards to the duration or severity of a given set of patients. For a given number of treatments with the same severity level, the combined impact of $a_{d,s}, b_{d,s}, c_{d,s}$ is decreasing in d and for the same waiting time is increasing in severity.

As an illustrative example, Figure 4.1(a) presents the hospital's utility function from treating up to 250 milder cases with waiting times ranging from 1 to 8 periods. The highest utility curve corresponds to patients treated within the same period (duration one); up until approximately 180 patients the hospital's utility is increasing, although after that, the hospital derives disutility from treating more patients with waiting time of one period. After the indicated threshold level, the marginal benefit from treating an extra patient becomes



(a) Different Durations (b) Different Severities

Figure 4.1: Illustrative Example: Hospital's Utility

negative, ensuring that not all of the hospital's resources are used up for the treatment of very short waiters, at the expense of very long waiters. Similarly, the second highest curve denotes the utility from treating patients with duration two, in which case the turning point happens at around 130 patients. It is clear that for the same number of treatments smaller duration is preferred. This is ensured by the $a_{d,s}, b_{d,s}, c_{d,s}$ and corresponds to the vertical differences across the utility curves. Similarly, as shown in Figure 4.1(b), for the same waiting time ($d = 1$), the hospital derives higher utility treating more severe cases relative to milder ones. In addition, the turning point for $s = 2$ is more to the right (around 210 patients) which indicates that the hospital would also prefer to treat them faster.

In the following subsection, we will allow for two extra functional forms of the utility function of the hospital $g(k_{d,s,t})$ and analyse their implications: (a) a monotonically increasing function with increasing rates (quadratic) and (b) a monotonically increasing function with decreasing rates (logarithmic).

On the cost side, the hospital is faced with a scale cost, as well as a cost specific to the duration and severity of each treatment. The scale cost reflects the capacity constraint of the hospital and thus refers to the overall number

of treatments irrespective of the waiting time or the severity of the treated. Denoting the full capacity number of patients as $\bar{k}$, the scale cost is assumed to take the following functional form:

$$c(k_t) = \tau(k_t - \bar{k})^2.$$

Deviations from full capacity are costly for the hospital. This holds the same for both under and over capacity[14]. However, note that at the optimum the hospital would prefer to be operating above its capacity limit, rather than with idle capacity, since for the same scale cost it is better to treat more patients rather than less[15]. The magnitude of this cost is given by parameter τ. A small τ implies that operating over capacity is relatively less costly for the hospital. Similarly, a higher value for τ suggests that it is more expensive for the hospital to operate when capacity constrained. In the later case, for example, it is relatively more expensive for the hospital to out-source equipment or personnel.

The default duration and severity specific cost, $c(k_{d,s,t})$, is set equal to:

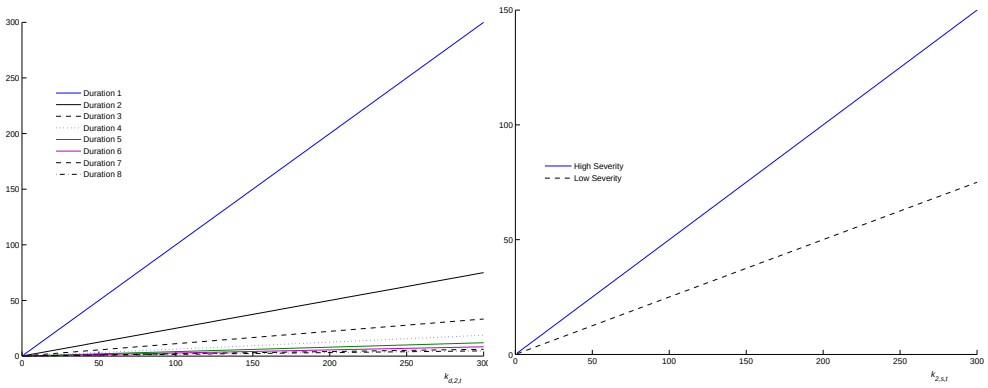
$$c(k_{d,s,t}) = \rho_{d,s} k_{d,s,t}.$$

That is, $c(k_{d,s,t})$ is linear in $k_{d,s,t}$ and $\rho_{d,s}$ is decreasing in d and increasing in s for the same number of treatments. Our specification implies that (i) for the same duration and severity, as the number of treatments increases the cost for those treatments increases linearly, (ii) for the same severity level and number of patients treated, the cost at lower durations is higher (i.e. $\frac{\partial \rho_{d,s}}{\partial d} < 0$) and (iii) for the same waiting time and number of patients treated, more severe cases cost most ($\frac{\partial \rho_{d,s}}{\partial s} > 0$). Subfigure 4.2(a) below shows the duration-specific cost for the first 8 durations and for up to 300 severe treatments. The highest cost

[14]Given the quadratic specification, the scale cost is identical for the same positive or negative deviation from $\bar{k}$.

[15]This holds as long as the budget is ample relative to the treatment-specific cost.

curve corresponds to treatments with one period of waiting time. In the same lines, the severity-specific part of the cost function is depicted in Figure 4.2(b) for fixed duration at 2. In the comparative statics below, we allow, amongst other things, for the duration and severity specific cost to have a quadratic specification.



(a) Different Durations (b) Different Severities

Figure 4.2: Illustrative Example: Hospital's Duration and Severity Specific Costs

Finally, the evolution of inflow is given by $x_t = Z_t - \theta E_{t-1}(d)$. As stated before, Z_t represents potential demand in terms of number of patients and θ the responsiveness of this potential demand to the expected waiting time. We always ensure that potential demand is greater than the hospital's capacity, $Z_t > \bar{k}$, so that the hospital is capacity constrained.

4.4 Comparative Statics I: No severity levels

We start with the simplest case in which patients are not differentiated by the severity of their condition and there is no waiting time target. The default parametarisation and the maximisation problem under the benchmark specification are depicted in Table 4.3.

Table 4.3: Benchmark functional specifications and parameters

$g(k_{d,t}) = a_d k_d^3 + b_d k_d^2 + c_d k_d + e$	Utility from treating k patients with duration d
where $a_d = -0.0002 + \frac{0.0001}{d}$, $b_d = 0.02 - \frac{0.01}{d}$, $c_d = 2 + \frac{5}{d}$ and $e = 0$	parameters of the cubic utility function
$c(k_{d,t}) = \rho_d k_{d,t}$	Cost from treatments at duration d
where $\rho_d = \frac{20}{d^2}$	parameter of the linear duration cost function
$c(k) = \tau(k - \bar{k})^2$	Scale cost of the total number of patients treated
where $\bar{k} = 900$	Hospital's capacity in terms of number of patients
$\tau = 10$	sensitivity of cost to deviations from full capacity $\bar{k}$
$B = 7000$	Hospital's budget
$Z = 1200$	Potential demand for healthcare
$\theta = 50$	Sensitivity of inflow to expected waiting time
q=36	Maximum allowed waiting time

Hospital's Maximisation Problem

$$\max_{\{k_d\}} \mathfrak{L} = \sum_d g(k_d) + \lambda\Big(B - \sum_d c(k_d) - c(k)\Big) + \sum_d v_d k_d + +\mu\left(Z - \theta E(d) - k\right)$$

For each k_h, $h = 1, 2, ..., q$, we have:

$$\frac{\partial \mathfrak{L}}{\partial k_h} = (3a_h k_h^2 + 2b_h k_h + c_h) - \lambda\left(\rho_h + 2\tau(k - \bar{k})\right) + v_h - \mu\Big(\theta\Big(h\frac{k - k_h}{k^2} - \sum_{d \neq h} d\frac{k_d}{k^2}\Big) + 1\Big) = 0$$

$$\frac{\partial \mathfrak{L}}{\partial v_h} = k_h \geq 0,\ v_h \geq 0 \quad \text{and} \quad v_h k_h = 0$$

$$\frac{\partial \mathfrak{L}}{\partial \lambda} = B - \sum_d c(k_d) - c(k) \geq 0,\ \lambda \geq 0 \quad \text{and} \quad \lambda\frac{\partial \mathfrak{L}}{\partial \lambda} = 0$$

$$\frac{\partial \mathfrak{L}}{\partial \mu} = Z - \theta E(d) - k = 0$$

Table 4.4 shows the number of patients treated from each duration and the waiting time distribution of the benchmark model. Based on those structural

parameters, the hospital treats 917 patients (17 patients above its capacity limit) at the steady state. The optimal maximum waiting time is 13 periods ($q^* = 13$). Thus, q^* is the 'bottom' of the list and the last patients on the waiting list were treated after having waited for 13 periods (entered the list 12 periods ago). Given that we are at the steady state, this is equivalent to saying that it takes 13 periods to clear the inflow of patients. Average duration is 5.65 periods. One can also see that the number of patients treated is decreasing in duration; the hospital treats 154.48 patients with duration one, 102.67 patients with duration two, 86 patients with duration three and so on.

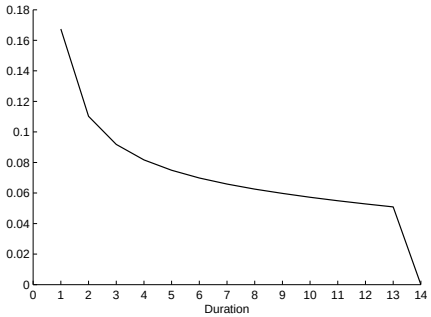
Table 4.4: Benchmark Model - Results

Duration	Optimal k_d	Survival	pf	Hazard Rate
0	0	1	0	
1	154.481	0.83156	0.16844	0.16844
2	102.677	0.71960	0.11196	0.13463
3	86.047	0.62578	0.09382	0.13038
4	76.759	0.54208	0.08370	0.13375
5	70.417	0.46530	0.07678	0.14164
6	65.574	0.39380	0.07150	0.15366
7	61.571	0.32667	0.06714	0.17048
8	58.066	0.26335	0.06331	0.19382
9	54.854	0.20354	0.05981	0.22711
10	51.768	0.14710	0.05645	0.27732
11	48.625	0.09408	0.05302	0.36044
12	45.237	0.04475	0.04932	0.52431
13	41.042	0	0.04475	1
k	917.119		$E(d)$	5.6576

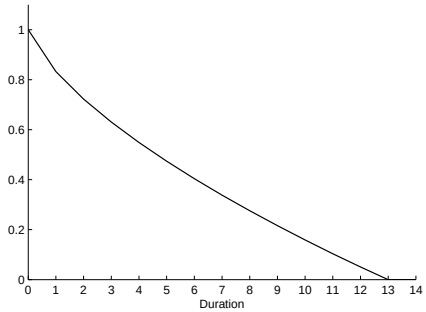
The mechanisms that drive the hospital to behave in such a way depend mainly on its utility, cost and inflow interactions. The hospital would prefer to treat as many patients as possible immediately (with duration one), however this comes at a higher cost. The vertical differences across utility curves and duration specific cost curves reflect this. Additionally, given the cubic specification assumed, the turning point in each utility curve for $d = 1, 2, ...q$ serves as a threshold for the amount of patients selected from each duration. In particular, this feature restrains the hospital from excessive 'front-loading' of treatments.

Another reason that restricts the hospital from treating too many patients up front is the impact of a small expected waiting time on future inflow. If the list is cleared quickly, then expected duration will be low and hence a higher number of patients will demand healthcare in the following period. Therefore, at the steady state, the hospital also takes into account the impact of its behaviour on the inflow of patients.

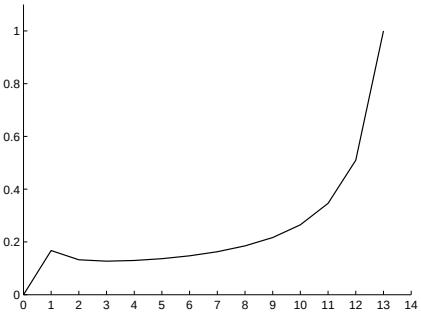
Figure 4.3 shows the graphical representations of (a) the probability function (b) survival function and (c) the hazard function for the benchmark model.



(a) Probability Function



(b) Survival Function



(c) Hazard Function

Figure 4.3: Benchmark Model - Graphs

The probability function, that is the instantaneous probability of receiving treatment, is declining and becomes zero for duration 14. This shape is due to the hospital's treating pattern, which admits more patients with shorter durations. In the second graph, the survival curve starts at one, as all patients

are waiting to be treated at duration zero; they are all part of the stock. As the hospital removes people off the list for treatment, the survival function is monotonically decreasing reaching zero at $d = 13$. The hazard curve exhibits a spike at $d = 1$, and after waiting period two, it increases monotonically until reaching unity. The observed decline between durations one and two it due to the largest proportion (0.168) of treatments taking place within the same period.

The shape of the benchmark theoretical survival curve matches the 'typical' empirical survival function (e.g. see Figure 2.4 in Chapter 2). That is, it is convex implying that patients are taken off the list as a decreasing function of waiting time (d)[16]. On the other hand, as already explained, the main difference among the theoretical and empirical hazard functions is due to the steady state condition. In general, the first spike is observed in both cases, while it is only the theoretical hazard that then continues increasing until 1.

4.4.1 Changes in the Structural Parameters of the Model

The following subsections present steady state comparative statics under changing parameters and/or functional specifications.

[16]The degree of convexity is higher (i.e. survival curve closer to the origin) with the introduction of severity levels.

Changes in the Utility function

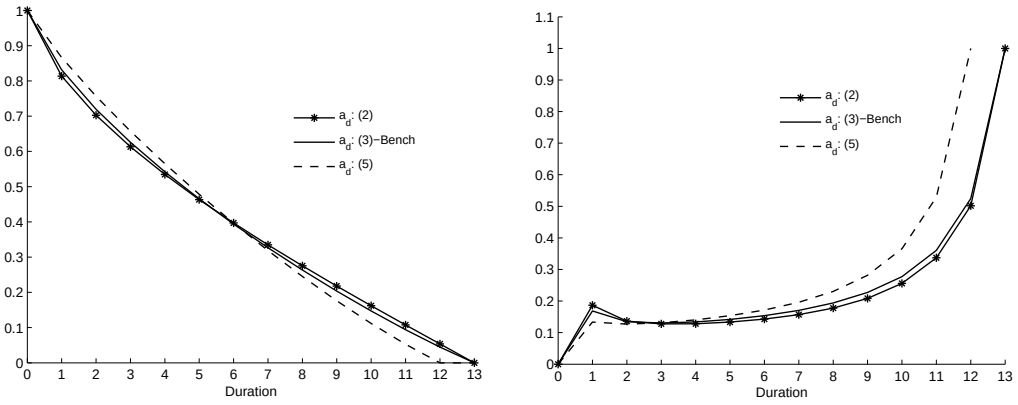
We start our analysis by changes in the parameters of the cubic utility specification. In particular, our first set of results focuses on increasing, in absolute terms, the size of $a_{d,s}$. This implies that the third order term in the utility function gains in importance which results in (i) the turning point of each $u(k_d)$ happens at a smaller number of treatments, that is, the maximum utility point shifts to the left for all $k_d, d = 1, 2, ..., q$ (ii) the utility level at each k_d gets smaller, i.e. all utility curves shift downwards and (iii) the $u(k_d)$ curves for $d = 1, 2, ..., q$ get closer to each other. Table 4.5 and Figure 4.4 illustrate the optimal number of patients treated from each duration and selected survival and hazard curves when a_d is changed. Recall that $a_d = -\frac{2}{10000} + \frac{1}{10000d}$ in the benchmark specification. In the example presented below (absolute) a_d is altered by changing the size of the positive term as follows: $\frac{1.5}{10000d}$, $\frac{1.2}{10000d}$, $\frac{1}{10000d}$, $\frac{0.6}{10000d}$, $\frac{0.3}{10000d}$.

As (absolute) a_d increases three patterns of the hospital's behaviour are apparent; the hospital treats less patients of the first duration, more patients of medium durations and less patients of relatively long durations. The overall number of treatments marginally increases, as average duration decreases. In addition, when moving from specification *(3)* to *(4)* and *(5)*, the long waiters are eliminated and thus the optimal waiting time to clear the list is reduced from 13 to 12 periods. A higher cubic term results in a more equal distribution of treated patients across durations and a quicker clearance waiting time. This is clearly seen when comparing case *(1)*, in which the cubic term is very weak, with case *(5)*. In *(1)* the hospital treats many patients 'up front', but then 46 patients are admitted for surgery in duration 36. In column *(5)*, on the other hand, patients are somehow more equally distributed in 12 durations.

Given the observed substitution of short and long waiters for more treatments at medium durations, the survival functions intersect (see Figure 4.4(a)).

Table 4.5: Changes in the cubic term of the utility function:

	(1)	*(2)*	*(3)* - Bench	*(4)*	*(5)*
$d \setminus a_d$	$-\frac{2}{10000}+\frac{1.5}{10000d}$	$-\frac{2}{10000}+\frac{1.2}{10000d}$	$-\frac{2}{10000}+\frac{1}{10000d}$	$-\frac{2}{10000}+\frac{0.6}{10000d}$	$-\frac{2}{10000}+\frac{0.3}{10000d}$
0	0	0	0	0	0
1	235.8466	170.9073	154.4790	131.6770	122.6675
2	112.1458	101.4629	102.6735	100.9167	101.0130
3	88.3879	82.1707	86.0451	88.7859	91.1851
4	76.5333	71.9129	76.7650	81.5564	84.7460
5	69.0047	65.2905	70.4137	76.3871	79.7426
6	63.6013	60.5982	65.5741	72.2755	75.4063
7	59.4561	57.1276	61.5751	68.7540	71.3787
8	56.1264	54.5171	58.0666	65.5722	67.4438
9	53.3747	52.5433	54.8552	62.5752	63.4168
10	51.0297	51.0949	51.7704	59.6394	59.1371
11	0	50.0854	48.6364	56.6653	54.3043
12	0	49.4302	45.2400	53.5491	48.3578
13	0	49.0816	41.0254	0	0
14	0	0	0	0	0
.	0	0	0	0	0
.	0	0	0	0	0
.	0	0	0	0	0
36	45.7799	0	0	0	0
k	911.286	916.222	917.119	918.353	918.798
$E(d)$	5.7743	5.6756	5.6576	5.6329	5.6240



(a) Survival Functions (b) Hazard Functions

Figure 4.4: Changes in a_d (Table 4.5)

Comparing the survival curves for specifications *(2)* and *(5)*, we notice that in the first case, the percentage of patients waiting more than two periods is about 70%, although for case *(5)* it is about 76%. Up until the intersection, the former

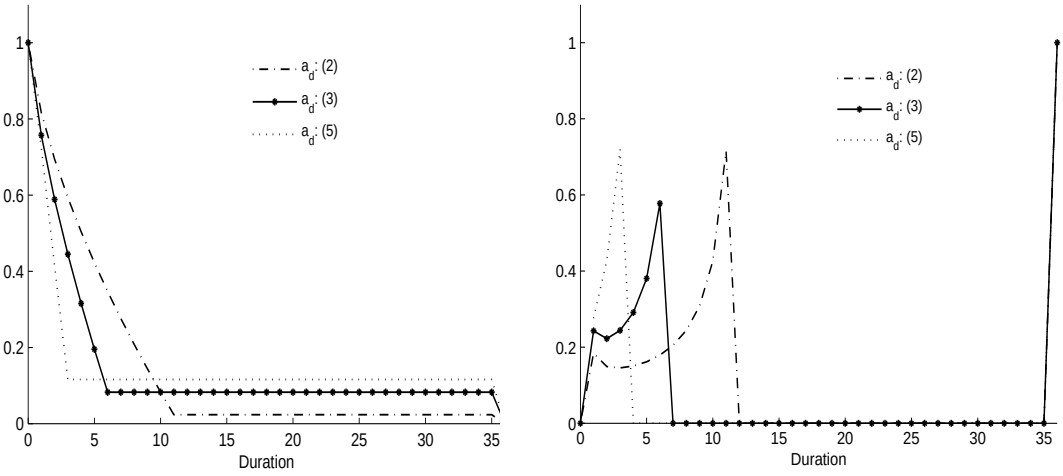
survival curve is closer to the origin implying that more patients are treated 'up front' (until duration 5.5). On the contrary, around 5% of patients are still waiting for treatment beyond 12 periods in *(2)*, but for case *(5)* no one is left untreated, since $q^* = 12$. With regard to the hazard functions, the most noticeable decrease between durations one and two is observed for case *(2)*, since it is in this case that the largest number of patients gets treated within the same period.

As the (absolute) magnitude of the cubic term gets weaker, the utility from treating patients in the first durations gets larger. Thus, the hospital strongly prefers to treat patients as quickly as possible. However, given its cost structure and budget, this can be achieved at the expense of very long waiters. This behaviour also ensures that future inflow is restrained (through higher average waiting times) and a steady state distribution is attainable.

Table 4.6: Changes in a_d - approaching zero

$d \setminus a_d$	*(1)* Bench	*(2)* $0.8a_d$	*(3)* $0.5a_d$	*(4)* $0.35a_d$	*(5)* $0.2a_d$
0	0	0	0	0	0
1	154.4812	168.4779	221.3891	263.0500	248.9558
2	102.6771	111.3854	153.4201	199.9500	279.4706
3	86.0469	92.4839	130.7895	177.9600	268.5102
4	76.7595	81.8446	118.0173	165.5900	0
5	70.4167	74.6426	109.3474	0	0
6	65.5741	69.3091	102.8531	0	0
7	61.5711	65.0947	0	0	0
8	58.0661	61.6785	0	0	0
9	54.8541	58.8270	0	0	0
10	51.7684	56.4072	0	0	0
11	48.6248	54.3782	0	0	0
12	45.2369	0	0	0	0
13	41.0424	0	0	0	0
14	0	0	0	0	0
.	0	0	0	0	0
.	0	0	0	0	0
.	0	0	0	0	0
36	0	21.5534	75.2341	97.1230	104.6517
k	917.1193	916.0826	911.0507	903.6730	901.5883
$E(d)$	5.6576	5.6783	5.7790	5.9265	5.9682

As shown in Table 4.6 and Figure 4.5, decreasing (absolute) a_d further away from the benchmark, the waiting time distributions become more unequal and the survival functions obtain distinct steps. For example, in specification *(5)*, 88% of patients are treated within the first three durations, and the rest leave the list after having waited the maximum possible time (36 periods).



(a) Survival Function (b) Hazard Function

Figure 4.5: As a_d approaches zero (Table 4.6)

Changes in the linear term of the utility function

In general, as c_d increases the turning point of the utility curves shifts towards the right for all $k_d, d = 1, 2, ..., q$, all utility curves shift upwards and the vertical distances between the $u(k_d)$ curves for different durations increase. Table 4.7 and Figure 4.6 illustrate the optimal number of patients treated from each duration and selected survival and hazard curves when c_d is changed. Recall that $c_d = 2 + \frac{5}{d}$ in the benchmark specification. As c_d increases the overall number of treatments decreases[17] and average waiting time goes up. The percentage of

[17]This does not hold for case *(1)*. Note that for this specification, the hospital is not using all of its budget.

Table 4.7: Changes in the linear term of the utility function:

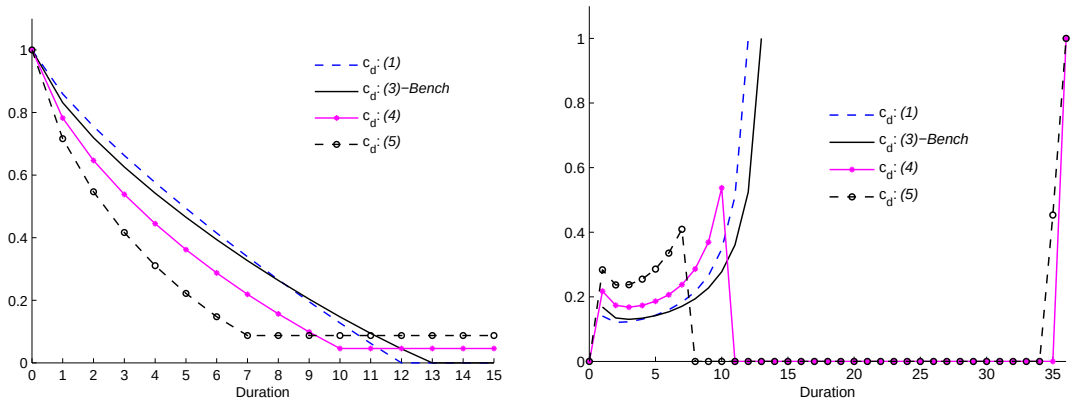
Optimal k_d at Steady State

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
$d \backslash c_d$	$0.45c_d$	$0.8c_d$	c_d	$2c_d$	$4c_d$	$5c_d$
0	0	0	0	0	0	0
1	128.3336	147.4386	154.4812	198.5575	257.4583	273.5851
2	94.6469	101.2871	102.6771	124.0844	153.9678	164.4531
3	84.6172	86.7711	86.0469	99.3437	117.8449	125.1002
4	79.0586	78.6071	76.7595	85.2973	96.3200	101.2245
5	75.1512	72.8971	70.4167	75.6574	80.6091	83.4923
6	72.0150	68.3509	65.5741	68.3352	67.4907	68.2875
7	69.2773	64.4164	61.5711	62.3771	54.7588	0
8	66.7484	60.7795	58.0661	57.2937	0	0
9	64.3171	57.2270	54.8541	52.7317	0	0
10	61.9047	53.5839	51.7684	48.4472	0	0
11	59.4361	49.5733	48.6248	0	0	0
12	56.8616	44.6784	45.2369	0	0	0
13	0	31.9083	41.0424	0	0	0
14	0	0	0	0	0	0
.	0	0	0	0	0	0
.	0	0	0	0	0	0
.	0	0	0	0	0	0
35	0	0	0	0	35.8577	41.2901
36	0	0	0	41.7161	43.2633	46.9872
k	912.3679	917.5189	917.1193	913.8414	907.5705	904.4201
$E(d)$	5.7526	5.6496	5.6576	5.7232	5.8486	5.9116

patients treated up front (short periods of wait) increases, while very long waiters start to accumulate. The lists are lengthened. The hospital, thus, creates more asymmetrical waiting time distributions which are characterised by more 'front loading', but at the same time by long tails to the right. On the contrary, when the linear term is quite small, we see a more balanced admissions pattern within 12-13 periods of wait.

Looking at the survival curves, specifications *(1)* and *(3)* intercept at around $d = 8$. It is obvious that as c_d increases the hospital admits quicker the short waiters by delaying the admittance of long waiters. In specifications *(4)* and *(5)* the survival curves have shifted greatly towards the origin (quicker admittance rates for durations 7 and 10 respectively), but a long tail persists that would

eventually reach zero at the maximum allowed duration ($d = 36$). Regarding the graph of the hazard functions, there is a small decline of the hazard between durations 1 and 2 (the largest number of patients are treated at period 1) after which it increases and reaches 1 at the relevant q^*. Note that the hazard curve for case *(5)* reaches 0 for duration 7, remains as such for another 28 periods and finally becomes 1 at period 36 when the last 43 patients are treated. Although not presented here, similar trade-offs are observed when we alter the quadratic term (b_d) of the utility function.



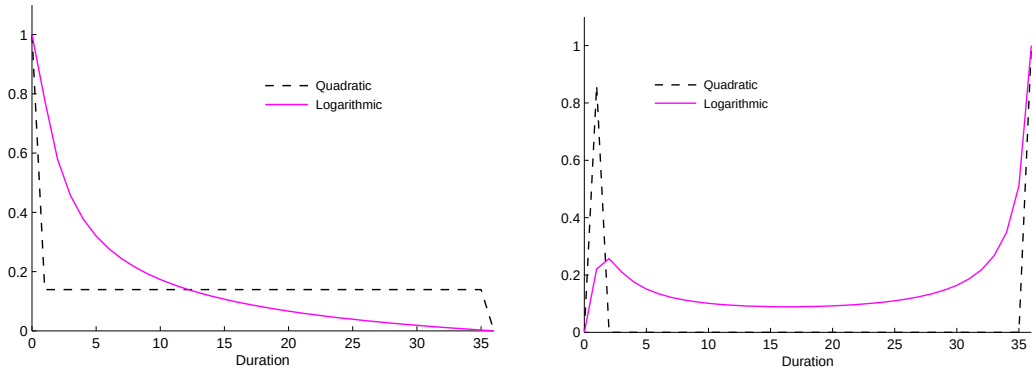
(a) Survival Function (b) Hazard Function

Figure 4.6: Changes in c_d (Table 4.7)

Hence, varying the hospital utility allow us to obtain two distinctive types of waiting lists attesting to the flexibility of the theoretical model developed here. In one case hospitals 'front load', treating many patients as quickly as possible at the detriment of a small fraction that is forced to wait for long periods or in the other, the hospital selects a smooth waiting list distribution where most people have to wait for more than 2 periods but no patients is forced to wait for long periods of time.

We now briefly present the steady state waiting time distributions in the cases where the utility function is (a) quadratic, $U(k_d) = b_d k_d^2 + c_d k_d + e$ and

(b) logarithmic, $U(k_d) = \gamma_d log(k_d) + g$, both depicted in Figure 4.7.



(a) Survival Function (b) Hazard Function

Figure 4.7: Quadratic and Logarithmic Utility specifications

With a quadratic utility function the steady state distribution becomes as follows: the majority of people are treated within the same period and the rest of people receive treatment at the largest possible duration (here 36). Why is this the case? Since the quadratic utility curves have no turning point and since $u(k_1)$ is always the highest, the hospital treats as many patients as possible with duration one (given the costs and the capacity it faces). However, once this is done, the remaining of patients are treated at the maximum possible waiting time, since this is the only way to maintain a steady state average waiting time and inflow. Consequently, the survival graph becomes a one step function, since 780 patients are treated with duration one, and the rest 126 after having waited for 36 periods. On the contrary, a logarithmic utility function results in a very smooth waiting time distribution, in which the hospital treats patients in each duration. Again, the number of treated patients is decreasing in d, with more treated up front, however, as the utility curves are now increasing at a decreasing rate (with no turning point), sufficient utility is obtained even when a small number of patients, k_d, is admitted from each d.

These two functional form assumptions serve as the two extremes of the

hospital behaviour discussed above, highlighting the trade-off in place. On the one hand, hospitals have an incentive to 'front load', treat as many patients in the first few periods. On other hand hospitals must ensure that they can deal with the current inflow under a control waiting list. Therefore, if the first incentive is strong enough (quadratic) survival functions become a step function otherwise when utility gains do not change as dramatically with duration, survival functions are very smooth (logarithmic).

Changes in the duration specific cost function

The benchmark duration-specific cost function is $\rho_d k_d$, with $\rho_d = 20/d^2$. The hospital's budget constraint has two distinct cost parts. The allocation of the budget is thus driven by (i) the number of patients that can be accommodated beyond its capacity, $\bar{k}$ (scale cost) and (ii) how quickly or slowly the patients (up until $\bar{k}$) can be taken off the list (duration specific cost). It is plausible to assume that a different cost structure and/or capacity imply a different budget. The benchmark budget (7000) has been set proportionally to those two costs (average unit cost (ρ_d) times capacity). Here, for the purpose of comparative statics, we will be changing the hospital's duration specific cost for a fixed budget.

In the first example (scenario A), we start by increasing ρ_d as follows: $\frac{20}{d^2}$, $\frac{40}{d^2}$, $\frac{60}{d^2}$, $\frac{80}{d^2}$, $\frac{120}{d^2}$. The cost is increased without altering the vertical differences across the cost curves for $d = 1, 2, ..., q$. That is, the cost of treating one patient with $d = 2$ is always $1/4$ of the cost for $d = 1$, the cost of one treatment with duration three is $1/9$ of the cost for $d = 1$ and so on. As shown in Table 4.8, increases in the duration related cost, force the hospital to treat a smaller overall number of patients. Since we are at the steady state, this translates both into a smaller outflow and inflow of patients and a higher average duration. When ρ_d is set to $\frac{60}{d^2}$ the hospital admits 900 patients, and beyond that

Table 4.8: Changes in Duration - specific cost, ρ_d - scenario A

$d \backslash \rho_d$	*(1)* $\frac{20}{d^2}$	*(2)* $40/d^2$	*(3)* $60/d^2$	*(4)* $80/d^2$	*(5)* $120/d^2$	*(6)* $140/d^2$
0	0	0	0	0	0	0
1	154.4790	126.2565	65.0407	32.0208	0	0
2	102.6735	96.3698	100.0166	107.1027	111.3901	66.3736
3	86.0451	83.8812	94.5169	102.4093	107.4771	113.4318
4	76.7650	76.2876	89.0702	96.7338	102.2487	116.4574
5	70.4137	70.8878	84.2797	91.3367	97.2365	112.9873
6	65.5741	66.6642	79.9049	86.1202	92.4303	107.1499
7	61.5751	63.1327	75.7371	80.8972	87.6852	99.8770
8	58.0666	60.0316	71.6095	75.4715	82.8624	91.2815
9	54.8552	57.1950	67.3731	69.6123	77.8273	80.9324
10	51.7704	54.4996	62.8597	62.9372	72.3969	67.4743
11	48.6364	51.8371	57.8218	54.5908	66.3120	41.2621
12	45.2400	49.1159	51.7281	39.5397	0	0
13	41.0254	46.1696	0	0	0	0
14	0	0	0	0	0	0
k	917.12	902.33	899.95	898.78	897.87	897.23
$E(d)$	5.6576	5.9534	6.0008	6.0246	6.0427	6.0554
$\sum c(k_d)/B$	0.5813	0.9923	0.999	0.9978	0.9935	0.9890

cost, it operates marginally bellow capacity. In effect all the budget is allocated to the duration specific cost (last row in Table 4.8). Regarding the waiting time distributions we mainly observe a reduction in the number (and proportion) of patients treated with very short waiting times. Since, the unit cost for 'up front' treatments is increased, the hospital postpones the quick admissions for later on. In fact, after specification *(2)* k_1 is drastically decreased and becomes zero for the last two columns. Due to this, we also observe a reverse in the number of treatments as duration increases. This behaviour is depicted in the survival functions, with the curvature shifting from convex to concave for short durations. The survival curves *(1)* and *(3)* of Figure 4.8 intersect around $d = 7$, with the benchmark survival being closer to the origin up until that point. For specification *(5)* less patients are (cumulatively) waiting for treatment after the intersection, and the list is cleared at a lower duration ($q^* = 11$).

In the following comparative exercise (scenario B), we compare the hospital's

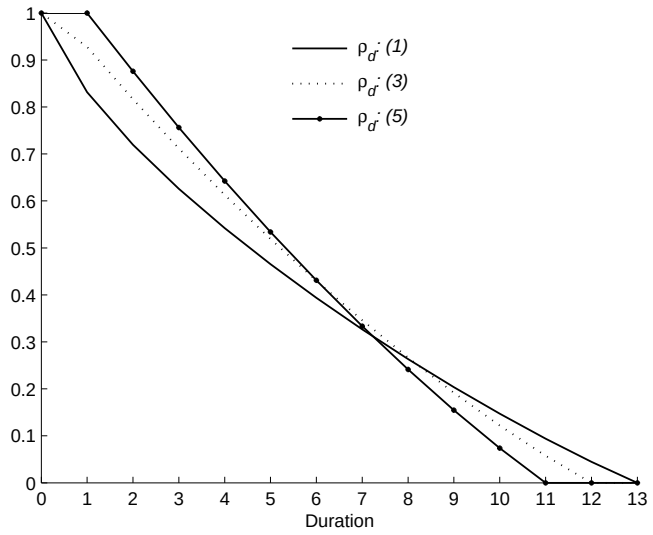


Figure 4.8: Changes in ρ_d - Scenario A (Table 4.8)

behaviour altering the coefficient of the duration related cost as such[18]:

$$\frac{20}{d^4}, \frac{20}{d^2}, \frac{20}{d}, \frac{20}{d^{0.7}}, \frac{20}{d^{0.65}}, \frac{20}{d^{0.6}}.$$

In this way, we change the vertical differences across the cost curves for different durations. Note that for $d = 1$, the unit cost of treatment is always the same for all specifications and equal to 20. As the power of d decreases, the vertical differences in the cost functions decrease as the $c(k_d)$ for $d = 2, 3, 4...$ get closer to the cost curve for $d = 1$ (i.e. the cost curves are now higher and get closer to the top one). Thus, when $\rho_d = \frac{20}{d^2}$, $c(k_d)$ decreases very fast as d increases, although in the case where $\rho_d = \frac{20}{d^{0.6}}$, the cost is decreasing slowly, which implies a relatively high unit cost for treating patients with medium, as well as short, durations. The unit cost for the first 10 durations for those two cases is presented below.

[18]Note that a steady state waiting time distribution cannot be obtained for powers lower than 0.6, unless the hospital's budget is increased.

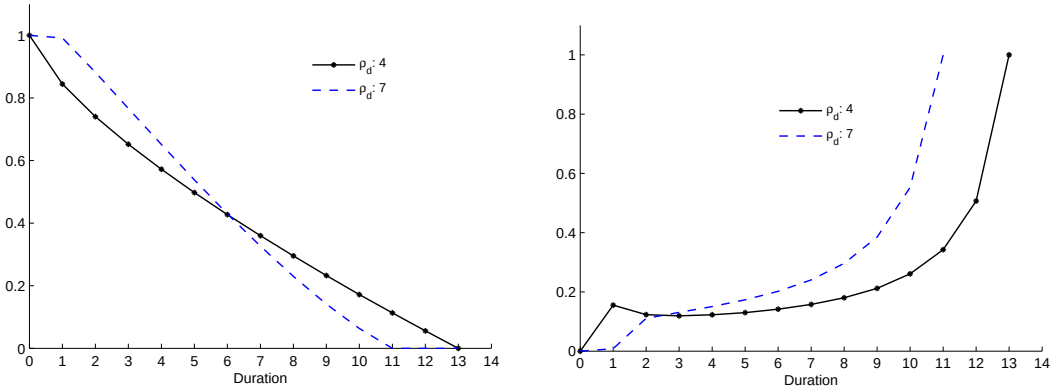
	Cost of one treatment for durations 1 to 10									
$\rho_d \backslash d$	1	2	3	4	5	6	7	8	9	10
$\frac{20}{d^2}$	20	5	2.22	1.25	0.80	0.56	0.41	0.31	0.25	0.20
$\frac{20}{d^{0.6}}$	20	13.20	10.35	8.71	7.61	6.83	6.22	5.74	5.35	5.02

Table 4.9: Changes in Duration Specific Cost, ρ_d - scenario B

$d \backslash c(k_d)$	$\underline{1}$ $\frac{20}{d^4}k_d$	$\underline{2}$ $\frac{20}{d^2}k_d$	$\underline{3}$ $\frac{20}{d}k_d$	$\underline{4}$ $\frac{20}{d^{0.8}}k_d$	$\underline{5}$ $\frac{20}{d^{0.7}}k_d$	$\underline{6}$ $\frac{20}{d^{0.65}}k_d$	$\underline{7}$ $\frac{20}{d^{0.6}}$
0	0	0	0	0	0	0	0
1	155.47	154.4812	150.0629	140.22	95.0348	57.0160	7.3750
2	103.82	102.6771	98.6871	93.864	84.9860	87.6120	98.5508
3	87.02	86.0469	82.3820	79.647	83.3582	92.0918	103.7143
4	77.636	76.7595	73.4491	72.077	81.8523	92.3116	103.7342
5	71.187	70.4167	67.5425	67.131	80.0004	90.6354	101.4706
6	66.238	65.5741	63.2104	63.535	77.7817	87.7783	97.7656
7	62.107	61.5711	59.8411	60.736	75.2015	83.9948	92.9022
8	58.438	58.0661	57.1134	58.437	72.2477	79.3210	86.8986
9	55.023	54.8541	54.8304	56.464	68.8785	73.6404	79.5329
10	51.681	51.7684	52.8665	54.711	64.9986	66.5824	70.1962
11	48.165	48.6248	51.1496	53.118	60.4339	57.0834	56.8056
12	44.177	45.2369	49.6465	51.663	54.7698	31.1057	0
13	38.349	41.0424	48.2915	50.261	0	0	0
14	0	0	0	0	0	0	0
k	919.31	917.12	909.07	901.8640	899.54	899.17	898.95
$E(d)$	5.6138	5.6576	5.8185	5.9628	6.0091	6.0165	6.0211
% Dur Cost	0.467	0.581	0.882	0.995	1.000	0.999	0.998

When the duration-specific cost increases in this way, the hospital treats less and less patients overall, reaching its capacity level while average waiting time increases. In addition, the duration required to clear the list gradually decreases from 13 to 11 periods and in general less patients are treated within the same period. In line with Scenario A, as the duration specific cost rises, the hospital is allocating almost all its budget to it. The last two columns of Table 4.9 exhibit a different admittance behaviour. The duration-specific cost curves are all close to the top one ($c(k_1)$), which implies that it is relatively more costly to treat significant amount of patients in the first few periods of wait. As a result, the hospital starts by treating only a few patients within one period (only 7.3 in the last case) and continues with increased number of treatments up until $d = 4$,

treating 14% less patients in the first 3 periods comparing to the benchmark case. After that, the decreasing feature of the cost structure induces the hospital to adopt the 'typical' admittance pattern in which the number of treatments decreases with duration. In Figure 4.9(a), survival curve $\underline{7}$ moves further away



(a) Survival Function (b) Hazard Function

Figure 4.9: Changes in ρ_d - Scenario B (Table 4.9)

from the origin until $d = 6$. This indicates a slower admittance rate relative to $\underline{4}$ for small waiting times. After the intersection, the admittance rates for higher ρ_d are increased. Due to an increasing number of treated patients for durations one to four, the survival function $\underline{7}$ starts off decreasing at a decreasing rate, and it then exhibits a change in curvature. Similarly, the hazard function is now monotonically increasing, with a noticeable increase between durations 1 and 2. Thus, the 'usual' spike in $d = 1$ is not observed in this case.

Once again, the model proves to be quite flexible in accounting for different patterns of waiting time distributions. In altering the relative cost of duration specific treatment we highlight the trade-offs the hospital faces. When costs of early treatments are relatively high, the incentive to 'front load' diminishes leading to an initially concave survival function. This pattern is strongly reinforced when a quadratic duration cost specification is used, as shown next.

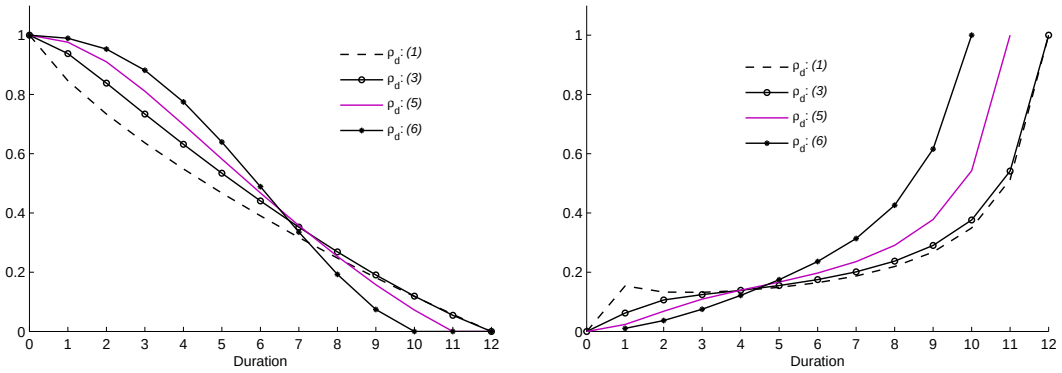
Quadratic Cost Function

We now turn to examine the case in which the duration specific cost of the hospital is quadratic, $c(k_{d,t}) = \rho_d k_{d,t}^2$. Due to the quadratic term, the share of the duration specific cost in the hospital's budget is considerably greater, thus, the budget is increased now to 30,000 and the unit cost is decreased to $1/d^2$. Selected comparative statics are depicted in Table 4.10 and Figure 4.10. With

Table 4.10: Changes in ρ_d - Quadratic Cost Function

$d \setminus \rho_d k_d^2$	*(1)* $\frac{1}{d^2}k_d^2$	*(2)* $\frac{2}{d^2}k_d^2$	*(3)* $\frac{4}{d^2}k_d^2$	*(4)* $\frac{6}{d^2}k_d^2$	*(5)* $\frac{7}{d^2}k_d^2$	*(6)* $\frac{8.5}{d^2}k_d^2$
0	0	0	0		0	0
1	141.99	101.32	55.640	29.469	20.833	9.1729
2	103.83	98.023	89.242	71.610	58.794	32.174
3	89.450	89.818	93.320	92.164	87.781	63.162
4	81.141	83.650	91.191	97.538	99.889	94.644
5	75.391	78.832	87.620	97.024	102.92	119.37
6	70.976	74.831	83.582	93.927	101.50	133.27
7	67.314	71.314	79.277	89.387	97.592	135.46
8	64.122	68.082	74.700	83.768	91.959	126.21
9	61.225	65.006	69.746	77.063	84.781	104.74
10	58.479	61.983	64.200	68.914	75.791	65.349
11	55.786	58.906	57.637	58.169	63.880	0
12	53.092	55.685	48.790	28.417	0	0
13	0	0	0	0	0	0
k	922.80	907.45	894.95	887.45	885.72	883.55
$E(d)$	5.544	5.851	6.101	6.251	6.286	6.329

a higher quadratic cost, the hospital treats a smaller overall number of patients at a higher average waiting time. Thus, as the unit cost progressively increases, lesser patients are treated quickly (with waiting of one or two periods) and from specification *(3)* and onwards the number of treatments is increasing in d for small durations. However, the clearance waiting time (q^*) remains the same until case *(4)*, and it actually decreases at high ρ_d levels. This could be attributed to the smaller number of overall treatments. The above mentioned admittance pattern is also reflected in the shapes of the survival and hazard functions. Apart from the first survival curve (dotted line) in Figure 4.10, all



(a) Survival Functions (b) Hazard Functions

Figure 4.10: Changes in ρ_d - Quadratic Cost Function (Table 4.10)

others start off further away from the origin and are concave for short durations. That is, the number of patients still waiting for treatment is decreasing, but at a decreasing rate. The shift in the curvature takes place at higher durations as ρ_d increases. The survival curves intersect. Thus, for instance, at $d = 3$, 63.6% of patients are still waiting to be treated in case *(1)*, while 88.1% are waiting in case *(6)*. However, at $d = 8$ (after the intersection) the reverse holds; 24.7% are still on the list for *(1)* against 19.2% for case *(6)*. Similarly, the hazard curves for specifications *(3), (5), (6)* are always increasing.

Concave survival functions are also observed empirically, especially at a disaggregate level. For example, in the specialty of ophthalmology (and in particular the operation of lens prosthesis) of Birmingham Hospital the concavity of the survival function is clear.

Changes in the budget of the hospital

Table 4.11 and Figure 4.11 illustrate the optimal number of patients treated from each duration and the relevant survival and hazard curves when the hospital has a higher budget, B, at its disposal. By increasing the budget of the hospital, we observe an increase in the total inflow (x_t) and outflow of patients

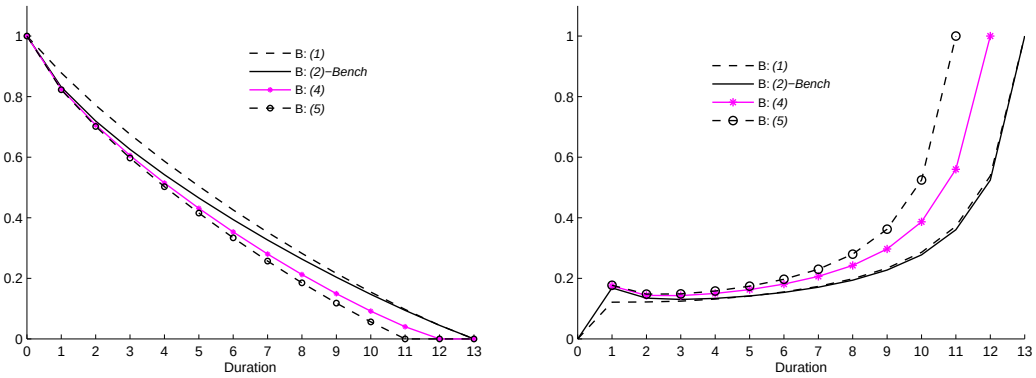
(k_t) and a quicker list clearance (q^* gets smaller). At the same time, the expected waiting time decreases. Ceteris paribus, with a bigger budget for elective surgery, the hospital has the opportunity to both treat above its capacity (scale) and to manage the (larger) list quicker.

Table 4.11: Changes in Budget - Optimal k_d at Steady State

	Lists						Proportions (PF)				
	(1)	*(2)*	*(3)*	*(4)*	*(5)*		*(1)*	*(2)*	*(3)*	*(4)*	*(5)*
$d\backslash B$	3150	7000	10500	20000	30000		3150	7000	10500	20000	30000
0	0	0	0	0	0		0	0	0	0	0
1	109.081	154.481	155.691	164.198	168.299		0.121	0.168	0.168	0.175	0.177
2	96.490	102.677	103.776	111.424	115.479		0.107	0.112	0.112	0.119	0.121
3	86.600	86.047	87.477	94.821	99.054		0.096	0.094	0.095	0.101	0.104
4	79.839	76.759	78.643	85.405	89.755		0.089	0.084	0.085	0.091	0.094
5	74.654	70.417	72.822	78.695	83.114		0.083	0.077	0.079	0.084	0.087
6	70.329	65.574	68.598	73.254	77.698		0.078	0.072	0.074	0.078	0.082
7	66.476	61.571	65.299	68.428	72.883		0.074	0.067	0.071	0.073	0.077
8	62.859	58.066	62.599	63.841	68.314		0.070	0.063	0.068	0.068	0.072
9	59.311	54.854	60.330	59.211	63.728		0.066	0.060	0.065	0.063	0.067
10	55.676	51.768	58.376	54.199	58.863		0.062	0.056	0.063	0.058	0.062
11	51.726	48.625	56.618	48.190	53.331		0.057	0.053	0.061	0.051	0.056
12	47.097	45.237	55.047	37.832	0		0.052	0.049	0.059	0.040	0.000
13	40.359	41.042	0	0	0		0.045	0.045	0.000	0.000	0.000
14	0	0	0	0	0		0	0	0	0	0
k	900.50	917.12	925.28	939.50	950.52	$E(d)$	5.99	5.65	5.49	5.21	4.98
% $c(k_d)$	0.999	0.581	0.392	0.218	0.149						

Moving from the benchmark budget, *(2)*, to *(4)* more patients with durations $d = 1, 2, ..., 10$ are admitted for treatment and the long waiters of duration 13 are eliminated. Despite the fact that the hospital now treats 23 more patients, it does so quicker ($q^* = 12$) and at a lower average waiting time. Comparing specifications *(3)* and *(4)*, where q^* stays the same at 12 periods, the hospital decides to treat a greater proportion of patients with $d = 1, 2, ..., 8$ while reducing the proportion of admitted people from the remaining periods. This behaviour reveals that increased budgets permit the hospital to reduce the number of people waiting a lot and admit more patients quickly. Increasing the budget beyond 30,000 is not affecting the steady state waiting time distribution. Given the fixed cost structure and capacity, the hospital cannot absorb

the excess funds (thus, the budget constraint holds strictly as an inequality). For the list to get shorter, we need to increase the hospital's capacity in line with the budget. This result indicates that policies aimed at improving hospital performance as regards waiting lists, must account for both types of investment, namely, monetary budget (flow) and capacity (stock).



(a) Survival Functions (b) Hazard Functions

Figure 4.11: Changes in the hospital's budget (Table 4.11)

The more funds the hospital has for elective surgery, the more the survival curves shift towards the origin without intersecting (Figure 4.11(a)). The admittance rates for treatment get greater throughout. The list also clears earlier, shown by the fact that they reach the x axis at 13, 12 and then 11 periods. Regarding the hazards, with the exception of a small decline between $d = 1$ and $d = 2$ (this is the case because, in all specifications, the hospital admits the larger number of patients for $d = 1$), they are all monotonically increasing and shifting to the left. As the list is gradually getting empty, the probability of treatment given still on the list gets bigger.

Changes in the scale capacity of the hospital

We continue our analysis with changes in the specification of the hospital's scale cost, $c(k) = \tau(k - \bar{k})^2$, both in terms of numbers of patients ($\bar{k}$) and the sensitivity to deviations from full capacity (τ).

Changes in capacity ($\bar{k}$)

As expected, steady state outflow (inflow) and expected waiting times are very responsive to the hospital's capacity. As $\bar{k}$ increases, the optimal waiting time to clear the list q^* decreases drastically, and the hospital treats more and more patients at shorter durations (Table 4.12). When capacity approaches the potential demand for health care (Z set at 1200), the distribution should eventually get eliminated[19].

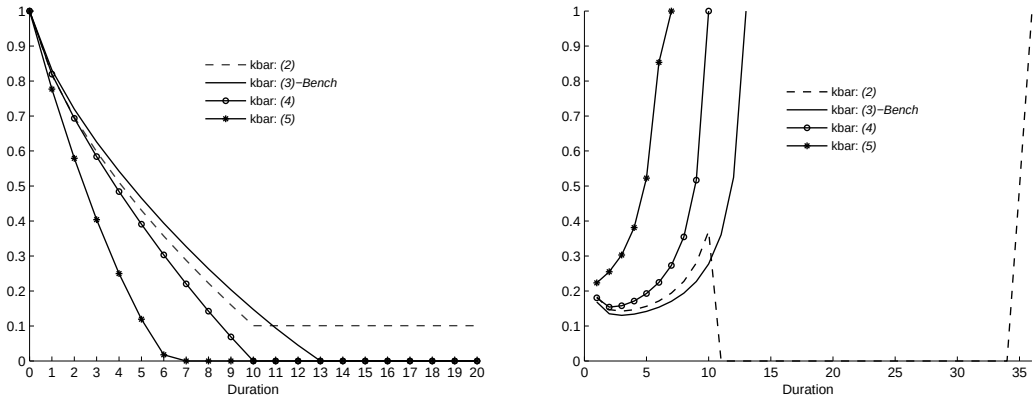
For relatively low capacity levels (below the benchmark case), long waiters at the bottom of the list are showing up. The hospital is heavily constrained, and given potential demand, attempts to restrain inflow by increasing expected waiting time, which for $\bar{k} = 700$ is almost 10 periods.

As shown in Figure 4.12, with increased capacity, the survival curves move closer to the origin and reach the $x-$axis quicker. Similar patterns are observed for the hazard curves. Specification *(2)* (dotted line) is an exception, since it stays to the left of the benchmark curve until they intersect at $d = 11$. The admittance rate is higher due to the smaller overall treatments (100 patients less).

[19] Given that the hospital's funds are in line with its high capacity and duration costs. Recall that here, for comparative statics reasons, the budget is fixed. That is why, when capacity is increased a lot, the hospital operates below capacity, since its budget is insufficient.

Table 4.12: Changes in Scale Cost: $\bar{k}$

$d\backslash\bar{k}$	*(1)* 700	*(2)* 800	*(3)-Bench* 900	*(4)* 980	*(5)* 1050
0	0	0	0	0	0
1	148.990	149.791	154.479	174.033	232.567
2	97.083	97.972	102.673	121.634	206.351
3	80.017	81.017	86.045	105.555	182.870
4	70.446	71.530	76.765	96.455	160.426
5	63.985	65.137	70.414	89.943	136.109
6	59.182	60.360	65.574	84.610	106.131
7	55.399	56.601	61.575	79.865	18.235
8	0	53.487	58.067	75.368	0
9	0	50.835	54.855	70.912	0
10	0	48.522	51.770	66.295	0
11	0	0	48.636	0	0
12	0	0	45.240	0	0
13	0	0	41.025	0	0
.	0	0	0	0	0
.	0	0	0	0	0
.	0	0	0	0	0
34	46.978	0	0	0	0
35	47.566	40.867	0	0	0
36	48.131	41.475	0	0	0
k	717.78	817.60	917.12	964.67	1042.69
E(d)	9.644	7.648	5.658	4.707	3.146



(a) Survival Functions (b) Hazard Functions

Figure 4.12: Changes in $\bar{k}$ (Table 4.12)

Changes in τ

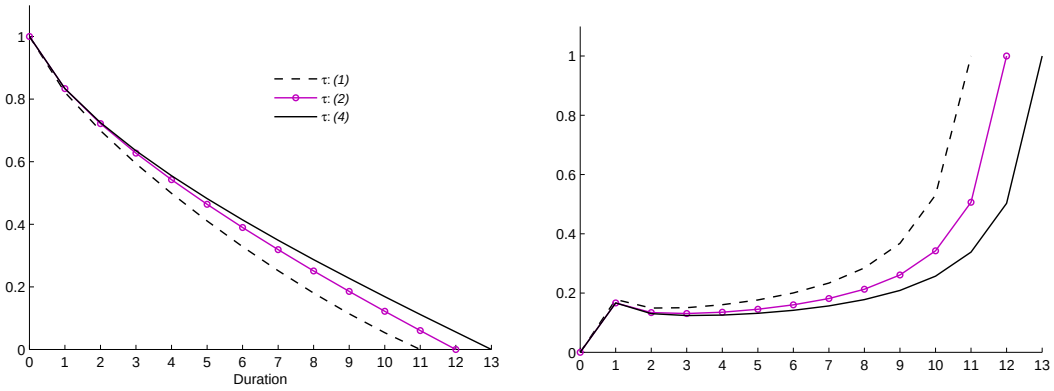
Recall that τ denotes how costly it is for the hospital when deviating from its full capacity. As τ increases, there is a decline in the overall number of treatments

Table 4.13: Changes in Scale Cost: τ

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
$d\backslash\tau$	0.5	5	Bench	50	100
	1	2	3	4	5
d	0.5	5	10	50	100
0	0	0	0	0	0
1	169.712	153.956	154.481	151.145	150.384
2	116.735	103.217	102.677	98.547	97.508
3	100.211	87.161	86.047	81.636	80.506
4	90.760	78.454	76.759	72.394	71.268
5	83.916	72.723	70.417	66.342	65.306
6	78.257	68.576	65.574	62.024	61.152
7	73.136	65.364	61.571	58.799	58.163
8	68.186	62.754	58.066	56.327	55.997
9	63.103	60.573	54.854	54.413	54.451
10	57.549	58.719	51.768	52.935	53.394
11	50.846	57.075	48.625	51.814	52.731
12	0	55.621	45.237	50.996	52.387
13	0	0	41.042	50.421	52.288
14	0	0	0	0	0
k	952.410	924.193	917.119	907.794	905.534
E(d)	4.952	5.516	5.658	5.844	5.889

and an expansion in both the clearance of the list and average waiting time, $E(d)$ (Table 4.13). Since it gets costlier for the hospital to operate above capacity, the overall number of treatments approaches 900 patients. Overall, a higher τ leads to a smaller number (and percentage) of patients treated quicker, at short durations up until $d = 7$ and there are more long waiters.

As depicted in Figure 4.13, the survival and hazard curves move proportionally further away from the origin.



(a) Survival Functions (b) Hazard Functions

Figure 4.13: Changes in τ (Table 4.13)

To sum up, when the budget or the scale cost structure of the hospital is altered, the corresponding response is an overall change in the admittance patterns. In particular, when the budget, capacity or the ability to operate above capacity are increased, the survival curves shift inwards, closer to the origin, and the hazard function shift upwards, reaching one faster. Therefore, the scale empirical differences observed when comparing the estimated survivals across different hospitals or specialities could be attributed to differences in the above mentioned parameters of the trust.

4.4.2 Waiting Time Targets

The waiting time target is imposed at an individual level, 'no patient should wait more than $\widehat{d}$ periods since he/she is added to the list', and is incorporated to our model in the cost function of the hospital, C_t, as an extra steep cost at the limit of the target. Denoting the targeted duration as $\widehat{d}$, then the extra cost faced by the hospital is:

$$\widetilde{c}(d) = \begin{cases} 0 & \text{if } d \leq \widehat{d} \\ \phi_d k_{d,s} & \text{if } d > \widehat{d} \end{cases}$$

Two are the important characteristics of our target structure: how restrictive the waiting target is (given by the level of $\widehat{d}$), and the magnitude of the cost/punishment when the target is breached ($\widetilde{c}(\widehat{d})$).

The first waiting time target we are introducing in the benchmark specification is a 12 period target; this indicates that no one should wait more than 12 periods until admittance for hospital treatment. In other words, the waiting time to clear the list should be $q^* = 12$, although without the target the benchmark steady state clearance waiting time is 13 periods (Column *(1)* in Table 4.14). For waiting times above $\widehat{d} = 12$ we start by assuming that ϕ_d is flat and independent of duration, meaning that the unit cost from breaching the target is the same irrespective of the duration at which the 'long' waiters are treated.

Table 4.14: Optimal k_d - Changes in the penalty of a target at 12 periods

	(1)-Bench	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
$d \setminus \phi_d$	0	5	10	20	30	50
0	0	0	0	0	0	0
1	154.4812	153.9767	153.4176	151.3251	150.9161	150.2007
2	102.6771	102.2808	101.8348	100.2629	99.9077	99.2741
3	86.0469	85.6767	85.2890	84.1744	83.8453	83.2729
4	76.7595	76.4430	76.1077	75.6485	75.3814	74.9038
5	70.4167	70.1717	69.9334	70.3173	70.1210	69.7870
6	65.5741	65.4184	65.2958	66.7112	66.6173	66.4662
7	61.5711	61.5332	61.5549	64.1887	64.2155	64.2793
8	58.0661	58.1807	58.3886	62.4078	62.5636	62.8640
9	54.8541	55.1606	55.6029	61.1587	61.4510	61.9989
10	51.7684	52.3209	53.0535	60.3044	60.7417	61.5388
11	48.6248	49.5179	50.6211	59.7590	60.3409	61.3825
12	45.2369	46.6442	48.2648	59.4528	60.1763	61.4559
13	41.0424	39.2652	36.7460	0	0	0
.	0	0	0	0	0	0
.	0	0	0	0	0	0
.	0	0	0	0	0	0
35	0	0	0	0	0	0
36	0	0	0	1.5387	1.0048	0
k	917.12	916.59	916.11	917.25	917.28	917.42
$E(d)$	5.658	5.668	5.678	5.655	5.654	5.652

As seen is Table 4.14, when the target is introduced and the penalty from

breaching it is not costly enough, the hospital starts by reducing the number of patients waiting at $d = 13$ (by increasing the admittance of patients in the durations before the targeted one). As the penalty from treating patients after having waited for 12 periods gets heavier, the hospital sets k_{13} equal to zero, but a couple of patients are left for treatment at the bottom of the list. Finally, for $\phi_d = 50$ and above, the target is met. Although the overall number of treatments and expected duration are very much in line with the pre-target levels, the two waiting time distributions are distinct. Comparing columns *(1)* and *(6)*, a trade-off between short and longer waiters is present; with a high penalty for treatments beyond the 12th period, the hospital is reducing the number of patients treated quickly, up until duration five, postponing the treatments for subsequent periods. The biggest increases in numbers of patients are observed in the periods right before the waiting time target. Finally, the 41 patients treated in *(1)* after waiting for 12 periods are eliminated in *(6)*.

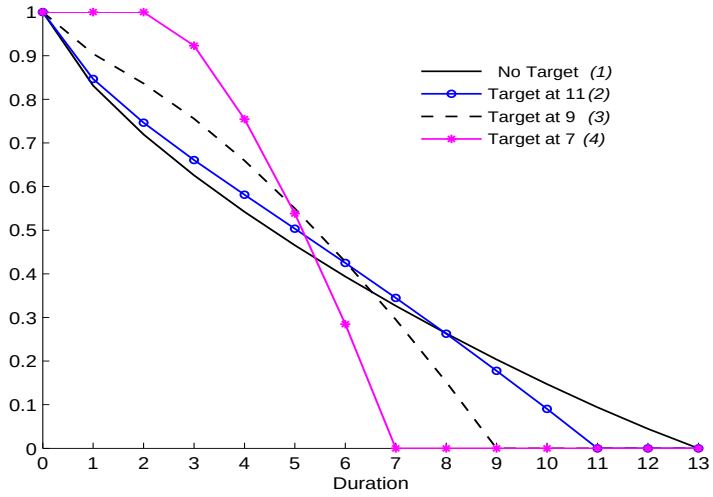
What happens when the target gets more stringent in terms of the waiting time? Table 4.15 depicts the hospital's response when the waiting time target is reduced to 11 periods. Here, the hospital is, *ceteris paribus*, asked to treat 86 patients quicker. The process of reaching target gets clearer. At low penalty levels, the hospital gradually reduces the number of patients previously waiting for more than 11 periods (k_{12}, k_{13}). As the penalty of breaching the target gets more sizeable, the hospital starts the process of eliminating the 'breached' waiters. However, this is done by creating a very long tail (treatments at the bottom of the list). This way, average waiting time is increased and overall treatments are lowered. The patients with $d = 13$ are eliminated first and then the ones that waited for 12 periods. When the cost becomes sufficiently high ($\phi_d = 350$), the longest possible waiters are eliminated and the target is reached.

Table 4.15: Optimal k_d - Changes in ϕ_d for a target at 11 periods

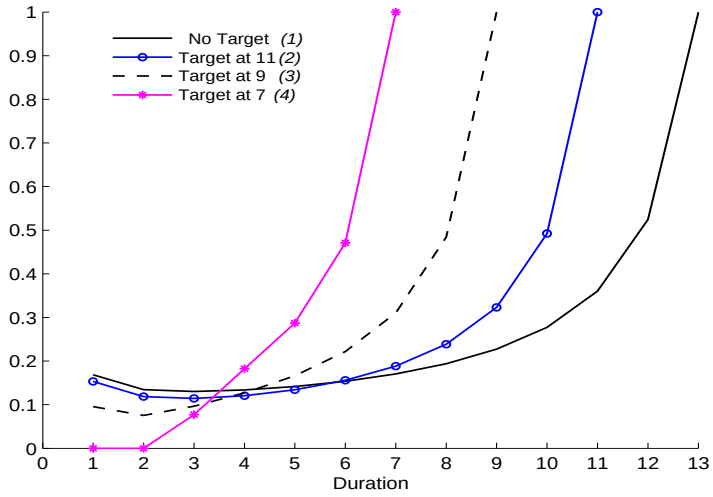
	(1)-Bench	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
$d\backslash\phi_d$	0	10	20	30	50	200	350
0	0	0	0		0	0	0
1	154.481	152.377	149.418	151.027	154.163	143.779	140.675
2	102.677	101.071	98.9741	100.713	103.551	95.3737	92.2389
3	86.0469	84.5882	82.7928	84.8542	87.8870	81.0117	78.4098
4	76.7595	75.5039	74.0738	76.4493	79.6936	74.6669	73.0670
5	70.4167	69.4379	68.5165	71.1991	74.5861	71.9036	71.5471
6	65.5741	64.9628	64.6670	67.6418	71.1136	71.0848	72.0047
7	61.5711	61.4334	61.8929	65.1367	68.6386	71.3932	73.5101
8	58.0661	58.5197	59.8721	63.3486	66.8300	72.3675	75.5618
9	54.8541	56.0346	58.4074	62.0776	65.4945	73.7357	77.8857
10	51.7684	53.8564	57.3600	61.1942	64.5093	75.3344	80.3324
11	48.6248	51.8920	56.6229	60.6094	63.7910	77.0638	82.8222
12	45.2369	43.9444	40.1972	40.2263	0	0	0
13	41.0424	41.0311	39.3670	0	0	0	0
.	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0
36	0	0	0	8.0352	14.5387	6.2070	0
k	917.119	914.652	912.161	912.512	914.797	913.921	918.055
$E(d)$	5.658	5.707	5.757	5434.572	5.704	5.722	5.639
%DurC	0.5813	0.693312	0.788721	0.776367	0.687224	0.723142	0.534304

The following graphs show the hospital's successful behaviour as the waiting time target gets more rigid for the same flat penalty set at $\phi = 2000$. We show the survival and hazard curves for the pre-target case together with the cases where the target is set at 11, 9 and finally 7 periods.

Once again we see how the hospital is managing to reach the ever restrictive targets. The trade-off between shorter and longer waiters is also apparent in the survival curves. Survival curves *(1)* and *(2)* have similar shapes and intercept at about $d = 8$ and. *(1)* and *(3)* intersect at about 6 periods, and the two trade off areas are already bigger. Up until $d = 6$, the cumulative admittance rate is higher for the pre-target case *(1)*, although the opposite occurs afterwards. For example, the percentage of patients waiting on the list for more than two periods is 72% for *(1)*, although it is about 83% for case *(3)*. On the contrary,



(a) Survival Functions



(b) Hazard Functions

Figure 4.14: Changes in the waiting time target - flat penalty

the percentage of patients still on the list after 8 periods is 26% for *(1)* and only 15.2% for *(3)*. In the latter, no one is waiting more than 8 periods. Moreover, in the three periods prior to the 8 period target, the hospital treats 394 patients or about 42% of the total admissions, but with no target the hospital treats half the amount. In survival curve *(4)* the waiting time target is set at 7 periods,

and the hospital manages its list, by having no treatments from durations one and two (i.e. the minimum waiting time for all patients is 3), 71 patients of waiting time three, 155 from $d = 4$, and the rest 698 patients treated in the last three periods prior to the targeted level. The curve starts off very far away from the origin and is decreasing at a decreasing rate until $d = 5$. The hazards curves move inwards as the targets get more stringent, reaching one at the corresponding target. In addition, they get steeper as the waiting time target is approached indicating the high admittances close to the target.

Therefore, at the steady state with the same resources, the hospital manages to eliminate the long waiters (i.e. patients previously treated after the set target) by reducing the amount of very short waiters and at the same time increasing the amount of medium waiters (increased treatments in the periods prior to the target). This 'manipulation' of the waiting time distribution is necessary so as to keep the steady state expected duration and overall number of treatments controlled. Note that for the targets set at 9 and 7 periods, overall admissions are increased and average waiting time is lowered, relative to the pre-target situation.

When the penalty structure is increasing in duration (for $d > \hat{d}, \phi_d = \alpha d$), the cost at longer waits is rising. Thus, the unit cost from breaching the target at, say, $d = 36$ is much bigger than at $d = 20$. In particular, when ϕ_d is set such that it matches the flat cost right after $\hat{d}$, we observe that, in the process of meeting the target, the hospital puts more effort in reducing the patients previously waiting straight after the targeted duration and has a smaller incentive to leave people for treatment at the end (36). In addition, the hospital meets the target 'easier', that is at a lesser penalty.

Let us consider now a case where a universal target is implemented to a set of hospitals differing with regards to the level of capacity. The penalty from breaching the target is increasing in waiting time (after $\hat{d}$). In particular, we

have $\phi_d = 20d$ and the target level is first set at 12 periods and then at 9. We focus on three cases where the capacity is 1050, 900 and 800 patients. The pre-target optimal lists are replicated in Table 4.16 together with the results from the two targets.

Table 4.16: Waiting Time Targets and different capacity

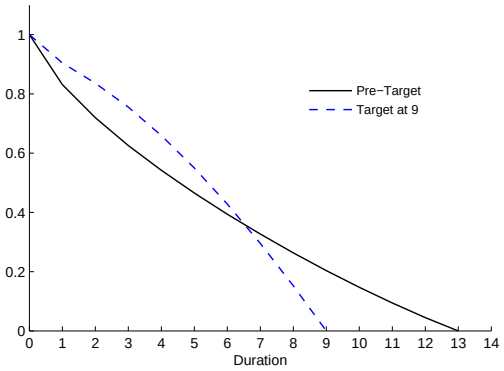
	Lists - No target			Lists - Target at 12			Lists - Target at 9		
	(1)	*(2)*-Bench	*(3)*	*(1)*	*(2)*-Bench	*(3)*	*(1)*	*(2)*-Bench	*(3)*
$d\backslash\bar{k}$	1050	900	800	1050	900	800	1050	900	800
0	0	0	0	0	0	0	0	0	0
1	232.567	154.479	149.791	232.567	150.200	104.740	232.567	87.878	0
2	206.351	102.673	97.972	206.351	99.275	43.919	206.351	62.780	0
3	182.870	86.045	81.017	182.870	83.269	0	182.870	74.383	0
4	160.426	76.765	71.530	160.426	74.915	0	160.426	88.451	0
5	136.109	70.414	65.137	136.109	69.786	54.032	136.109	101.064	68.137
6	106.131	65.574	60.360	106.131	66.458	65.657	106.131	112.273	136.502
7	18.235	61.575	56.601	18.235	64.279	74.860	18.235	122.371	174.333
8	0	58.067	53.487	0	62.863	82.737	0	131.605	203.640
9	0	54.855	50.835	0	61.992	89.743	0	140.133	228.412
10	0	51.770	48.522	0	61.540	96.095	0	0	0
11	0	48.636	0	0	61.396	101.960	0	0	0
12	0	45.240	0	0	61.450	107.420	0	0	0
13	0	41.025	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0
35	0	0	40.867	0	0	0	0	0	0
36	0	0	41.475	0	0	0	0	0	5.489
k	1042.69	917.12	817.60	1042.69	917.42	821.16	1042.69	920.94	816.51
$E(d)$	3.146	5.658	7.648	3.146	5.652	7.577	3.146	5.581	7.670

Clearly, in the case with ample capacity (1050) the pre-target list is cleared in 7 periods and the targets have no impact. In case *(2)* both the 12 period and 9 period targets are met. In particular, with the more restrictive target, the hospital ends up treating a few patients more (relative to the pre-target distribution) and at a slightly lower average waiting time. The majority of the patients are again treated with waiting times just prior to the targeted level, and apart from $d = 1$, the number of treated patients is increasing in duration.

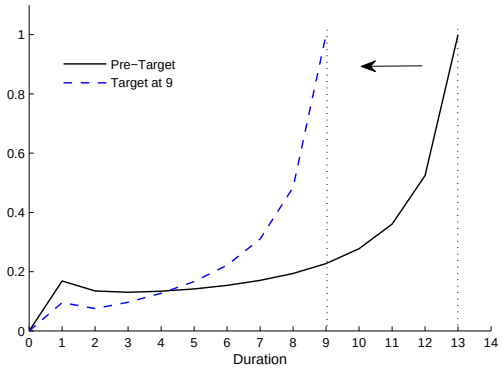
When the hospital has the capacity to treat 800 patients, it only meets the 12 period target, by decreasing the short waiters (durations 1-5), eliminating the very long waiters (35 and 36 periods of wait), and increasing the admissions in medium durations. However, the hospital fails to reach the 9 period target (with 5.5 patients treated after 36 periods of wait).

The first two subfigures in Figure 4.15 show the survival and hazard curves for the hospital with capacity of 900 patients, when we move from the pre-target distribution to a 9-period target. The waiting time target is achieved while the after target survival curve's curvature has changed (to concave). The hazard function moves leftwards (attaining unity at the corresponded targeted period), and it gets steeper prior to the target. The last two subfigures show the case of a hospital with capacity of 800 patients. The waiting time target is not achieved, however the steady state response of the hospital is apparent. The majority of patients (99%) are treated with maximum wait of 9 periods, and this is only attainable with decreasing the number of short-waiters substantially. Similarly, the spike of the hazard function moves inwards and close to the targeted waiting time.

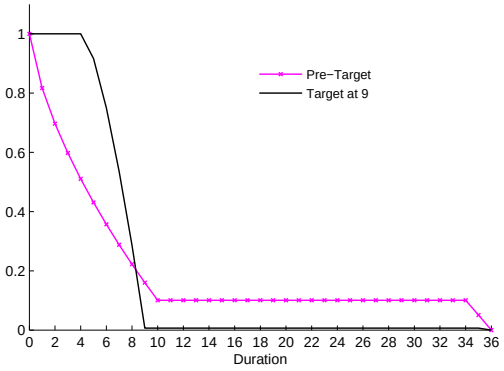
The introduction of the waiting time targets allows us to theoretically explore the waiting list management of the hospital, when such an important policy shift is implemented, as in the case of NHS since 1999-2000. Other things being equal, the waiting time target can be achieved by reducing both the long and short waiters, at the expense of higher admittance rates in medium durations (prior to the target). This distinct pattern is also observed to a higher or lesser degree in particular hospitals analysed in Chapter 3. Figure 3.22 compared the survival curves of nine hospitals at a pre-target year (1997) and an after-target year (2005, target set at 6 months). More than half of the hospitals exhibit the trade-off between short and long waiters. In the same lines, the response of the hazards curves is also similar; the peaks of the estimated hazard



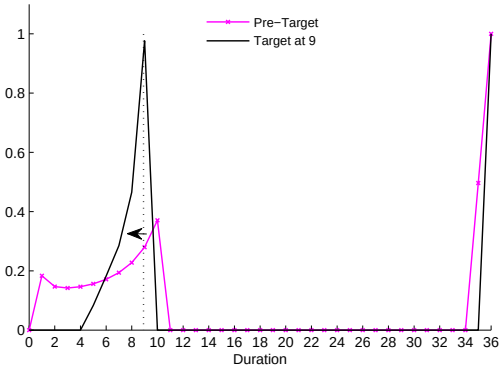
(a) Survival Functions - $\bar{k} = 900$



(b) Hazard Functions - $\bar{k} = 900$



(c) Survival Functions - $\bar{k} = 800$



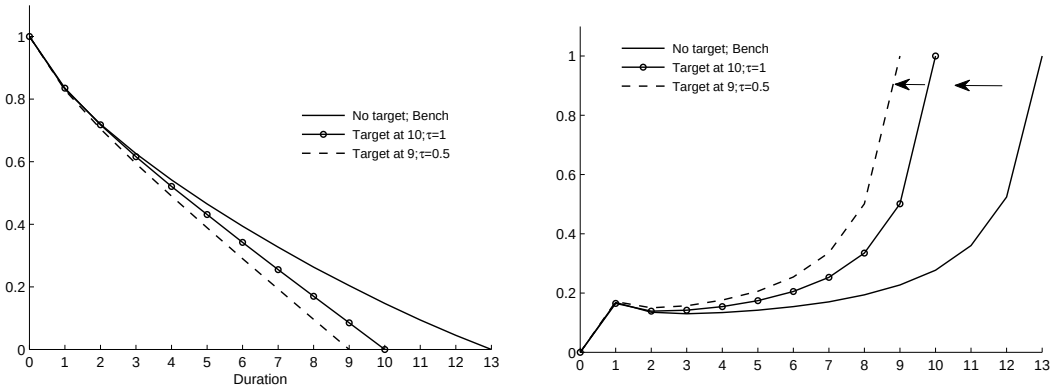
(d) Hazard Functions - $\bar{k} = 800$

Figure 4.15: Survival and Hazard Functions - Impact of Targets to different capacity (Table 4.16)

functions are moving inwards following the introduced targets.

Another empirically observed response to the waiting time target is an overall inwards shift of the survival curve towards the origin, which implies an improvement in admittance rates throughout the distribution. This pattern can be replicated here only if we allow for the budget or capacity parameters to change at the same time. Indicatively, we present a case below where in conjunction with the target, τ is decreasing. Hence, as the waiting time gets stricter the hospital's access to outsourcing is ameliorated.

As one can see from Figure 4.16, the survival curves shift inwards in line



(a) Survival Functions (b) Hazard Functions

Figure 4.16: Waiting time targets and changes in τ

with the tougher waiting time targets. Therefore, since it is less costly for the hospital to deviate from its capacity (by outsourcing for example private hospitals/clinics), the institutional target is achieved by increased admittances at any given duration.

4.5 Comparative Statics II: Severity Levels

Patients are now differentiated according to the level of severity of their health condition. We have two types of severity ($s = 1$, low) and ($s = 2$, high). This allows the hospital some degree of prioritisation of the list. We start with the parametarisation presented in the following table.

Table 4.17: Parameters specification with two levels of severity

$g(k_{d,s,t}) = a_{d,s}k_d^3 + b_{d,s}k_d^2 + c_{d,s}k_d$	Utility from treating k patients with duration d and severity s
where for the case of low severity:	parameters of the cubic utility function for low severity
$a_{d,1} = -0.0002 + \frac{0.0001}{d}$ $b_{d,1} = 0.02 - \frac{0.01}{d}$ $c_{d,1} = 2 + \frac{5}{d}$	
and for the case of high severity:	parameters of the cubic utility function for high severity
$a_{d,2} = (-0.0002 + \frac{0.0001}{d}) * 0.9$ $b_{d,2} = 0.02 - \frac{0.01}{d}$ $c_{d,2} = 3 + \frac{5}{d}$	
$c(k_{d,s,t}) = \rho_{d,s}k_{d,s,t}$	Cost from treatments at duration d and severity s
where $\rho_{d,1} = \frac{20}{d^2}$	parameters of the linear duration and severity cost function
and $\rho_{d,2} = \frac{30}{d}$	
$c(k) = \tau(k - \bar{k})^2$	Scale cost of the total number of patients treated
where $\bar{k} = 900$	Hospital's capacity in terms of number of patients
$\tau = 10$	sensitivity of cost to deviations from full capacity $\bar{k}$
$B = 13500$	Hospital's budget
$Z = 1200$	Potential demand for healthcare
$\theta = 50$	Sensitivity of inflow to expected waiting time
p = 0.7	Proportion of the milder treatments ($s = 1$)
q=36	Maximum allowed waiting time

The more severe cases have a higher utility gain, but at the same time they are more costly (for any given d)[20]. Given the magnitude of those two trading-off forces and in relation as well to the costs and gains of the milder cases, the hospital admits for surgery the severe cases (30% of the overall

[20]Given that the treatment specific cost has now increased (due to the high severity patients), the budget of the hospital is appropriately adjusted.

Table 4.18: Steady State List with Severities

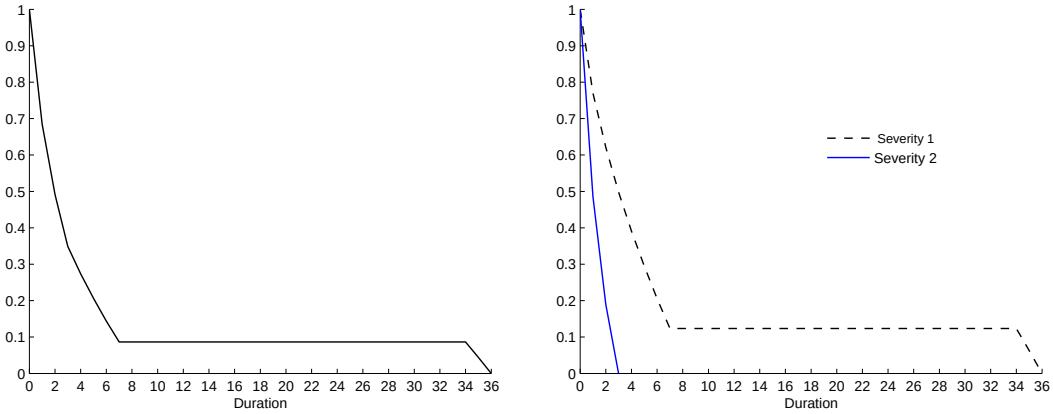
Duration	list1	list2	Agg List
0	0	0	0
1	147.922	141.765	289.686
2	96.075	81.802	177.877
3	78.818	52.199	131.016
4	68.982	0	68.982
5	62.223	0	62.223
6	57.045	0	57.045
7	52.847	0	52.847
8	0	0	0
.	0	0	0
.	0	0	0
.	0	0	0
35	39.219	0	39.219
36	40.322	0	40.322
k	643.45	275.765	919.218
$E(d)$	7.3044	1.6752	5.6156

treatments) much quicker ($q^* = 3$ and average duration at 1.6). At the same time the hospital also treats less severe cases (in a pattern similar to before) up until $d = 7$ and then the last 80 patients are treated at the end (after 35 and 36 periods of wait). The overall number of treatments is 919 and the overall average waiting time is 5.6 periods, although milder patients wait on average much more than more serious cases (Table 4.18).

Thus, the hospital prioritises the more severe cases. However, given that this entails a higher cost for the quicker treatment of the more severe patients ($c(k_{d,2})$), some of the milder cases are prolonged until the maximum possible duration, due to the budget the hospital has available.

As shown in Figure 4.17, the survival curve for the more severe is very close to origin, decreasing quite steeply and reaches zero after only three periods of wait. On the other hand, the survival function for the milder cases is further away from the origin throughout, decreasing much slower until $d = 7$, after which point it flattens, until the maximum duration of 36. The aggregate survival curve still displays the same long right tail, however, we also observe

a change in the rate of decrease. Thus, the cumulative admittance rates are relatively larger for the first three durations, and then slow down after that period of wait. In other words, the survival's (decreasing) slope is much steeper before $d = 3$ relative to after.

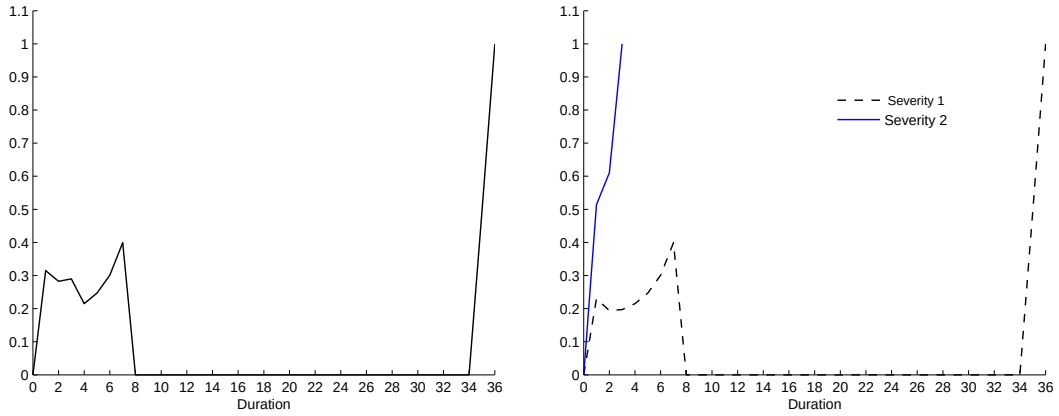


(a) Aggregate Survival Function (b) Survival Functions for $s = 1, 2$

Figure 4.17: Survival Functions with two Severity Levels (Table 4.18)

The majority of the treatments take place within the same period ($d = 1$), thus, the aggregate hazard curve decreases between $d = 1$ and 2. In addition, since the more severe cases are treated within the first three periods, we observe a second drop in the hazard function between durations 3 and 4. After that the conditional probability of being treated keeps increasing until duration 7, drops to zero for the next 28 periods and finally reaches one at duration 36.

As shown in Figure 4.18, the hazard curves for the two levels of severities are quite distinct. The hazard for $s = 2$ is quite steep, attaining one (i.e. clearance of the list) in three periods, while the hazard for $s = 1$ attains unity only in duration 36. The aggregate hazard curve is now more 'volatile' relative to the case with no severities. Apart from the typical first spike at $d = 1$ (since again the largest percentage of patients are treated after one period of wait), the hazard is decreased substantially at $d = 4$, due to the clearance of the more



(a) Aggregate Hazard Function (b) Hazard Functions for $s = 1, 2$

Figure 4.18: Hazard Functions with two severity levels (Table 4.18)

severe patients.

4.5.1 Changes in the Structural Parameters of the Model

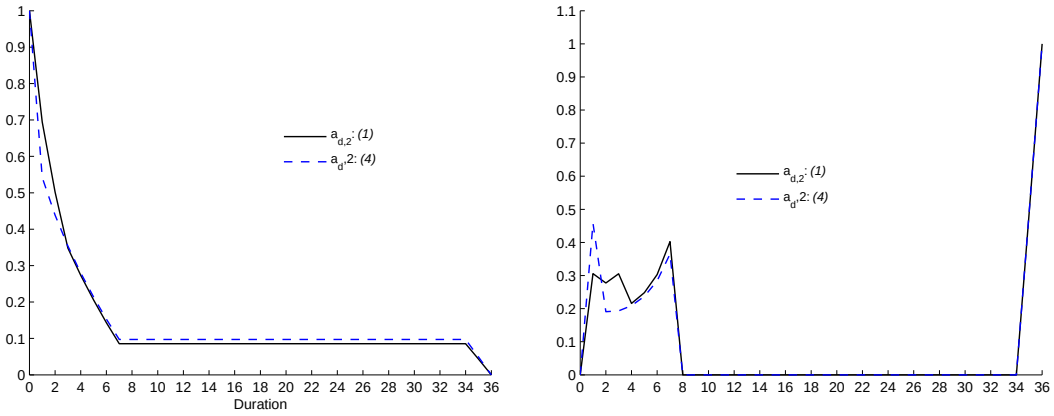
Changes in the utility function

Suppose that the gain the hospital obtains from treating more severe cases increases, by lowering (absolute) $a_{d,2}$. Recall that such a change implies that the turning point for $U(k_{d,2})$ moves to the right and the utility level at each $k_{d,2}$ gets larger (the utilities shift upwards).

As depicted in Table 4.19, the hospital has a higher incentive to treat the more severe patients quicker. Therefore, as (absolute) $a_{d,2}$ decreases, the list for $s = 2$ clears faster and in specification *(4)* all severe cases are treated within the same period (no waiting time distribution). In doing so, the hospital decreases a little bit the overall patients treated, coming mainly from the milder cases, while the aggregate average waiting time rises. In addition, although marginally, the treatment of patients of severity one is pushed further down (admitted for surgery slower).

Table 4.19: Increasing the Utility for the More Severe Patients

	Aggregate Lists				Lists - Severity 1				Lists - Severity 2			
	(1)	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*	*(1)*	*(2)*	*(3)*	*(4)*
$d \backslash a_{d,2}$	1.1	Bench	0.5	0.2	1.1	Bench	0.5	0.2	1.1	Bench	0.5	0.2
0	0	0	0	0	0	0	0	0	0	0	0	0
1	281.020	289.690	331.430	418.050	148.190	147.920	146.770	144.290	132.830	141.760	184.660	273.760
2	177.280	177.880	185.760	94.306	96.217	96.077	95.378	94.306	81.067	81.802	90.386	0
3	140.940	131.020	78.128	77.275	78.939	78.817	78.128	77.275	61.999	52.199	0	0
4	69.113	68.983	68.250	67.442	69.113	68.983	68.250	67.442	0	0	0	0
5	62.357	62.222	61.425	60.592	62.357	62.222	61.425	60.592	0	0	0	0
6	57.216	57.046	56.178	55.320	57.216	57.046	56.178	55.320	0	0	0	0
7	53.013	52.846	51.857	50.931	53.013	52.846	51.857	50.931	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0
35	38.791	39.221	41.413	43.880	38.791	39.221	41.413	43.880	0	0	0	0
36	39.932	40.320	42.374	44.748	39.932	40.320	42.374	44.748	0	0	0	0
k	919.662	919.228	916.815	912.544	643.768	643.452	641.773	638.784	275.896	275.761	275.046	273.76
$E(d)$	5.6067	5.6156	5.6635	5.7492	7.2625	7.3044	7.5214	7.7846	1.7433	1.6752	1.3286	1



(a) Aggregate Survival Functions (b) Aggregate Hazard Functions

Figure 4.19: Changes in $a_{d,2}$ - Increasing Utility for High Severity patients (Table 4.19)

Looking at the survival and hazard functions (Figure 4.19), the most noticeable changes between specifications *(1)* and *(4)* are observed during the first 3 durations, since it is that part of the overall distribution that is changing, driven by the quicker admittance of the severe cases.

We now consider the case where the utility specification for the milder cases is also altered. In particular, we assume that the cubic term for $s = 1$ is increased (in absolute terms) reducing the hospital's incentive to treat a lot of less severe patients up front. Table 4.20 and Figure 4.20 compare the steady state waiting time distribution for the benchmark specification *(1)* and the one in which $a_{d,1} = 2(-0.0002 + 0.0001/d)$, and $a_{d,2} = 0.5(-0.0002 + 0.0001/d)$, denoted *(2)*.

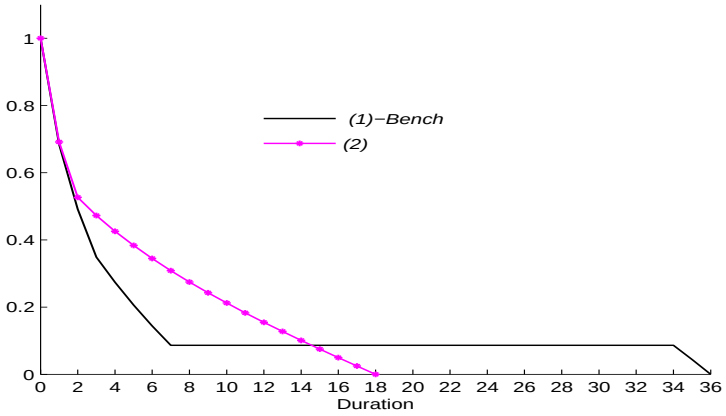
Comparing *(1)* to *(2)*, the incentive to 'front load' the more severe cases in increased and at the same time the incentive to balance out the list of the milder cases in strengthened. Thus, what we observe here is that the list for $s = 2$ gets cleared quicker (from $q_2^* = 3$ to 2 periods), while the patients with $s = 1$ are admitted for treatment in a much smoother way, relative to *(1)*. There are no long waiters at the end, but their inflow clears in 18 periods.

Table 4.20: Changes in both $a_{d,1}$ and $a_{d,2}$

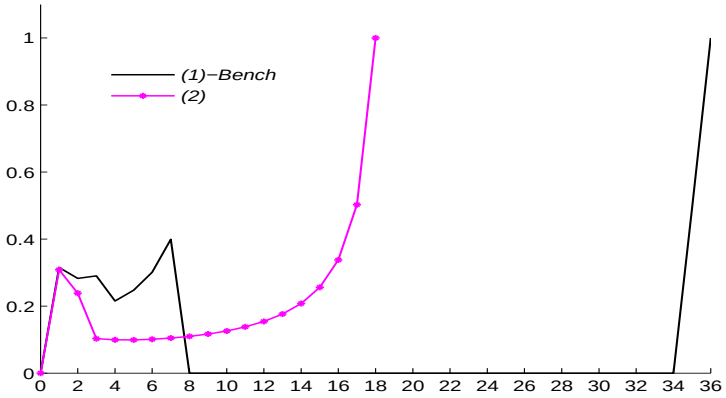
	Aggregate Lists		Lists - Severity 1		Lists - Severity 2	
	(1)	*(2)* $a_{d,1} = 2a_d$	*(1)*	*(2)* $a_{d,1} = 2a_d$	*(1)*	*(2)* $a_{d,1} = 2a_d$
$d \backslash a_{d,s}$	Bench	$a_{d,2} = 0.5a_d$	Bench	$a_{d,2} = 0.5a_d$	Bench	$a_{d,2} = 0.5a_d$
0	0	0	0	0	0	0
1	289.6864	283.7640	147.9216	97.8804	141.7648	185.8836
2	177.8773	151.7322	96.0751	61.5906	81.8022	90.1416
3	131.0161	49.7792	78.8176	49.7792	52.1985	0
4	68.9825	43.2234	68.9825	43.2234	0	0
5	62.2233	38.8548	62.2233	38.8548	0	0
6	57.0450	35.6430	57.0450	35.6430	0	0
7	52.8470	33.1528	52.8470	33.1528	0	0
8	0	31.1466	0	31.1466	0	0
9	0	29.4858	0	29.4858	0	0
10	0	28.0912	0	28.0912	0	0
11	0	26.9146	0	26.9146	0	0
12	0	25.9215	0	25.9215	0	0
13	0	25.0845	0	25.0845	0	0
14	0	24.3821	0	24.3821	0	0
15	0	23.8025	0	23.8025	0	0
16	0	23.3500	0	23.3500	0	0
17	0	23.0009	0	23.0009	0	0
18	0	22.7549	0	22.7549	0	0
19	0	0	0	0	0	0
.	0	0	0	0	0	0
.	0	0	0	0	0	0
.	0	0	0	0	0	0
35	39.2189	0	39.2189	0	0	0
36	40.3221	0	40.3221	0	0	0
k	919.2186	920.084	643.453	644.0588	275.7656	276.0252
$E(d)$	5.61563	5.5983	7.3044	7.4291	1.6752	1.3266

Overall the changes in average waiting time or total number of treatments are minimal. Despite those sizeable differences, note that the aggregate steady state outflows are not changing a lot for $d = 1, 2$ (first two columns of Table 4.20). That is, the two opposing drives (more severe patients but less milder patients treated quicker) offset each other at the aggregate level. This result also shows in Figure 4.20(a), where, the two aggregate survival curves are diverging from one another after $d = 2$. However, beyond that point, the two curves deviate noticeably and intersect at duration 15. Cumulatively, the benchmark case has

a higher admittance rate (until $d = 15$), but specification *(2)* clears the list much faster (at $q^* = 18$).



(a) Survival Functions



(b) Hazard Functions

Figure 4.20: Changes in utilities of both milder and severe cases

With regards to the hazard structures, specification *(2)* exhibits a clear spike at $d = 1$, it then decreases until $d = 3$, and after that in keeps on increasing steadily until the overall list is emptied at waiting period 18. In contrast, the hazard of the benchmark, while with more fluctuations, stays on top of *(2)* until waiting time 7. It then drops to zero and reaches one at the maximum possible period.

Changes in the Duration and Severity Specific Cost

We continue by analysing the optimal behaviour of the hospital when the duration and severity specific cost is altered. We concentrate in changing the power of $\rho_{d,s}$, which affects the vertical distances among the cost functions, making sure that the cost function of the more severe is still higher than the one of the milder cases. Table 4.21 depicts selected results, as $\rho_{d,s}$.

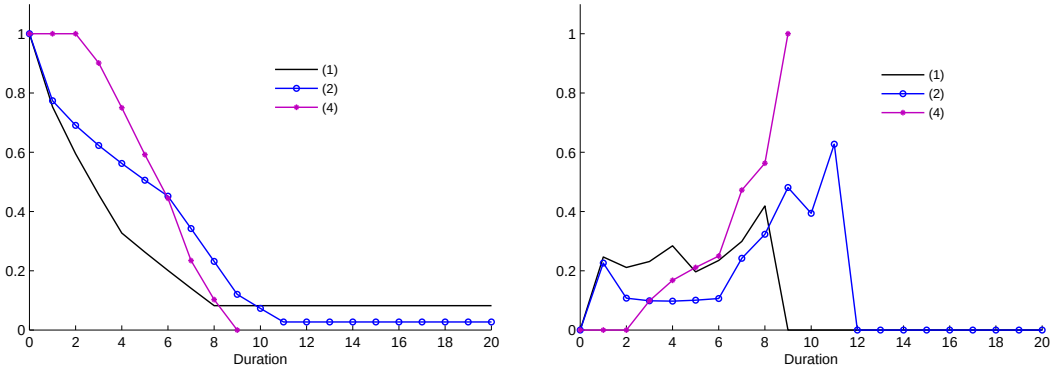
	(1)	(2)	(3)	(4)
$\rho_{d,1}$:	$\frac{20}{d^{0.4}}$	$\frac{20}{d^{0.3}}$	$\frac{20}{d^{0.25}}$	$\frac{20}{d^{0.2}}$
$\rho_{d,2}$:	$\frac{30}{d^{0.6}}$	$\frac{30}{d^{0.4}}$	$\frac{30}{d^{0.3}}$	$\frac{30}{d^{0.3}}$

As the treatment specific cost increases, we observe that a lesser number of patients are treated up front, at the aggregate level. In specification (4) admittance for elective surgery starts only after two periods of wait. Overall number of treatments and aggregate average waiting time remain unchanged. The reduction in very short waiters is reflected in both severity levels, although it is much more sizeable for $s = 2$ (see last three columns of Table 4.21). At the same time, the very long (mild) waiters of specification (1) are at first substituted by a smaller amount of more severe long waiters (2), and then eliminated, as the distribution of treated concentrates in the middle. It is important to notice the interplay in the admittance rates between the two severity levels. The hospital prefers treating more severe cases quicker, but this comes at an ever growing cost. Moving from (1) to (2), both individual average waiting times increase. However, from specification (2) onwards, there is a trade-off in the 'delaying' of patients: In (3) milder patients are waiting on average much less than severe patients; In (4) this gap is reduced.

Table 4.21: Changes in the Duration and Severity Cost Structure

	Aggregate Lists				Lists Severity 1				Lists Severity 2			
$d\backslash\rho_{d,s}$	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
0	0	0	0	0	0	0	0	0	0	0	0	0
1	221.946	203.341	118.364	0	126.016	120.564	118.364	0	95.930	82.776	0	0
2	143.144	74.876	76.765	0	80.778	74.876	76.765	0	62.367	0	0	0
3	123.846	61.272	65.948	88.558	67.645	61.272	65.948	88.558	56.201	0	0	0
4	117.018	54.644	61.072	135.745	61.363	54.644	61.072	135.745	55.655	0	0	0
5	57.794	50.825	58.024	142.274	57.794	50.825	58.024	142.274	0	0	0	0
6	55.606	48.360	55.558	132.753	55.606	48.360	55.558	132.753	0	0	0	0
7	54.181	98.402	104.538	188.628	54.181	46.604	53.157	109.245	0	51.798	51.381	79.383
8	53.176	99.716	104.226	118.688	53.176	45.185	50.521	20.536	0	54.531	53.705	98.152
9	0	100.359	102.119	92.084	0	43.902	47.319	0	0	56.457	54.801	92.084
10	0	42.617	98.066	0	0	42.617	42.929	0	0	0	55.137	0
11	0	41.168	54.828	0	0	41.168	0	0	0	0	54.828	0
12	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0
.	0	0	0	0	0	0	0	0	0	0	0	0
34	0	24.445	0	0	0	0	0	0	0	24.445	0	0
35	37.985	0	0	0	37.985	0	0	0	0	0	0	0
36	35.813	0	0	0	35.813	0	0	0	0	0	0	0
k	900.510	900.025	899.507	898.731	630.357	630.017	629.655	629.111	270.153	270.007	269.852	269.619
$E(d)$	5.990	6.000	6.010	6.025	7.586	8.225	4.715	5.159	2.265	4.715	9.031	8.047

Figure 4.21 depicts the survival and hazard curves for specifications (1), (2) and (4)[21]. As $\rho_{d,s}$ increases, we see the trade-offs between shorter and longer waiters in subfigure (a). Originally, much more patients were treated within the first 8 waiting periods, but a few were left for the end. In (2) more patients are still waiting to be treated (until $d = 10$), but less are waiting until the end. Finally, in (4), the curvature becomes concave (until approximately duration 6) and then switches to convex. In all three survival curves, we observe changes in the rate of change of the slope. With regards to the hazard curves, apart from (4) which monotonically increases, the other two exhibit three spikes, as the consequence of the change in the second derivative of survival curves (1) and (2).



(a) Aggregate Survival Functions (b) Aggregate Hazard Functions

Figure 4.21: Changes in the cost structure of both milder and severe cases

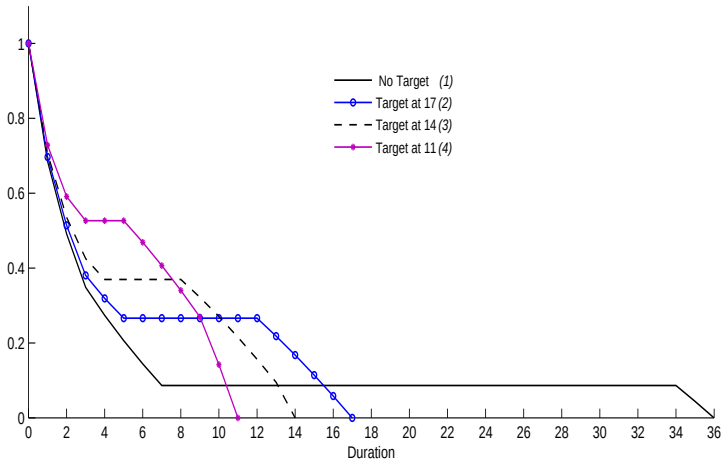
4.5.2 Waiting Time Targets with Severity Levels

As already discussed in Section 4.4.2, the waiting time targets are introduced at an individual level, and the hospital is faced with a penalty proportional to the number of patients and/or the periods after which the target level is surpassed.

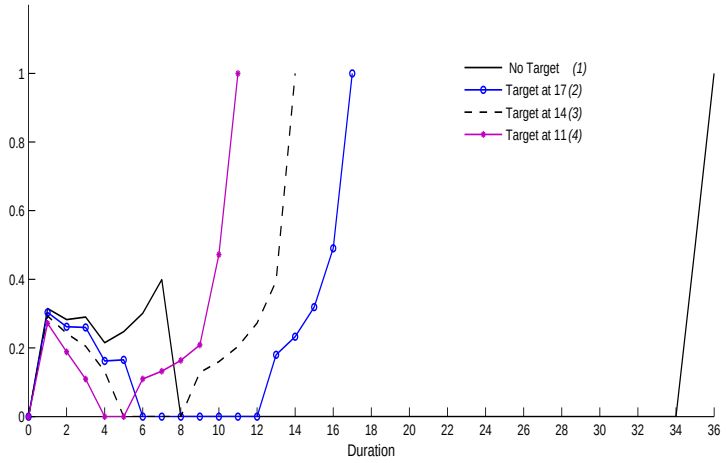
Recall that the benchmark parametarisation with severities has a long right

[21]The time span on Figure 4.21 has been reduced to 20 periods of wait, so as to have a clearer picture of the graphs. That is why, hazards (1) and (2) are not reaching one.

tail (since 80 of the milder patients are treated in periods 35 and 36). Table 4.22 and Figure 4.22, present the hospital's optimal response for a target starting at 17 periods of wait and decreasing to 14 and finally 11 periods of wait. The penalty from breaching the target is set at 400 irrespective of the severity level. All targets are met with the overall number of treatments and aggregate waiting



(a) Aggregate Survival Functions



(b) Aggregate Hazard Functions

Figure 4.22: Changes in the waiting time targets

time remaining almost at the same pre-target levels. Similarly to the case with no severities, at the aggregate level, the hospital's efforts to avoid breaching

the institutional target entail trading-off between short and long ('breached') waiters. As confirmed by Figure 4.22, aggregate survival functions reach zero at the corresponding targets, but their intersections suggest the above mentioned trade-offs. The hazard curves move inwards, reaching one at the ever restrictive targets, with much variation and different peaks during the first few durations.

However, having a closer look at the admission patterns across severities provides us with more insight on the hospital's behaviour. The first target (17 periods of wait) is achieved by manipulating the treatment of the milder cases almost exclusively; the admittance rates of $s = 2$ remain virtually unchanged. On the contrary, when the more restricted target of 14 periods in imposed, the average waiting time for $s = 2$ is increased, while for $s = 1$ decreases. For $s = 1$ both short and long waiters are decreased (at the expense of medium waiters), although in the case of $s = 2$ short waiters are substituted for long waiters. Thus, although before the introduction of the target all severe cases were treated within the first three periods, now 27 patients are postponed treatment until the targeted waiting time ($d = 14$). The average waiting time for $s = 2$ increases from 1.68 (pre-target) to 2.85 periods of wait. This pattern gets considerably evident when the target is set at 11 periods; 38% of the more severe patients are treated just before the target (d=10 and 11). Their waiting time further increases to 4.84.

Table 4.22: Impact of Targets in the Presence of Severity Levels

	No Target			Target at 17 periods			Target at 14 periods			Target at 11 periods		
$d\backslash\hat{d}$	All	List 1	List 2	All	List 1	List 2	All	List 1	List 2	All	List 1	List 2
0	0	0	0	0	0	0	0	0	0	0	0	0
1	289.686	147.922	141.765	279.054	139.205	139.849	270.253	135.023	135.230	250.734	129.223	121.511
2	177.877	96.075	81.802	167.889	86.679	81.210	157.796	82.305	75.491	126.812	77.432	49.381
3	131.016	78.818	52.199	122.974	68.006	54.969	101.125	63.136	37.988	59.752	59.752	0.000
4	68.982	68.982	0.000	56.756	56.756	0.000	51.490	51.490	0.000	0.000	0.000	0.000
5	62.223	62.223	0.000	48.475	48.475	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	57.045	57.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	53.341	53.341	0.000
7	52.847	52.847	0.000	0.000	0.000	0.000	0.000	0.000	0.000	57.233	57.233	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	61.429	61.429	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	43.754	43.754	0.000	65.530	65.530	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	47.535	47.535	0.000	117.450	69.453	**47.997**
11	0.000	0.000	0.000	0.000	0.000	0.000	50.935	50.935	0.000	131.366	73.161	**58.206**
12	0.000	0.000	0.000	0.000	0.000	0.000	54.066	54.066	0.000	0.000	0.000	0.000
13	0.000	0.000	0.000	44.050	44.050	0.000	56.957	56.957	0.000	0.000	0.000	0.000
14	0.000	0.000	0.000	46.777	46.777	0.000	87.316	59.658	**27.659**	0.000	0.000	0.000
15	0.000	0.000	0.000	49.193	49.193	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	0.000	0.000	0.000	51.421	51.421	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	0.000	0.000	0.000	53.502	53.502	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
35	39.219	39.219	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
36	40.322	40.322	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
k	919.22	643.45	275.77	920.09	644.06	276.03	921.23	644.86	276.37	923.65	646.55	277.09
$E(d)$	5.62	7.30	1.68	5.60	7.27	1.69	5.58	6.74	2.85	5.53	5.82	4.84

These results provide evidence of trade-offs across both the duration and the severity of the patients on the list. Elimination of the long-waiters (which were milder cases) is achieved by delaying the admittance of short-waiters (which corresponded to more severe cases). In other words, our theoretical model suggests that targets change the order by which the hospital takes patients off the list so as to influence the speed of patients treated. This is achieved by substituting less urgent cases for more urgent cases. The introduction of stricter targets distorts clinical priorities. The hospital, thus, is considering both the target and the severity of patients when admitting patients for treatment. It is important, however, to stress that this result is observed as a comparative statics exercise in which the introduction of the target is the only factor we allow to change. As discussed in Section 4.4.2 this result would be different if the hospital, together with the introduction of the waiting time target, was also allocated with more operational resources (increased budget, capacity or outsourcing).

4.6 Concluding Remarks

In the final chapter of this thesis, we develop a theoretical model of the supply side for healthcare provision. Two are the distinct features of our theoretical framework (i) the dynamic element of the model and (ii) the derivation of the entire optimal waiting time distribution of patients treated at the steady state. The focus of the chapter is on, first, matching important empirical patterns of the hospitals, and second identifying possible factors that can explain the observed patterns and the differences among them.

Our theoretical model proves to be quite flexible in accounting for different patterns of waiting time distributions. Differences in the parameters of the

utility of the hospital account for two distinct admissions patterns: (i) the hospital 'front loads', treating many patients are quickly as possible, at the expense however of a small fraction that waits for long. Thus, more emphasis is put on increased short waits, and (ii) the hospital prefers a 'smoother' waiting list distribution where patients receive treatment more gradually, but no one waits extensively; the emphasis is on the long waits. When the treatment specific cost is different, we observe again differences in the shape of the survival curves, but now the curvature is altered. When the cost for quick treatment is increased, the survival curves exhibit concave parts. Finally, changes in the resources allocated to elective surgery (budget, capacity, cost of operating above capacity) produce changes in the instantaneous admissions rate for the whole distribution, thus we observe shifts in the positing of the survival (and hazard) curves.

The introduction of waiting targets provides additional insight of the hospitals' response to such a policy shift. Other things being equal, the hospital manages to eliminate the long waiters (i.e. patients previously treated after the set target) by reducing the amount of very short waiters and at the same time increasing the amount of medium waiters (increased treatments in the periods prior to the target). This trade-off between shorter and longer waiters has also been empirically confirmed at several levels in Chapters 2 and 3. In addition, increased admission rates throughout the waiting list can be attained when, together with the target, more spending for elective surgery is allowed.

The identified essential patterns of the hospital's behaviour are also apparent with the introduction of different severity levels. What is important to highlight here is that the aggregate survival curves and hazard curves get 'richer' shapes, matching the empirical estimated distributions in Chapters 2 and 3 closer. Survival curves display various changes in the rate of change of the (decreasing) slope, and hazard functions are more volatile, showing more

spikes. This could suggest that in practice some form of prioritisation is also taking place, since, the extra variation is a result of interplays not only over the patients waiting time, but also across their level of severity (more versus less urgent cases). Furthermore, when allowing for differences in the severity of the patients' health status, the introduction of strict waiting time targets affects the prioritisation of the list.

The theoretical model succeeds in replicating the distinct waiting time distributions observed empirically and at the same time provides valuable insight on the potential factors that may explain the distinct behaviour patterns.

CHAPTER 5

Epilogue

Much attention has been devoted by academics and policy makers to the question of whether the policy measure of targets has been meeting its objective. In other words, did hospitals meet the targets? Is there anybody waiting more than the corresponding maximum target set by the government? However, in accordance with this query should lie another one aiming at examining the ways employed by hospitals to abide to national standards. Thus besides seeking an answer to the before mentioned question, questions such as 'How did hospitals meet the targets?' 'How are they managing the lists and waiting times of their patients?' and 'What are the steps taken to improve performance?' are of similar importance. From both a policy and academic perspective it would be essential to comprehend the actions taken to manage the process of elective admissions, in an attempt to learn from the successful trusts and at the same time avoid poor management from the under-performing institutions.

This thesis investigated these exact issues by *unwinding* the whole waiting time distribution of patients. Our main contribution was exposing the great

level of variability in waiting time distributions and implied admission tactics by hospital, specialty and operative procedure. Furthermore, we found that the wait distributions tend to change over time. The results from the theoretical model successfully replicated the distinct waiting time distributions observed empirically and at the same time provided valuable insights on the potential factors that may explain these distinct behavioural patterns.

Previous studies have shown that hospitals have eliminated the extremely long waits for elective treatment. The evidence we presented confirms this positive effect of targets on hospitals' performance. Not only did we find evidence that waiting times have indeed decreased, but as a result of the more detailed view of the waiting time distributions after the estimation of the survival and hazard functions, we were able to detect details in their patterns. In particular, we found that increasing probabilities of admission expressed as peaks in the hazard curves coincided with the prevailing waiting time target and that the introduction of shorter targets coincided with a reduction in the waiting time at which the new peak occurred. This implies that providers tend to increase their effort as the target approaches and decrease it after the target.

As seen in Chapter 3, the patterns of survival rates differed substantially by hospital and by doctor. In particular, we observed a significant variation in both the shape and scale of their survival curves (plasticity). Some curves, while retaining the same curvature, exhibited abrupt changes in the magnitude of the slope, while others altered from convex to concave or vice versa. These distinct differences in the slope (magnitude or sign) of the survival functions, corresponded to spikes in the hazard curves. Additionally, we found evidence of survival curves that move closer to the origin and others that shift rightwards, in a parallel way or not. The different shapes reflect differences in the second order derivative of the removal rate of individuals from the list and variation in the shifts implies that hospitals admit all patients with a slower or quicker

rate compared to others.

Trends in hazard curves also differed markedly between various sets of hospitals and consultants. We observed trusts/doctors with notable peaks of high intensity, others with very short wider peaks and finally some with constant hazard rates expressed as straight lines. Lastly, we reported cases in which hazards were illustrated as monotonically increasing probabilities of admission. These results indicate differences in the management of the lists, in the decision process and admission criteria, even when we control for particular characteristics of the list (type of hospital, operation, and even by physician).

In the final chapter of the thesis, we managed to incorporate the waiting time distribution of patients within the utility maximisation model of the hospital. As a result, our model proved useful in interpreting the different empirical patterns. Differences in the parameters of the utility of the hospital, the treatment specific cost and the hospital resources accounted for distinct admissions patterns and hazard rates. With the introduction of waiting time targets, other things being equal, the hospital managed to eliminate the long waiters by reducing the amount of short waiters (trade-off). The introduction of different severity levels led to the appearance of 'richer' shapes of the survival and hazard curves, matching the empirical estimated distributions in Chapters 2 and 3 closer. Furthermore, when allowing for differences in the severity of the patients' health status, the introduction of strict waiting time targets affects the prioritisation of the list.

Finally, future research can be directed towards a more thorough empirical investigation on the influence of hospital characteristics on waiting times for elective surgery. These were not included in the regression analysis performed in Chapter 2 as they were not part of HES data. However, as Chapter 3 indicated, there is scope in undertaking a systematic empirical analysis on the impact of supply-side factors at a hospital level on the observed waiting time

distributions. Moreover, the theoretical model could be expanded to include the demand side in a more explicit manner (e.g. considering alternative treatments (private hospitals), the possibility of decay of health status while waiting or the gatekeeping role of GPs in determining 'effective' demand). A very interesting but at the same time challenging path for future research is also the derivation and exploration of the hospital's admission patterns off the steady state (off-equilibrium behaviour of providers).

Bibliography

Allison, P. (1984). *Event history analysis: Regression for longitudinal event data*. 46, Sage Publications, Incorporated.

Alvarez-Rosete, A., Bevan, G., Mays, N. and Dixon, J. (2005). Effect of diverging policy across the NHS. *BMJ*, **331** (7522), 946–950.

Appleby, J., Boyle, S., Devlin, N., Harley, M., Harrison, A. and Locock, L. (2005a). *Sustatining Reductions in Waiting Times: Identifying Successful Stategies. Final report to the Department of Health*. Tech. rep.

—, —, —, —, — and Thorlby, R. (2005b). Do english NHS waiting times targets distort treatment priorities in orthopaedics? *Journal of Health Services Research and Policy*, **10** (3), 167–72.

— and Coote, A. (2002). *Five-year health check: A review of Government health policy 1997-2002*. King's Fund.

— and Devlin, N. (2004). *Measuring success in the* NHS*: Using patient-assessed health outcomes to manage the performance of health care providers*. Tech. rep.

—, Robinson, R., Ranade, W., Little, V. and Salter, J. (1990). The use of markets in the health service: The NHS reforms and managed competition. *Public Money & Management*, **10** (4), 27–33.

Barros, P. P. and Olivella, P. (2005). Waiting lists and patient selection. *Journal of Economics & Management Strategy*, **14** (3), 623–646.

Barzel, Y. (1974). A theory of rationing by waiting. *Journal of Law and Economics*, **17** (1), 73–95.

Besley, T., Bevan, G. and Burchardi, K. (2008). Accountability and incentives: The impacts of different regimes on hospital waiting times in england and wales. *London School of Economics Working Papers, University of London*, pp. 1–20.

—, — and — (2009). *Naming & Shaming: The impacts of different regimes on hospital waiting times in England and Wales.* Centre for economic policy research.

— and Ghatak, M. (2003). Incentives, choice, and accountability in the provision of public services. *Oxford Review of Economic Policy*, **19** (2), 235–249.

—, Hall, J. and Preston, I. (1998). Private and public health insurance in the UK. *European Economic Review*, **42** (3-5), 491–497.

—, — and — (1999). The demand for private health insurance: Do waiting lists matter? *Journal of public economics*, **72** (2), 155–181.

Bevan, G. and Hood, C. (2006a). Health policy: Have targets improved performance in the english NHS? *BMJ: British Medical Journal*, **332** (7538), 419.

— and — (2006b). What's measured is what matters: Targets and gaming in the english public health care system. *Public administration*, **84** (3), 517–538.

Bird, S. M., Cox, S. D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005). Performance indicators: Good, bad, and ugly. *Journal of the Royal Statistical Society Series A*, **168** (1), 1–27.

Brekke, K., Siciliani, L. and Straume, O. (2008). Competition and waiting times in hospital markets. *Journal of Public Economics*, **92** (7), 1607–1628.

Brouwer, W., van Exel, J., Hermans, B. and Stoop, A. (2003). Should I stay or should I go? Waiting lists and cross-border care in the Netherlands. *Health Policy*, **63** (3), 289–298.

Burge, P., Devlin, N., Appleby, J., Rohr, C. and Grant, J. (2005). London patient choice project evaluation. *RAND Corporation*.

Burgess, S. and Ratto, M. (2003). The role of incentives in the public sector; Issues and Evidence. *Oxford Review of Economic Policy*, **19** (2), 285–300.

Cohen, M., Naylor, C., Basinski, A., Ferris, L., Llewellynthomas, H. and Williams, J. (1992). Small-area variations-what are they and what do they mean. *Canadian Medical Association J.*, **146**, 467–470.

Collett, D. (2003). *Modelling survival data in medical research*, vol. 57. Chapman and Hall/CRC.

Cox, D. and Oakes, D. (1984). *Analysis of survival data*. Chapman and Hall.

Croxson, B., Propper, C. and Perkins, A. (2001). Do doctors respond to financial incentives? UK family doctors and the GP fundholder scheme. *Journal of Public Economics*, **79** (2), 375–398.

Cullis, J. G. and Jones, P. R. (1976). Some economics of hospital waiting lists in the NHS. *Journal of Social Policy*, **5** (03), 239–64.

— and — (1986). Rationing by waiting lists: An implication. *American Economic Review*, **76** (1), 250–56.

—, — and Propper, C. (2000). Waiting lists and medical treatment: Analysis and policies. In A. J. Culyer and J. P. Newhouse (eds.), *Handbook of Health*

Economics, *Handbook of Health Economics*, vol. 1, *23*, Elsevier, pp. 1201–1249.

Dimakou, S., Parkin, D., Devlin, N. and Appleby, J. (2009). Identifying the impact of government targets on waiting times in the NHS. *Health care management science*, **12** (1), 1–10.

Dixon, H. and Siciliani, L. (2009). Waiting-time targets in the healthcare sector: How long are we waiting? *Journal of Health Economics*, **28** (6), 1081–1098.

Dixon, S. (2004). Trends in waiting time to date and total time waited: Are the sources compatible? *Health Statistics Quarterly*, **24**, 23–29.

Eibich, P. and Ziebarth, N. R. (2013). Analyzing regional variation in health care utilization using (rich) household microdata. *Health Policy*.

Elandt, R. and Johnson, N. (1999). *Survival models and data analysis*. John Wiley & Sons.

Ellis, R. (1998). Creaming, skimping and dumping: Provider competition on the intensive and extensive margins. *Journal of Health economics*, **17** (5), 537–556.

Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics*, **5** (2), 129–151.

Farnworth, M. (2003). A game theoretic model of the relationship between prices and waiting times. *Journal of health economics*, **22** (1), 47–60.

Fleming, T. and Harrington, D. (1991). *Counting processes and survival analysis*, vol. 8. Wiley Online Library.

Folland, S. and Stano, M. (1990). Small area variations: A critical review of propositions, methods, and evidence. *Medical Care Research and Review*, **47** (4), 419–465.

Forrest, C. (2003). Primary care gatekeeping and referrals: Effective filter or failed experiment? *Bmj*, **326** (7391), 692–695.

Gérvas, J., FERNA, M. and STARFIELD, B. (1994). Primary care, financing and gatekeeping in Western Europe. *Family practice*, **11** (3), 307–317.

Glied, S. (2000). Managed care. In A. J. Culyer and J. P. Newhouse (eds.), *Handbook of Health Economics*, *Handbook of Health Economics*, vol. 1, *13*, Elsevier, pp. 707–753.

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

González, P. (2005). On a policy of transferring public patients to private practice. *Health Economics*, **14** (5), 513–527.

Gravelle, H., Dusheiko, M. and Sutton, M. (2002). The demand for elective surgery in a public system: Time and money prices in the UK national health service. *Journal of Health Economics*, **21** (3), 423–449.

—, Smith, P. and Xavier, A. (2003a). Performance signals in the public sector: The case of health care. *Oxford Economic Papers*, **55** (1), 81–103.

—, — and — (2003b). Waiting lists and waiting times: A model of the market for elective surgery. *Oxford Economic Papers*, pp. 81–103.

Hadorn, D. and Holmes, A. (1997). The New Zealand priority criteria project. Part 1: Overview. *BMJ*, **314** (7074), 131.

Hamblin, R., Harrison, A. and Boyle, S. (1998). Waiting lists. The wrong target. *Health Serv J.*, **108**, 28–31.

Hanning, M. (1996). Maximum waiting-time guarantee an attempt to reduce waiting lists in Sweden. *Health policy*, **36** (1), 17–35.

Harrison, A. and Appleby, J. (2005). *The War on Waiting for Hospital Treatment: What Has Labour Achieved and what Challenges Remain*. King's Fund.

Hauck, K. and Street, A. (2007). Do targets matter? A comparison of English and Welsh national health priorities. *Health Economics*, **16** (3), 275–290.

Hoel, M. and Saether, E. M. (2003). Public health care with waiting time: The role of supplementary private health care. *Journal of Health Economics*, **22** (4), 599–616.

Hosmer, D., Lemeshow, S. and May, S. (2011). *Applied survival analysis: regression modeling of time to event data*, vol. 618. Wiley-Interscience.

Iversen, T. (1993). A theory of hospital waiting lists. *Journal of Health Economics*, **12** (1), 55–71.

— (1997). The effect of a private sector on the waiting time in a National Health Service. *Journal of Health Economics*, **16** (4), 381–396.

Johnson, D. H. (1999). The Insignificance of Statistical Significance Testing. *Journal of Wildlife Management*, **63** (3), 763–772.

Kalbfleisch, J. and Prentice, R. (1980). *The statistical analysis of failure time data*. John Wiley & Sons.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53** (282), 457–481.

Le Grand, J. (1991). Quasi-markets and social policy. *The Economic Journal*, **101** (408), 1256–1267.

—, Mays, N. and Mulligan, J. (1998). *Learning from the* NHS *internal market*. King's Fund.

Levy, A., Sobolev, B., Hayden, R., Kiely, M., J.M., F. and Schechter, M. (2005). Time on wait lists for coronary bypass surgery in British Columbia, Canada, 1991 - 2000. *BMC Health Services Research*, **5** (22).

Lindsay, C. M. and Feigenbaum, B. (1984). Rationing by Waiting Lists. *American Economic Review*, **74** (3), 404–17.

MacCormick, A. and Parry, B. (2003). Waiting time thresholds: Are they appropriate? *ANZ Journal of Surgery*, **73** (11), 926–928.

Martin, S., Rice, N., Jacobs, R. and Smith, P. (2007). The market for elective surgery: Joint estimation of supply and demand. *Journal of Health Economics*, **26** (2), 263–285.

— and Smith, P. (1999). Rationing by waiting lists: An empirical investigation. *Journal of Public Economics*, **71** (1), 141–164.

— and — (2003). Using panel methods to model waiting times for National Health Service surgery. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **166** (3), 369–387.

Miller, R., Gong, G. and Muñoz, A. (1981). *Survival analysis*. Wiley New York.

Mullen, P. (1992). *Waiting Lists and the* NHS *Review: Reality and Myths.* HSMC Research Report 29, Birmingham University, Health Services Management Centre.

Olivella, P. (2002). Shifting public-health-sector waiting lists to the private sector. *European Journal of Political Economy*, **19** (1), 103–132.

Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. and Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. Analysis and Examples. *British journal of cancer*, **35** (1), 1.

Propper, C. (1990). Contingent Valuation of Time Spent on NHS Waiting Lists. *Economic Journal*, **100** (400), 193–99.

— (1995). Agency and incentives in the NHS internal market. *Social Science & Medicine*, **40** (12), 1683–1690.

—, Burgess, S. and Gossage, D. (2007a). Competition and quality: Evidence from the NHS internal market 1991-1999. *The Economic Journal*, **118** (525), 138–170.

—, — and Green, K. (2004). Does competition between hospitals improve the quality of care? Hospital death rates and the NHS internal market. *Journal of Public Economics*, **88** (7), 1247–1272.

— and Söderlund, N. (1998). Competition in the NHS internal market: An overview of its effects on hospital prices and costs. *Health economics*, **7** (3), 187–197.

—, Sutton, M., Whitnall, C. and Windmeijer, F. (2007b). *Did 'Targets and Terror' Reduce Waiting times in England for Hospital Care?* The Cen-

tre for Market and Public Organisation 07/179, Department of Economics, University of Bristol, UK.

—, —, — and — (2008a). Did targets and terror reduce waiting times in England for hospital care? *The B.E. Journal of Economic Analysis & Policy*, **8** (2), 5.

—, —, — and — (2008b). *Incentives and Targets in Hospital Care: Evidence from a Natural Experiment*. The Centre for Market and Public Organisation 08/205, Department of Economics, University of Bristol, UK.

—, —, — and — (2010). Incentives and targets in hospital care: Evidence from a natural experiment. *Journal of Public Economics*, **94** (3-4), 318–335.

— and Wilson, D. (2003). The use and usefulness of performance measures in the public sector. *Oxford review of economic policy*, **19** (2), 250–267.

—, — and Söderlund, N. (1998). The effects of regulation and competition in the NHS internal market: The case of general practice fundholder prices. *Journal of Health Economics*, **17** (6), 645–674.

Siciliani, L. (2005). Does more choice reduce waiting times? *Health Economics*, **14** (1), 17–23.

— (2006). A dynamic model of supply of elective surgery in the presence of waiting times and waiting lists. *Journal of Health Economics*, **25** (5), 891–907.

— and Hurst, J. (2003). *Explaining Waiting Times Variations for Elective Surgery Across* OECD *Countries*. OECD Health Working Papers 7, OECD Publishing.

— and — (2005). Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 oecd countries. *Health policy*, **72** (2), 201–215.

—, Stanciole, A. and Jacobs, R. (2009). Do waiting times reduce hospital costs? *Journal of Health Economics*, **28** (4), 771–780.

Smith, P. (1990). The use of performance indicators in the public sector. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 53–72.

— (1993). Outcome-related performance indicators and organizational control in the public sector1. *British Journal of Management*, **4** (3), 135–151.

Sobolev, B., Brown, P. and Zelt, D. (2000). Variation time spent waiting list elective vascular surgery: A case study. *Clinical and Investigative Medicine*, **23** (4), 227–238.

—, — and — (2001). Modeling and Analysis of Multistate Access to Elective Surgery. *Health Care Management Science*, **4** (2), 125–132.

Sobolev, B. G., Brown, P. M., Zelt, D. and FitzGerald, M. (2005). Priority waiting lists: Is there a clinically ordered queue? *Journal of Evaluation in Clinical Practice*, **11** (4), 408–410.

Wennberg, J. and Gittelsohn, A. (1973). Small area variations in health care delivery a population-based health information system can guide planning and regulatory decision-making. *Science*, **182** (4117), 1102–1108.

Weon, B. and Je, J. (2011). Plasticity and rectangularity in survival curves. *Scientific reports*, **1**.

— and — (2012). Trends in scale and shape of survival curves. *Scientific Reports*, **2**.

Windmeijer, F., Gravelle, H. and Hoonhout, P. (2005). Waiting lists, waiting times and admissions: an empirical analysis at hospital and general practice level. *Health economics*, **14** (9), 971–985.

Worthington, D. (1987). Queueing models for hospital waiting lists. *Journal of the Operational Research Society*, pp. 413–422.

— (1991). Hospital waiting list management models. *Journal of the Operational Research Society*, pp. 833–843.

Xavier, A. (2003). Hospital competition, GP fundholders and waiting times in the UK internal market: The case of elective surgery. *International Journal of Health Care Finance and Economics*, **3** (1), 25–51.

Yates, J. (1987). *Why are We Waiting?: An Analysis of Hospital Waiting-lists*. Oxford University Press Oxford.

Official Reports

Commission for Health Improvement (2003), NHS Performance Ratings. Acute Trusts, Specialist Trusts, Ambulance Trusts 2002/2003, London. (http://www.chi.nhs.uk/Ratings/Downloads/CHIAcuteF.pdf)

Department of Health (1997a). Changing the Internal Market. Executive Letter, (97) 33.

Department of Health (1997b). The New NHS: modern, dependable, London: HMSO.

Department of Health (2000). The NHS plan: A plan for investment. A plan for reform, London: The Stationery Office, 2000: Cmd 4818-I.

Department of Health. (2003). Choice of Hospital. Guidance for PCTs, NHS Trusts and SFAs on offering patients choice of where they are treated.

Department of Health. (2005). Choice at six months: Good practice. Gateway No4843.

Department of Health (2001), NHS Performance Ratings. Acute Trusts 2000/01, London.
http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/
PublicationsPolicyAndGuidance/DH_4003181)

Department of Health (2002), NHS Performance Ratings. Acute Trusts, Specialist Trusts, Ambulance Trusts, Mental Health Trusts 2001/02, London.

(http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/
PublicationsPolicyAndGuidance/DH_4002706)

Healthcare Commission (2004), NHS performance ratings 2003/2004, London.
(http://ratings2004.healthcarecommission.org.uk/Downloads/4662_HC_ratings.pdf)

Healthcare Commission (2005), NHS performance ratings 2004/2005, London.
(http://ratings2005.healthcarecommission.org.uk/Downloads/MoreInformationPageDocs/
Performance_ratings.pdf)

National Audit Office. Inpatient and Outpatient waiting in the NHS. Report by the Comptroller and Auditor General. HC 221 Session 2001-2002: 26 July 2001. National Audit Office: TSO, London.

National Audit Office. Inappropriate adjustments to NHS waiting lists. Report by the Comptroller and Auditor General. HC 452 Session 2001-2002: 19 December 2001.