CURING THE QUEUE

Maartje Elisabeth Zonderland

Dissertation Committee

Chairman & Secretary	prof. dr. ir. A.J. Mouthaan
Promotor	prof. dr. R.J. Boucherie
Assistant-promotors	dr. F. Boer
	dr. N. Litvak
Members	prof. dr. J.H. van Bockel
	prof. dr. N.M. van Dijk
	dr. ir. E.W. Hans
	prof. dr. J.L. Hurink
	prof. dr. M. Lambrecht
	prof. dr. D.A. Stanford

This research has been partly funded by Leiden University Medical Center, Leiden, the Netherlands.

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Center for Telematics and Information Technology (No. 11-214, ISSN 1381-3617)
Beta Research School for Operations Management and Logistics (No. D146)
Center for Healthcare Operations Improvement and Research

This dissertation was edited with TeXnicCenter and typeset with LaTeX. The graphics were created using Microsoft Excel, Powerpoint and Visio, and Adobe Acrobat Pro. The models were coded in Waterloo Maple (Chapters 3 and 6), Borland Delphi (Chapter 4), and The MathWorks MATLAB (Chapters 5–7, and 9).

Printed by Gildeprint Drukkerijen, Enschede, the Netherlands

ISBN 978-90-365-3306-5

CURING THE QUEUE

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 27 januari 2012 om 14:45 uur

door

Maartje Elisabeth Zonderland
geboren op 24 januari 1982
te Warnsveld

Dit proefschrift is goedgekeurd door de promotor,
prof. dr. Richard J. Boucherie

en de assistent-promotoren,
dr. Fredrik Boer en dr. Nelly Litvak

Voor mijn familie

Voorwoord

Het schrijven en vooral het afronden van dit proefschrift was niet mogelijk geweest zonder de hulp van een aantal personen. Om te beginnen wil ik mijn promotor, Richard Boucherie, bedanken. Richard, met je schijnbaar oneindige energie en enthousiasme, niet alleen voor het onderzoek maar ook voor de implementatie van de resultaten in de praktijk, ben je voor mij een voorbeeld geweest. Ik hoop dat wij nog eens samen zullen werken in de toekomst.

Dan mijn assistent promotoren, Fred Boer en Nelly Litvak. Fred, dankzij jou ben ik in het LUMC blijven werken na mijn afstuderen. Jij hebt me steeds met de juiste mensen in aanraking gebracht en gezorgd dat ik de praktijk niet uit het oog verloor. Nelly, je wiskundig inzicht is fenomenaal. Door jou ben ik mijn formules echt gaan begrijpen.

De leden van mijn commissie, allereerst Erwin Hans, Hajo van Bockel en David Stanford: Erwin, met jou is het ooit allemaal begonnen. Eerst bij DOBP op de (vroege) woensdagochtend, later weer tijdens mijn afstuderen. Je enthousiasme is aanstekelijk. Hajo, het is een eer dat je in mijn commissie hebt plaats genomen. Bedankt voor je voortdurende belangstelling in mijn onderzoek. David, I highly enjoyed our cooperation - first at CRHE in Toronto and later at UWO in London. Thank you for the Sushi lunches. Ook de overige commissieleden, Johann Hurink, Marc Lambrecht, Nico van Dijk, Ton Mouthaan, en de vervangend voorzitter, Pieter Hartel, wil ik bedanken voor hun bijdrage aan mijn promotie.

Mijn LUMC collega's van het bureau bedrijfsvoering Divisie 1, en dan met name Ben Nijman, Leontine den Dijker, Jos Gubbi, Patty Verhoeven en Paula van der Hilst. Ben, het was echt top om voor en met jou te werken en te zien dat de organisatie veranderde. Leontine en Jos, bedankt voor alle gezelligheid op de kamer en de liters thee die jullie voor mij hebben gezet. Patty, bedankt voor alle queries die je voor me hebt gedraaid. Paula, bedankt voor alle Leidse gezelligheid, (juist) ook buiten het werk. Verder wil ik ook de artsen en verpleging van het LUMC bedanken voor de prettige samenwerking in diverse projecten.

Op de UT: mijn SOR en CHOIR collega's. Vanaf het begin heb ik mij bij de SOR leerstoel erg thuis gevoeld. Bedankt voor alle gezelligheid en het leuke contact. Alle CHOIR AIO's: het is erg leuk geweest om in een groep jonge enthousiaste mensen te werken. Bedankt voor alle gezelligheid op vrijdag en tijdens congressen. Tevens wil ik de studenten

noemen die ik heb mogen begeleiden tijdens het afronden van hun Bachelor of Master studie: Siebe Brinkhof, Jurjen Tjoonk, Daphne Looije, Mik Schous, Astrid Stallmeyer, Nicole Havinga en Thomas Schneider.

Ook wil ik alle coauteurs van de publicaties die de basis vormen voor dit proefschrift bedanken. In het bijzonder: Ad Vletter voor het verzamelen en interpreteren van de data benodigd voor Hoofdstuk 3, Nikky Kortbeek voor Hoofdstuk 4: na een wat naïeve planning in het begin is het nu toch echt bijna af. Ahmad Al-Hanbali voor de hulp met de analyse van het model in Hoofdstuk 5, en Judith Timmer voor Hoofdstuk 6: in het begin was het even wennen maar het uiteindelijke resultaat mag er zijn. Carmen Vleggeert-Lankamp voor Hoofdstuk 7 en 8: onze samenwerking heeft geresulteerd in twee leuke hoofdstukken en tot nu toe één publicatie. Ik heb bewondering voor je inzet en enthousiasme. Anouk Streeder voor het bijhouden van de data voor Hoofdstuk 8 en alle hulp bij het verwerken ervan. Michael Carter: thank you for giving me the opportunity to visit CRHE in October and November of 2010. Our cooperation lead to the basis of what later would be Chapter 9 of this dissertation. Van een Hoofdstuk 10 is het helaas niet meer gekomen, maar ik wil Daisy Koks toch bedanken voor alle moeite om dit onderzoek van de grond te krijgen en af te ronden.

En dan ten slotte mijn familie, schoonfamilie en vrienden. In het bijzonder wil ik mijn ouders bedanken voor hun steun en vertrouwen. Daarnaast mijn paranimfen Kirstin van Lijden en Nienke Nijhof: het is goed om deze vier jaar met jullie samen af te ronden. Ook Judith Suurenbroek en Ingeborg van Gessel horen in dit rijtje thuis; helaas mocht ik geen vier paranimfen meenemen. Lieve Rens, de tweede helft van mijn promotie met jou was vele malen leuker dan de eerste helft zonder jou. Ik hoop dat er nog vele jaren zullen komen.

Laag-Soeren,
December 2011

Contents

Part I

Introduction

Chapter 1

Challenges in Modern Healthcare Delivery

1.1 Introduction

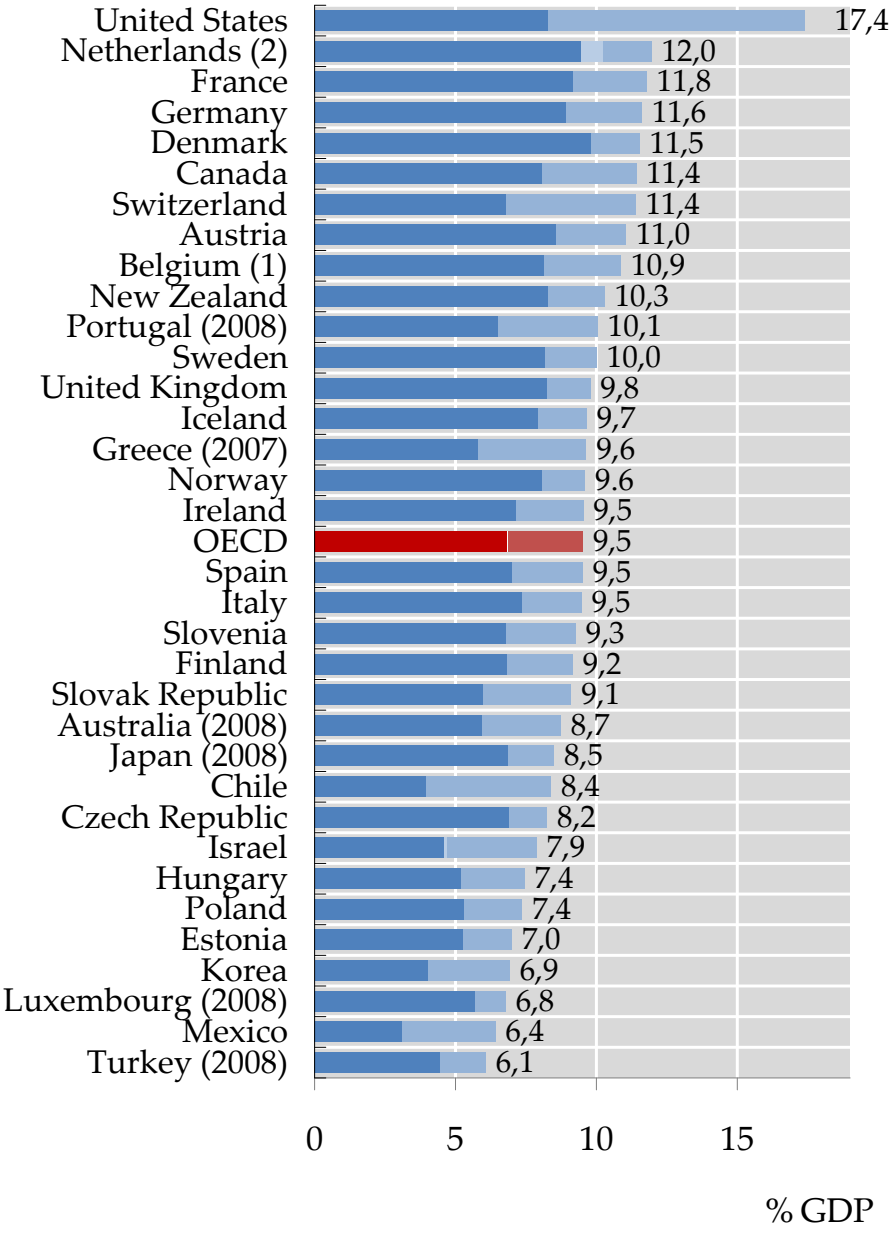
In 2006, the Dutch government dramatically reformed the healthcare sector and the underlying financial system [198]. The main purposes of the reforms were to decrease costs and improve efficiency. More freedom for care providers and patients was introduced: care providers were allowed to employ commercial initiatives and could make (limited) choices regarding the patient groups they would like to treat; patients could more or less freely choose where they wanted to be treated. Since inhabitants of the Netherlands are obliged to buy health insurance, the government decided to give the health insurers a major role in enforcing the new paradigm of market thinking in the Dutch healthcare system.

In terms of quality and efficiency, the Dutch healthcare system performs about average compared to other western countries [169]. An aging population, increased use of technology and a society demanding a higher quality and accessibility of care, are among others reasons that healthcare costs in developed countries consume a larger part of the Gross Domestic Product (GDP) every year (see Figure 1.1). The Netherlands is one of the countries whose healthcare system faces immense financial challenges, now and in the future.

Since the financial funds and thus the supply of healthcare is finite, policy makers have to ration care and make choices on how to distribute physical, human, and monetary resources. Such choices also have to be made at the hospital level (e.g., which patient groups will be treated in this hospital), and on a departmental level (e.g., which patient gets which available bed). An extra challenge involved with an aging population is that the total working population, and thus the number of healthcare professionals decreases, while the part of the population that requires care increases. With the current

hospital efficiency levels it will be difficult, if not impossible, to provide an appropriate level of care for the sick and the elderly in the coming decades.

Figure 1.1: Total health expenditure as share of GDP, 2009. The dark gray (first) part of the bar in the chart represents the public share, while the lighter gray (second) part of the bar represents the private share. (1): Total expenditure excluding investments. (2): In the Netherlands, it is not possible to distinguish clearly the public and private share for the part of health expenditures related to investments. Source: OECD Health Data 2011 [144]



1.2 Curing the Queue

"Managers make resource allocation decisions, but doctors decide what the hospital does with those resources" [39]. Even though this statement is ten years old, it is still the status quo. The interests of doctors and managers will eventually be conflicting at some point. While doctors focus on treating each individual patient as well as they possibly can, managers also focus on optimal usage of resources. One can imagine that this easily leads to ethical dilemma's; what if the treatment of a single cancer patient costs 100K Euros, while five other patients suffering from cardiovascular disease can be treated for 20K Euros each? Should a single patient with a mean length of stay (LOS) of 20 days be admitted at an inpatient ward, or should four patients with a mean LOS of 5 days be admitted sequentially instead?

These dilemma's easily show the difficult decisions doctors and healthcare managers have to make. It is however very common in hospitals to avoid explicit decisions on resource allocation and capacity distribution and to react on ad-hoc basis to problems that occur. Sometimes this is accompanied with very undesirable system outcomes (e.g., patients canceled for surgery several times, unused (scarce) time at outpatient clinics, extremely long waiting times).

The models we present in this dissertation allow for a quantification of consequences of capacity distribution decisions. The item that is distributed can either be time, or another kind of resource such as staffed beds. Since each nurse has a limited amount of time during a working day, this is ultimately also a time distribution problem. With the models a clear and succinct understanding of the problem, its possible solutions, and implications of these solutions can be obtained. Of course, the decision is then still not easy. But hopefully doctors and managers then have a profound idea of what they are actually deciding upon.

Hospital departments often function as separate islands, and have their own, sometimes conflicting interests. A low level of integration with other departments is common [74, 75]. It comes at no surprise that many efficiency improvement studies also focus on single departments [189]. However, departments may have a significant influence on each other [129]. This is (partly) recognized by the increasing popularity of care pathways. In a care pathway, care is optimized for patients with identical characteristics (e.g., symptoms, disease, age, etc.). All steps in the care process (for example outpatient consultation, diagnostic testing, surgery, hospitalization, and so on) are meticulously described and planned. An adverse consequence of prioritizing patients in a care pathway is a suboptimal care process for regular patients.

Together with care pathways, techniques from operations management and operations research, such as lean, theory of constraints, six sigma and simulation [209], have gained increased attention in the last decade. Even though the results in the (mostly theoretical) studies are usually promising and show room for efficiency gain, most techniques from industry are not directly applicable [136] and careful study is required to choose the

right technique.

In this dissertation, which consists of three parts, we study several problems that are related to the management of healthcare and the cure of disease. In every chapter a hospital capacity distribution problem is analyzed using operations research techniques. An immediate consequence of rationing resources is the expansion of queues, and it comes at no surprise that the usage of queuing theory to study healthcare problems has increased in the last years. This is not only visible in the operations research journals (see for example [29, 80, 156, 190, 213]) but also from the medical journals (e.g., [69, 135, 185, 188, 212]).

1.3 Stochastic Operations Research in Healthcare

The mathematical field of operations research, or decision science, has emerged from military applications in the first half of the 20th century. In most operations research problems a complex decision needs to be made, where several constraints and interests of various stakeholders need to be taken into account. Since the end of last century, complex decision problems emerging from the healthcare sector have gained increased attention from operations researchers. Well developed areas include benchmarking of healthcare facilities using data envelopment analysis [94], nurse rostering [33], operating room planning and scheduling [38], appointment scheduling in outpatient clinics [40], and simulation studies to improve patient flow [102]. We suggest [99] for a structured review of the literature.

In stochastic operations research, problems are studied that involve decision making under uncertainty. This basically means that at least one parameter or variable in the problem is random. In most cases a probability distribution is used to account for the stochasticity. Since life involves many uncertainties, one can imagine that techniques from stochastic operations research are very well applicable to model real life problems, for instance from the healthcare domain. The field of stochastic operations research includes queuing theory, Markov decision theory and game theory, which are the three techniques that are used in this dissertation to tackle the complex healthcare problems we came across.

Queuing theory, which analyzes waiting times and service levels in service systems, originated from telecom problems. It is the most invoked approach in this dissertation, rather than simulation, which is another, widely used, approach to analyze healthcare problems (see [26, 102]). One of the advantages of simulation modeling compared to queuing modeling is the possibility to take into account any desired system characteristic. This is at the same time also one of the major drawbacks of this method, since one might get lost in the details and lose sight of the real problem. In order to perform a simulation study, a large amount of data and computation time is required [47], which makes it very time consuming. Performing a mathematical analysis gives the modeler a

fundamental insight in the problem. In this dissertation we show, among other things, the added value of queuing theory in the complex process of decision making in healthcare.

1.4 Applied Research Environment

The majority of the research presented in this dissertation is inspired by logistical challenges faced by Leiden University Medical Center (LUMC). The LUMC is situated in the historic city of Leiden, and serves together with eight other general hospitals a community of around two million people in an urban area in the south-west of the Netherlands. The main focus of the LUMC is top clinical and highly specialized care. It is the smallest and oldest of the eight academic hospitals in the Netherlands and employs around 7,000 people. For 2010, almost 500,000 outpatient clinic visits, more than 200,000 diagnostic procedures, over 10,000 surgeries were registered, and the average inpatient LOS was 6.4 days. The major patient flows and their dimensions are given in Figure 1.2. The level as to which the research findings have been implemented in a hospital setting varies and is summarized in Table 1.1. Since the models that are developed are of generic nature, they can be directly applied to represent another hospital than LUMC.

Table 1.1: Level of implementation in LUMC (if not mentioned otherwise) per dissertation Chapters 3 – 9

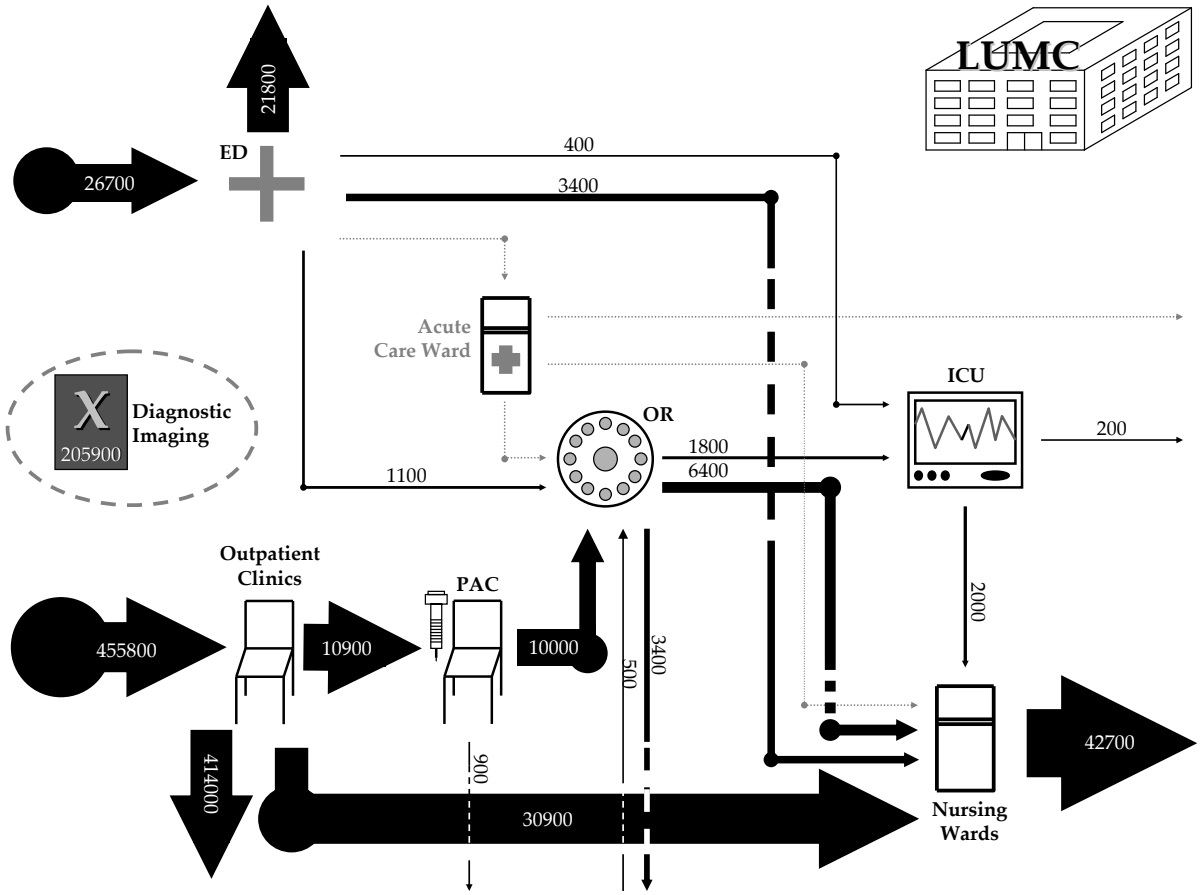
Chapter	Level of implementation
3	Findings completely implemented
4	Implementation studies at AMC and LUMC
5,6	Theoretical
7,8	Partially implemented
9	Theoretical

1.5 Structure of this Dissertation

This dissertation consists of three parts. In Figure 1.3 is shown how the chapters relate to the hospital departments as also shown in Figure 1.2.

Part I serves as an introduction, and consists of this chapter and Chapter 2, *Queuing Networks in Healthcare Systems.* In this chapter we describe how queuing theory, and networks of queues in particular, can be invoked to model, study, analyze and solve healthcare problems. We describe important classical queuing results, especially meant to provide medical professionals with a theoretical background on the techniques used

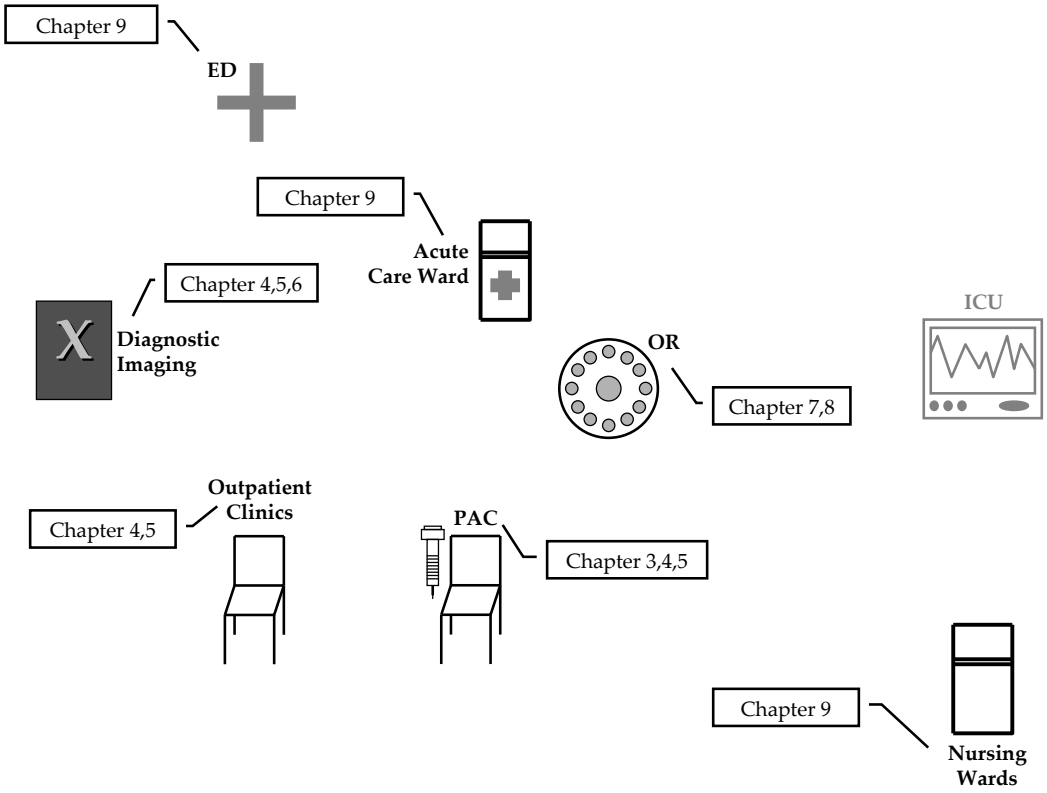
Figure 1.2: LUMC patient flow, based on 2010 data. The size of the arrow indicates the magnitude of the flow. The gray colored arrows are fictitious, since the Acute Care Ward opened in 2011. Abbreviations: ED – Emergency Department; ICU – Intensive Care Unit; OR – Operating Rooms; PAC – Preanesthesia Evaluation Clinic. Data source: LUMC Management Information System



in this thesis. We also provide a review of the literature on queuing networks in healthcare.

Part II consists of four chapters, and is devoted to challenges faced by outpatient clinics and diagnostic facilities. Chapter 3, *Redesign of the PAC*, studies the reorganization of an outpatient clinic. We demonstrate how the involvement of essential employees combined with applications of mathematical techniques to support the decision making process results in a successful intervention. The setting is the preanesthesia evaluation clinic of a university hospital, where patients consult several medical professionals, either on walk-in or appointment basis. We use queuing theory to model the initial set-up of the clinic and possible alternative designs. With the queuing model, possible improvements in efficiency are investigated. Key points in the intervention are the rescheduling of appointments and the reallocation of tasks.

Figure 1.3: Relationship of dissertation chapters with LUMC departments



Outpatient clinics and diagnostic facilities show an increased acceptance of unscheduled patient arrivals to improve accessibility. The methodology we present in Chapter 4, *Designing Cyclic Appointment Systems*, keeps waiting time at the facility for unscheduled patients below an acceptable level, while controlling the access time for scheduled patients. Formally, the access time is defined as the time between an appointment request and the appointment date, where the time scale is usually in days or weeks. Waiting time is defined as the time between the patient's arrival at a hospital facility and the start of the consultation and/or treatment, where the time scale is usually in minutes or hours. The method developed in this chapter consists of two separate but iteratively linked models, one for the day process that governs scheduled and unscheduled arrivals on the day and one for the access process of scheduled arrivals. A blueprint for the appointment schedule, consisting of the number of appointments to plan per day and the moment on the day to schedule the appointments, is calculated iteratively using the outcomes of the two models. Herein, the waiting and access times are balanced.

Chapter 5, *Appointments for Care Pathway Patients*, is motivated by the increasing popularity of care pathways in outpatient clinics. It is not uncommon that patients complete a significant part of the path in one day. Given the vast number of hospital facilities the patient has to visit, hospitals aim to optimize the flow of these patient groups by priori-

tizing them in the appointment planning process. As a result, regular patients who are not in a care pathway may experience increased waiting times. We develop a queuing model that allows for finding a trade-off between the accessibility for patients from the care pathway and waiting time for regular patients at an outpatient clinic.

In Chapter 6, *Allocation of MRI Scan Capacity*, we consider an MRI scanning facility run by a Radiology department. Several medical departments compete for capacity and have private information regarding their demand for scans. The fairness of the capacity allocation by the Radiology department depends on the quality of the information provided by the medical departments. We employ a generic Bayesian Game approach that stimulates the disclosure of true demand (truth-telling), so that capacity is allocated fairly.

Part III consists of three chapters and considers challenges that evolve when urgent and elective patient flow are mixed. Chapter 7, *Planning and Scheduling of Semi-Urgent Surgeries*, studies the trade-off between cancellations of elective surgeries due to semi-urgent surgeries, and unused operating room (OR) time due to excessive reservation of OR time for semi-urgent surgeries. Semi-urgent surgeries, to be performed soon but not necessarily today, pose an uncertain demand on available hospital resources, and interfere with the planning of elective patients. For a highly utilized OR, reservation of OR time for semi-urgent surgeries avoids excessive cancellations of elective surgeries, but may also result in unused OR time, since arrivals of semi-urgent patients are unpredictable. First, using a queuing theory framework, we evaluate the OR capacity needed to accommodate the incoming semi-urgent surgeries. Second, we introduce another queuing model that enables a trade-off between the cancellation rate of elective surgeries and unused OR time. Third, based on Markov decision theory, we develop a decision support tool that assists the scheduling process of elective and semi-urgent surgeries.

Using the methodology presented in Chapter 7, part of the OR capacity of the Neurosurgery department at LUMC was allocated to semi-urgent surgeries. In Chapter 8, *Implementation Study: Neurosurgery Planning*, we study the implementation process and the effect of dedicating OR slots to semi-urgent surgeries on elective patient cancellations and OR utilization.

Chapter 9, *The Emergency Observation and Assessment Ward*, is based on a project which started during a working visit to the University of Toronto in October-November 2010, and was finished during a working visit to the University of Western Ontario in June 2011. A recent development to reduce Emergency Department (ED) crowding and increase urgent patient admissions is the opening of an Emergency Observation and Assessment Ward (EOA Ward). At these wards urgent patients are temporarily hospitalized until they can be transferred to an inpatient bed. We present an overflow model to evaluate the effect of employing an EOA Ward on elective and urgent patient admissions. We conclude this dissertation with an epilogue that reviews the most important results and provides an outlook for the future.

Chapter 2

Queuing Networks in Healthcare Systems

2.1 Introduction

In this chapter we describe how queuing theory, and networks of queues in particular, can be invoked to model, study, analyze and solve healthcare problems. We describe important theoretical queuing results, give a review of the literature on the topic, and suggest directions for future research. For further reference, the book chapter [78] provides an overview of queuing theory applications in healthcare.

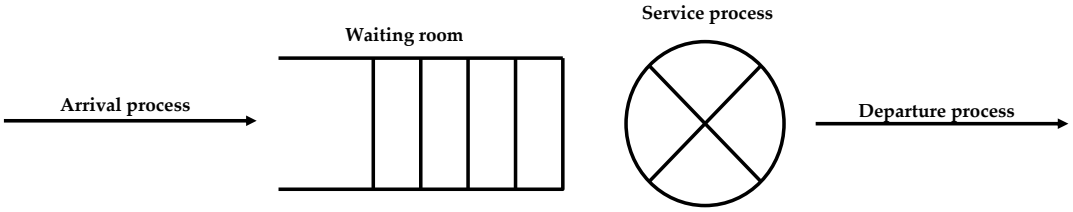
2.1.1 Some General Queuing Concepts in a Healthcare Setting

A queue can generally be characterized by its arrival and service processes, the number of servers, and the service discipline. The arrival process is specified by a probability distribution that has an arrival rate associated with it, which is usually the mean number of patients that arrives during a time unit (e.g., minutes, hours or days). A common choice for the probabilistic arrival process is the Poisson process, in which the inter-arrival times of patients are independent and exponentially distributed.

The service process specifies the service requirements of patients, again using a probability distribution with associated service rate. A common choice is the exponential distribution, which is convenient for obtaining analytical tractable results. The number of servers in a healthcare setting may represent the number of doctors at an outpatient clinic, the number of MRI scanners at a diagnostic department, and so on. The service discipline specifies how incoming patients are served. The most common discipline is First Come First Serve (FCFS), where patients are served in order of arrival. Other examples are briefly addressed in Subsection 2.2.2. Some patients may have priority over other patients. This can be such that the service of a lower priority patient is interrupted

when a higher priority patient arrives (preemptive priority), or the service of the lower priority patient is finished first (non-preemptive priority).

Figure 2.1: A simple queue



Typical measures for the performance of the system include the mean sojourn time, $\mathbb{E}[W]$, the mean time that a patient spends in the queue and in service. The sojourn time is a random variable as it is determined by the stochastic arrival and service processes. The mean waiting time, $\mathbb{E}[W^q]$, gives the mean time a patient spends in the queue waiting for service. How $\mathbb{E}[W]$ and $\mathbb{E}[W^q]$ are calculated depends, among other things, on the choice for the arrival and service processes, and is given for several basic queues in Subsection 2.2.2.

Kendall's Notation

All queues in this chapter are described using the so-called Kendall notation: $A/B/s$, where **A** denotes the arrival process, **B** denotes the service process, and **s** is the number of servers. There are several extensions to this notation, see for example [202]. Clearly, there are many distinctive cases of queues:

$M/M/1$: The single-server queue with Poisson arrivals and exponential service times. The M stands for the Markovian or Memoryless property.
$M/D/1$: The single-server queue with Poisson arrivals and Deterministic service times.
$M/G/1$: The single-server queue with Poisson arrivals and General (i.e., not specified) service time distribution.

Other arrival processes may also apply: consider for example the $D/M/1$, $G/M/1$ and $G/G/1$ queue. All of the forms above also exist in the case of multiple servers ($s > 1$).

The load of the queue is defined as the mean utilization rate per server, which is the amount of work that arrives on average per time unit, divided by the amount of work the queue can handle on average per time unit. Suppose our server is a single doctor in an outpatient clinic, then the load specifies the fraction of time the doctor is working. The load, ρ, equals the amount of work brought to the system per time unit, i.e. the

patient arrival rate, λ, multiplied by the mean service time per patient, $\mathbb{E}[S]$:

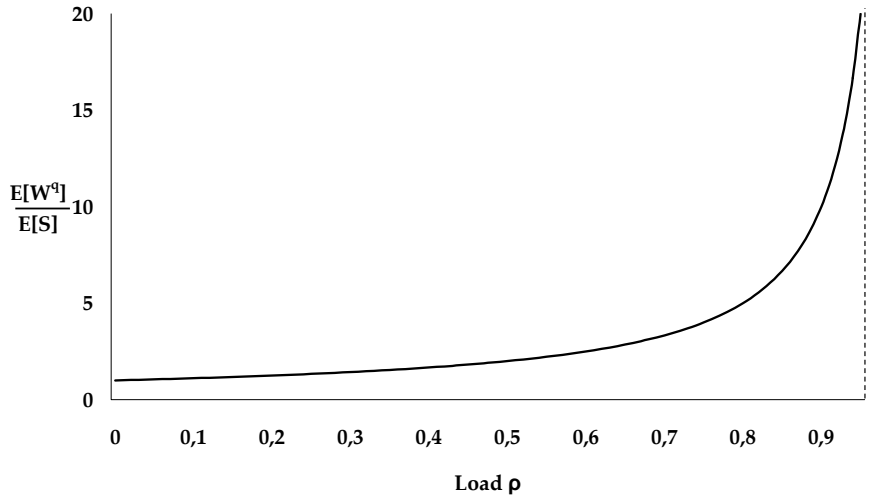
$$\rho = \lambda \mathbb{E}[S]. \tag{2.1}$$

The load is the fraction of time the server, working at unit rate, must work to handle the arriving amount of work. It is required that $\rho < 1$ (in other words, the server should work less than 100 percent of the time). If $\rho \geq 1$, then on average more work arrives at the queue than can be handled, which inevitably leads to a continuously growing number of patients in the queue waiting for service, i.e., an unstable system. Only when the arrival and service processes are deterministic (i.e., the inter-arrival and service times have zero variance), may the load equal 1. The mean waiting time, $\mathbb{E}[W^q]$, increases with load ρ. As an illustration, consider a single-server queue with Poisson arrivals and general service times (the so-called $M/G/1$ queue), with mean $\mathbb{E}[S]$ and squared coefficient of variation (scv) c_S^2, which is calculated by dividing the variance by the squared mean. For this queue, the relationship between ρ and $\mathbb{E}[W^q]$ is characterized by the Pollaczek-Khintchine formula [48]:

$$\mathbb{E}[W^q] = \mathbb{E}[S] \frac{\rho}{1-\rho} \frac{1+c_S^2}{2}, \tag{2.2}$$

In Figure 2.2 the relation is shown graphically for $c_S^2 = 1$. We see that the mean wait-

Figure 2.2: The relationship between load ρ and mean waiting time $\mathbb{E}[W^q]$ for the $M/M/1$ queue with Poisson arrivals and exponential service times

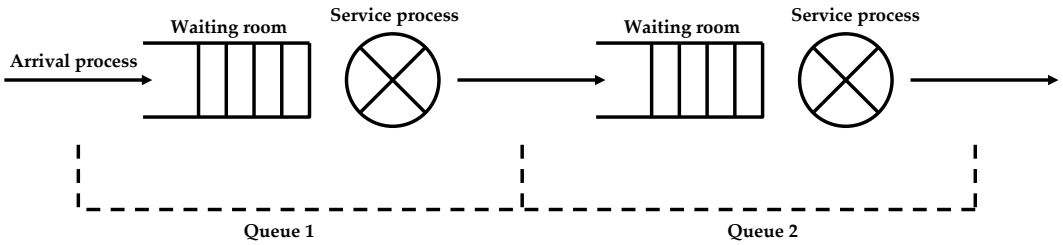


ing time increases with the load. When the load is low, a small increase therein has a minimal effect on the mean waiting time. However, when the load is high, a small increase has a tremendous effect on the mean waiting time. As an illustration, increasing

the load from 50% to 55% increases the waiting time by 10%, but increasing the load from 90% to 95% increases the waiting time by 100%! This explains why a minor change (for example a small increase in the number of patients) can result in a major increase in waiting times as sometimes seen in outpatient clinics. Formulas such as (2.2) allow for an exact and fast quantification of the relationships between (influencable) parameters and system outcomes. Queuing theory is a very valuable tool to identify bottlenecks and to calculate the effect of removing them.

We conclude this subsection with a basic queuing network: the $M/M/1$ tandem queue. In this network we have two queues with exponential service, which are placed in series. Patients arrive at the first queue according to a Poisson process with rate λ. When the service at the first queue is completed, the patient is routed immediately to the second queue. Upon service completion at this queue, the patient leaves the system. At both queues the service discipline is FCFS, and there is an infinite waiting room (see Figure 2.3). It can be shown that the mean sojourn time in the entire system, $\mathbb{E}[W]$, is

Figure 2.3: The $M/M/1$ tandem queue



just the sum of the mean sojourn times of the two queues when considered separately, which is $\mathbb{E}[W_j]$ for queue j:

$$\mathbb{E}[W] = \mathbb{E}[W_1] + \mathbb{E}[W_2], \tag{2.3}$$

since the departure process from each queue has the same characteristics as its input process. This remarkable result can be generalized to larger networks of queues, as is shown in Subsection 2.3.1.

2.1.2 Queuing Networks in Healthcare

When patients share and use multiple resources, a queuing network usually arises. Consider, for example, a patient that visits the Orthopedic outpatient clinic and then needs to have an X-ray at Radiology; or the surgical patient who is operated in the OR, then cared for at the Intensive Care Unit (ICU) and subsequently cared for in a nursing ward. The formulation and analysis of these queuing network models is usually not straightforward. This likely explains why (discrete-event) simulation [121] is a commonly used

approach to analyze healthcare problems. Simulation models are robust in terms of the setting they can represent, however they are very time consuming to develop and require a vast amount of data (-analysis). Also, the resulting model is, with a few exceptions, not generic and thus not suitable to represent other problems or organizations other than the one it was built for.

In this chapter we describe how queuing theory, and networks of queues in particular, can be invoked to study, analyze and solve healthcare problems. In Sections 2.2 and 2.3 we provide an introduction to the theory of queues and queuing networks. In Section 2.4 we give a review of the literature on the topic. In the last section we suggest directions for further research. Given the numerous modeling opportunities of queuing networks, many difficult healthcare problems can, and hopefully will, be solved in the future. The literature references on applications of queuing theory in healthcare are included in the categorized ORchestra bibliography [145], provided by research institute CHOIR from the University of Twente, Enschede, the Netherlands.

2.2 Single Queues

In this section we discuss several basic queues. We start by introducing the Poisson process, which is a basic element in many queuing systems. We then proceed to the building blocks for the networks: the individual queues.

2.2.1 The Poisson Process

As mentioned in Subsection 2.1.1, the Poisson process is commonly used to model the arrival of customers to a queue, and in general to model independent arrivals from a large population. As an example, consider patient arrivals at an ED. They originate from a large population (the demographic area surrounding the hospital) and usually arrive independently. The probability that an arbitrary person has an urgent medical problem is very small. Then the arrival process tends to a Poisson process.

The Poisson process is common in real world processes and has many interesting and very useful properties for analysis. For example, the number of ticks a Geiger counter records is a Poisson process. This example also indicates that merging or splitting Poisson processes independently results in Poisson processes, as this corresponds to joining two lumps of radioactive material or breaking one lump into parts. Or, for the population example, ED arrivals from a population subgroup (men, women, children, ...) are also Poisson.

For a Poisson process, the time between two successive arrivals is exponentially distributed. A very important property of the exponential distribution is that it is memoryless: the probability that the inter-arrival time exceeds $u + t$ time units, given that

it already has exceeded u time units, equals the probability that the inter-arrival time exceeds t time units. Mathematically, a random variable X that has an exponential distribution satisfies:

$$\mathbb{P}\left(X > u + t | X > u\right) = \mathbb{P}\left(X > t\right), \quad \forall u, t \geq 0. \tag{2.4}$$

We may also rephrase this property as: what happens in the future is independent of what happened in the past. Because of this Markovian or memoryless property, the complexity of analyzing systems with this property significantly reduces, as we show in the subsequent subsections.

Little's Law

The simple relationship $\mathbb{E}[L] = \lambda \mathbb{E}[W]$, presented in 1961 by J.D.C. Little [127], is known as Little's Law. It relates the mean number of patients in the queue, $\mathbb{E}[L]$, the average arrival rate, λ, and the mean time the patient spends in the queue, $\mathbb{E}[W]$.

A common intuitive reasoning for obtaining Little's Law is the following. Suppose patients pay 1 Euro for each time unit they spend in the queue. On average, the queue receives $\mathbb{E}[L]$ Euro per time unit, since there are on average $\mathbb{E}[L]$ patients present in the queue. Alternatively, if each patient would pay upon entering the queue for its entire time spent in the queue, a patient would on average have to pay $\mathbb{E}[W]$ to finance the entire stay. Since each time unit on average λ patients enter the queue, the amount received by the queue per time unit then equals $\lambda \mathbb{E}[W]$. Both methods of payment must result in the same benefit for the queue, thus $\mathbb{E}[L] = \lambda \mathbb{E}[W]$. The formal proof actually follows the lines of this reasoning. It is remarkable that Little's Law requires only mild assumptions on the system in equilibrium, and is valid irrespective of the number of servers, distribution of the arrival and service processes, queuing and service order. Thus Little's Law applies to many types of queues.

2.2.2 Basic Queues

We introduce the most commonly used queues: single and multi-server queues with Poisson arrivals and exponential or general service times. Unless mentioned otherwise, we consider the FCFS service discipline and queues with infinite capacity for waiting patients.

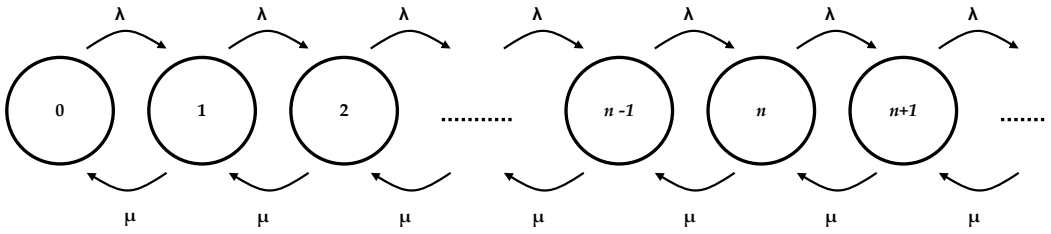
The $M/M/1$ Queue

In an $M/M/1$ queue, patients arrive according to a Poisson process with rate λ and exponentially distributed service requirement with mean service time $\mathbb{E}[S]$. The service rate per unit time is $\mu = \frac{1}{\mathbb{E}[S]}$, the number of patients that would be completed per time unit

when the system would continuously be serving patients. As denoted in Section 2.1.1, the load of the queue is $\rho = \lambda\mathbb{E}[S]$, where it is required that $\rho < 1$, that is, the amount of work brought into the queue should be less than the rate of the server. The number of patients present in the queue at time t, i.e., those waiting in line and in service, is obtained from Markov chain analysis.

Let $N(t)$ record the number of patients in the system at time t. Then $\mathbf{N} = (N(t), t \geq 0)$ is a Markov chain with state space $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, arrival rate λ, which is the rate at which a transition occurs from a state with n patients to a state with $n + 1$ patients, and departure rate μ from state n to state $n - 1$. We are interested in the probability P_n that

Figure 2.4: Transition rates in the $M/M/1$ queue



at an arbitrary point in time in statistical equilibrium the system contains n patients[1]:

$$P_n = \lim_{t\to\infty} \mathbb{P}(N(t) = n). \tag{2.5}$$

The probability P_n also reflects the fraction of time that the system contains n patients. The total probability may be seen as an amount of fluid of total volume 1 that is distributed over the states of the Markov chain and flows from state to state according to the transition rates (for the $M/M/1$ queue the arrival and departure rates). The system is in statistical equilibrium when these flows out of state n balance the flows into state n for each state n, $n = 0, 1, 2, \ldots$ (see Figure 2.4). Mathematically, this is expressed as:

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ (\lambda + \mu)P_1 &= \lambda P_0 + \mu P_2, \\ (\lambda + \mu)P_2 &= \lambda P_1 + \mu P_3, \\ &\vdots \end{aligned} \tag{2.6}$$

and in general:

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ (\lambda + \mu)P_n &= \lambda P_{n-1} + \mu P_{n+1} \quad \text{for} \quad n > 0. \end{aligned} \tag{2.7}$$

[1]We consider the system in statistical equilibrium only, as is a standard approach in classical queuing theory. For the $M/M/1$ queue, relaxation or convergence to equilibrium usually occurs fast. See [79] for a discussion on the validity of equilibrium analysis.

Since P_n is a probability, the summation of all probabilities P_n, $n = 0, 1, \ldots$, should equal unity:

$$\sum_{n=0}^{\infty} P_n = 1. \tag{2.8}$$

Using equation (2.7) and this additional property, we derive the queue length distribution P_n:

$$\begin{aligned} P_0 &= 1 - \rho, \\ P_n &= (1-\rho)\rho^n \quad \text{for} \quad n > 0. \end{aligned} \tag{2.9}$$

Note that P_0, also called the normalization constant, denotes the probability that there are zero patients present, but also the fraction of time the queue is empty. Further, ρ is the probability there are one or more patients present, and the fraction of time the queue is busy.

The PASTA Property

In a queuing system with Poisson arrivals, the probability that an arriving patient finds n patients in the queue is equal to the fraction of time the queue contains n patients. This property is referred to as PASTA, or Poisson Arrivals See Time Averages [203].

Usually, queuing systems with non-Poisson arrival processes do not conform to this property. For example, consider the $D/D/1$ queue with deterministic inter-arrival and service times. Time is equally distributed in slots of length one, and the service time is half a slot. Suppose that at the start of each time slot a patient arrives (so the inter-arrival time is one slot). Then the queue is empty upon arrival for all patients, while half of the time the queue contains one patient.

The mean number of patients in the queue, $\mathbb{E}[L]$, including those in service, is given by:

$$\mathbb{E}[L] = \sum_{n=0}^{\infty} nP_n = \frac{\rho}{1-\rho}. \tag{2.10}$$

Since ρ is the mean utilization rate of the server, the mean number of patients waiting, $\mathbb{E}[L^q]$, equals:

$$\mathbb{E}[L^q] = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}. \tag{2.11}$$

Using Little's Law, the relationship between the mean number of patients in the queue, $\mathbb{E}[L]$, and the mean sojourn time, $\mathbb{E}[W]$, can be explicitly quantified as follows [127]:

$$\mathbb{E}[L] = \lambda \mathbb{E}[W]. \tag{2.12}$$

This also holds for the relationship between the mean number of patients waiting for service, $\mathbb{E}[L^q]$, and the mean waiting time in the queue, $\mathbb{E}[W^q]$:

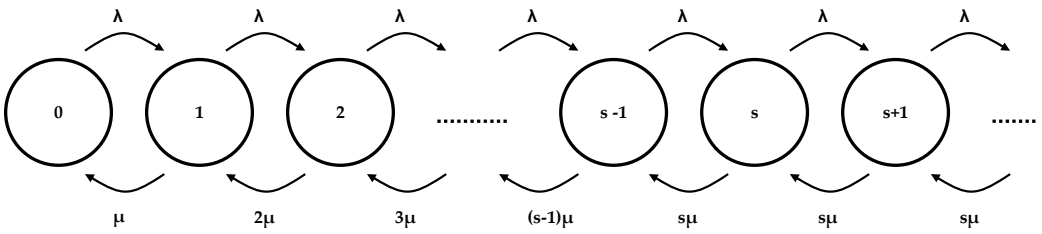
$$\mathbb{E}[L^q] = \lambda \mathbb{E}[W^q]. \tag{2.13}$$

Note that the equilibrium distribution and performance measures are characterized by the single parameter ρ and can be calculated in a straightforward manner. As we will see in the subsequent subsections, this is more involved for more complicated queuing systems.

The $M/M/s$ Queue

The $M/M/s$ queue is the multi-server variant of the $M/M/1$ queue. Patients arrive with rate λ, each patient is served by one server and a patient waits in queue when all servers are occupied. There are s servers so that the maximum service rate of the queue is $s\mu$, where μ is the service rate of the individual servers. If the number of patients in the queue, n, is less than the number of servers, s, the service rate equals $n\mu$ (see the transition rate diagram in Figure 2.5). Again it is required that the amount of work that

Figure 2.5: Transition rates in the $M/M/s$ queue



arrives per time unit (ρ) is less than the maximum service rate, i.e., $\rho = \lambda \mathbb{E}[S] < s$. The equilibrium distribution is obtained from:

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ (\lambda + n\mu) P_n &= \lambda P_{n-1} + (n+1)\mu P_{n+1} \qquad \text{for} \quad n < s, \\ (\lambda + s\mu) P_n &= \lambda P_{n-1} + s\mu P_{n+1} \qquad \text{for} \quad n \geq s. \end{aligned} \tag{2.14}$$

Thus

$$P_n = \frac{\rho^n}{m(n)} P_0, \quad \text{where} \quad m(n) = \begin{cases} n! & \text{for} \quad 0 \leq n < s, \\ s^{n-s} s! & \text{for} \quad n \geq s. \end{cases} \tag{2.15}$$

Invoking the normalization condition (2.8), we obtain:

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{\rho^n}{n!} + \frac{\rho^s}{s!} \frac{s}{s-\rho} \right)^{-1} . \tag{2.16}$$

For $s = 1$, equations (2.15)–(2.16) reduce to the queue length distribution for the $M/M/1$ queue (2.9). The probability P_s deserves special attention; this is the fraction of time all servers are occupied, and because of the PASTA property, also the fraction of arriving patients that find all servers occupied. Thus the probability that a patient will be served immediately upon arrival is $1 - \sum_{n=s}^{\infty} P_n = \sum_{n=0}^{s-1} P_n$, and the probability that a patient has to wait is $\sum_{n=s}^{\infty} P_n$. The latter probability can be calculated using the Erlang-C formula [84]:

$$P_{s^+} = \mathbb{P}(n \geq s) = \frac{\rho^s}{s!} \frac{s}{s-\rho} P_0. \tag{2.17}$$

There are several Erlang-C calculators available online to compute P_{s^+}, see e.g. [70] and [197]. The mean number of patients waiting for service is:

$$\mathbb{E}[L^q] = \sum_{n=s+1}^{\infty} (n-s) P_n = \frac{\rho}{s-\rho} P_{s^+} . \tag{2.18}$$

By applying Little's Law we find the mean waiting time:

$$\mathbb{E}[W^q] = \frac{\mathbb{E}[L^q]}{\lambda} . \tag{2.19}$$

The mean sojourn time is then obtained by adding the mean service time to the mean waiting time:

$$\mathbb{E}[W] = \mathbb{E}[S] + \mathbb{E}[W^q]. \tag{2.20}$$

The mean number of patients in the queue can be calculated by adding the mean number of patients in service, ρ, to the mean number of patients waiting [84]:

$$\mathbb{E}[L] = \rho + \mathbb{E}[L^q]. \tag{2.21}$$

The $M/M/s/s$ Queue

The $M/M/s/s$ queue, or Erlang loss queue, is different from the $M/M/s$ queue in that it has no waiting capacity. Thus when all servers are occupied, patients are blocked and lost (i.e., they leave and do not come back). This type of queue is very useful when modeling healthcare systems with limited capacity, where patients are routed to another facility when all capacity is in use. Examples are nursing wards and the ICU. Figure 2.6 gives the transition rates for this queue. We obtain:

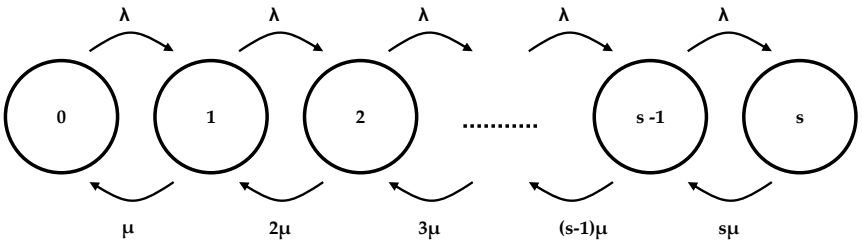
$$\begin{aligned} \lambda P_0 &= \mu P_1 \\ (\lambda + n\mu)P_n &= \lambda P_{n-1} + (n+1)\mu P_{n+1} \qquad \text{for} \quad 0 < n < s \\ \lambda P_{s-1} &= s\mu P_s, \end{aligned} \tag{2.22}$$

with solution:

$$P_n = \frac{\rho^n/n!}{\sum_{i=0}^{s} \rho^i/i!} \quad \text{for} \quad 0 \leq n \leq s, \quad \text{where} \quad \rho = \lambda \mathbb{E}[S]. \tag{2.23}$$

Surprisingly, (2.23) also holds for general service times (the $M/G/s/s$ queue) and is thus insensitive to the service time distribution [84]. The probability that all servers are

Figure 2.6: Transition rates in the $M/M/s/s$ queue



occupied, is often called the blocking probability, and is given by:

$$P_s = \frac{\rho^s/s!}{\sum_{i=0}^{s} \rho^i/i!}. \tag{2.24}$$

Formula (2.24) is often referred to as the Erlang loss formula, or Erlang-B [84]. For large s, the direct calculation of P_s by using (2.24) often introduces numerical problems. The following stable recursion exists where these problems are avoided [211].

Recursion for Erlang-B

Step 1.
Set $X_0 = 1$.

Step 2.
For $j = 1, \ldots, s$ compute

$$X_j = 1 + \frac{jX_{j-1}}{\rho}. \tag{2.25}$$

Step 3.
The blocking probability P_s is given by

$$P_s = \frac{1}{X_s}. \tag{2.26}$$

Another option is to use one of the Erlang-B calculators available online, see e.g. [150] and [197]. The performance measures are given by:

$$\mathbb{E}[L] \;\; = \rho\,(1 - P_s)\,, \quad \mathbb{E}[W] \;\; = \mathbb{E}[S]. \tag{2.27}$$

As we have seen in this subsection, the computation of the blocking probabilities can be quite involved. The infinite server, or $M/M/\infty$ queue, is often used to approximate the $M/M/s/s$ queue for a large number of servers. In this queue, upon arrival each patient obtains his own server. The queue length has a Poisson distribution with parameter ρ, where $\rho = \lambda\mathbb{E}[S]$, and is thus given by

$$P_n^\infty \;\; = \;\; \frac{\rho^n}{n!}P_0, \quad \text{where} \quad P_0^\infty = e^{-\rho}. \tag{2.28}$$

The blocking probability for the system with s servers is approximated by [187]:

$$P_s \approx \sum_{n \geq s}^{\infty} P_n^\infty. \tag{2.29}$$

Queues with General Arrival and/or Service Processes

For the $M/M/s$ queue a single parameter suffices to calculate the queue length distribution and related performance measures. However, assuming exponentiality of the distributions involved in a queuing process is not always a valid choice. When the coefficient of variation is not close to 1 (the value for the exponential distribution) other

probability distributions should be used to obtain reliable outcomes, since the variance of the inter-arrival and service times has strong influence on the performance measures.

Results for non-exponential systems are scarce and are often characterized via the scv, c^2. In general, when the scv increases, the variability in the related queuing system also increases. In this subsection we will focus on results for mean waiting times. Additional results are given in the books [84], [187] and [203]. The software package QtsPlus that accompanies [84] supports the calculation of many relevant performance measures, is free available online [159] and implemented in MS Excel, but also has an open source variant.

For the $M/G/1$ queue the Laplace-Stieltjes transform for the waiting time distribution is known. From this result, we obtain the Pollaczek-Khintchine formula [48] that characterizes the waiting time in the single-server queue with Poisson arrivals and general service times:

$$\mathbb{E}[W^q] = \mathbb{E}[S]\frac{\rho}{1-\rho}\frac{1+c_S^2}{2}, \tag{2.30}$$

where c_S^2 denotes the scv of the service time. The mean sojourn time for the $G/M/1$ queue is:

$$\mathbb{E}[W] = \frac{\mathbb{E}[S]}{1-\sigma}\ , \tag{2.31}$$

where σ is the unique root in the range $0 < \sigma < 1$ of the following equation:

$$\sigma = \bar{A}(\mu - \mu\sigma), \tag{2.32}$$

with $\bar{A}$ the Laplace-Stieltjes transform of the inter-arrival time and $\mu = \frac{1}{\mathbb{E}[S]}$ [203]. For the $G/G/1$ queue the following approximation solution is often used [187]:

$$\mathbb{E}[W^q] \approx \mathbb{E}[S]\frac{\rho}{1-\rho}\frac{c_A^2+c_S^2}{2}, \tag{2.33}$$

where c_A^2 denotes the scv of the arrival process. This result includes the $G/M/1$ queue and is exact for the $M/G/1$ queue.

It is hard to determine the exact effect of using the exponential distribution to represent a non-exponential process. As a rule of thumb, we suggest that as long as the actual variance is below that of the exponential distribution, then the exponential distribution provides a conservative estimate. In other words, the calculated expectations of the queue length and waiting times will over-estimate the actual values. Such a conservative estimate is for instance useful when a strategic decision that does not involve a lot of detail needs to be made.

For the mean waiting time in the $G/G/s$ queue the following approximation is very useful [84]:

$$\mathbb{E}[W^q] \approx \mathbb{E}[W^q_{(M/M/s)}]\frac{c_A^2 + c_S^2}{2}, \tag{2.34}$$

where $\mathbb{E}[W^q_{(M/M/s)}]$ denotes the mean waiting time in the $M/M/s$ queue with identical λ and μ. In [84] lower and upper bounds on $\mathbb{E}[W^q]$ are also provided. Using the results for $\mathbb{E}[W^q]$, Little's Law can be applied to determine the mean number of patients in the queues mentioned in this subsection.

Service Disciplines

So far, we have only discussed the FCFS service discipline. Other options are Processor Sharing (PS) and Last Come First Serve (LCFS). We will elaborate on queuing networks with these kind of queues in Subsection 2.3.2.

In the processor sharing service discipline, all arriving patients are immediately served, thus there is no queuing. A single server is shared equally among patients, where each patient class may have its own service requirement. For the $M/M/1 - PS$ queue the queue length distribution, P_n, is identical to that of the $M/M/1 - FCFS$ queue (2.9). Intuitively, this can be explained as follows. The server works at rate μ, and when there are n patients in the queue, an individual patient is served with rate $\frac{\mu}{n}$. However, since n patients are served simultaneously, the overall completion rate is still μ ($\frac{\mu}{n} \cdot n = \mu$). Since the patient arrival rate equals λ, the flow in and out of the queue is identical to that of the $M/M/1 - FCFS$ queue.

The $M/M/1 - LCFS$ queue with preemptive resume can be seen as a stack, for instance of patient files, where a single server (the doctor) works on the top item of the stack. Whenever a new item is added, the server immediately starts working on this item. However, when the server returns to the previous item, it resumes service (i.e., the queue is work conserving). The queue length distribution is again given by (2.9), where the same argument holds as for the $M/M/1 - PS$ queue.

Miscellaneous Queuing Results

In this subsection we briefly mention a couple other queuing results. Some of the results that can be obtained for $G/G/1$ queues are exact, but do not transfer to queuing networks. In particular, the equilibrium distribution at arrival instants in the $G/M/1$ queue is:

$$P_n = (1 - \sigma)\sigma^n, \tag{2.35}$$

with σ defined as in (2.32).

The equilibrium distributions of the $M/G/1$ queue and the $G/M/1$ queue at arrival epochs have a geometric form. The queue length distribution of the $M/G/1$ queue at departure epochs can be obtained using the theory of matrix geometric queues. At arbitrary (non arrival or departure epochs) the equilibrium distribution of these queues is not available in amenable form. To further characterize the equilibrium distribution of these queues, we introduce the class of so-called phase type distributions [120]. A distribution is of phase-type if it can be represented as a continuous time Markov chain on the phases such that the chain remains in a phase during an exponential time and jumps from phase to phase according to transition probabilities, see [120] for details.

It is interesting to observe that each probability distribution that attains positive values, only, can be approximated arbitrarily closely by a phase-type distribution. Using phase-type distribution for respectively the service time and inter-arrival time distribution, the equilibrium distributions for the $M/Ph_r/1$ and $Ph_s/M/1$ queues are available in closed form. For these queues, the state description requires the number of patients n and the phase of the service or inter-arrival times r resp. s. The equilibrium distribution is obtained in closed form:

$$P_n = P_0 R^n, \quad n = 0, 1, 2, \ldots, \tag{2.36}$$

where P_0 and P_n are r resp. s vectors over the phases of the service or inter-arrival times and R is an $r \times r$ or $s \times s$ matrix over these phases. The result generalizes to the $Ph_r/Ph_s/1$ queue where P_0 and P_n become rs vectors recording the joint phases of inter-arrival and service times. Although the form (2.36) is geometric, obtaining the matrix R is quite involved and goes beyond the scope of this chapter, see [119] for details. We specifically mention this queue since phase-type distributions are common in healthcare. For example the LOS in geriatric care has been modeled using phase-type distributions [66].

Instead of joining the queue, patients may be impatient and leave the queue before service. When this happens upon arrival, it is called balking. When patients leave after waiting some time, it is referred to as reneging. In the $M/M/s/s$ queue it is assumed that patients who are blocked are lost to the system. When blocked and/or impatient patients return to the queue after some time, we have a retrial queue [84].

In this subsection we have considered only queues with a single class of patients. When more than one patient class arrives at the queue, and classes have priority over one another, we have a priority queue [203]. In the case of preemptive priority, the service of the low priority patient is interrupted immediately when a higher prioritized patient arrives. Afterwards, the service of the low priority patient is resumed (work conserving) or may have to start allover again (work is lost). In the case of non-preemptive priority, a patient that is already in service is completed first.

Vacation queues are a generalization of the $M/G/1$ queue, where the server may take a vacation (i.e., becomes idle for a certain amount of time), also when there are patients in the queue [203]. A generalization of the vacation queue is the polling model, where

a single server visits multiple queues [182]. In this chapter we restrict our focus to networks of queues with continuous availability.

2.3 Basic Queuing Networks

Now that we have defined the building blocks, we can proceed to queuing networks. We start with networks of exponential queues with either a single or multiple servers.

2.3.1 Networks of Exponential Queues

Tandem Networks

Consider a tandem network of J queues that are placed in series. All queues have infinite waiting room, a single-server, and the service requirement at queue j, $j = 1, \ldots, J$, has an exponential distribution with mean service time $\mathbb{E}[S_j]$. Patients arrive at queue 1 according to a Poisson process with rate λ. Upon service completion at queue j the patient routes to queue $j + 1$, $j = 1, \ldots, J - 1$, and finally departs from queue J.

From Burke's theorem [34] it follows that the departure process of a queue with Poisson arrivals and exponential service times, is again a Poisson process with the same rate as the arrival process, and that departures from queue 1 before time t_0 are independent of the queue length of queue 1 at time t_0. This fundamental result indicates that the queue length at time t_0 in queue 1 and queue 2 are statistically independent. Hence, for the tandem queue of Figure 2.3,

$$P(n_1, n_2) = \mathbb{P}(N_1 = n_1, N_2 = n_2) = (1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2}, \quad n_1, n_2 \geq 0, \tag{2.37}$$

where $\rho_1 = \lambda\mathbb{E}[S_1]$, $\rho_2 = \lambda\mathbb{E}[S_2]$, and N_j is the random queue length at queue j in equilibrium. Continuing this argument, for a tandem network of J queues, we obtain the so-called product-form solution [187]:

$$P(n_1, \ldots, n_J) = \prod_{j=1}^{J}(1 - \rho_j)\rho_j^{n_j}, \quad \text{where} \quad \rho_j = \lambda\mathbb{E}[S_j]. \tag{2.38}$$

This elegant result leads us to Open Jackson Networks with general patient routing.

Open Jackson Networks

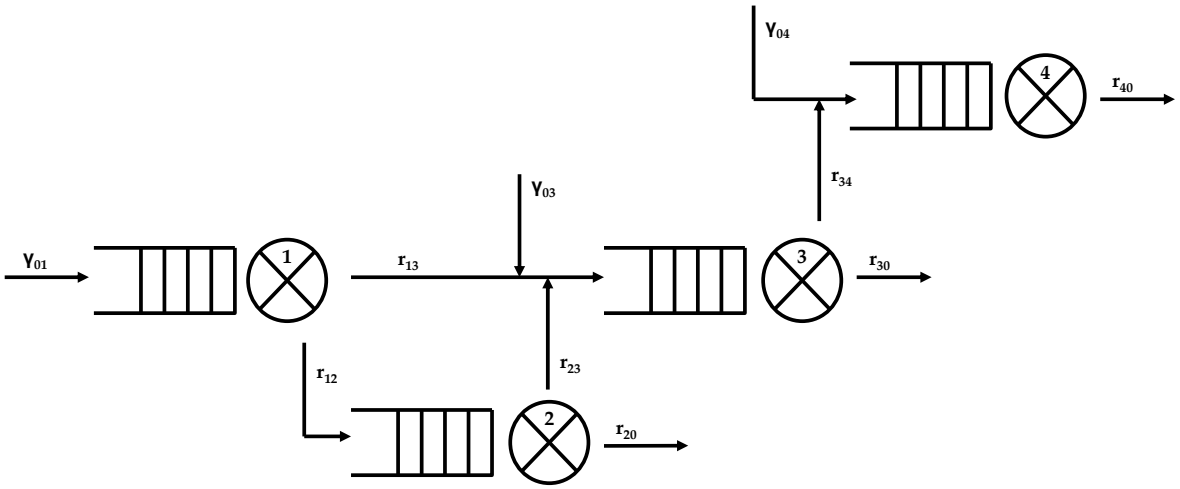
We now consider a network consisting of J single-server queues. The external arrival process at queue j, $j = 1, \ldots, J$, is Poisson distributed with rate γ_j, $\gamma_j \geq 0 \, \forall j$. Each queue

j has an exponentially distributed service requirement with mean service time $\mathbb{E}[S_j]$. Patients are routed from queue i to queue j with state independent routing probability r_{ij}, $0 \leq r_{ij} \leq 1$, i.e., a fraction r_{ij} of patients served at queue i routes to queue j. The parameter r_{i0} denotes the fraction of patients leaving the network at queue i. The total arrival rate λ_j at queue j is given by:

$$\lambda_j = \gamma_j + \sum_{i=1}^{J} \lambda_i r_{ij}, \quad j = 1, \ldots, J, \tag{2.39}$$

and is composed of the arrivals to queue j from outside and inside the network. A queuing network with these characteristics is called an Open Jackson Network, named after James R. Jackson who first studied its properties in 1957 [100]. In Figure 2.7 an example of an Open Jackson Network is given. According to Jackson's Theorem [100],

Figure 2.7: An example of an Open Jackson Network with four queues and patient routing from queues 1→2, 1→3, 2→3, and 3→4. External arrivals occur at queue 1, 3, and 4; departures occur at queue 2, 3, and 4



the product-form solution for this type of network is given by:

$$P(n_1, \ldots, n_J) = \prod_{j=1}^{J} (1 - \rho_j)\rho_j^{n_j}, \quad n_j \geq 0, \quad j = 1, \ldots, J, \quad \text{where} \quad \rho_j = \lambda_j \mathbb{E}[S_j]. \tag{2.40}$$

The Power of Jackson's Theorem

From Jackson's theorem it follows that per queue only a single parameter, ρ_j, is required for the calculation of $P(n_1, \ldots, n_J)$. Consequently, only J parameters are required to analyze the entire network! This result is surprising since usually many parameters are required to characterize a probability distribution. Note that the product form expression states that the queues lengths are independent random variables at a specific point in time. This does not imply that the queue length processes are independent.

Since the queues in the network act as if they are independent $M/M/1$ queues, the performance measures are easy to compute:

$$\mathbb{E}[L_j] = \frac{\rho_j}{1-\rho_j}, \quad \mathbb{E}[W_j] = \frac{\mathbb{E}[L_j]}{\lambda_j}. \tag{2.41}$$

The mean sojourn time for an arbitrary patient can be calculated using Little's Law:

$$\mathbb{E}[W] = \frac{\sum_{j=1}^{J} \mathbb{E}[L_j]}{\sum_{j=1}^{J} \gamma_j}. \tag{2.42}$$

Note that this is not equal to $\sum_{j=1}^{J} \mathbb{E}[W_j]$, since patients may not visit all queues in the network or visit some queues several times.

Jackson's result can be extended to the multi-server case. We obtain:

$$\begin{aligned} P(n_1, \ldots, n_J) &= \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{m(n_j)} P_{0j}, \quad \text{where} \quad \rho_j = \lambda_j \mathbb{E}[S_j], \\ m(n_j) &= \begin{cases} n_j! & \text{for} \quad 0 \le n_j < s_j, \\ s_j^{n_j - s_j} s_j! & \text{for} \quad n_j \ge s_j, \end{cases} \end{aligned} \tag{2.43}$$

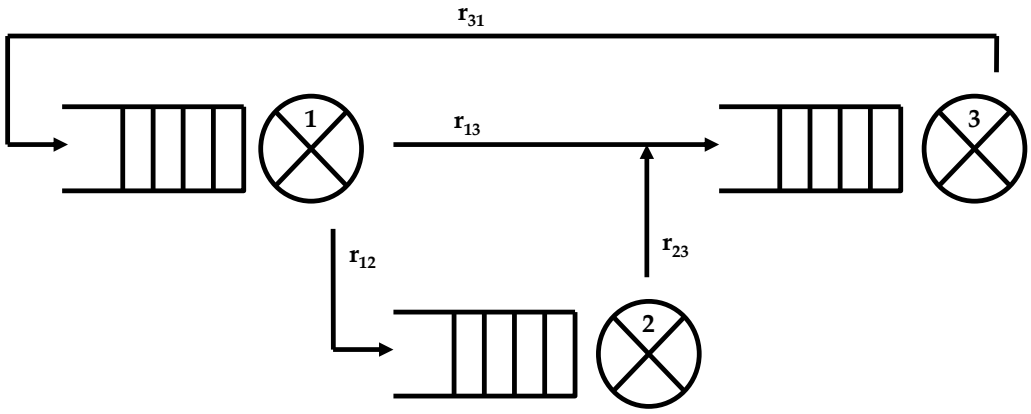
and $s_j \geq 1$ for $j = 1, \ldots, J$. The normalization constant P_{0j} is given by

$$P_{0j} = \left(\sum_{n_j=0}^{s_j-1} \frac{\rho_j^{n_j}}{n_j!} + \frac{\rho_j^{s_j}}{s_j!} \frac{s_j}{s_j - \rho_j} \right)^{-1}. \tag{2.44}$$

Closed Jackson Networks

A Jackson Network where the external arrival rates $\gamma_j = 0 \; \forall j$ and the departure probabilities $r_{i0} = 0 \; \forall i$, is a called a Gordon-Newell or Closed Jackson Network, since patients do not enter or leave (see Figure 2.8). The finite number N of patients that is present in

Figure 2.8: An example of a Closed Jackson Network with three queues and patient routing from queues 1→2, 1→3, 2→3, and 3→1



the network is continuously routed among J queues according to the state independent routing probabilities r_{ij}. For the single-server case we obtain a product-form solution [77]:

$$P(n_1, \ldots, n_J) = B(N)^{-1} \prod_{j=1}^{J} \rho_j^{n_j}, \quad \text{where} \quad \sum_{j=1}^{J} n_j = N. \tag{2.45}$$

In this formula $B(N)$ is called the normalization constant. In the open network variant, the expression $\prod_{j=1}^{J}(1-\rho_j)$ is actually the normalization constant and easy to compute. In the closed network variant, $B(N)$ is given by:

$$B(N) = \sum_{\sum_{j=1}^{J} n_j = N} \prod_{j=1}^{J} \rho_j^{n_j}. \tag{2.46}$$

Calculating $B(N)$ can be quite cumbersome, even for small N. Buzen's algorithm [36] is very helpful in this case and works as follows.

Buzen's Algorithm

Step 1.
Define

$$G_j(k), \quad \text{where} \quad j = 0, \ldots, J \quad \text{and} \quad k = 0, \ldots, N, \tag{2.47}$$

with initial values

$$G_1(k) \;=\; \rho_1^k, \quad G_j(0) = 1. \tag{2.48}$$

Step 2.
Recursively compute

$$G_j(k) = G_{j-1}(k) + \rho_j G_j(k-1). \tag{2.49}$$

Step 3.
The normalization constant is given by:

$$B(N) = G_J(N). \tag{2.50}$$

Buzen's algorithm can also be used to compute other performance measures of interest. The marginal probability that n_j patients are present at queue j is given by:

$$P(n_j) = B(N)^{-1} \rho_j^{n_j} \left(G_J(N - n_j) - \rho_j G_J(N - n_j - 1) \right). \tag{2.51}$$

The mean number of patients present at queue j is given by:

$$\mathbb{E}[L_j] = \sum_{n_j=1}^{N} \rho_j^{n_j} B(N)^{-1} G_J(N - n_j). \tag{2.52}$$

The Closed Jackson Network can also be extended to the multi-server case. The product-form solution is then given by:

$$P(n_1, \ldots, n_J) = B(N)^{-1} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{m(n_j)}, \tag{2.53}$$

where $\sum_{j=1}^{J} n_j = N$, $m(n_j)$ is given by (2.43), and

$$B(N) = \sum_{\sum_{j=1}^{J} n_j = N} \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{m(n_j)}. \tag{2.54}$$

For the multi-server case $B(N)$ can also be calculated using Buzen's algorithm.

In a closed single-server Jackson network the mean waiting time and mean number of patients at queue j can be calculated without evaluating $B(N)$ [84]. This algorithmic approach is called Mean Value Analysis (MVA). We present the basic algorithm, but MVA has been extended to many other queuing systems, see [3].

MVA Algorithm

Step 1.
Set $\lambda_1 = 1$ and solve the traffic equations:

$$\lambda_j = \sum_{i=1}^{J} \lambda_i r_{ij}, \quad j = 1, \ldots, J. \tag{2.55}$$

Step 2.
Define $L_j(0) = 0$ for $j = 1, \ldots, J$.

Step 3.
For $n = 1, \ldots, N$, calculate

$$\begin{aligned} W_j(n) &= (1 + L_j(n-1))\,\mathbb{E}[S_j], \quad j = 1, \ldots, J, \\ \nu_1(n) &= \frac{n}{\sum_{j=1}^{J} \lambda_j W_j(n)}, \\ \nu_j(n) &= \nu_1(n)\lambda_j \quad j = 2, \ldots, J, \\ L_j(n) &= \nu_j(n) W_j(n), \quad j = 1, \ldots, J. \end{aligned} \tag{2.56}$$

Step 4.
The mean waiting time at queue j is given by:

$$\mathbb{E}[W_j] = W_j(N). \tag{2.57}$$

The mean number of patients at queue j is given by:

$$\mathbb{E}[L_j] = L_j(N). \tag{2.58}$$

2.3.2 Networks of Queues with General Arrival and/or Service Processes

As said, the few exact results that exist for general queues cannot be transferred to general queuing networks. However, many of the approximation results are. In this subsection we describe three types of networks that have an exact solution for the queue length

distribution, namely networks with fixed routing, BCMP networks, and loss networks. We conclude with the Queuing Network Analyzer (QNA). This is a generalization of MVA for networks of $G/G/s$ queues.

Networks with Fixed Routing

All of the queuing networks we have discussed so far employ Markovian routing. This means that after departure, patients are routed to other queues or leave the network with a certain probability. This excludes fixed routes in which patients follow a prescribed path.

Consider a network in which each patient class has its own route. The route of patient class k, $k = 1, \ldots, K$, is given by the sequence of queues to visit before leaving the system [104]:

$$r(k,1), r(k,2), \ldots, r(k,H(k)). \tag{2.59}$$

So in stage h, $h = 1, \ldots, H(k)$, patient class k visits queue $r(k,h)$. Note that one queue may appear multiple times in the route. Using this notation enables to include patients that visit the same queue multiple times, but have a different destination depending on the times the queue has been visited. An example route for a patient class could be $3 \to 2 \to 3 \to 4$, where queue 2 is visited after the patient departs from queue 3 the first time, and queue 4 is visited after the patient departs from queue 2 the second time. This type of queuing network can be seen as a set of intertwined tandem networks (Subsection 2.3.1). Each patient class is routed through its own tandem network of queues, and different patient classes may meet each other at one of the queues.

Let γ_k denote the arrival rate of patient class k. As a consequence of fixed routes, the arrival rate of patient class k at stage h to queue $r(k,h)$ equals the arrival rate of the patient class to the network. In order to be able to determine how many patients of class k being in stage h of their route, are present at queue j, we have to record the position in the queue for each individual patient. We introduce some additional notation. Let $k_j(\ell)$ denote the class of the patient that holds position ℓ in queue j, and let $h_j(\ell)$ denote the stage the patient is currently in. Then $c_j(\ell) = (k_j(\ell), h_j(\ell))$ gives the type of this patient. Since a patient may visit one queue several times, his type potentially gives more information than his class. The state of queue j is given by the vector $c_j = (c_j(1), \ldots, c_j(n_j))$, and $C = (c_1, \ldots, c_J)$ gives the state of the queuing network. Now if we define the parameter $\alpha_j(k,h)$ as follows:

$$\alpha_j(k,h) = \begin{cases} \nu_k & \text{if} \quad r(k,h) \equiv j, \\ 0 & \text{otherwise}, \end{cases} \tag{2.60}$$

where ν_j is given by $\lambda_j \mathbb{E}[S_j]$, and a_j is the load of queue j:

$$a_j = \sum_{k=1}^{K} \sum_{h=1}^{H(k)} \alpha_j(k,h), \tag{2.61}$$

then the marginal queue length distribution of the number of patients of class k, $k = 1, \ldots, K$, present at queue j, is given by:

$$P_j(c_j) = B_j^{-1} \prod_{\ell=1}^{n_j} \alpha_j\left(k_j(\ell), h_j(\ell)\right), \quad \text{where} \quad B_j = \sum_{n=0}^{\infty} a_j^n. \tag{2.62}$$

The queue length distribution for the entire queuing network is then given by:

$$P(C) = \prod_{j=1}^{J} P_j(c_j). \tag{2.63}$$

The queue length distribution of the number of patients at the queues in the network is given by:

$$P(n_1, \ldots, n_J) = \prod_{j=1}^{J} (1-\nu_j)\nu_j^{n_j}. \tag{2.64}$$

Note that this result does not discriminate among patient classes. Even though the notation required can be quite cumbersome, networks with fixed routing introduce substantial modeling flexibility.

BCMP Networks

If each queue j in a network of J queues is one of the following types:

1. $M/M/s - FCFS$
2. $M/G/1 - PS$
3. $M/G/1 - LCFS$ preemptive resume
4. $M/G/\infty$,

an exact solution exists and the network is a BCMP network. It is named after the authors Baskett, Chandy, Muntz and Palacios, who described it in 1975 [15]. The network may be open or closed with multiple patient classes, and employ Markovian or fixed routing. In the case of an open network, the external arrival rates to the queues are Poisson. For notational convenience, we give the product-form solution for a BCMP network with Markovian routing and a single patient class. In this case the queue length distribution is given by:

$$P(n_1, \ldots, n_J) = B(N) \prod_{j=1}^{J} P_j(n_j), \tag{2.65}$$

where $B(N)$ is the normalization constant such that $\sum_N P(n_1, \ldots, n_J) = 1$, and $P_j(n_j)$ is the equilibrium distribution for queue j, $j = 1, \ldots, J$. If queue j is of type 1:

$$\begin{aligned} P_j(n_j) &= \frac{\rho_j^{n_j}}{m(n_j)} P_j(0), \quad \text{where} \\ m(n_j) &= \begin{cases} n_j! & \text{for} \quad 0 \leq n_j < s_j, \\ s_j^{n_j - s_j} s_j! & \text{for} \quad n_j \geq s_j, \end{cases} \quad \text{and} \\ P_j(0) &= \left(\sum_{n_j=0}^{s_j-1} \frac{\rho_j^n}{n_j!} + \frac{\rho_j^s}{s_j!} \frac{s_j}{s_j - \rho_j} \right)^{-1}. \end{aligned} \tag{2.66}$$

If queue j is of type 2 or 3:

$$\begin{aligned} P_j(n_j) &= \rho_j^{n_j} P_j(0), \quad \text{where} \\ P_j(0) &= 1 - \rho_j. \end{aligned} \tag{2.67}$$

If queue j is of type 4:

$$\begin{aligned} P_j(n_j) &= \frac{\rho_j^{n_j}}{n_j!} P_j(0), \quad \text{where} \\ P_j(0) &= e^{-\rho_j}. \end{aligned} \tag{2.68}$$

Note that the four queue types include the service disciplines we discussed in Subsection 2.2.2. For BCMP networks the queue length distributions for these service disciplines are insensitive to the service requirement distribution, that is, only the mean service times are required to obtain the equilibrium distribution (2.65).

Loss Networks

A loss network is the multi-dimensional generalization of the Erlang loss queue (Subsection 2.2.2). In a loss network, patients simultaneously claim at least one server in at least one queue. When upon arrival at the network one of the designated queues is full, the patient is blocked and lost. Note that this kind of queuing network shows an analogy with some hospital processes. For instance, a patient that needs to be admitted to the ICU after surgery, will not be operated on when there is no ICU bed available. Thus the patient simultaneously claims an operating room and an ICU bed. If either one is not available, the surgery will not commence.

For a loss network handling K patient classes, the queue length distribution of the number of patients of class k, $k = 1, \ldots, K$, is given by [105, 210]:

$$\begin{aligned} P(n_1, \ldots, n_K) &= B(S)^{-1} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!}, \quad \text{where} \quad n \in \mathcal{S}(S), \\ \mathcal{S}(S) &= \{n \in \mathbb{N}_0, \sum_{k=1}^{K} A_{jk} n_k \leq s_j\}, \\ B(S) &= \sum_{n \in \mathcal{S}(S)} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!}, \quad \rho_k = \lambda_k \mathbb{E}[S_k], \end{aligned} \tag{2.69}$$

with λ_k the arrival rate to the network of patients of class k, $\mathbb{E}[S_k]$ the mean sojourn time in the network, s_j the number of servers at queue j and A_{jk} the number of servers a patient of class k claims at queue j. Loss networks are insensitive to the sojourn time distribution. Various algorithms and approximations exist to obtain blocking probabilities [105, 210].

The Queuing Network Analyzer

Despite the fact that many real world problems do not exhibit exponential service times, open Jackson networks have been used in numerous applications, often with good results. However, to analyze networks of general queues, the Queuing Network Analyzer (QNA) is a better alternative. The QNA was developed in 1983 by Ward Whitt [199] for approximate analysis of open networks of $G/G/s$ queues with FCFS service discipline. There are several variations on the QNA, also known as reduction or decomposition methods (see [35]). In this subsection we summarize the basic QNA algorithm.

QNA Algorithm

Step 1.
Calculate the aggregate arrival rates at queue j, λ_j:

$$\lambda_j = \gamma_j + \sum_{i=1}^{J} \lambda_i r_{ij}. \tag{2.70}$$

Step 2.
Calculate the load of a server at queue j, ϕ_j:

$$\phi_j = \frac{\lambda_j \mathbb{E}[S_j]}{s_j}. \tag{2.71}$$

Step 3.
Calculate the flow from queue i to queue j, λ_{ij}:

$$\lambda_{ij} = \lambda_i r_{ij}, \tag{2.72}$$

and the fraction of arrivals at queue j that come from queue i, q_{ij}:

$$q_{0j} = \frac{\gamma_j}{\lambda_j}, \quad q_{ij} = \frac{\lambda_{ij}}{\lambda_j}, \tag{2.73}$$

where q_{0j} denotes the fraction of external arrivals to queue j.

Step 4.
Calculate the scv for the arrival process at queue j, $c_{A,j}^2$:

$$\begin{aligned} c_{A,j}^2 &= a_j + \sum_{i=1}^{J} c_{A,i}^2 b_{ij}, \quad \text{with} \\ a_j &= 1 + w_j \left[(q_{0j} c_{0j}^2 - 1) + \sum_{i=1}^{J} q_{ij}((1 - r_{ij}) + r_{ij}\phi_i^2 x_i) \right], \end{aligned} \tag{2.74}$$

where c_{0j}^2 is the scv of the external arrival process at queue j, and

$$x_i = 1 + \frac{1}{\sqrt{m_i}}(\max(c_{S,i}^2, \frac{1}{5}) - 1), \tag{2.75}$$

with $c_{S,i}^2$ the scv of the service process at queue i. We have

$$\begin{aligned} b_{ij} &= w_j q_{ij} r_{ij}(1 - \phi_i^2), \quad w_j = \left[1 + 4(1 - \phi_j)^2(\eta_j - 1)\right]^{-1}, \quad \text{and} \\ \eta_j &= \left[\sum_{i=0}^{J} q_{ij}^2 \right]^{-1}. \end{aligned} \tag{2.76}$$

Step 5.
The mean waiting time at queue j, $\mathbb{E}[W_j]$, is given by

$$\mathbb{E}[W_j] = \mathbb{E}[W_{M/M/s}]\frac{c_{A,j}^2 + c_{S,j}^2}{2}. \tag{2.77}$$

The calculations involved with the QNA are usually straightforward and can be done by hand. However, when the parameters need to be changed often, we suggest using a spreadsheet program such as MS Excel. QtsPlus [159] also supports the analysis of general queuing networks. Even though the QNA has proved to be very useful, other approximation methods give better results when the network is highly congested (see [35] for further reference).

2.3.3 State of the Art in Networks of Queues

Queuing theory traces back to Erlang's historical work for telephony networks in 1909 [27]. The simplicity and fundamental flavor of Erlang's famous expressions, such as his loss formula for an incoming call in a circuit switched system to be lost (see Subsection 2.2.2) has remained intriguing, and has motivated the development of results with similar elegance and expression power for various systems modeling congestion and competition over resources.

A second milestone was the evolution of queuing theory into queuing networks as motivated by the product form results for manufacturing systems in the nineteen fifties obtained by Jackson [100]. These results revealed that the queue lengths at nodes of a network, where customers route among the nodes upon service completion in equilibrium can be regarded as independent random variables, that is, the equilibrium distribution of the network of nodes factorizes over (is a product of) the marginal equilibrium distributions of the individual nodes as if in isolation, see Subsection 2.3.1. These networks are nowadays referred to as Jackson networks.

A third milestone was inspired by the rapid development of computer systems and brought the attention for service disciplines such as the Processor Sharing discipline introduced by Kleinrock in 1967 [110]. More complicated multi-server nodes and service disciplines such as First Come First Served, Last Come First Served and Processor Sharing, and their mixing within a network have led to a surge in theoretical developments and a wide applicability of queuing theory, see Subsection 2.3.2.

Queuing networks have obtained their place in both theory and practice. New technological developments such as Internet and wireless communications, but also advancements in existing applications such as manufacturing and production systems, public transportation, and logistics, have triggered many theoretical and practical results. The questions arising in health care will no doubt again lead to a surge in the development of queuing theoretical results and applications.

Queuing network theory has focused on both the analysis of complex nodes, and the interaction between nodes in networks. Many textbooks and handbooks include or are devoted to queuing theory. Basic level textbooks include [180, 202], and more advanced handbooks are [11, 84, 110, 111, 141, 165, 187, 203]. The state of the art in the mathematical theory for queuing networks is described in the handbook [24]. Topics treated include:

- A general theory for product form equilibrium distributions far beyond those for Jackson and BCMP networks.
- Monotonicity and comparison results that allow analytical bounds on performance measures for networks that slightly deviate from Jackson or BCMP type networks.
- Fluid and diffusion limits that aim at analyzing networks in the regimes dominated by the mean or the variances of the underlying processes such as service times and inter arrival times.
- Computational results that are far more general than the queuing network analyzer of Subsection 2.3.2.

In the last chapter an application of networks of queues in healthcare is presented, indicating that many available theoretical results for networks of queues are waiting to be disclosed for application in healthcare.

2.4 Examples of Healthcare Applications

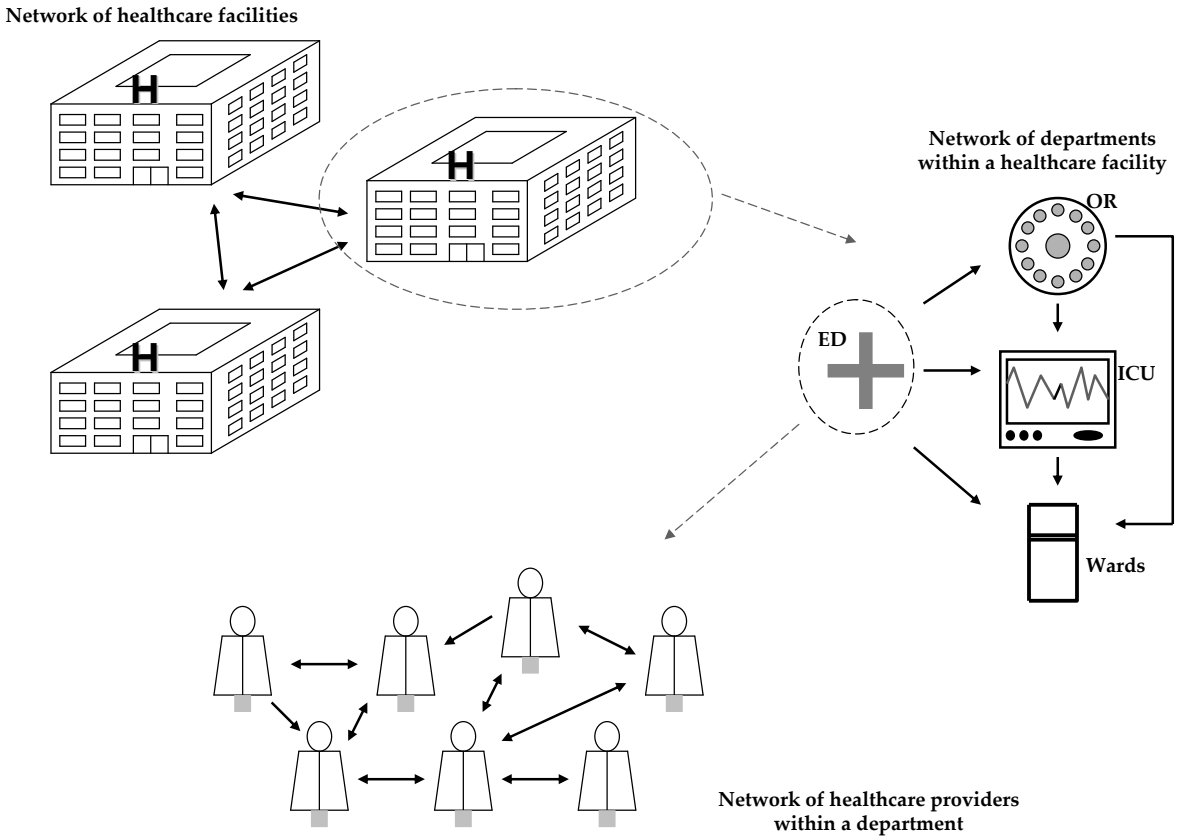
As we have seen in the previous section, for some queuing networks that consist of only exponential queues analytical solutions are available. When either the arrival or service process is non-exponential, approximation methods are usually required. In this section we provide several references to healthcare examples that involve queuing networks. For examples that involve single queues, we refer to [78]. Generally speaking, three

Table 2.1: Categorization of references

Network type	Exponential networks	General networks
Healthcare facilities	[9, 8, 20, 112, 122]	[1]
Departments within a facility	[45, 46, 58, 147]	[54]
Healthcare providers within a department	-	[4, 47, 101, 212]
Miscalleneous	[43, 208]	-

types of healthcare networks have been studied using queuing network topologies. We distinguish between networks of healthcare facilities, networks of departments within a facility, and networks of healthcare providers within a department (see Figure 2.9).

Figure 2.9: Different types of networks in healthcare



Using this network classification, and the distinction among exponential and general networks, the references provided in this section can be categorized as presented in Table 2.1.

2.4.1 Applications of Exponential Networks

Modeling a healthcare network with exponential queues gives a lot of insight into the structural behavior, such as bottlenecks. The modeling power of these networks is greatest when many of the details on patient behavior are not yet specified, but randomness is an essential part of the behavior of the system, i.e., at the strategical level of allocation of capacity, facilities and resources.

Facility Location and Bed Blocking Problems

One of the earliest developments in this area is given in [20], where a network of $M/M/s/s$ queues is combined with an algorithm to determine the optimal location of burn care facilities in the state of New York. The resulting system of equations can be solved, but due to computational difficulties only for a small number of facilities and beds. This type of network is further studied in [147]. The latter paper involves an example where patients are routed through a network of operative and post-operative units (such as the OR, ICU and nursing wards), and may experience bed-blocking when the next unit on the route operates at full capacity. Also in this model the numerical computations remain problematic when there are numerous units and beds. The relationship between the OR and bed availability on the ICU is further studied in [58], where the authors use a loss network to determine the blocking probability for surgical patients caused by a lack of ICU beds. The bed blocking problem is also considered in [112], where the flow of psychiatric patients within a network of healthcare facilities is considered. A relatively simple steady-state analysis results in a product-form solution. The capacity planning problem for neonatal units (how many cots to place at each care unit) is analyzed in [9] using a loss network model. The implementation of the solution is described in [8].

Patient Flow

Modeling patient flow has received limited attention [189]. Patient flow between different hospital departments is studied in two papers by the same author. In [45] the patient flow from the ED to the ICU and nursing wards is studied using an open Jackson Network. The same methodology is used in [46] to analyze flow of obstetric patients. Patient flow within a care facility is studied from another perspective in [43] and [208]. In these papers, different phases in the care trajectory of a patient are considered. While in [43] a closed queuing network is used, in [208] the model is extended to a semi-open queuing network with a capacity constraint (the maximum number of patients that can be admitted).

Clinical Capacity Problem

Patients with renal failure are considered in [122]. These patients either receive dialysis at a clinic, or when their condition worsens, (temporarily) hospitalized. A multi-class open queuing network with two queues (the clinic and the hospital respectively) is used to determine the clinic's capacity and the maximum number of patients to be admitted into the clinic, given that patients do not use clinic resources when they are hospitalized.

2.4.2 Applications of General Networks

When a higher level of detail is required, for example when networks of healthcare providers withing a department are studied, models with general queues are of more value.

Organization of Acute Care

The organization of acute care is studied in [47] and [101]. In [47] an ED is modeled with a multi-class open network of $M/G/s$ queues. The main purpose of this model is to determine the required ED capacity needed to achieve service targets such as waiting time and overflow probabilities. In [101] the same kind of network is used to model an urgent care center, which is basically an outpatient clinic that delivers ambulatory urgent care to relieve pressure from the ED. The main goal of this model is to determine whether parallelization of tasks in the patient's care trajectory has a positive effect on the patient's LOS at the urgent care center.

Other Applications

In [54] hospital departments and their interdepartmental relationships are modeled as a network with $G/G/s$ queues. Analysis of the network gives relevant information such as utilization rates and mean waiting times for each queue, and also allows for exploring the impact of service interruptions, aggregating patient flows, and determining the optimal number of patients in a clinic session. Another application is the recent outbreaks of viruses, such as the H1N1 influenza virus, which call for a rapid response of the authorities. In [1] the authors show how a queuing network can help to plan emergency mass dispensing and vaccination clinics. In [4] and [212] two outpatient clinics are studied using the Queuing Network Analyzer. The papers demonstrate how queuing networks can be of added value when performing bottleneck analysis.

2.5 Challenges and Directions for Future Research

In the last decade the number of healthcare problems that have been studied using a queuing network approach has increased tremendously. Except for [4] and [20], all of the references included in Section 2.4 were published in the years 2000-2010. In this final section we point out a few directions for future research. We distinguish between mathematical challenges: healthcare problems for which appropriate queuing network models have not yet been developed, and healthcare challenges: healthcare problems which have not been studied yet, but could be studied with the queuing techniques available in literature.

2.5.1 Mathematical Challenges

The mathematical challenges mainly lie in the modeling aspect. One example is the development of models for networks of care providers who perform several tasks in parallel, in sequence, and sometimes even in a mixed form. Polling models [182] could be of interest here. Also, clinics where patients have to (re-) visit specific care providers in a network of care queues still involve modeling complications. However, re-visiting occurs often in reality (consider for example the complex network of multiple care providers in ED treatment).

The application of time inhomogeneous models that capture the time-dependent arrival patterns of patients has attained only limited attention, see for example [83]. Introducing time inhomogeneity in queuing networks is a tremendous challenge. Related is the development of computationally efficient methods that explicitly take into account opening hours of healthcare facilities.

2.5.2 Healthcare Challenges

Healthcare professionals in a couple of fields are familiar with the possibilities of mathematical decision support techniques in general and queuing theory in particular. As we have seen in Section 2.4, modeling networks of healthcare facilities, departments and care providers has received some attention. However, capturing the complex relationships between hospital departments has proved to be quite involved. The relationship studied is usually that with a downstream department [189], while that with upstream departments is not considered, even though it can be of significant influence.

Our aging population requires more and more care, which has to be delivered with limited resources. Rationing care and the consequences thereof has therefore become an important research topic. Decisions regarding which patient class will be offered what type of care are inevitable. The influence of these decisions on other patient classes, regarding accessibility and other important matters, should be studied in detail. Moreover, the dimensioning of healthcare facilities, not only in the number of beds required, but also regarding care that will be offered to certain patient classes only, will become increasingly important.

This chapter has provided a thorough theoretical background on networks of queues and examples of how networks of queues may be used to model, analyze and solve health care problems. In that respect, often, the theory has to be amended or extended. We are confident that this contribution has made health care professionals increasingly aware of the possibilities and opportunities queuing networks have to offer to tackle the challenges they are facing, now and in the future.

Part II

Challenges for Outpatient Clinics and Diagnostic Facilities

Chapter 3

Redesign of the PAC

3.1 Introduction

In the past two decades, it has become common practice to provide preoperative screening in an outpatient clinic setting [49, 125, 155]. Lee [123] was the first to outline the concept of the preanesthesia evaluation clinic (PAC). He stated that the purpose of the preoperative screening process is 'to examine and treat the patient, so that he will arrive in the operating theatre as strong and as healthy as possible', a definition that still adequately defines the process. Today many hospitals operate a PAC [155]. An accurately performed screening reduces the risk of cancellation on the day of surgery due to the physical condition of the patient [67], increases the rate of same-day admissions and reduces peri-operative morbidity, resulting in decreased costs and increased quality of care [109, 149].

Congestion is a common phenomenon in outpatient clinics [55, 62, 86]. Patients arriving for a preoperative screening are usually not categorized and therefore the consultation time needed per patient is difficult to estimate. This increases the complexity of the PAC organization as compared with a regular outpatient clinic. In our PAC at Leiden University Medical Center, patient waiting times and LOS (in this case the total duration of one clinic visit) were initially significantly shorter than in a comparable clinic [64], but these increased dramatically after introduction of an electronic patient data management system, since together with the information system additional administrative activities were introduced. Also, the workload of the staff increased, leading to multiple complaints about work stress. The prolonged waiting times, together with the low level of job satisfaction for clinic employees, called for an evaluation of alternative clinic designs. The aim of this study was to explore possibilities for a more efficient operation of our PAC organization. Since all patient movements within the PAC were logged, we chose to use mathematical techniques to analyze performance.

The major advantage of mathematical modeling is the possibility to execute a thorough analysis of a system, while having no impact on the system itself. Using our mathe-

matical model, we investigated the effect of various designs on selected performance measures, such as patient LOS and staff utilization rate (the fraction of time clinic staff is occupied with patient related activities). One of the alternative designs we considered was regarded as superior to the initial design by the clinic staff. This design was implemented at our PAC in 2007. Following the intervention, an unexpected increase of 16% in patient visits in the first quarter of 2008 occurred. However, this did not cause a significant increase in waiting times, and in addition resulted in a decrease of employee costs per patient. Furthermore, the time needed to approve a patient for surgery decreased, and employee satisfaction increased. This chapter describes the redesign process and provides directions for other PAC managers.

The present study is based on a queuing modeling approach. Simulation is a more common approach in this area. Already in 1952, Bailey used Monte-Carlo Simulation to analyze appointment systems for outpatient clinics [14]. Since then, simulation has been used extensively for the study of outpatient clinics. Within the scope of the PAC, simulation was used to analyze the capacity needed to shorten the waiting list [64] and to study the design of appointment systems for the PAC to minimize patient waiting times [55]. The queuing modeling approach we employ in this study requires only mean and variance of consultation times and patient arrival processes, and no further assumptions on the underlying distributions. Due to the careful analysis required prior to the formulation of the equations used in the model, a robust insight in the underlying relationships of the system is obtained. As can be seen in Section 3.5, the queuing model presented in this chapter consists of several related formulas that can be entered into a spreadsheet. It enables a bottleneck analysis of the processes at the clinic and can easily be adjusted so that it represents one of the alternative designs considered in the redesign process. It is also possible to adjust the model so that it represents a preanesthesia clinic at another hospital. Applications of queuing theory in outpatient clinic settings are scarce. The majority of papers published on this matter are covered by Preater in his extensive bibliography on queues in health [156].

3.2 Methods

3.2.1 Initial service of the PAC

The study was performed at a university hospital preanesthesia evaluation clinic, with approximately 6000 patient visits annually. A majority of patients were seen on walk-in basis (about 70%), and the remaining on appointment basis. Walk-in patients arrived directly from surgical outpatient clinics within the hospital. Only ASA I or II patients were evaluated on walk-in basis, since for ASA III or IV patients more time for patient contact and additional information from other specialists was often required (see [7] for more information on the ASA classification system). It was clinic policy to maximize

the number of walk-in patients, although at the same time these patients posed an uncertain demand on clinic resources. Although less than 10% of patients were classified ASA III or IV and therefore required an appointment, 30% of all patients were given an appointment. When walk-in patients were deferred to an appointment, it was usually because of overcrowding in the waiting room.

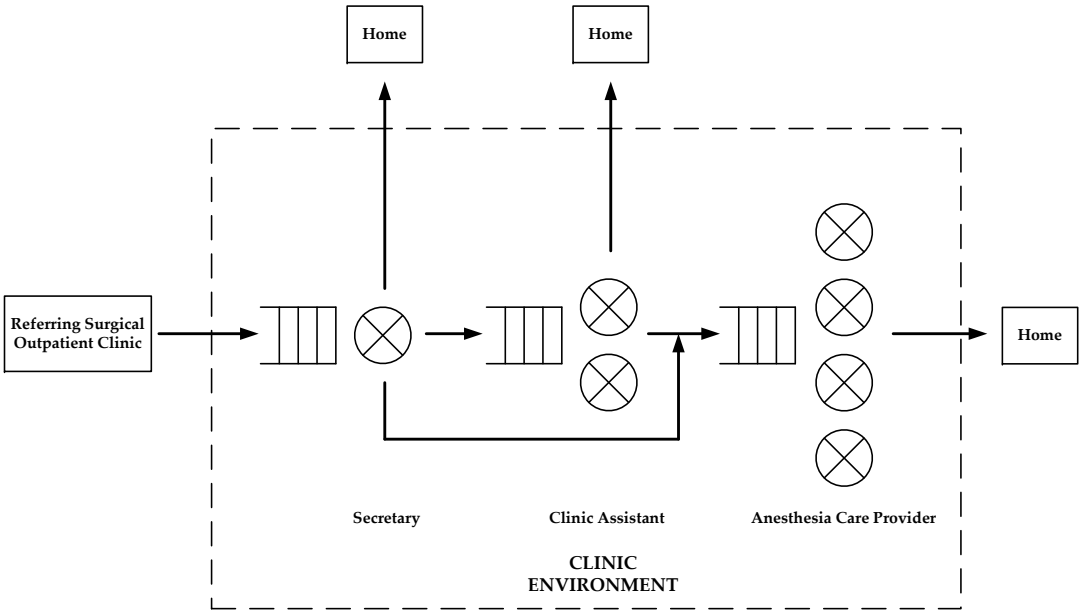
3.2.2 Resources and Tasks

The clinic was run by the department of Anesthesiology, with four anesthesia care providers attending: one staff anesthesiologist, two residents, and a nurse practitioner, supported by a secretary and two clinic assistants. The screening process consisted of at most three steps: an intake at the secretary and two separate contacts with the clinic assistant and anesthesia care provider respectively. All patients would see the secretary and anesthesia care provider, only adults were seen by the clinic assistant. Patients returned to the waiting room between visiting each care provider (see also the diagram in Figure 3.1). Based on a form completed by the referring specialist, the secretary decided whether the patient could be assessed immediately or during an appointment at a later time. Since the secretary is not equipped to make decisions regarding the medical status of the patient, this procedure resulted occasionally in patients receiving an appointment they did not need and vice versa. If the patient received an appointment, the time interval, usually one or two weeks, was used for back-office activities to complete the patient's file. Walk-in patients were approved for surgery by the anesthesia care provider during their visit. The staff anesthesiologist performed the back-office activities, consisting mostly of processing additional patient information that was required to finish the case of appointment patients. Because the staff anesthesiologists also served as backup manpower for the front-office activities, they experienced significant work stress. Furthermore, the anesthesia care providers were unhappy because complicated cases had to be finalized by an anesthesia care provider who had not seen the patient initially, which ultimately may result in an incomplete understanding of the medical condition of the patient [73].

3.2.3 Using Queuing Theory to Analyze PAC Performance

The initial and alternative designs were compared with a multi-class open queuing network model (for a detailed description see Section 3.5). An advantage of this queuing model is that only the first two moments (mean and standard deviation) of the arrival and service time distributions are needed in the calculations. This allows usage of all possible types of distributions, including empirical distributions. For the comparison two performance measures were calculated with the queuing model, namely patient LOS and employee utilization rate (ρ). In the recent work by Jiang and Giachetti [101], the authors briefly describe a survey held at their outpatient clinic. From the survey it

Figure 3.1: Diagram of clinic operations



followed that patients considered the waiting time, being an important contributor to the LOS, as very important and unsatisfactory long. Other aspects, such as the consultation with the anesthesia care provider and the clinic assistant, also contribute to the patient's contentment on the clinic visit [63]. Employee utilization rate, ρ, and the patient's waiting time to see this specific employee, $\mathbb{E}[W^q]$, are related (see equation (2.2)). Knowledge of the utilization rate is essential, since increasing this factor when it is already close to one, by increasing either the arrival rate or the service time, will result in a considerable increase of the waiting time.

3.2.4 Intervention

All parties involved felt that the situation at the PAC required an intervention. A working group was formed with representatives of all PAC employees. The working group discussed the initial (i.e., the in place) design, and developed four alternative designs, which are described in the subsequent paragraphs. When discussing the initial design, the working group identified all relevant activities at the PAC and characterized the order of these activities in the initial design in several flow charts. Ultimately the working group decided upon the planned design from the presented alternatives. Again, the order of all activities in the new design was documented in several flow charts and medical protocols. The queuing model results were used to guide the decision making process and enabled a numerical comparison of the initial and alternative clinic designs.

Alternative Design 1: Clinic Assistant Selects at Front Desk The clinic assistants were convinced that many patients with an actual ASA III or IV score were assigned an erroneous ASA I or II score by the secretary. These mis-categorized patients were then handled on walk-in basis and consumed too much time in the office of the anesthesia care provider, resulting in congestion in the waiting room. They suggested that one of the clinic assistants should take over part of the front desk task from the secretary, while the other clinic assistant performs measurements and blood sampling.

Alternative Design 2: Treat all Patients on Appointment Basis Demand for an outpatient clinic's services can be divided into two components: controlled (appointment patients) and uncontrolled (walk-in patients) demand [164]. In the initial set-up most ASA I or II patients were seen as walk-in patients. In the second alternative all patients are deferred to an appointment, since a clinic with an appointment-only system will always provide a better service level with respect to patient waiting times, than a clinic that allows walk-in arrivals [55].

Alternative Design 3: Reschedule Appointments Rising [164] suggested to schedule appointments such that they complement walk-in arrivals. This results in a more homogeneous arrival pattern throughout the day. In the PAC under study the number of walk-in arrivals was significantly lower in the early morning and on Friday afternoon. In this alternative all appointments are scheduled in these periods.

Alternative Design 4: Regroup Employee Tasks and Amend Patient Flows In this alternative the secretary accepts all patients; therefore all patients are seen by the clinic assistant on their first visit. Clinic assistants are provided with protocols to aid in the decision whether the patient can be seen immediately based on the extent of co-morbidity, contacts with medical specialists and the requirement to obtain additional medical information prior to the visit to the anesthesiologist. If the patient requires additional testing, these tests are immediately performed and/or requested and the patient is deferred to an appointment, scheduled when all additional information is available. Consequently, the patient can be approved for surgery when the appointment takes place.

3.3 Results

3.3.1 Model Input

Data from all PAC visits recorded in the first quarter of 2007 was used to obtain input parameters for the queuing model ($n = 1492$). For the analysis, patients were divided in three separate classes: (1) children (<16 years old), (2) adult patients ASA Score I or II, and (3) adult patients ASA Score III or IV. This classification was chosen since children and adults have a different routing, moreover the three classes can be distinguished with respect to how much time each requires in consultation with the anesthesia care provider. An advantage of this classification is that it is similar to that used by clinic

staff. Arrival rates for each patient class, and mean and standard deviation of the contact time with the clinic assistant and anesthesia care provider were determined. Not all registered contacts had complete data and therefore the records of 1293 patients (87%) could be used for the latter part of the data analysis. The time patients spent with the secretary was not recorded and therefore we used an estimate. The secretary was often consulted by co-workers who inquire after the approval status of a particular patient, either by phone or in person at the reception desk. The secretary was also responsible for dealing with patient inquiries, either on the phone or in person. The anesthesia care providers were often consulted by co-workers, the inquiries usually concerning their other professional responsibilities. We estimated that regarding the time available for direct contact with patients visiting the clinic for a consult, the secretary lost 50% and the anesthesia care providers lost 33% due to these interferences. The values were obtained by direct observation and interviewing the employees. Even though the aforementioned tasks do not directly contribute to the patient's visit at the clinic, they need to be done and are part of the job in our hospital organization.

The number of arrivals per patient class was used to determine the distribution of patients among classes. We found that the majority of patients arrived between 10 AM and 4 PM. Hence we focused our analysis on this interval and calculated the arrival rate (3.73 patients/hour) by using patient arrivals recorded during this interval. We observed that within this period, patients from all classes arrived in a homogeneously distributed manner. This corresponds with the scv (see Subsection 2.1.1) of the arrival process being equal to 1 for all patient classes. The arrivals of patients that were immediately deferred to an appointment were not recorded. Assuming that all appointment patients make their appointment at the reception desk, we calculated the arrival rate of non-admitted patients by multiplying the admitted patient arrival rate by the appointment percentage for each patient class. A summary of input data is given in Table 3.1. Senior clinic staff members discussed and carefully checked all parameter values; additionally they discussed and approved the queuing model design.

Table 3.1: Summary of input data; mean service time $\mathbb{E}[S]$ is in minutes

Patient class	N	App. %	Arrival rate (pt/hr)	Secretary $\mathbb{E}[S]$; SD$[S]$	Clin. ass. $\mathbb{E}[S]$; SD$[S]$	Anes. care prov. $\mathbb{E}[S]$; SD$[S]$
Children	274	15	0.79	5.00; 5.00	-	24.30; 20.64 (n =274)
Adults ASA I&II	902	25	2.60	5.00; 5.00	10.71; 8.97 (n =711)	27.24; 17.26 (n =902)
Adults ASA III&IV	117	78	0.34	5.00; 5.00 (n =86)	16.31; 14.20 (n =117)	52.05; 25.50
Deferred to appointment	-	-	1.04	2.50; 2.50	-	-

3.3.2 Comparison of Initial Design and Alternatives

Using the queuing model, we compared each alternative design with the initial design. If necessary, input parameters were adjusted (see the list below for the modifications per alternative and Section 3.5 for an explanation of the parameters itself).

Alternative 1 One clinic assistant moves to the secretary station ($s_2 = 1$), no disturbance during welcoming of patients ($e_1 = 1$).

Alternative 2 The secretary gives all patients an appointment the first time they arrive at the PAC, thus the arrival rate increases ($\zeta_1 = 3.73$). We assume that appointment patients arrive on time; i.e., patients are assumed to arrive on appointment basis with fixed and identical inter-arrival times, so as to analyze the maximal possible benefit of an appointment scheme. Hence the standard deviation of the inter-arrival time equals zero for all patient classes
($c^2_{A,2,1} = c^2_{A,3,1} = c^2_{A,4,1} = 0$).

Alternative 3 Appointments are rescheduled outside the interval 10AM to 4PM, and therefore the fraction of patients with an appointment is removed from the arrival rates ($\zeta_2 = 1.93$, $\zeta_3 = 0.07$, $\zeta_4 = 0.67$).

Alternative 4 No patients are deferred to an appointment by the secretary ($\zeta_1 = 0$). Consultation time at the secretary decreases with 2.5 min, because part of tasks are reallocated to clinic assistants ($E[S_{r,1}] = 2.50$); consultation times at clinic assistants increase with these 2.5 min and with an additional 2.5 min needed to determine upon additional testing ($E[S_{2,2}] = 15.71$, $E[S_{3,2}] = 21.31$, we assumed that the ratio between expectation and variance of the contact time at the clinical assistants, and therefore the scv, remained constant). We assume that appointment patients arrive on time. Therefore, the standard deviation of the inter-arrival time equals 0, which results in an scv equal to 0 ($c^2_{A,5,1} = c^2_{A,6,1} = c^2_{A,7,1} = 0$).

The performance measures we chose to compare were mean patient LOS (the total duration of one clinic visit) and employee utilization rate. The initial design could be characterized by a long mean patient LOS, caused by prolonged waiting times at the secretary and later in the process, prior to the contact with the anesthesia care provider (Table 3.2). These two care stations also had high utilization rates. Comparing the performance measures of the initial design to those of the alternative designs lead to the conclusion that all alternative designs, except alternative 2 (treat all patients on appointment basis) would result in a better overall performance. Once the model results were available, the working group was consulted to make a decision on the next step to take in the redesigning process. It was apparent to all members that the initial design could not be maintained. The first alternative of relocating one clinic assistant to the secretary's station was not regarded as a valuable alternative, since the expected decrease in patient LOS was minimal. Furthermore, patient waiting time at the remaining clinic

Table 3.2: Results of analytical model; $\mathbb{E}[W^q]$ and patient LOS (LOS) (for the most common group of walk-in ASA I/II patients) are in minutes

Design	Secretary ρ; $\mathbb{E}[W^q]$	Clin. ass. ρ; $\mathbb{E}[W^q]$	Anes. care prov. ρ; $\mathbb{E}[W^q]$	Patient LOS
Initial	0.68; 19.20	0.28; 0.60	0.67; 9.60	77.35
Alternative 1	0.34; 2.40	0.56; 12.60	0.67; 9.00	71.95
Alternative 2	0.90; 54.00	0.28; 0.60	0.67; 9.60	107.15
Alternative 3	0.51; 9.60	0.18; 0.60	0.45; 1.80	59.95
Alternative 4	0.38; 3.00	0.40; 2.40	0.67; 9.60	62.95
Alternative 3+4	0.30; 2.04	0.40; 2.63	0.44; 1.60	54.22

assistant increased substantially, which was also undesirable. Based on the predicted increase in patient waiting time at the secretary in alternative 2, which was caused by all patients having to make an appointment first, and since introducing an appointment-only system was regarded as patient unfriendly (in the sense of one-stop shopping) by the working group, alternative 2 was eliminated. The working group members decided to implement alternative 3 and 4, so that advantages of both alternatives were included. The effects of combining alternatives 3 and 4 were again studied with the queuing model (Table 3.2). The queuing model predicted that this intervention would also result in an improvement. Supported by the results, all working group members were convinced that implementing a combination of the two alternatives would yield a better overall performance of the clinic.

3.3.3 Effect of Intervention

The new design was implemented in the summer of 2007. We compared actual measured times of total patient LOS before and after the intervention. To minimize seasonal influences and to allow for learning effects, we used data from both the first quarter of 2007 and 2008. Before the intervention, only one clinic assistant was present on Fridays. Since the intervention involved scheduling the majority of appointments on Fridays, one additional clinic assistant shift was now required. This caused an increase in total employee capacity from 7.20 FTE (total costs: 109K Euros/quarter) to 7.87 FTE (total costs: 116K Euros/quarter, +6%). Before the intervention, the total LOS as obtained from the measurements over a 90 day period (01/01/2007 - 03/31/2007) was on average 70.0 minutes for the entire patient group (95% CI : [62.8; 77.1]). After the intervention, the total LOS as obtained from the measurements over a 91 day period (01/01/2008 - 03/31/2008) was on average 77.0 minutes for the entire patient group (95% CI: [70.6; 83.3]). Although the total LOS did not increase significantly, longer contact and waiting times at the clinic assistant were measured (95% CI of increase in contact times: [5.5; 7.6], 95% CI of waiting times in 2007: [8.6; 25.1], 95% CI of waiting times in 2008: [25.8; 30.3]). Recall that not all patients see the clinic assistant and therefore the increase in to-

tal patient LOS was less. The contact and waiting times at the anesthesia care provider did not increase significantly (95% CI of increase in contact times: [-1.5; 1.5], 95% CI of waiting times in 2007: [19.6; 26.9], 95% CI of waiting times in 2008: [19.8; 28.1]). In the first quarter of 2008, 1737 patient contacts were registered during the opening hours of the clinic, an increase of 245 patients (+16%) compared to the first quarter of 2007. Dividing the total personnel costs by the number of patients for both quarters, we see that personnel costs decreased from 73 to 67 Euros per patient (-8%). The percentage of patients seen on walk-in basis increased from 72% in 2007 to 81% in 2008. Furthermore, in 2008 the anesthesiologist needed 6.8 days to decide upon approving the patient for surgery, compared to 7.9 days in 2007 (95% CI: [-0.3; 2.3]). The staff anesthesiologists were responsible for finalizing the status of those patients for which new information was obtained in the days or weeks after the patient had visited the PAC. After the intervention this task was minimal (less than 30 minutes), as for most patients all relevant information was available prior to the first visit to the attending anesthesia care provider.

3.3.4 Validity of the Model

The average LOS of the most common group of patients (walk-in patients with ASA Score I or II) measured at the clinic in the first quarter of 2007 (70.6 min) was slightly less than predicted with the queuing model (77.4 min, -9%), and thus the queuing model provided a conservative but close prediction of system behavior. When comparing the average LOS for the same patient group measured in the first quarter of 2008 (77.9 min) with the model's prediction (62.2 min) we see that the model underestimated the LOS with 25% (see Table 3.3). However, we found that in the new clinic design, the secretary was not able to halve her consultation time, since her remaining tasks required more time than expected prior to the intervention. If we incorporate this in the queuing model, and use the original consultation time, we come to a LOS equal to 90.4 min (-14%), and the queuing model again gives a conservative estimate. The validity of the model outcomes highly depends on the parameter values.

3.4 Discussion

We demonstrated a queuing modeling approach that enables a fast and robust analysis of PAC performance. The methodology can be applied to other preoperative screening clinics as well. Given the queuing model results, the PAC was redesigned. This process consisted of two parts, namely the re-scheduling of appointments to the early morning and Friday, and the reassignment of tasks from the secretary to the clinic assistants. As a consequence, all patients were seen on their walk-in visit by the clinic assistant. Patients requiring more contact time with the anesthesia care provider or back-office activities

were deferred to an appointment by the clinic assistant, scheduled when all required information was available. Literature about the re-design of hospital care is extensive [65]. However, the literature on re-design of outpatient and preanesthesia evaluation clinics is limited. Some studies are dedicated to the design of appointment systems [55], others concentrated largely on waiting times and patient satisfaction [64, 92]. The concept of re-design by reallocating tasks at the outpatient clinic has received less attention.

A limitation of this study is that all outcomes of the queuing model were calculated under the assumption of steady state behavior. The system under study will never reach this equilibrium state, due to inhomogeneous patient arrivals and restrictive opening hours. We used the queuing model solely for comparison purposes and not for prediction of actual patient LOS and utilization rates, which further strengthened our confidence in the followed approach.

The model enabled us to analyze the effect of increased pressure on the clinic. As mentioned in the results section, patient arrivals had increased with 16% in the first quarter of 2008, compared to the same period in 2007. Nevertheless, empirical analysis showed that patient LOS had only increased slightly. The model shows that the rise in patient arrivals would have resulted in a tremendous increase in patient LOS and employee utilization rate, if we had not changed the design of our PAC (Table 3.3). Under the 2008 data the initial design operates under high pressure, with an increase in LOS of 53%, due to the 16% increase in patient arrivals. In the implemented design, due to increased efficiency, the system operates under modest pressure, with an increase in LOS of only 15% (Table 3.3). This is in line with the relationship given in Formula (2.2), indicating the typical relation between waiting time and load. By organizing the processes at the clinic more efficiently, we reduced the load. Therefore, the increase in patient arrivals did cause an increase in the load but only a slight increase in waiting time, and patient LOS.

The majority of patients visiting our PAC are seen on a walk-in basis. Since patients

Table 3.3: Results of analytical model with 2008 data (arrival rates: children 0.87, adults ASA I/II 3.32, adults ASA III/IV 0.41, deferred to appointment 0.88); time is in minutes

Design	Secretary ρ; $\mathbb{E}[W^q]$	Clin. ass. ρ; $\mathbb{E}[W^q]$	Anes. care prov. ρ; $\mathbb{E}[W^q]$	Patient LOS
Initial	0.81; 38.24	0.35; 1.46	0.83; 30.38	118.03
Alternative 3+4	0.37; 2.80	0.50; 4.75	0.63; 6.70	62.20

have the opportunity to go straight from the surgical outpatient clinic to the PAC, they are often able to finalize the entire preoperative preparation within one hospital visit (one-stop shopping), avoiding multiple hospital visits. However, walk-in outpatient clinics are notoriously more difficult to handle in terms of optimizing waiting times for patients and peak pressures for anesthesia care providers. Dexter [55] states that the best service walk-in PACs can provide will always be worse than appointment PACs. The walk-in PAC requires more resources to have acceptable waiting times for patients

[64], since more slack is required to deal with unexpected peaks in patient arrival. Appointment systems on the other hand deal with peaks in demand for PAC services by building waiting lists. To allow for patients that need to be seen with some urgency, these appointment-only outpatient clinics will usually have some unplanned time slots (or add-on manpower). At the PAC under study, we use a system that allows both walk-in and appointment patients. The decrease of back-office activities enabled the anesthesia care providers to dedicate more time to patient contact. This explains how 16% more patients could be seen without an increase in the number of anesthesia care providers.

3.5 The Queuing Model

To identify bottlenecks in the PAC's operations, the clinic was modeled as a multi-class open queuing network (see Figure 3.1). There were three patient classes: children, adults eligible for direct (walk-in) screening, and adults requiring an appointment because of their (more severe) health status. The PAC queuing network has three separate (connected) queues, where the employees act as servers. Patients only enter the PAC through the secretary queue, but may leave the system at any queue. The PAC queuing network was analyzed using a decomposition method, based on the QNA (see Subsection 2.3.2). This method consists of three steps. We first summarize our approach and then provide a detailed description of the model with the corresponding formulas.

First, the multi-class network is reduced to a single class network. This is done by aggregating all patient flows that enter a queue. Then the workload ρ is calculated for each queue. This already gives significant and valuable information; recall that ρ is a measure for the fraction of time employees are busy. In the next step, the single class open queuing network is analyzed, where the mean contact time and scv of the joint arrival and service processes at the three queues are deduced. In the final step the mean waiting time per queue is calculated, using the variables that were derived in step 1 and 2.

The PAC queuing network consists of three queues. The secretary queue is a single-server queue whereas the clinic assistant and anesthesia care providers are represented by multi-server queues. Patients enter the queuing network via the secretary queue and depart the system from any of the queues. Furthermore, if upon arrival at a queue an employee is available patients are served immediately; otherwise they join the queue and are treated on first come first serve basis. We use an approximate decomposition method [19] that is based on the QNA to analyze the model. The model we will present here is more involved than the initial QNA formulation as given in Subsection 2.3.2. Practical situations can usually not be directly translated into an existing model. Instead, the theory has to be amended and extended to represent reality. We will describe in detail the changes we have made to the QNA algorithm.

First we introduce some notation. We denote by k a patient class, where $k = 1$ are patients deferred to an appointment by the secretary, $k = 2$ adults with ASA I or II, $k = 3$

adults with ASA III or IV, and $k = 4$ are children. To evaluate the alternative clinic design, we also introduce $k = \{5, 6, 7\}$ to represent patients (adults with ASA I or II, adults with ASA III or IV, and children, respectively) who return for their appointment. We denote a queue with j, where $j = 1, 2, 3$ resp. represents the secretary, clinic assistant and anesthesia care provider.

Step 1.
The aggregated arrival rates at queue j are:

$$\lambda_1 = \sum_{k=1+d}^{4+3d} \gamma_k, \quad \lambda_2 = \sum_{k=2}^{3} \gamma_k, \quad \lambda_3 = \sum_{k=2}^{4} (1 - da_k)\gamma_k + d \sum_{k=5}^{7} \gamma_k, \tag{3.1}$$

where γ_k is the arrival rate of patient class k at queue 1, and a_k is the fraction of patients of class k who are deferred to an appointment in the alternative clinic design. Since all patients in the alternative design are seen by the clinic assistant during their first visit at the PAC, the secretary does not defer patients to an apopintment, and patient class $k = 1$ does not exist anymore. Also, the index $k = \{5, 6, 7\}$ only exist when the alternative clinic design is evaluated. We therefore introduce the binary variable d, which equals 1 if the alternative clinic design is evaluated and 0 otherwise.

Step 2.
The load per patient class per server for queue 1,2, and 3 is:

$$\begin{array}{rcll} \phi_{1,k} & = & \gamma_k \mathbb{E}[S_{k,1}] \frac{1}{e_1 s_1} & \text{for} \quad k = \{1 + d, \ldots, 4 + 3d\}, \\ \phi_{2,k} & = & \gamma_k \mathbb{E}[S_{k,2}] \frac{1}{s_2} & \text{for} \quad k = \{2, 3\}, \\ \phi_{3,k} & = & \gamma_k \mathbb{E}[S_{k,3}] \frac{1}{e_3 s_3} + d(1 - a_k)\gamma_k \mathbb{E}[S_{k,3}] \frac{1}{e_3 s_3} & \text{for} \quad k = \{2, \ldots, 4 + 3d\}, \end{array} \tag{3.2}$$

where $\mathbb{E}[S_{k,j}]$ is the mean service time for patient class k at queue j. Since the secretary is often consulted by other patients and co-workers while handling a patient at the reception desk, an effective capacity e_1, $0 < e_1 \leq 1$, is taken into account when calculating the mean time a patient spends at this queue. The anesthesia care provider is often disturbed, but not while treating patients and therefore the effective capacity, e_3, $0 < e_3 \leq 1$, is only used in calculating the load. These effective capacities are calculated by using direct observations and interviews with the employees. The number of servers (i.e. employees) at queue j equals s_j. Adding the load over all patient classes gives the aggregated load per server of queue j, $j = 1, 2, 3$:

$$\phi_1 = \sum_{k=1+d}^{4+3d} \phi_{1,k}, \quad \phi_2 = \sum_{k=2}^{3} \phi_{2,k}, \quad \phi_3 = \sum_{k=2}^{4+3d} \phi_{3,k}. \tag{3.3}$$

For stability it is required that $\phi_j < 1$ for all queues j.

Step 3.
The flow from queue 1 to queue 2 or 3 and from queue 2 to queue 3 is given by:

$$\lambda_{1,2} = \frac{\sum_{k=2}^{3}(1-da_k)\gamma_k}{\lambda_1}, \quad \lambda_{1,3} = \frac{\sum_{k=4}^{4+3d}(1-da_k)\gamma_k}{\lambda_1}, \quad \lambda_{2,3} = \frac{\sum_{k=2}^{3}(1-da_k)\gamma_k}{\lambda_2} \tag{3.4}$$

The fraction of arrivals at queue 3 that come from queue 1 or 2 is given by (note that $q_{1,2} = 1$):

$$q_{1,3} = \frac{\sum_{k=4}^{4+3d}(1-da_k)\gamma_4}{\lambda_3}, \quad q_{2,3} = \frac{\sum_{k=2}^{3}(1-da_k)\gamma_k}{\lambda_3}. \tag{3.5}$$

Step 4.
The arrival process at queue 1 has scv, $c^2_{A,1}$:

$$c^2_{A,1} = w_1 \sum_{k=1+d}^{4+3d} Q_{k,1} c^2_{A,k,1} + 1 - w_1, \tag{3.6}$$

where $c^2_{A,k,1}$ is the scv of the arrival process of patient class k at queue 1, and

$$w_1 = \left(1 + 4(1-\phi_1)^2(\eta_1 - 1)\right)^{-1}, \quad \eta_1 = \frac{\lambda_1^2}{\sum_{k=1+d}^{4+3d}\gamma_k^2}, \quad Q_{k,1} = \frac{\gamma_k}{\lambda_1}. \tag{3.7}$$

The mean service time, $\mathbb{E}[S_1]$ and scv at queue 1, $c^2_{S,1}$, are:

$$\mathbb{E}[S_1] = \frac{\sum_{k=1+d}^{4+3d}\gamma_k \mathbb{E}[S_{k,1}]}{\lambda_1}, \quad c^2_{S,1} = \frac{\sum_{k=1+d}^{4+3d}\gamma_k \mathbb{E}^2[S_{k,1}](c^2_{S,k,1}+1)}{\lambda_1 \mathbb{E}^2[S_1]} - 1, \tag{3.8}$$

where $c^2_{S,k,j}$ is the scv of the service time for patient class k at queue j. The arrival process at queue 2 has scv, $c^2_{A,2}$:

$$c^2_{A,2} = \lambda_{1,2} c^2_{D,1} + 1 - \lambda_{1,2}, \tag{3.9}$$

where $c^2_{D,1}$ is the scv of the departure process at queue 1. Queue 2 has mean service time, $\mathbb{E}[S_2]$, and scv, $c^2_{S,2}$:

$$\mathbb{E}[S_2] = \frac{\sum_{k=2}^{3}\gamma_k \mathbb{E}[S_{k,2}]}{\lambda_2}, \quad c^2_{S,2} = \frac{\sum_{k=2}^{3}\gamma_k \mathbb{E}^2[S_{k,2}](c^2_{S,k,2}+1)}{\lambda_2 \mathbb{E}^2[S_2]} - 1. \tag{3.10}$$

The arrival process at queue 3 has scv, $c^2_{A,3}$:

$$\begin{aligned}
c^2_{A,3} &= w_3(q_{2,3}c^2_{2,3} + q_{1,3}c^2_{1,3}) + 1 - w_3, \quad \text{with} \\
w_3 &= \left(1 + 4(1-\phi_3)^2(\eta_3 - 1)\right)^{-1}, \quad \eta_3 = \left(q^2_{2,3} + q^2_{1,3}\right)^{-1}, \\
c^2_{1,3} &= \lambda_{1,3}c^2_{D,1} + 1 - \lambda_{1,3}, \quad c^2_{2,3} = (1-d)c^2_{D,2} + d(\lambda_{2,3}c^2_{D,2} + 1 - \lambda_{2,3}), \\
c^2_{D,2} &= 1 + (1-\phi_2^2)(c^2_{A,2} - 1) + \frac{\phi_2^2}{\sqrt{s_2}}(c^2_{S,2} - 1),
\end{aligned} \tag{3.11}$$

where $c^2_{2,3}$ is the scv of the patient flow from queue 2 to queue 3, $c^2_{1,3}$ the scv of the patient flow from queue 1 to queue 3, and $c^2_{D,2}$ is the scv of the departure process at queue 2. Queue 3 has mean service time, $\mathbb{E}[S_3]$, and scv, $c^2_{S,3}$:

$$\begin{aligned}
\mathbb{E}[S_3] = &\quad \frac{\sum_{k=2}^{4}(1-da_k)\gamma_k\mathbb{E}[S_{k,3}]}{\lambda_3} + d\sum_{k=5}^{7}\gamma_k\mathbb{E}[S_{k,3}], \\
c^2_{S,3} = &\quad \frac{\sum_{k=2}^{4}(1-da_k)\gamma_k\mathbb{E}^2[S_{k,3}](c^2_{S,k,3}+1) + \sum_{k=5}^{7}\gamma_k\mathbb{E}^2[S_{k,3}](c^2_{S,k,3}+1)}{\lambda_3\mathbb{E}^2[S_3]} - 1.
\end{aligned} \tag{3.12}$$

Step 5.
We are interested in the waiting times for patients per queue and the load per employee at each queue. The latter is given by the aggregated load derived in step 1, while the mean waiting times are obtained by using the scv and mean service time calculated in step 2. The mean waiting time, $\mathbb{E}[W^q_j]$, is equal for all patient classes.

$$\begin{aligned}
\mathbb{E}[W^q_1] &= \frac{c^2_{A,1} + c^2_{S,1}}{2}\frac{\phi_1}{1-\phi_1}\frac{\mathbb{E}[S_1]}{e_1}, \\
\mathbb{E}[W^q_j] &= \frac{c^2_{A,j} + c^2_{S,j}}{2}\mathbb{E}[W^q_{j(M/M/s)}], \quad \text{where} \\
\mathbb{E}[W^q_{j(M/M/s)}] &= G_j^{-1}\frac{(s_j\phi_j)^{s_j}}{s_j!}\frac{\mathbb{E}[S_j]}{s_j(1-\phi_j)^2}, \\
G_j &= \sum_{n=0}^{s_j-1}\frac{(s_j\phi_j)^n}{n!} + \frac{(s_j\phi_j)^{s_j}}{(1-\phi_j)s_j!} \quad \text{for} \quad j = 2, 3.
\end{aligned} \tag{3.13}$$

Patient LOS for each patient class can now be calculated by adding the mean waiting and LOS of all care queues the patient calls at on his visit to the PAC.

Chapter 4

Designing Cyclic Appointment Schedules

4.1 Introduction

Developing appointment schedules for service facilities that process both scheduled and unscheduled arrivals is challenging, as it requires planning and scheduling on different time scales. A well-designed appointment system comprises an efficient day appointment schedule and provides timely access. This chapter is motivated by challenges faced by hospital outpatient clinics that serve patients on a walk-in basis. Most of these clinics also have a limited number of appointment slots. There are various organizational (e.g., fixed slots for patients in a care pathway, patients with long travel time to the hospital, children) and medical (e.g., local anesthesia or contrast fluid required) reasons to give a patient an appointment. In this chapter, we introduce a method to design appointment schedules for such facilities.

Advantages of a walk-in system are a higher level of accessibility and more freedom for patients to choose the date and time of their hospital visit. Disadvantages are a possible highly variable demand and as a consequence low utilization and high waiting time. The advantage of an appointment system is that workload can be dispersed, while it has the disadvantage of a potentially long access time. Since prolonged access times result in a delay of treatment, deterioration of health condition is a serious risk [140]. Allowing patients to walk in effectively reduces access times to zero, and thus increases quality of care. Additionally, healthcare facilities typically aim to guarantee a certain service level with respect to the access time for patients with an appointment. The challenge in a mixed system is thus to balance access time for appointment patients and waiting time for walk-in patients. To achieve this, we develop a methodology that schedules appointments when the expected walk-in demand is low. To smoothen the system, in periods of high demand part of the walk-in patients is offered an appointment at a later moment. Of course, this is undesirable since it increases access time and may involve an

additional clinic visit. Walk-in demand [10, 47] and demand for appointments requests [201] are often cyclic; therefore, we develop a cyclic appointment schedule. Appointment scheduling has received considerable attention in the literature, as opposed to the development of models that relate access and waiting time [85].

The methodology incorporates unscheduled and scheduled arrivals and maximizes the number of unscheduled patients served on the day of arrival, while satisfying a prespecified access time norm for scheduled patients. We model the unscheduled arrivals with a stochastic non-stationary arrival process and incorporate balking behavior. The scheduled patients have priority, may not show up, and appointment requests are assumed to arrive according to a cyclic pattern. To account for the cyclic arrivals, the appointment schemes we develop are also cyclic, where the cycle is a repeating sequence of days. The cycle length can, for instance, be a week or a month. The cyclic appointment schedule (CAS) specifies a capacity cycle (the maximum number of patients that can be scheduled on each day of the cycle) and a day schedule (the maximum number of patients to be scheduled per time slot on each day). Access time and waiting time are measured on different time scales, since access time is counted between days and waiting time during a day. To facilitate the two time scales, our approach consists of decomposing the appointment planning process and the service process during the day. For both processes we propose an analytical evaluation model. The first model determines the access time for scheduled patients for any given capacity cycle. The second model determines the mean number of unscheduled patients that cannot be seen on the day of arrival. The two models are linked by an iterative algorithm that stops when the CAS is found in which the fraction of unscheduled patients seen on the day of arrival is maximized, given that the restriction on the access time is satisfied. A numerical example of a small problem instance demonstrates the potential of the methodology. In this example complete enumeration is applied to find optimal day schedules. Our future research will be aimed at incorporating heuristics to quickly find (close to) optimal day schedules, so that larger problem sizes can be tackled. Finding an optimal day schedule is not straightforward and a field of research on its own [40, 85].

In many service facilities customers are requested to make an appointment. There is a substantial body of literature focusing on the design of appointment systems. Healthcare is the most prevalent application area and hence also most considered in the literature (see the surveys [40] and [85]). Appointment systems can be regarded as a combination of two distinct queuing systems. The first queuing system concerns customers making an appointment and waiting until the day the appointment takes place. The second queuing system concerns the process of a service session during a particular day. We denote these two queuing processes as the 'access process' and the 'day process'. The remainder of this section provides an overview of the literature relevant for the present work and is structured as follows: (1) appointment scheduling, (2) access time models, and (3) integrating the access process and the day process.

4.1.1 Appointment Scheduling

Appointment scheduling concerns designing blueprints for day-appointment schedules with typical objectives as minimizing customer waiting time, and maximizing resource utilization or minimizing resource idle time. A large part of the literature focuses on scheduling a given number of appointments on a particular day [23, 103, 126, 130]. The extent to which various aspects that impact the performance of an appointment schedule are incorporated varies, such as customer punctuality [124], customers not showing up ('no-shows') [93, 103], lateness of the server at the start of a service session [130], service interruptions [124] and the variance of service duration [93].

Research techniques employed in appointment scheduling can be divided in analytical and simulation-based approaches, of which the latter is most widely applied [40]. In the day process we aim for an analytical approach, namely finite time Markov chain analysis. Related examples with healthcare applications are [23, 103, 126, 153] and [91], although these references do not consider unscheduled customers. Often, a homogeneous customer population is assumed [52]. Some studies however, focus on service systems with various customer types. Differentiation between customer types is identified as a consequence of distinct service requirements [23, 22, 42, 107, 196]. Also, distinct priority levels may be a reason for patient type differentiation. An example can be found in [151], where service slots are pre-marked for various scheduled customer classes. In this chapter, customer type differentiation arises from distinct arrival processes.

The effect of mixed arrival processes is studied in [81, 113] and [174]. Here, scheduled outpatients, unscheduled inpatients and emergency patients are taken into account. Patients without an appointment are either emergency patients who require non-preemptive priority or inpatients available for 'call-in' at any time during the day. These unscheduled patients are assumed to arrive according to an equal arrival rate throughout the day. In our case, we consider walk-in patients without priority who cannot be called in during the day. Moreover, we consider non-stationary arrivals to incorporate the expected peak behavior of walk-in demand. Studies that do incorporate non-priority unscheduled arrivals similar to the unscheduled arrivals in this chapter are [10, 41, 42, 117, 163, 177, 179]; however, in all cases a simulation approach is employed. Also, these studies do not incorporate balking behavior of unscheduled customers.

4.1.2 Access Time Models

As our approach consists of a decomposition, isolated access time models are also of interest. The access process we consider is discrete-time and cyclical in both the arrival and service processes. Various access time models based on continuous-time queuing models are available. Examples are the $M(t)/M/s(t)$ queue [82] and the adapted $M/M/s$ queue that models time-dependent demand [79]. The latter method is also applied to a healthcare problem in [83]. To preserve the discrete-time nature we take as starting

point the generating function approach for slotted queuing models in discrete time [32]. A survey on discrete-time queuing systems is presented in [30]. Models to evaluate the length of hospital waiting lists are introduced in [204], and further studied in for example [76]. In these models homogeneous appointment request arrivals are assumed. In polling models, multiple queues are served by one server in cyclic order (see [181] for an overview). However, cyclic arrival rates and cyclic service capacity have not yet been incorporated in polling models.

4.1.3 Linking the Access and Day Process

Only a few examples jointly consider the access and day process. In [22] and [108], appointment schedules ranging over a horizon of several days are evaluated. The aim is to minimize the patient's waiting and the doctor's idle time, but the patient's access time is not studied in detail. In [162] the authors propose a two time scale model for the ED – Ward patient flow. The fast time scale of the ED is modeled by a continuous time Markov chain, while the slower time scale of the wards is modeled by a discrete time Markov chain. The advanced (or open) access methodology [140] also considers two time scales. With advanced access, a clinic leaves a fraction of appointment slots vacant for patients that request an appointment on the same day or within a couple of days. As many patients as possible are scheduled on the day they make an appointment request. One should determine the optimal ratio between the reserved capacity for long-term and same-day appointments [60]. This principle is slightly adapted in [131], where the demand for short term appointments is distributed over several days, to smooth the daily load of the system. The aim of the advanced access methodology is to minimize access time ("do today's work today"). Note that in an advanced access clinic patients do announce themselves in advance and make a (same-day) appointment, contrary to the type of unscheduled patients we consider, who just show up. Models that study the advanced access methodology usually focus on capacity distribution [60, 160, 161].

Formulating a model to design an appointment schedule considering two time scales is usually done using simulation techniques (e.g., [115]). An analytic approach is presented in [152], where the effect of capacity allocation among competing patient classes on access time targets is studied using techniques from Markov decision modeling and mathematical programming. An approach related to ours, although without the presence of walk-in patients, is given in [53]. The authors consider a service facility, and first develop a vacation queuing system to determine the access time. Subsequently an appointment system is developed that calculates the waiting time at the facility.

This chapter is organized as follows. In Section 4.2, we give an introduction to the methodology and provide a formal problem description. Sections 4.3-4.5 present the access and day process evaluation models and the algorithm. Section 4.6 describes the numerical example, followed by the discussion in Section 4.7.

4.2 Formal Problem Description

This section defines all modeling assumptions, defines the CAS, formally states the research goal and gives an overview of the proposed approach. Since our approach is generically applicable, we also present the methodology in the generic terms: a facility that serves scheduled and unscheduled patients.

4.2.1 Assumptions

A facility consisting of R resources is operational during T time slots of length h, during each day in a cycle of D days. Two types of patients have to be served: scheduled and unscheduled patients. Service takes one time slot. Scheduled patients are given a specific date and time immediately when an appointment is requested. In addition, when the facility is temporarily congested, unscheduled patients are also offered an appointment: if the service of an unscheduled patient cannot start within g time slots after arrival, the patient will leave the facility and an appointment will be planned for another day. We will refer to such patients as deferred unscheduled patients, or just deferred patients. The first available appointment slot for scheduled and deferred patients is always the next day at the earliest. All appointments, both scheduled patients and deferred unscheduled patients, are scheduled according to a First Come First Served (FCFS) principle.

We assume a non-stationary Poisson process for the arrivals of appointment requests, with $\lambda^1, \ldots, \lambda^D$ the arrival rates for different days in the cycle. Next, during each day in the cycle, we assume a non-stationary Poisson arrival process for unscheduled patient arrivals, with slot-dependent arrival rates: χ_t^d for day $d = 1, \ldots, D$ and time slot $t = 1, \ldots, T$. Table 4.1 summarizes the notation introduced in this section.

4.2.2 Cyclic Appointment Schedule

To balance the non-stationarity at both the daily and cyclic (i.e. weekly, biweekly or monthly) level, we aim to design an appointment schedule that is cyclic. We introduce the CAS, $C = (C^1, \ldots, C^D)$, with $C^d = (c_1^d, \ldots, c_T^d)$, where c_t^d specifies the maximum number of patients that may be scheduled in slot t on day d. To find an adequate appointment schedule, we propose a decomposition. First, we introduce the concept of a capacity cycle, $K = (k^1, \ldots, k^D)$, where k^d prescribes the maximum number of patients to schedule for day d in the cycle. Second, given the capacity cycle K, the day plan is specified. In order to match the capacity cycle K, the day plan C^dshould be such that $k^d = \sum_{t=1}^{T} c_t^d$.

4.2.3 Goal

An effective strategy balances (1) the opportunities for unscheduled patients to be served on the same day without long waiting time and (2) for scheduled patients to be served within an acceptable access time. To this end, we define the best policy as the cyclic appointment schedule in which the mean fraction of unscheduled patients served on the day of arrival, F, is maximized, while for scheduled patients the access time service level, $S(y)$, defined as the percentage of patients that is served within y days, is above a pre-specified norm $S^{norm}(y)$. The value of the vector $(y, S^{norm}(y))$ is chosen by the facility.

Table 4.1: Notation introduced in Section 4.2

Symbol	Description
R	Number of resources
T	Number of time slots during a day
t	Time slot index, $t = 1, \ldots, T$
h	Length of a time slot
D	Cycle length in days
d	Day index, $d = 1, \ldots, D$
g	Patience of an unscheduled patient, given in the number of slots a patient is willing to wait
λ^d	Initial appointment request arrival rate on day d
χ_t^d	Unscheduled patient arrival rate on day d during time interval $(t-1, t]$
c_t^d	Maximum number of appointments to schedule in slot t on day d
C^d	Appointment schedule on day d, $C^d = (c_1^d, \ldots, c_T^d)$
C	Cyclic appointment schedule, $C = (C^1, \ldots, C^D)$
k^d	Maximum number of appointments to schedule on day d
K	Capacity cycle, $K = (k^1, \ldots, k^D)$
F	$\mathbb{E}$[Fraction of unscheduled patients to serve at day of arrival during one cycle]
$S(y)$	Access time service level: fraction of patients with access time not greater than y
$(y, S^{norm}(y))$	Access time service level requirement: fraction of patients with access time not greater than y is at least $S(y)$
ϕ^d	Distribution of the number of deferred patients on day d
γ^d	Total appointment request arrival distribution on day d
ν^d	Expected number of deferred patients on day d

4.2.4 Approach

The best CAS is determined by employing an iterative algorithm that effectively utilizes our decomposition of the CAS in the capacity cycle and the day plan. In each iteration, first, capacity cycles are generated with at most $R{\cdot}T$ appointments per day, for which the access time service level norm will be satisfied. All patients requesting an appointment are taken into account –thus both scheduled patients and deferred unscheduled patients. We derive the distribution of the number of deferred unscheduled patients ϕ^d, so that the distribution of the total number of appointment requests on day d is the sum of a Poisson distribution with parameter λ^d and the distribution ϕ^d. To assess whether specific capacity cycles with arrival distribution γ^d satisfy the access time norm, $S^{norm}(y)$, a cyclic slotted queuing model is proposed (Model I, presented in Section 4.3). Next, for each capacity cycle generated in the first step, the best day schedule is determined. Given the queue length probabilities resulting from Model I and the unscheduled patient arrival rates, χ_t^d, for each day the k^d appointments are distributed over the T time slots, such that the number of deferred unscheduled patients is minimized. To achieve this, a Markov reward model is presented (Model II, Section 4.4), which is used to calculate the performance of a specific day schedule. Then, the capacity cycle that achieves the lowest mean number of deferred unscheduled patients over the entire cycle is chosen as the best cycle. If the mean numbers of deferred unscheduled patients ν^d, did not change significantly since the last iteration, the algorithm stops. If not, the entire process is repeated. A detailed description of the algorithm is given in Section 4.5.

4.3 Model I: Access Time Evaluation

In this section, a cyclic slotted queuing model is presented that allows for an evaluation of the access time for scheduled patients, given an arbitrary capacity cycle. To this purpose, we focus on the backlog, B^d, at the start of each day d. We define the backlog as the number of patients for which a request for an appointment has already been made, while the appointment itself has not yet taken place. We formulate a Lindley type equation to characterize the backlog, and use a probability-generating function approach to derive expressions for the distribution of the backlog at the start of each day in the cycle. From the backlog distribution, we will derive the access time distribution. A summary of the notation used in this section is given in Table 4.2.

4.3.1 Lindley Type Equation

Consider day d. During the day, a maximum number of patients, k^d, is served, and a number of new patients arrives, A^d. At the start of day d, there is a backlog, B^d. Since it is not possible to make an appointment on the day of arrival itself, the backlog at the start of the next day equals the backlog on day d minus the number of patients served

Table 4.2: Notation introduced in Section 4.3

Symbol	Description
B^d	Backlog at start of day d
$P_{B^d}(z)$	Generating function of B^d
A^d	Number of appointment requests arriving at day d
a_j^d	Appointment request arrival probabilities, $\mathbb{P}\left(A^d = j\right)$
$P_{A^d}(z)$	Generating function of A^d
π_j^d	Stationary backlog probabilities, $\mathbb{P}\left(B^d = j\right)$
k	Total number of available appointment slots in a capacity cycle, $k = \sum_d k^d$
$\mathbb{E}[W^d]$	$\mathbb{E}$[Access time for an appointment request arriving at day d]
$\mathbb{E}[W]$	$\mathbb{E}$[Access time for an arbitrary appointment request]

on day d plus the number of patients that arrived on day d. This can be formalized in the following Lindley type equation:

$$B^{d+1} \;=\; (B^d - k^d)^+ + A^d, \tag{4.1}$$

where $(x)^+ = x$ if $x > 0$, and 0 otherwise.

4.3.2 Generating-Function Approach

Using an approach based on generating functions [32], we derive expressions for the distribution of the backlog at the start of each day in the cycle. The transition probabilities for going from state $B^d = i$ to state $B^{d+1} = i'$ are given by:

$$\mathbb{P}\left(B^{d+1} = i' | B^d = i\right) = \begin{cases} \mathbb{P}\left(A^d = i'\right) & \text{if} \quad i - k^d \leq 0 \\ \mathbb{P}\left(A^d = i' - i + k^d\right) & \text{if} \quad i - k^d > 0. \end{cases} \tag{4.2}$$

Let π_j^d denote the stationary probability that at the start of day d, the backlog equals j patients. Furthermore, let a_j^d denote the probability that $A^d = j$. Note that the underlying probability distribution does not necessarily has to be Poisson. The stationary probabilities can be computed recursively, under the condition that the capacity for scheduled patients is larger than the average demand, i.e. $\sum_d \mathbb{E}[A^d] < \sum_d k^d$, since otherwise we would be dealing with an unstable system. For $d = 1, \ldots, D, j \geq 0$ we obtain:

$$\pi_j^{d+1} \;=\; a_j^d \sum_{i=0}^{k^d-1} \pi_i^d + \sum_{q=0}^{j} a_{j-q}^d \pi_{k^d+q}^d \,. \tag{4.3}$$

We multiply both sides of (4.3) with z^j, where $|z| \leq 1$, and z^j denotes z raised to the power j, as opposed to index d in π_j^d, a_j^d and k^d. The summation of both sides of the

resulting equation over j yields the probability-generating function for π^{d+1}:

$$\sum_{j=0}^{\infty} \pi_j^{d+1} z^j = \sum_{j=0}^{\infty} \left(a_j^d \sum_{i=0}^{k^d-1} \pi_i^d + \sum_{q=0}^{j} a_{j-q}^d \pi_{k^d+q}^d \right) z^j. \tag{4.4}$$

From this we obtain:

$$P_{B^{d+1}}(z) = \sum_{j=0}^{\infty} \pi_j^{d+1} z^j = P_{A^d(z)} z^{-k^d} P_{B^d(z)} + P_{A^d(z)} z^{-k^d} \sum_{i=0}^{k^d-1} \pi_i^d \left(z^{k^d} - z^i \right). \tag{4.5}$$

Rearranging terms and changing the order of summation leads to the probability generating function of B^d, $P_{B^d}(z)$:

$$P_{B^d}(z) = \frac{\sum_{i=1}^{D} \sum_{q=0}^{k^{d+D-i}-1} (z^{k^{d+D-i}} - z^q) \pi_q^{d+D-i} \left[\prod_{s=d}^{d+D-i-1} z^{k^s} \prod_{r=0}^{i-1} P_{A^{d+D-r-1}}(z) \right]}{\prod_{g=1}^{D} z^{k^g} - \prod_{h=1}^{D} P_{A^h}(z)},$$

where, since we consider days in a repeating cycle, we define:

$$d := \begin{cases} D, & d \bmod D = 0 \\ d \bmod D, & \text{otherwise.} \end{cases} \tag{4.6}$$

The generating functions uniquely determine the stationary probabilities $\pi_j^d, j = 0, \ldots, k^d - 1$. To calculate these probabilities, we build upon the approach given in [2]. Define k as the total number of available appointment slots in a capacity cycle, i.e. $k = \sum_{d=1}^{D} k^d$. Then, the denominator of $P_{B^d}(z)$ has $k-1$ zeros inside the unit disk; this can be shown by using Rouché's theorem [110]. All generating functions, including $P_{B^d}(z)$, are bounded for $|z| \leq 1$, and therefore the zeros of the denominator are also zeros of the numerator [32]. Thus we obtain $k-1$ equations, and use $P_{B^d}(1) = 1$ to secure the last equation. The $k-1$ zeros of the denominator of $P_{B^d}(z)$ can be found by solving:

$$\prod_{g=1}^{D} z^{k^g} - \prod_{h=1}^{D} P_{A^h}(z) = 0. \tag{4.7}$$

The solutions of (4.7) also represent zeros of the numerator. Together with the normalizing equation $P_{B^d}(1) = 1$, $P_{B^d}(z)$ is completely defined for $d = 1, \ldots, D$. Note that now only the backlog probabilities for $j = 0, \ldots, k^d - 1$, have been derived. The remaining backlog probabilities are calculated directly using (4.3).

4.3.3 Performance Measures

The access time distribution can be directly derived from the backlog probabilities, since appointment requests are served according to the FCFS principle. The FCFS service order and the impossibility of making an appointment request for the day of arrival results in an access time of at least one day. Several performance measures can be derived. Of particular interest are the probability distribution of the access time, the mean access time and the access time service level.

The Probability Distribution of the Access Time

First we derive the conditional access time probability that the access time for a client arriving at day d exceeds y days, given that the backlog at the start of day d equals b clients. As argued, for $y = 0$, we have that

$$\mathbb{P}[W^d > y | B^d = b] = 1 \quad \forall b. \tag{4.8}$$

For $y > 0$, we have that

$$\mathbb{P}[W^d > y | B^d = b] = \begin{cases} 1 & \text{if} \quad b \geq \sum_{i=0}^{y} k^{d+i} \\ \frac{\sum_{j=s+1}^{\infty} (j-s) \cdot \mathbb{P}[A^d = j]}{\mathbb{E}[A^d]} & \text{otherwise,} \end{cases} \tag{4.9}$$

where s represents the number of patients arrived on day d that will be served within y days:

$$s = \min \left\{ \sum_{i=1}^{y} k^{d+i}, \sum_{i=0}^{y} k^{d+i} - b \right\}. \tag{4.10}$$

We can explain formula (4.9) as follows. First, when the backlog b outnumbers the available capacity in y days, the conditional probability that the access time exceeds y days equals 1. Otherwise, all arrivals beyond the number s will wait for more than y days. There are $j - s$ such arrivals. Then, the probability that the access time for a client arriving at day d exceeds y days, equals

$$\mathbb{P}[W^d > y] = \sum_{b=0}^{\infty} \mathbb{P}[W^d > y | B^d = b] \cdot \mathbb{P}[B^d = b]. \tag{4.11}$$

The Expected Access Time

Analogously, the mean access time for an appointment request that arrives on day d is computed with:

$$\mathbb{E}[W^d|B^d = b] \;=\; \sum_{y=0}^{\infty} \mathbb{P}[W^d > y|B^d = b] \tag{4.12}$$

and thus

$$\mathbb{E}[W^d] \;=\; \sum_{b=0}^{\infty} \mathbb{E}[W^d|B^d = b] \cdot \mathbb{P}[B^d = b] \tag{4.13}$$

and

$$\mathbb{E}[W] = \sum_{d=1}^{D} \mathbb{E}[W^d] \frac{\mathbb{E}[A^d]}{\sum\limits_{q=1}^{D} \mathbb{E}[A^q]}. \tag{4.14}$$

The Access Time Service Level

Using the access time probability distribution, we determine the fraction of scheduled patients for which the access time does not exceed y. We define this as follows:

$$S(y) = \sum_{d=1}^{D} \left(1 - \mathbb{P}[W^d > y]\right) \frac{\mathbb{E}[A^d]}{\sum\limits_{q=1}^{D} \mathbb{E}[A^q]}. \tag{4.15}$$

4.4 Model II: Day Process Evaluation

In this section, we present a model to evaluate the performance of a single day in the CAS. Recall that the CAS consists of a capacity cycle , $K = (k^1, \ldots, k^D)$, that prescribes the maximum number of patients that can be scheduled for day d. Using model I, we were able to evaluate the access time performance of a given capacity cycle. Below, we evaluate the day process of a given appointment schedule, by formulating a Markov reward process. Note that although the day appointment schedule, C^d, is open for scheduling appointments, there may be less backlog than the $k^d = \sum_t c_t^d$ available appointment slots. Therefore, we introduce the notation $\widetilde{C}^d$ to represent the realized day planning, which is the schedule we evaluate. Now, $\widetilde{C}^d = (\widetilde{c}_1^d, \ldots, \widetilde{c}_T^d)$ expresses the actually utilized appointment slots. Since appointments are planned on a FCFS basis, the

realized appointment day schedule $\widetilde{C}^d$ will always be a truncated version of day schedule C^d. Of course, unoccupied appointment slots can be used for unscheduled patients. Since we will consider the day performance on a day-by-day basis, in the remainder of this section we drop the superscript d for notational convenience. Table 4.3 provides a summary of the notation introduced in this section.

Table 4.3: Notation introduced in Section 4.4

Symbol	Description
$\widetilde{C}$	Realized schedule under CAS C, $\widetilde{C} = (\widetilde{C}^1, \ldots, \widetilde{C}^D), \widetilde{C}^d = \left(\widetilde{c}_1^d, \ldots, \widetilde{c}_T^d\right)$
q	$\mathbb{P}$(No-show of a scheduled patient)
e_t	Number of slots available for unscheduled patients in the next g intervals after time t
$p_t^s(s)$	$\mathbb{P}$(Number of scheduled patients arriving at the start of slot $t = s$)
$p_t^u(u)$	$\mathbb{P}$ (Number of unscheduled patients arriving during interval $(t-1, t] = u$)
$\mathbb{P}\left[(s,u)_{t+1} \mid (k,l)_t\right]$	Transition probability from state (t,k,l) to state $(t+1,s,u)$
$Q_t(s,u)$	$\mathbb{P}$ (Number of scheduled, unscheduled patients waiting at start of slot $t = s, u$)
ν_t	$\mathbb{E}$[Number of deferred patients in time interval $(0,t]$]
ν	$\mathbb{E}$[Total number of deferred patients]
ϕ_t	Distribution of the number of deferred patients in time interval $(t-1,t]$
ϕ	Distribution of the total number of deferred patients

4.4.1 Assumptions

For clarity of presentation, some of the assumptions introduced in Section 4.2 are repeated. During one day the facility of R resources is operational during T intervals of length h. Two types of patients have to be served: scheduled and unscheduled patients. Service always takes one time slot of length h. At the beginning of each time slot, a service can start. If there are both scheduled and unscheduled patients, scheduled patients are given priority. Overtime is not allowed. Scheduled patients arrive on time, according to the schedule $\widetilde{C}$. In addition, we allow for no-shows, that is, the probability that a scheduled patient actually arrives at the facility equals $1 - q$, so that q represents the probability that a patient does not show up.

Unscheduled patients arrive at the facility according to an inhomogeneous Poisson process with slot-dependent arrival rate χ_t. If the service of an unscheduled patient cannot start within g time slots after arriving, the patient will leave the facility and an appointment will be planned for another day. We assume that the facility has no pre-knowledge about potential no-shows. Therefore, an unscheduled patient arriving during interval $(t-1, t]$ will stay if and only if the number of unscheduled patients already waiting is

strictly smaller than the minimum number of service slots during the upcoming g intervals that are not utilized by scheduled patients. The number of time slots anticipated to be available for unscheduled patients during the upcoming g intervals is denoted by e_t:

$$e_t = \sum_{j=t}^{\min\{t+g-1,T\}} (R - \widetilde{c}_j). \tag{4.16}$$

4.4.2 States and Transition Probabilities

The state of the system is denoted by the tuple (t, s, u), which specifies that at the beginning of time slot t, s scheduled and u unscheduled patients are present. Let $p_t^s(s)$ denote the probability that s scheduled patients arrive at the beginning of time slot t. Since each no-show is assumed to occur independently, these probabilities are calculated as follows:

$$p_t^s(s) = \begin{cases} \binom{\widetilde{c}_t}{s}(1-q)^s(q)^{\widetilde{c}_t - s}, & 0 \leq s \leq \widetilde{c}_t \\ 0, & s > \widetilde{c}_t. \end{cases} \tag{4.17}$$

Let $p_t^u(u)$ denote the probability that u unscheduled patients arrive during time interval $(t-1, t]$. As specified, $p_t^u(u)$ is Poisson distributed with slot dependent parameter χ_t. Note that χ_1 represents the arrival rate of unscheduled patients that arrive before the opening time of the facility. Furthermore, note that any distribution function p_t^u can be used in the day process evaluation model. Therefore, for model I the assumption of a Poisson arrival process is not strictly required.

Let $\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right]$ denote the transition probability of jumping from state (t, v, w) to $(t+1, s, u)$. Below we specify these transition probabilities for all possible events. In Figure 4.1, the state space for an arbitrary time slot t is displayed in which the seven different possible events (a)-(g) are indicated. The events can be separated in three groups: first, cases (a)-(c) in which no scheduled patient is served ($v = 0$), second, cases (d) and (e) in which both scheduled and unscheduled patients are served ($v < R$), and third, cases (f) and (g) in which only scheduled patients are served ($v \geq R$). In the expressions below, $\mathbb{1}_A$ represents the indicator function; $\mathbb{1}_A = 1$ if condition A is satisfied, and 0 otherwise.

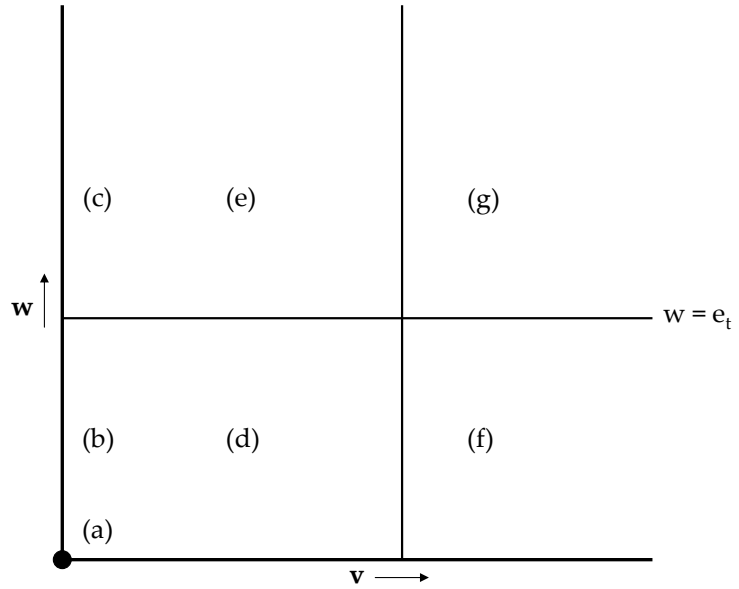
Case (a). $v = w = 0$; no patient served:

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p_{t+1}^s(s)p_{t+1}^u(u).$$

Case (b). $v = 0, 0 < w \leq e_t$; unscheduled patient(s) served:

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p_{t+1}^s(s)p_{t+1}^u(u - w + \min\{R, w\})\mathbb{1}_{(u \geq w - \min\{R,w\})}.$$

Figure 4.1: Day process state space and events



Case (c). $v = 0, w > e_t$; unscheduled patient(s) served, unscheduled patient(s) deferred:

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p^s_{t+1}(s)p^u_{t+1}(u - e_t + R)\mathbb{1}_{(u \geq e_t - R)}.$$

Case (d). $v < R, w \leq e_t$; scheduled and unscheduled patient(s):

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p^s_{t+1}(s)p^u_{t+1}(u - w + \min\{(R-v), w\})\mathbb{1}_{(u \geq w - \min\{(R-v),w\})}.$$

Case (e). $v < R, w > e_t$; scheduled and unscheduled patient(s) served, unscheduled patient(s) deferred:

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p^s_{t+1}(s)p^u_{t+1}(u - e_t + R - v)\mathbb{1}_{(u \geq e_t - R + v)}.$$

Case (f). $v \geq R, w \leq e_t$; scheduled patient(s) served:

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p^s_{t+1}(s - v + R)p^u_{t+1}(u - w)\mathbb{1}_{(s \geq v - R)}\mathbb{1}_{(u \geq w)}.$$

Case (g). $v \geq R, w > e_t$; scheduled patient(s) served, unscheduled patient(s) deferred:

$$\mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right] = p^s_{t+1}(s - v + R)p^u_{t+1}(u - e_t)\mathbb{1}_{(s \geq v - R)}\mathbb{1}_{(u \geq e_t)}.$$

4.4.3 Performance Measures

Let $Q_t(s,u)$ denote the probability that at the start of slot t there are s scheduled and u unscheduled patients present. $Q_t(s,u)$ can be calculated as follows:

$$Q_1(s,u) = p_1^s(s) \cdot p_1^u(u). \tag{4.18}$$

For $t = 2, ..., T$:

$$Q_{t+1}(s,u) = \sum_v \sum_w Q_t(v,w) \mathbb{P}\left[(s,u)_{t+1} \mid (v,w)_t\right]. \tag{4.19}$$

The mean number of deferred patients, $\nu = \nu_T$, is calculated accordingly:

$$\nu_1 = \sum_{s=0}^{\infty} \sum_{u=e_1+1}^{\infty} (u - e_1) \cdot Q_1(s,u). \tag{4.20}$$

For $t = 2, ..., T$:

$$\nu_t = \nu_{t-1} + \sum_{s=0}^{\infty} \sum_{u=e_t+1}^{\infty} (u - e_t) \cdot Q_t(s,u). \tag{4.21}$$

The distribution of the number of deferred patients, ϕ, can be calculated as follows. For $t = 1, \ldots, T$:

$$\phi_t(j) = \begin{cases} \sum_{s=0}^{\infty} \sum_{u=0}^{e_t} Q_t(s,u), & j = 0 \\ \sum_{s=0}^{\infty} Q_t(s, e_t + j), & j > 0, \end{cases} \tag{4.22}$$

and

$$\phi = \phi_1 * \ldots * \phi_T, \tag{4.23}$$

where $*$ denotes the discrete convolution function.

4.5 Algorithm: Finding a Balance

The algorithm presented in this section links the access and the day process. Models I and II are used iteratively to maximize the number of unscheduled patients served during the day of arrival, given the pre-specified access time service level norm. As mentioned before, unscheduled patients that cannot be served within g time slots receive an appointment. The algorithm determines the optimal size of this group of deferred patients by gradually increasing its size during each iteration. Table 4.4 summarizes the notation presented in this section.

In the first iteration, the mean number of deferred patients is set to zero. Then, the best scheduling cycle (using Model I) with accompanying appointment schedule (using Model II) is determined, given the appointment request arrival processes with rate λ^d and that of unscheduled patient arrivals with rate χ_t^d. The distribution of the number of deferred patients on day d in iteration n is denoted by $\phi^d(n)$, and the mean by $\nu^d(n)$. To account for the patients that were deferred, the distribution of appointment request arrivals, $\gamma^d(n)$, is in the next iteration set to

$$\gamma^d(n) = P(\lambda^d) * \phi^d(n-1), \tag{4.24}$$

where $P(\lambda^d)$ denotes the Poisson distribution with parameter λ^d. As such, the appointment requests generated by deferred patients are taken into account on the day of occurrence in the previous iteration. Then, a new best policy is calculated. As more appointment slots are reserved, this may result in more deferred patients than in the previous iteration. This iterative procedure is repeated until on each day in the cycle, a balance is found between the anticipated extra demand for appointments from deferred unsched-

Table 4.4: Notation introduced in section 4.5

Symbol	Description
n	Iteration counter
$\phi^d(n)$	Distribution of the number of deferred patients on day d in iteration n
$\nu^d(n)$	Expected number of deferred patients on day d in iteration n
$\gamma^d(n)$	Total appointment request arrival distribution on day d in iteration n
ϵ	Precision of the algorithm's stop criterion
$K(n_f)$	Capacity cycle option f consisting of $(k^1(n_f), \ldots, k^D(n_f))$ in iteration n
$C(n_f)$	The best CAS given capacity cycle $K(n_f)$
$\bar{\pi}_j^d(n_f)$	The probability that in iteration n under capacity cycle $K(n_f)$ j appointment reservations are utilized by appointments on day d
$\nu_C^*(n_f)$	$\mathbb{E}$[Total number of deferred patients in iteration n under capacity cycle $K(n_f)$ and CAS C]
$\nu_{C^d\|j}^d(n_f)$	$\mathbb{E}$ [Number of deferred patients on day d in iteration n under capacity cycle $K(n_f)$ and CAS C when j appointment slots are utilized]

uled patients (which was $\nu^d(n-1)$) and the realized deferred unscheduled patients (which is $\nu^d(n)$); expressed formally, the algorithm terminates if, for some small ϵ,

$$|\nu^d(n) - \nu^d(n-1)| < \epsilon. \tag{4.25}$$

It is important to note that we aim for balance on a day-by-day basis. Balance just on a cycle basis ($|\sum_d \nu^d(n) - \nu^d(n-1)| < \epsilon$) is not sufficient, since only in the case that $\left|\nu^d(n) - \nu^d(n-1)\right| < \epsilon$, $d = 1, \ldots, D$, it is guaranteed that the appointment requests of deferred patients are as anticipated. Only then we can assure that in the access time calculations, we account for the deferred patients on the day they occur, since the access time calculations that use $\phi^d(n-1)$, based upon which the capacity cycle is designed, are still valid for $\phi^d(n)$ in this case.

We now specify the procedure used to find an optimal policy within each iteration. First, by applying Model I, all capacity cycles fulfilling the specified access time service level norm are generated. So, given $\gamma^d(n)$, all capacity cycles $K = (k^1, \ldots, k^D)$ satisfying $S^{norm}(y)$ are generated. Suppose that m different capacity cycles satisfy the norm, then denote these options for iteration n by $K(n_f) = (k^1(n_f), \ldots, k^D(n_f))$, $f = 1, \ldots, m$. From these options, the best capacity cycle is selected, which is the capacity cycle that minimizes the mean number of deferred patients. To do this, for each scheduling cycle option $K(n_f)$, the best CAS $C(n_f)$ is determined. The best CAS's are determined by applying Model II as follows.

First, observe that although in a capacity cycle $K(n_f)$ there are $k^d(n_f)$ appointment slots reserved on day d, not all of these reserved slots are necessarily utilized by scheduled patients. Since appointments are planned according to the FCFS principle, we know from Model I the queue length probability vectors, $\pi^d(n_f)$, which also give the probabilities of utilizing the first j out of the $k^d(n_f)$ reservations under capacity cycle $K(n_f)$. Let us denote these probabilities by $\bar{\pi}^d_j(n_f)$:

$$\bar{\pi}^d_j(n_f) = \begin{cases} \pi^d_j(n_f), & j = 0, \ldots, k^d(n_f) - 1 \\ \sum_{q=k^d(n_f)}^{\infty} \pi^d_q(n_f), & j = k^d(n_f) \end{cases}. \tag{4.26}$$

By evaluating each day appointment schedule for $d = 1, \ldots, D$, $f = 1, \ldots, m$ and $j = 0, \ldots, k^d(n_f)$, the best CAS is determined for each capacity cycle $K(n_f)$, so by complete enumeration. Denote the mean total number of deferred patients in cycle $K(n_f)$ under appointment schedule C by $\nu_C(n_f)$. With $\nu^*(n_f)$ defined as the mean total number of deferred patients in cycle $K(n_f)$, under the best CAS the best cyclic appointment schedules are those that minimize:

$$\nu^*(n_f) = \min_C \nu_C(n_f) = \min_C \sum_{d=1}^{D} \sum_{j=0}^{k^d(n_f)} \bar{\pi}^d_j(n_f)\, \nu^d_{C^d|j}(n_f), \tag{4.27}$$

where $\nu^d_{C^d|j}(n_f)$ denotes the mean number of deferred patients on day d under capacity cycle $K(n_f)$ and cyclic appointment schedule C, if j appointment slots are utilized by scheduled patients. Note that $C^d|j$ is a truncated version of C^d, in exactly the same way that $\widetilde{C}^d$ was defined in Section 4.4. Now, the final step is to select the capacity cycle, $K(n_f)$, and accompanying CAS, which is the CAS with the lowest mean number of deferred patients, namely:

$$\nu^*(n) = \min_f \nu^*(n_f), \quad f^*(n) = arg\min_f \nu^*(n_f), \quad C^*(n) = arg\min_C \nu_C(n_{f^*}). \tag{4.28}$$

Figure 4.2 displays the complete algorithm in pseudo code.

Figure 4.2: The algorithm

Step 1: *specify input*	*Specify:* $R, T, D, g, q, S^{norm}(y), \epsilon$; $\forall d : \lambda^d$; $\forall d, t : \chi^d_t$.
Step 2: *initialize algorithm*	$n := 1$; $\forall d : \nu^d(1) := 0, \gamma^d(1) := P(\lambda^d)$.
Step 3: *determine feasible cycles*	*Given* $\gamma^d(n)$, *determine all* $K(n_f), f = 1, \ldots, m,$ *such that* $S(y) \geq S^{norm}(y)$. $\forall f, d$: *store* $\pi^d(n_f)$.
Step 4: *choose best cycle*	*Determine* $\nu^*(n)$, $f^*(n)$ *and* C^*.
Step 5: *assess current solution*	*If* $\forall d : \lvert\nu^d(n) - \nu^d(n-1)\rvert < \epsilon$, *then stop, else proceed to step 6.*
Step 6: *adjust deferrals*	$\forall d : \nu^d(n+1) := \nu^d(n), \phi^d(n+1) := \phi^d(n)$, $\gamma^d(n+1) := P(\lambda^d) + \phi^d(n+1)$; $n := n + 1$ *and return to step 3.*

Convergence

For the system to be stable we require that $\sum_d \lambda^d + \sum_d \sum_t \chi^d_t < R \cdot T$, so that total demand does not exceed capacity. In addition, we would like to determine the conditions under which the algorithm will converge. Therefore, first observe that since the unscheduled patient arrival rate χ^d_t is fixed and the first iteration starts with no deferred patients, i.e. $\nu^d(0) = 0$, in each iteration it is not possible to choose the CAS such that $\sum_d \nu^d(n) < \sum_d \nu^d(n-1)$. The total mean number of deferred patients $\sum_d \nu^d(n)$ is thus monotonically non-decreasing. Also, if the access time norm $S^{norm}(y)$ is set such that it can be satisfied if all patients are planned, we ensure that in each iteration it is possible to find feasible capacity cycles, i.e. capacity cycles for which $S(y) \geq S^{norm}(y)$. However,

convergence of the algorithm is not assured. Although not likely for practical instances, it cannot be guaranteed that the algorithm does not run into the situation that it keeps jumping between points for which the total mean number of deferred patients does not change, but without day-by-day balance, i.e. $\left|\sum_d \nu^d(n) - \nu^d(n-1)\right| < \epsilon$, and not $|\nu^d(n) - \nu^d(n-1)| < \epsilon$, for all d. If such a case occurs, an additional rule to act as a tie-breaker is required. We extensively tested the algorithm by evaluating fifteen different instances (see Section 4.6).

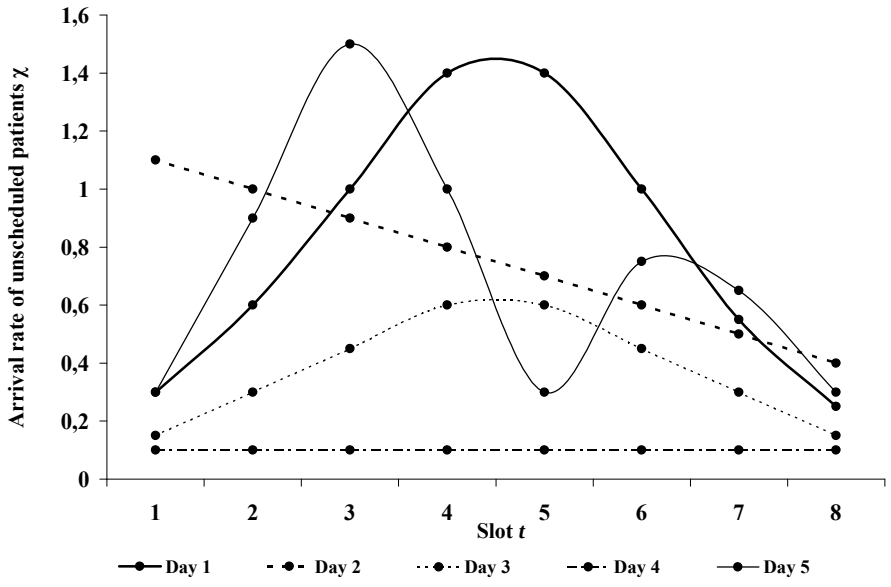
4.6 Numerical Experiments

We tested the algorithm on fifteen scenarios, each with different characteristics. To illustrate the methodology, we present in this section the results of one of the numerical experiments.

4.6.1 Input Parameters

We consider a facility with one resource, and a cycle length of $D = 5$ days, where each day consists of $T = 8$ slots. The initial demand per day for appointment requests is given by $(\lambda^1, \ldots, \lambda^5) = (5, 0, 2, 0, 7)$. The arrival rates of unscheduled patients χ_t^d are given in Table 4.5. These arrival rates are chosen such that different days in the cycle represent different unscheduled arrival patterns, as also illustrated by Figure 4.3.

Figure 4.3: Graphical representation of the appointment request arrival rates per slot per day



The access time service level norm is set such that 95% of the patients that are eventually scheduled are served within two cycles or less, $(y, S^{norm}(y)) = (10, 0.95)$. Furthermore, we assume that unscheduled patients are willing to wait for a maximum of two time slots, i.e. $g = 2$, and for computational convenience we assume that the number of deferred patients on day d, ϕ^d, is Poisson distributed. For simplicity, we also assume that all scheduled patients show up, i.e. $q = 0$. The stop criterion of the algorithm applies the threshold $\epsilon = 0.0001$. Table 4.6 provides an overview of the input parameters. Note that the total mean demand for scheduled patients per cycle is 14, and the total mean demand for unscheduled patients per cycle is 22. Since there are $D \cdot T = 40$ time slots available within a cycle, the utilization of the system is 90%.

Table 4.5: Unscheduled patient arrival rates per slot per day

χ_t^d	t								
d	1	2	3	4	5	6	7	8	Total
1	0.30	0.60	1.00	1.40	1.40	1.00	0.55	0.25	6.50
2	1.10	1.00	0.90	0.80	0.70	0.60	0.50	0.40	6.00
3	0.15	0.30	0.45	0.60	0.60	0.45	0.30	0.15	3.00
4	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.80
5	0.30	0.90	1.50	1.00	0.30	0.75	0.65	0.30	5.70

Table 4.6: Overview of the input parameters

Parameter	Description	Value
D	Cycle length	5
T	Number of time slots	8
$\lambda^1, \ldots, \lambda^5$	Appointment request arrival rates	$5, 0, 2, 0, 7$
$(y, S^{norm}(y))$	Service level norm	$(10, 0.95)$
g	Patience of unscheduled patients	2
q	No-show probability	0
ϵ	Algorithm precision	0.0001

4.6.2 Execution of the Algorithm

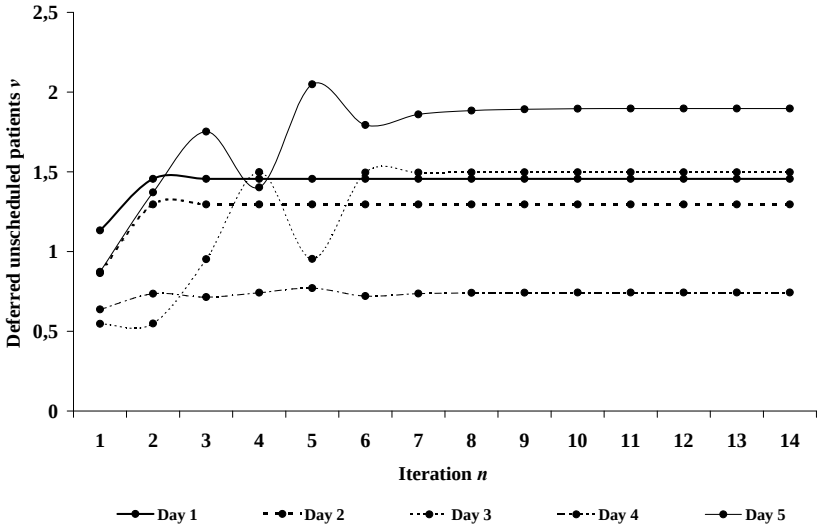
The algorithm was executed and the results obtained after each iteration are displayed in Table 4.7. In the first iteration the number of deferred unscheduled patients is positive on each day of the cycle, $\nu^d(1) > 0$. The total number of deferred patients is $\sum_d \nu^d(1) = 4.055$. Therefore, the deferred patients are added to the scheduled arrival stream and a new iteration is started. This procedure is repeated until after iteration 14, balance is obtained for each day, i.e. $|\nu^d(n) - \nu^d(n-1)| < \epsilon$. In Figure 4.4 and 4.5 we see that the total number of deferred patients is monotonically non-decreasing, while deferrals on the day level are both increasing and decreasing. The fluctuations are substantial in the first iterations and the system stabilizes already after six iterations.

This behavior is also reflected by the dynamics of the capacity cycles. The total number of reserved slots for appointment slots develops as follows: $(16, 19, 21, 21, 21, 22, \ldots, 22)$. Again, although the total number of reserved slots $\sum_d k^d$ is monotonically non-decreasing, for a specific day k^d may also decrease. For example, the capacity cycles of iterations 3–5 all have a total capacity of 21, but the capacity cycle obtained in the third iteration is changed in iteration 4 so that one appointment is shifted from day 5 to day 3. This change is reversed in iteration 5. The final capacity cycle is already obtained in iteration 6. The only purpose of iteration 7–14 is to obtain the desired balance in the daily deferrals. Note that this is a direct result of the magnitude of ϵ. If ϵ had been set larger, the algorithm would have stopped earlier.

Table 4.7: Results per iteration step of the algorithm

Iteration	Day	Tot. app. req. rate	Deferral rate		Difference	Capacity cycle	CAS
n	d	γ^d	$\nu^d(n-1)$	$\nu^d(n)$	$\lvert\nu^d(n-1)-\nu^d(n-1)\rvert$	k^d	C^d
1	1	5	0	1.133	1.133	1	(1,0,0,0,0,0,0,0)
	2	0	0	0.865	0.865	1	(1,0,0,0,0,0,0,0)
	3	2	0	0.547	0.547	4	(1,1,0,1,0,0,1,0)
	4	0	0	0.637	0.637	8	(1,1,1,1,1,1,1,1)
	5	7	0	0.873	0.873	2	(1,1,0,0,0,0,0,0)
2	1	6.133	1.133	1.456	0.323	2	(1,1,0,0,0,0,0,0)
	2	0.865	0.865	1.296	0.431	2	(1,0,0,0,0,0,1,0)
	3	2.547	0.547	0.549	0.002	4	(1,1,0,1,0,0,1,0)
	4	0.637	0.637	0.736	0.099	8	(1,1,1,1,1,1,1,1)
	5	7.873	0.873	1.371	0.498	3	(1,1,0,0,0,0,1,0)
3	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	2.549	0.549	0.952	0.403	5	(1,1,1,0,0,1,0,1)
	4	0.736	0.736	0.715	0.021	8	(1,1,1,1,1,1,1,1)
	5	8.371	1.371	1.752	0.381	4	(1,1,0,0,0,1,1,0)
4	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	2.952	0.952	1.498	0.546	6	(1,1,1,0,1,0,1,1)
	4	0.715	0.715	0.742	0.027	8	(1,1,1,1,1,1,1,1)
	5	8.752	1.752	1.402	0.350	3	(1,1,0,0,0,0,1,0)
5	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	3.498	1.498	0.954	0.544	5	(1,1,1,0,0,1,0,1)
	4	0.742	0.742	0.771	0.029	8	(1,1,1,1,1,1,1,1)
	5	8.402	1.402	2.049	0.647	4	(1,1,0,0,1,0,1,0)
6	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	2.954	0.954	1.495	0.541	6	(1,1,1,0,1,0,1,1)
	4	0.771	0.771	0.721	0.050	8	(1,1,1,1,1,1,1,1)
	5	9.049	2.049	1.794	0.255	4	(1,1,0,0,0,1,1,0)
		⋮				⋮	
14	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	3.497	1.497	1.497	0.000	6	(1,1,1,0,1,0,1,1)
	4	0.743	0.743	0.743	0.000	8	(1,1,1,1,1,1,1,1)
	5	8.897	1.897	1.897	0.000	4	(1,1,0,0,0,1,1,0)

Figure 4.4: Graphical representation of the evolution of the deferral rates per day

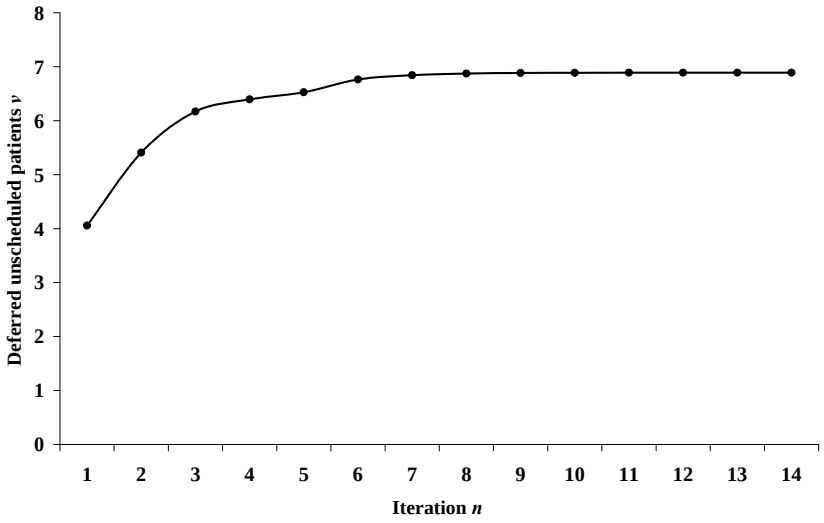


4.6.3 The Resulting CAS

Table 4.8 presents the final results for the numerical example. The percentage of unscheduled patients served on the day of arrival is 69%, so $F = 0.69$. This fraction is composed by fractions $F^1, \ldots, F^D$ that differ from day to day ($F^d = (\sum_t \chi_t^d - \nu^d)/\sum_t \chi_t^d$). For example, since day 4 is a quiet day with respect to unscheduled patient arrivals, it is completely filled with appointments. Only if no appointment request is made in one of the reserved slots, an unscheduled patient can be served. Apparently, it pays off to serve on average only 7% of the unscheduled patients directly on day 4 in the cycle. This is a result of the fact that only 3.6% of the unscheduled patients arrive on day 4, and that accordingly appointments are preferably planned on this day. The deferred unscheduled patients stream per day and the mean number of unscheduled patients served on the day of arrival are displayed in Table 4.8, which also reflects that on day 4 a small amount of unscheduled patients is directly served but also relatively few patients are deferred. The realized service level $S(10) = 0.962$ is well above the defined service level norm of 0.95.

The resulting capacity cycle is $K = (2, 2, 6, 8, 4)$, with corresponding day schedules which we discuss one-by-one below. Note that to achieve the service level norm it is required to reserve a buffer capacity of 1.11 to account for variability in appointment request arrivals, since 22 appointment slots are reserved while the average total number of patients to schedule within a cycle is $\sum_d (\lambda^d + \nu^d) = 14 + 6.89 = 20.89$. Apparently, the service level norm is achieved with only 5% buffer capacity, thus reserved capacity for appointments can be used efficiently. The realized mean load per day, denoted by $L^1, \ldots, L^D$, is a result of the capacity cycle, the probabilities that the reserved appointment slots are utilized by appointment requests and the mean number of un-

Figure 4.5: Graphical representation of the evolution of the total deferral rate



scheduled patients served on day of arrival $\sum_t \chi_t^d - \nu^d$. It turns out that the load is balanced throughout the cycle where each day has a realized load between 6.7 and 7.7. Finally, we discuss the resulting day schedules, to explain the moments on which the appointments are planned (see also Figure 4.6).

Day 1, $C^1 = (1, 1, 0, 0, 0, 0, 0, 0)$**.** Although the lowest unscheduled arrival rate occurs at end of the day, the appointments are planned at the beginning of the day. Since unscheduled patients are willing to wait 2 time slots, a peak in arrivals has an impact until two slots afterwards. If appointments were planned at the end of the day, there is no possibility to serve arriving unscheduled patients, while when planning appointments at slots at the beginning of the day, early unscheduled arrivals can be served in the third time slot.

Day 2, $C^2 = (1, 0, 0, 0, 0, 0, 1, 0)$**.** Again, the tendency to plan appointments early shows up. But, the drop in unscheduled arrivals is such that it is worthwhile to plan one appointment at the end of the day. However, again although the lowest arrival rate occurs in the latest time slot, the appointment is planned one slot before, to be able to serve an unscheduled patient arriving during interval $(T-3, T-1]$.

Day 3, $C^3 = (1, 1, 1, 0, 1, 0, 1, 1)$**.** The demand for unscheduled patients is relatively low. Therefore, only two slots are left open in which no appointment is planned. These are planned during the peak hours of unscheduled arrivals. However, the open slots are not planned consecutively, so to spread the possibilities for unscheduled patient service.

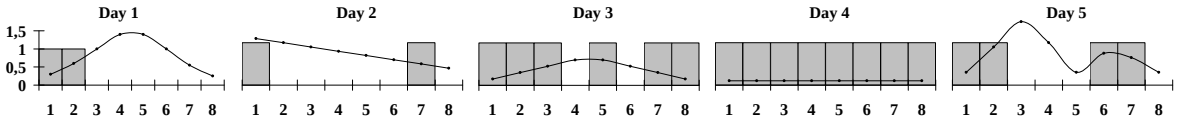
Day 4, $C^4 = (1, 1, 1, 1, 1, 1, 1, 1)$**.** As described before, this is a quiet day for unscheduled patients. Therefore, all slots are reserved for scheduled patients. However, note that not always are all reserved slots used for appointments; in 88% of the cases all reserved slots on day 4 are utilized for scheduled patients.

Table 4.8: End results for the case study

Indicator	Description	Value
F	Fraction unscheduled directly served	0.69
$F^1, \ldots, F^5$	Daily fraction unscheduled directly served	$0.78, 0.78, 0.50, 0.07, 0.67$
$S(10)$	Service level scheduled patients	0.962
$\nu^1, \ldots, \nu^D$	Deferral rate per day	$1.46, 1.30, 1.50, 0.74, 1.90$
$\sum_t \chi_t^1 - \nu^1, \ldots, \sum_t \chi_t^D - \nu^D$	Unscheduled patient service rate per day	$5.04, 4.70, 1.50, 0.06, 3.80$
$L^1, \ldots, L^D$	Realized utilization per day	$7.04, 6.70, 7.48, 7.71, 7.06$
K	Capacity cycle	$(2, 2, 6, 8, 4)$
C^1	CAS day 1	$(1, 1, 0, 0, 0, 0, 0, 0)$
C^2	CAS day 2	$(1, 0, 0, 0, 0, 0, 1, 0)$
C^3	CAS day 3	$(1, 1, 1, 0, 1, 0, 1, 1)$
C^4	CAS day 4	$(1, 1, 1, 1, 1, 1, 1, 1)$
C^5	CAS day 5	$(1, 1, 0, 0, 0, 1, 1, 0)$

Day 5, $C^4 = (1, 1, 0, 0, 0, 1, 1, 0)$**.** The appointments are planned around the unscheduled arrival peaks. It is remarkable that the two later appointments do not occur exactly during the off-peak hours but later, which can also be explained by the aforementioned delayed impact of unscheduled arrival peaks.

Figure 4.6: The CAS versus the unscheduled patient arrival rates per slot



The final conclusion is that the resulting CAS and its performance is the outcome of the complex interaction between the scheduled patient arrival rates λ^d, the unscheduled patients arrival patterns χ_t^d and the service level requirement $S^{norm}(y)$. For example, if $S^{norm}(y)$ is set tighter, it is to be expected that the resulting capacity cycle more closely resembles the total arrival rates for appointment requests $\gamma^d, d = 1, \ldots, D$. Also, since there would be less flexibility to spread the appointments, the fraction of unscheduled patients served on the day of arrival, F, would decrease.

4.7 Discussion

In this chapter we have outlined a methodology to develop an appointment schedule for facilities with scheduled and unscheduled arrival streams. The methodology consists of two separate models, one to evaluate the access and the other to evaluate the day process. The two models are linked by an iterative algorithm. An advantage of this

modular approach is that the models and the algorithm can be updated separately, so that a high level of flexibility is obtained.

This chapter focused on developing a methodology that incorporates the key characteristics of a mixed system and an effective communication between the two time scales of the access process and day process. Achieving numerical efficiency will be our next challenge. For the problem instance in Section 4.6, the CAS was found using complete enumeration. Our work is currently aimed at incorporating heuristics so that larger, more realistic instances can be evaluated. The model structure of the day process suggests that local search techniques are worth exploring (see e.g. [23, 22, 103].

Some extensions can readily be incorporated in our approach. Management is free to choose the service level norm for the access time. As such, the resulting appointment schedules can be compared for several service levels. Also, different choices for the time patients are willing to wait could be studied or overbooking to anticipate for no-shows. Furthermore, the access time for scheduled patients and the fraction of unscheduled patients who cannot be served on the day of arrival are outcomes of model I and model II respectively, and serve as input for the algorithm. Of course, other model outcomes could be chosen as well. Finally, to incorporate for example planned maintenance of a service facility, the number of available slots in the day process can easily be amended by closing slots. Worthwhile to consider would also be to introduce stochastic service times and variability in the number of slots patients are willing to wait in the day process. This might be a better reflection of reality, in particular in healthcare applications. Last but not least, our focus will be on practical issues in the implementation of the methodology in healthcare settings in Leiden University Medical Center and Academic Medical Center Amsterdam.

Chapter 5

Appointments for Care Pathway Patients

5.1 Introduction

Care pathways have gained popularity in the healthcare sector the last two decades [6]. A care pathway is a management tool to organize multidisciplinary care for patients with identical characteristics (i.e., disease symptoms, diagnosis, age, etcetera). The care pathway specifies the steps in the care process [5] and routes patients along a pre-defined path of care providers and diagnostic facilities. Patients may complete a significant part of the path in one day. Given the vast number of hospital facilities incorporated in the path, planning is usually involved and hospitals tend to prioritize these patients. It is therefore not uncommon that slots are reserved for care pathway patients in an otherwise walk-in clinic. Examples are for instance found at diagnostic services, such as Radiology outpatient clinics (X-ray, CT) and blood withdrawal facilities. When these facilities are highly utilized (>85%), reserving a few slots for care pathway may lead to a significant increase of the waiting time of walk-in patients (recall the Pollaczek-Khintchine formula (2.2)).

In this chapter we translate the above problem setting to a queuing model. The hospital facility decides on the number of slots that is reserved for care pathway patients. The model then enables a trade-off between the delay for walk-in patients and the probability that the number of slots reserved for the care pathway patients is not sufficient.

The service and hospitality industry is quite familiar with policies where a part of the (unscheduled) customer stream is diverted and scheduled on a later moment on the day. This concept is also known as virtual queuing (see e.g., [57, 157]). Probably the most famous organization that employs virtual queuing is Walt Disney, that uses for the most popular attractions in its theme parks the FastPass system [59]. Park guests decide upon arrival at an attraction whether they want to join the waiting line, or get a ticket (the 'FastPass'), that gives them a time-frame to return and enter the attraction

without waiting. To avoid a large number of no-shows and long waiting time for the non-FastPass guests, it is only allowed to possess a FastPass ticket for one attraction at the same time. The queuing system behind FastPass is analyzed in [116]. However, in the FastPass system park guests are supported by information on the state of both the regular and FastPass queue (i.e., the waiting time in the regular queue and the come back time for the FastPass ticket) and decide upon arrival which queue they want to join. In this chapter, the two patient types originate from separate arrival processes (walk-in or care pathway) that determine their type and thus the queuing discipline. We have found no evidence that the particular reservation discipline we consider has been studied before.

The remainder of this chapter is organized as follows. In the next section we describe our queuing model, followed by the analysis in Section 5.3. In Section 5.4 we provide a couple of numeric examples, and we conclude with the discussion in Section 5.5.

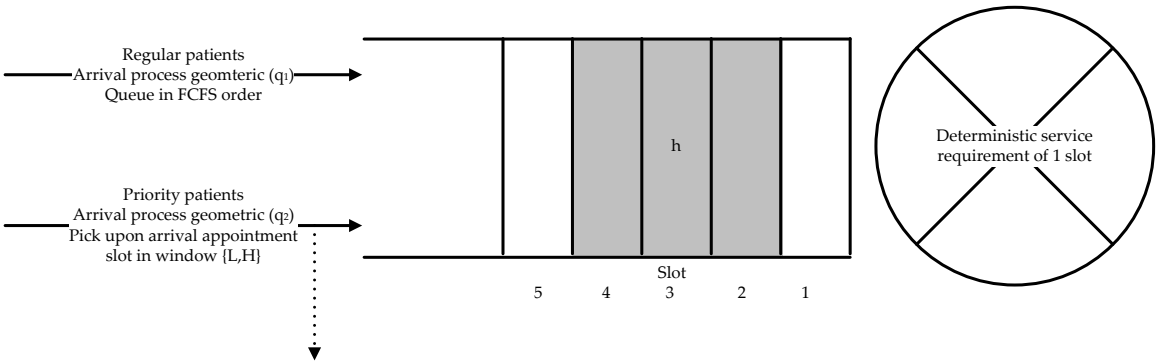
5.2 Model

For ease of notation we refer to the walk-in patients as regular patients, and to the care pathway patients as priority patients.

5.2.1 Assumptions

We consider a hospital facility which serves regular and priority patients. Both patient types have a deterministic service time requirement of 1 slot and arrive according to a geometric arrival process with arrival probability q_1 and q_2 respectively. The regular patients queue in FCFS order, while a priority patient picks, with probability p_h, upon arrival an appointment h slots later, $L \leq h \leq H$, where $1 \leq L \leq H < \infty$. When the desired slot is already taken by another priority patient, the newly arrived priority patient proceeds to slot $h-1$,...,L, until a slot is found that has not yet been taken by a priority patient. When all slots in the window that precede h are taken, the priority patient is blocked and lost. If the slot taken by the priority patient is occupied by a regular patient, then the regular patient is shifted to the first higher slot that is not taken by a priority patient. If this slot is non-empty as well, the regular patient that was occupying this slot is shifted upwards to the first slot not taken by a priority patient, and so on. Note that h equals the maximum number of slots the new priority patient has to wait until his service commences. It can readily be observed that the service facility can be modeled as a discrete-time single server queue serving priority and regular patients. Regular patients join the back of the queue. Priority patients select the last slot in the interval $(L, \ldots, h)$. Regular patients are shifted to higher queue positions when a priority patient takes their position (see Figure 5.1). The slot pick probability p_h can follow any discrete probability distribution. While the priority patients do not 'see' regular

Figure 5.1: The $G/D/1$ queue with appointments, appointment window $(L, \ldots, H) = (2, 3, 4)$ and $h = 3$.



patients, the regular patients may experience significant delay when a priority patient joins the queue. If there is a priority patient on the first queue position at the moment of a service completion, this patient is served. Otherwise, a regular patient will be served. If there are no regular patients in the queue, the server is idle (even though there may be a priority patient on a slot position higher up in the queue).

5.2.2 Matrix Structure

The transitions in the appointment window at the end of each time slot are independent of the number of regular patients present. We therefore first define a submatrix with the 1-step transition probabilities for priority patients. Then we define submatrices for the 1-step transition probabilities of regular patients, which do depend on the state of the priority patient appointment window. Finally, we combine these matrices into one transition probability matrix.

Priority Patient Transition Probability Submatrix D

We define an appointment vector $\mathbf{v}$ of length H, specifying which slots contain priority patient appointments. At most one priority patient can claim an appointment slot, so $\mathbf{v} = (v_1 \ldots v_H)$, where v_h is a binary variable, equal to 1 when slot h is reserved by a priority patient and 0 otherwise. Note that the appointment vector $\mathbf{v}$ is of length H, while the appointment window is of length $H-L+1$. Even though the slots $(1, \ldots, L-1)$ in the appointment window can no longer be chosen by priority patients, they possibly contain appointments and thus should be taken into account in the analysis. At the end of each time slot $\mathbf{v}$ is updated; new appointments are added and existing appointments are moved forward one slot. There are 2^H possible combinations for $\mathbf{v}$: when $H = 4$, $\mathbf{v}$ can for example be equal to $(0000), (0101), (1101)$, and so on. It follows immediately

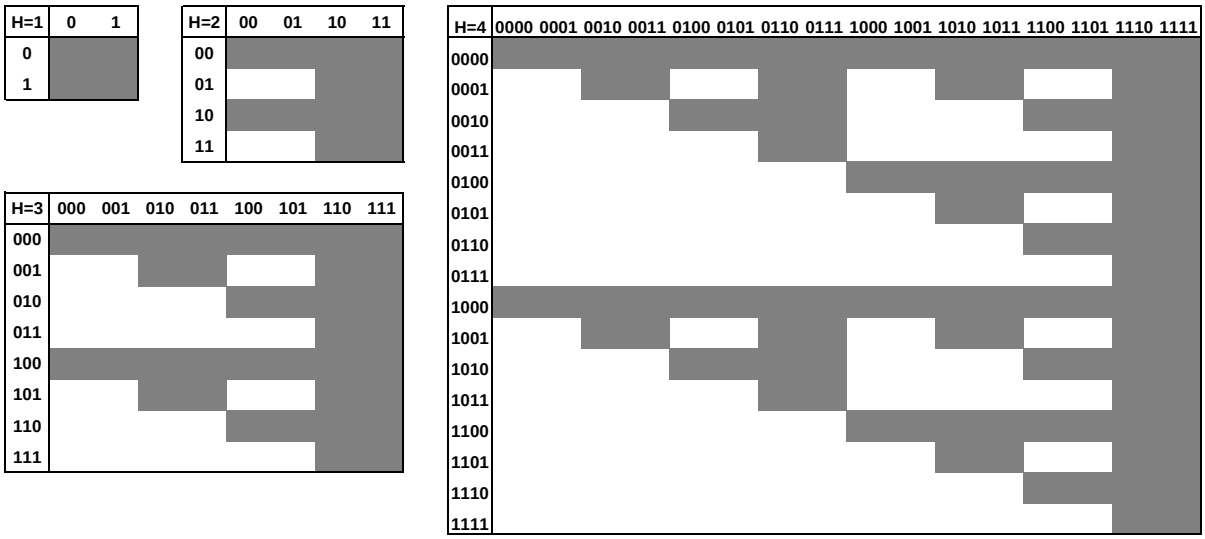
that the 1-step transition probability submatrix, D, has size $2^H \times 2^H$. Deriving D can be quite cumbersome for $H > 2$. We therefore present an algorithm to simplify this process. Alternatively, D can be computed numerically using Monte Carlo simulation.

Algorithm for computation of D

Step 1. Initialization
1a. Create the 2^H possible appointment combinations and order them lexicographically.
1b. Create an (empty) matrix of size $2^H \times 2^H$, where the rows and columns represent the 2^H lexicographically ordered possible combinations for **v** at time slot t and $t+1$ respectively.

Step 2. Creating the Block Structure
The possible shifts in **v** at the end of each time slot lead to a unique submatrix structure. Since at the end of each time slot the appointments are advanced one slot, all vectors with a 1-entry (an appointment) on position x, $x > 1$, will not have a possible transition to a vector with a 0-entry (no appointment) one position to the left, i.e., on position $x - 1$. Also, since appointments on the first position will be removed from **v** in the next shift, the submatrix' structure is identical for the first and second 2^{H-1} rows. Figure 5.2 shows the repetition in the structure of D for $H = \{1, ..., 4\}$. In fact, for $H > 3$ the upper-left block of four rows and eight columns is repeated each four rows down and eight columns to the right.

Figure 5.2: Structure of D for $H = \{1, \ldots, 4\}$



Step 3. Calculating the Required Number of Arrivals N
For each possible transition a certain number of priority patient arrivals, N, is required. It follows that for $H > 3$ the upper-left 4×8 building block is filled with the number of required arrivals, as given in Figure 5.3, and each repetition to the right, the required number of arrivals is raised

by one. When the first entry of $\mathbf{v}$ in the column of D equals 1, a minimum number of arrivals is required to make this transition (denoted in Figure 5.3 with $N = n+$). When the first entry of $\mathbf{v}$ equals 0, an exact number of arrivals is required to make this transition ($N = n$). For example, see Figure 5.3. For the transition from (1000) to (0111) exactly 3 arrivals are required, but for the transition from (0001) to (1011) at least 2 (2+) arrivals are required. Not only the structure of the upper-left building block is identical for $H > 3$, but also the required number of arrivals (as given in Figure 5.3) remains the same.

Figure 5.3: Required number of arrivals in D for $H = 4$

H=4	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0000	0	1	1	2	1	2	2	3	1+	2+	2+	3+	2+	3+	3+	4+
0001			0	1			1	2			1+	2+			2+	3+
0010					0	1	1	2					1+	2+	2+	3+
0011							0	1							1+	2+
0100									0+	1+	1+	2+	1+	2+	2+	3+
0101											0+	1+			1+	2+
0110													0+	1+	1+	2+
0111															0+	1+
1000	0	1	1	2	1	2	2	3	1+	2+	2+	3+	2+	3+	3+	4+
1001			0	1			1	2			1+	2+			2+	3+
1010					0	1	1	2					1+	2+	2+	3+
1011							0	1							1+	2+
1100									0+	1+	1+	2+	1+	2+	2+	3+
1101											0+	1+			1+	2+
1110													0+	1+	1+	2+
1111															0+	1+

Step 4. Adapting the Blocks for $L > 1$
If $L > 1$, the slots $(1, ..., L-1)$ cannot be claimed by priority patients. This changes the structure of D: the blocks are halved $L - 1$ times. In the left half of the remaining part of the block n arrivals are required, while in the right half n or more arrivals are required (see Figure 5.4 for an example with $H = 3$ and $L = \{1, 2, 3\}$).

Step 5. Calculating the Transition Probabilities
In the last step of the algorithm we need to calculate the transition probabilities $\mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1})$ that fill the gray cells in D (in all white cells, no transition is possible and $\mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1}) = 0$). Recall that we use N to denote the number of required arrivals as given in D. The transition probabilities are multinomial distributed and given by:

$$\mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1}) = \begin{cases} 0 & \text{if} \quad \mathbf{v}^t \not\to \mathbf{v}^{t+1}, \\ \sum_{j=N}^{J} b_j \sum_{\substack{k_L,\ldots,k_H \\ \sum_{h=L}^{H} k_h = j}} \binom{j}{k_L, \ldots, k_H} p_H^{k_H} \cdots p_L^{k_L} & \text{otherwise,} \end{cases} \tag{5.1}$$

where $J = N$ if $N = n$ and ∞ if $N = n+$, b_j is the geometric probability that j priority patients

Figure 5.4: Structure and required number of arrivals in D for $H = 3$ and $L = \{1, 2, 3\}$

L=1	000	001	010	011	100	101	110	111
000	0	1	1	2	1+	2+	2+	3+
001			0	1			1+	2+
010					0+	1+	1+	2+
011							0+	1+
100	0	1	1	2	1+	2+	2+	3+
101			0	1			1+	2+
110					0+	1+	1+	2+
111							0+	1+

L=2	000	001	010	011	100	101	110	111
000	0	1	1+	2+				
001			0+	1+				
010					0	1	1+	2+
011							0+	1+
100	0	1	1+	2+				
101			0+	1+				
110					0	1	1+	2+
111							0+	1+

L=3	000	001	010	011	100	101	110	111
000	0	1+						
001			0	1+				
010					0	1+		
011							0	1+
100	0	1+						
101			0	1+				
110					0	1+		
111							0	1+

arrive in a time slot, given by:

$$b_j = (1 - q_2)q_2^j, \tag{5.2}$$

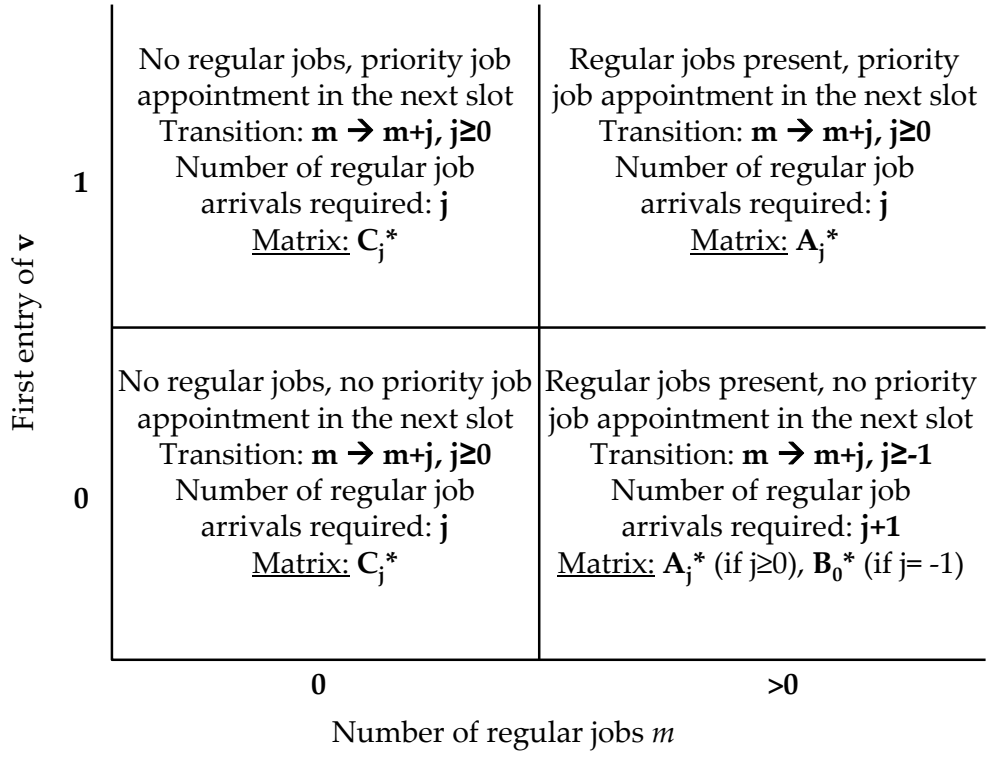
and p_h is the slot pick probability. The distribution of the j arrivals over the slots is denoted by $k_L, \ldots, k_H$, and for each slot $h = (L, \ldots, H)$ the following should hold to ensure the j arrivals are distributed over the slots such that $\mathbf{v}^{t+1}$ is obtained:

$$\begin{aligned}
&\text{If} \quad (v_h^{t+1} - v_{h+1}^t) = 0 \quad \text{for} \quad h = (L, \ldots, H-1), \quad \text{or} \quad v_H^{t+1} = 0, \\
&\text{then} \quad k_h = 0, \quad \text{and} \quad \sum_{i=h+1}^{H} k_i = \sum_{i=h+1}^{H-1} (v_i^{t+1} - v_{i+1}^t) + v_H^{t+1} \quad \text{for} \quad h = (L, \ldots, H-1). \\
&\text{If} \quad (v_h^{t+1} - v_{h+1}^t) = 1 \quad \text{for} \quad h = (L, \ldots, H-1), \quad \text{or} \quad v_H^{t+1} = 1, \\
&\text{then} \quad \sum_{i=h}^{H} k_i \geq \sum_{i=h}^{H-1} (v_h^{t+1} - v_{h+1}^t) + v_H^{t+1} \quad \text{for} \quad h = (L, \ldots, H-1), \quad \text{and} \quad k_H \geq 1.
\end{aligned} \tag{5.3}$$

Regular Patient Transition Probability Submatrices A^*, B^*, and C^*

While D is the same for all possible priority patient transitions, the regular patient transition probability submatrices, which contain the probabilities for transitions in the number of regular patients present, m, depend on the appointment vector **v**. Since we consider 1-step transitions, only the first entry of **v** is of interest. Three submatrices, A^*, B^*, and C^*, can be identified, which one to apply depends on m and **v** (see Figure 5.5). The submatrices given all have size $2^H \times 2^H$ and are constructed as follows. Define

Figure 5.5: Applicability of regular patient submatrices



u and **w** as vectors of length 2^H. The first 2^{H-1} entries of **u** are equal to q_1, and the second 2^{H-1} entries of **u** are equal to 1. The first 2^{H-1} entries of **w** are equal to 1, and the second 2^{H-1} entries of **w** are equal to 0. Furthermore, define **e** as the vector of ones, also of length 2^H. Then we obtain:

$$\begin{aligned} A_j^* &= a_j A^* \quad \text{where} \quad A^* = \mathbf{u}^T \times \mathbf{e}, \\ B_0^* &= a_0 B^* \quad \text{where} \quad B^* = \mathbf{w}^T \times \mathbf{e}, \\ C_j^* &= a_j C^* \quad \text{where} \quad C^* = \mathbf{e}^T \times \mathbf{e}. \end{aligned} \tag{5.4}$$

Since the arrival process of regular patients is geometrically distributed, the probability a_m that m regular patients arrive in a time slot is given by:

$$a_m = (1 - q_1)q_1^m, \quad m \geq 0. \tag{5.5}$$

The Combined Transition Probability Matrix P

The priority and regular patient arrival processes are independent, and therefore we can multiply D element wise with A^*, B^*, and C^*, i.e., every (m, n)-entry of D is multiplied with the (m, n)-entry of A^*, B^*, and C^*, in order to obtain the transition probability matrix P with elements A_j, B_0, and C_j, $j \geq 0$. Each entry of P is a matrix in itself of size $2^H \times 2^H$, and represents the state transition $(m_t, \mathbf{v}^t) \rightarrow (m_{t+1}, \mathbf{v}^{t+1})$.

$$P = \begin{pmatrix} C_0 & C_1 & C_2 & \cdots & C_m & \cdots \\ B_0 & A_0 & A_1 & \cdots & A_{m-1} & \cdots \\ 0 & B_0 & A_0 & \cdots & A_{m-2} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \cdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Note that A_j can also be written as $a_j \bar{A} D$, where $\bar{A}$ is the diagonal matrix with the elements of **u** on the diagonal. The same holds for B_0, which can be written as $b_0 \bar{B} D$, where $\bar{B}$ is the diagonal matrix with the elements of **w** on the diagonal, and for C_j, which can be written as $c_j \bar{C} D$, where $\bar{C}$ is the diagonal matrix with the elements of **e** on the diagonal.

5.3 Analysis

The matrix P shows similarities with the transition probability matrix for the $M/G/1$ queue embedded at departure moments (see [142] for further reference). An overview of discrete time queuing systems can be found in [30]. Several priority disciplines have been studied for discrete time queuing models, but these are usually related to the non-preemptive [195] or preemptive resume priority disciplines [183]. In [194] a different, but related, service discipline is considered, where a slot is reserved for regular patients at the end of the queue. In the case of high load traffic from priority patients, it is then guaranteed that regular patients receive service as well.

5.3.1 Stability of the Queue

In order for the queue to be stable, the mean load, ρ, should be less than one. Since the service time is 1 slot, ρ equals the sum of the mean number of regular patient arrivals

per slot and the accepted priority patients per slot:

$$\rho = \frac{q_1}{1-q_1} + (1-\mathbb{P}_{B_2})\frac{q_2}{1-q_2} < 1, \tag{5.6}$$

It is necessary that $q_1 < \frac{1}{2}$, but not sufficient since the number of accepted priority patients per slot also depends on the blocking probability for priority patients, $\mathbb{P}_{B_2}$. The latter is calculated as follows. A priority patient is accepted when the slot h, picked with probability p_h, is still available, or if not, when one of the slots $(L, \ldots, h-1)$ is still available. The blocking probability for priority patients is therefore given by:

$$\mathbb{P}_{B_2} = 1 - p_h \cdot \sum_{\substack{v^t \to v^{t+1}: \\ \sum_{i=L}^{h} v_i^{t+1} < h}} \mathbb{P}(\mathbf{v}^t \to \mathbf{v}^{t+1}) \, . \tag{5.7}$$

5.3.2 Vector Generating Function of Equilibrium Probabilities $\pi(m, \mathbf{v})$

We derive the vector generating function of the equilibrium probability $\pi(m, \mathbf{v})$ for the number of regular patients present, m, and the realization of the appointment vector, $\mathbf{v}$. For notation purposes, denote $\pi(m, \mathbf{v})$ by the vector π_s, where $s = (0, 1, \ldots)$. Using the property $\Pi P = \Pi$, we obtain:

$$\pi_s = \pi_0 C_s + \sum_{i=1}^{s} \pi_i A_{s-i} + \pi_{s+1} B_0 \quad \text{for} \quad s \geq 1, \quad \text{and} \tag{5.8}$$

$$\pi_0 = \pi_0 C_0 + \pi_1 B_0, \quad \text{where} \quad \sum_{s=0}^{\infty} \pi_s \mathbf{e}^T = 1. \tag{5.9}$$

Define the vector generating function for π_s, $P_{\Pi}(z)$, as:

$$P_{\Pi}(z) = \sum_{s=0}^{\infty} \pi_s z^s. \tag{5.10}$$

Furthermore, define:

$$A(z) = \sum_{s=0}^{\infty} A_s z^s, \quad \text{and} \quad C(z) = \sum_{s=0}^{\infty} C_s z^s. \tag{5.11}$$

Multiplying both sides of (5.8) with the scalar z^s, where $|z| \leq 1$, and summing the result for $s = (0, \ldots, \infty)$, we obtain:

$$\sum_{s=0}^{\infty} \pi_s z^s = \sum_{s=0}^{\infty} \pi_0 C_s z^s + \sum_{s=1}^{\infty} \sum_{i=1}^{s} \pi_i A_{s-i} z^s + \sum_{s=0}^{\infty} \pi_{s+1} B_0 z^s, \tag{5.12}$$

and it follows that:

$$P_{\Pi}(z) \;=\; \pi_0 C(z) + P_{\Pi}(z)A(z) - \pi_0 A(z) + B_0 z^{-1} P_{\Pi}(z) - \pi_0 B_0 z^{-1}\,. \tag{5.13}$$

Multiplication of (5.13) with z and rearranging terms gives:

$$P_{\Pi}(z)\left[zI - zA(z) - B_0\right] \;=\; \pi_0\left[zC(z) - zA(z) - B_0\right]. \tag{5.14}$$

5.3.3 Mean Number of Regular Patients Present

We derive the mean number of regular patients in the queue, $\mathbb{E}[L_R]$, by following the analysis from [142], pp. 143-148. Let $z = 1$. First we list the relations we already have.

$$\begin{aligned}
\mathbb{E}[L_R] &= P'_{\Pi}(1)\mathbf{e}^T \\
\mathbb{E}[L_R]\pi^{\infty} &= P'_{\Pi}(1)\mathbf{e}^T\pi^{\infty} \\
P_{\Pi}(1) &= \pi^{\infty} \\
P_{\Pi}(1)\mathbf{e}^T &= 1 \\
A(1) + B_0 &= C(1) = D \\
D\mathbf{e}^T &= \mathbf{e}^T\,,
\end{aligned} \tag{5.15}$$

where π^{∞} is the vector with the equilibrium probabilities of the number of priority patients in the queue, which can be obtained from $\pi^{\infty}D = \pi^{\infty}$. The first derivative of (5.14) with respect to z is:

$$\begin{aligned}
P'_{\Pi}(z)\;& \left[zI - zA(z) - B_0\right] + P_{\Pi}(z)\left[I - A(z) - zA'(z)\right] \\
&= \pi_0\left[C(z) + zC'(z) - A(z) - zA'(z)\right].
\end{aligned} \tag{5.16}$$

For $z = 1$, it follows that:

$$P'_{\Pi}(1)\left[I - D\right] + \pi^{\infty}\left[I - A(1) - A'(1)\right] = \pi_0\left[C(1) + C'(1) - A(1) - A'(1)\right]. \tag{5.17}$$

Denote $\left[I - D + \mathbf{e}^T\pi^{\infty}\right]$ by U and $\left[I - \frac{1}{1-q_1}\bar{A}D\right]$ by K. Furthermore, note that $\left[\frac{1}{1-q_1}D - \frac{1}{1-q_1}\bar{A}D\right]$ is equal to $\bar{B}D$. By adding $P'_{\Pi}(1)\mathbf{e}^T\pi^{\infty} = \mathbb{E}[L_R]\pi^{\infty}$ we obtain:

$$\begin{aligned}
P'_{\Pi}(1)\;& \left[I - D + \mathbf{e}^T\pi^{\infty}\right] + \pi^{\infty}\left[I - \bar{A}D - \frac{q_1}{1-q_1}\bar{A}D\right] \\
&= \mathbb{E}[L_R]\pi^{\infty} + \pi_0\left[D + \frac{q_1}{1-q_1}D - \bar{A}D - \frac{q_1}{1-q_1}\bar{A}D\right] \\
\Rightarrow & \\
P'_{\Pi}(1)\;& \left[I - D + \mathbf{e}^T\pi^{\infty}\right] + \pi^{\infty}\left[I - \frac{1}{1-q_1}\bar{A}D\right] \\
&= \mathbb{E}[L_R]\pi^{\infty} + \pi_0\left[\frac{1}{1-q_1}D - \frac{1}{1-q_1}\bar{A}D\right].
\end{aligned} \tag{5.18}$$

From Theorem 5.1.3 in [106] it follows directly that the matrix U is invertible. We then have that $\pi^\infty U^{-1} = \pi^\infty$ and thus:

$$P'_\Pi(1) \quad = \mathbb{E}[L_R]\pi^\infty + \pi_0 \bar{B} D U^{-1} - \pi^\infty K U^{-1}. \tag{5.19}$$

Multiplying with ^{T} it follows that:

$$\pi_0 \bar{B} D \mathbf{e}^T \quad = \quad \pi^\infty K \mathbf{e}^T. \tag{5.20}$$

By taking the second derivative of (5.14) with respect to z, setting $z = 1$ and multiplying with $\mathbf{e}^T$ we obtain:

$$\begin{aligned} P''_\Pi(1) \quad & [I - D]\,\mathbf{e}^T + 2P'_\Pi(1) K \mathbf{e}^T \\ & = \pi^\infty \left[\frac{2q_1}{(1-q_1)^2} \bar{A} \mathbf{e}^T \right] + \pi_0 \left[\frac{2q_1}{1-q_1} \bar{B} D \mathbf{e}^T \right]. \end{aligned} \tag{5.21}$$

Since $P''_\Pi(1)\,[I - D]\,\mathbf{e}^T = 0$ we get:

$$P'_\Pi(1) K \mathbf{e}^T = \frac{q_1}{(1-q_1)^2} \pi^\infty \bar{A} \mathbf{e}^T + \frac{q_1}{1-q_1} \pi_0 \bar{B} D \mathbf{e}^T. \tag{5.22}$$

Now we combine (5.19) and (5.22) to obtain an expression for $\mathbb{E}[L_R]$:

$$\begin{aligned} & \mathbb{E}[L_R]\pi^\infty K \mathbf{e}^T \\ & = \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T + \pi_0 \bar{B} D \left[\frac{q_1}{1-q_1} I - U^{-1} K \right] \mathbf{e}^T \\ & = \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T + \pi_0 \bar{B} D U^{-1} \left[\frac{3q_1 - 1}{1-q_1} \mathbf{e}^T + (\mathbf{e}^T - \mathbf{w}^T) \right] \\ & = \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T + \pi_0 \bar{B} \left[\frac{3q_1 - 1}{1-q_1} \mathbf{e}^T + D U^{-1} (\mathbf{e}^T - \mathbf{w}^T) \right]. \end{aligned} \tag{5.23}$$

Using (5.20) this simplifies to:

$$\mathbb{E}[L_R]\pi^\infty K \mathbf{e}^T \quad = \quad \pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K + \frac{2q_1}{1-q_1} K \right] \mathbf{e}^T - \pi_0 \bar{B} D U^{-1} \mathbf{w}^T \,, \tag{5.24}$$

and

$$\mathbb{E}[L_R] \quad = \quad \left[\pi^\infty \left[\frac{q_1}{(1-q_1)^2} \bar{A} + K U^{-1} K \right] \mathbf{e}^T - \pi_0 \bar{B} D U^{-1} \mathbf{w}^T \right] \left[\pi^\infty K \mathbf{e}^T \right]^{-1} + \frac{2q_1}{1-q_1} \,. \tag{5.25}$$

The second and higher moments of $\mathbb{E}[L_R]$ can be computed using the same approach.

In expression (5.25) there is still an unknown, π_0. We suggest two approximations for π_0 and thus for $\mathbb{E}[L_R]$. Since the load for regular patients is high and therefore the probability that the server is idle while there are priority patients in the queue is low, the first approximation is obtained by $\pi_0 = (1-\rho)\pi^\infty$. The second approximation is to set $\pi_0 \bar{B} D U^{-1} \mathbf{w}^T = \mathbf{0}$. We use simulation (see Table 5.1) to determine which of the two approximations is most accurate in terms of the parameter values of our problem setting, i.e., a high load for regular patients ($q_1 = 0.45$) and a low to moderate load for priority patients ($q_2 = 0.10$). The slot pick probability p_h is uniform distributed. We also give the load ρ_S that follows from the simulation.

Table 5.1: Comparing the values of $\mathbb{E}[L_R]$ that follow from the simulation and approximations

Case	L	H	ρ_S	$\mathbb{E}[L_R]$			$\lvert\delta\rvert$ with sim.	
				Sim.	Approx. 1	Approx. 2	Approx. 1	Approx. 2
1	1	1	0.9171	10.1	10.1	10.9	0.0	0.8
2	1	3	0.9250	11.0	11.2	12.0	0.2	1.0
3	1	5	0.9270	11.5	11.5	12.3	0.0	0.8
4	3	3	0.9171	9.9	10.1	10.9	0.2	1.0
5	3	5	0.9250	11.2	11.2	12.0	0.0	0.8
6	5	5	0.9170	10.0	10.2	10.9	0.2	0.7

The mean number of regular patients in the queue in the simulation, $\mathbb{E}[L_R]_S$, was calculated by simulating a period of 100,000 slots (so that there would be $\approx$ 10,000 priority patient arrivals), preceded by a warm-up period of 1,000 slots. When in run n,

$$\left| \frac{\sum_{i=1}^{n} \mathbb{E}[L_R]_{S,i}}{n} - \frac{\sum_{i=1}^{n-1} \mathbb{E}[L_R]_{S,i}}{n-1} \right| < \epsilon, \tag{5.26}$$

the simulation would stop. For ϵ a value of e^{-1} was chosen, which corresponds in the case of ten minute slots to an error margin of one minute. We see that the first approximation is more accurate with a maximum error in the six test cases of 0.2 (2 minutes).

5.3.4 Mean Waiting Time for Regular Patients

Even though the regular patients may experience additional delay when a priority patient takes their spot, the mean waiting time for regular patients, $\mathbb{E}[W_R]$, can still be calculated using Little's law. This is because the queuing discipline for the regular patients is FCFS and therefore the order in the queue for regular patients does not change when a priority patient arrives and picks a slot in the appointment window. The mean waiting time is therefore equal to the sojourn time (which is calculated using the mean number of regular patients present, $\mathbb{E}[L_R]$, and the mean throughput of regular patients

per slot, ρ_1, minus one slot):

$$\mathbb{E}[W_R] = \frac{\mathbb{E}[L_R]}{\rho_1} - 1, \quad \text{where} \quad \rho_1 = \frac{q_1}{1 - q_1}\,. \tag{5.27}$$

5.4 Results

To generate the results presented in this section, we use the first (most accurate) approximation of $\pi_0 = (1-\rho)\pi^\infty$. We use the same parameter values as in the previous section, i.e., $q_1 = 0.45$, $q_2 = 0.10$.

5.4.1 The Effect of the Size and Position of the Appointment Window

In Table 5.2 we see the effect of the size and position of the appointment window on the waiting time for regular patients, $\mathbb{E}[W_R]$, and the blocking probability for priority patients, $\mathbb{P}_{B_2}$. As is also apparent from Figure 5.6, $\mathbb{E}[W_R]$ increases and $\mathbb{P}_{B_2}$ decreases when the appointment window becomes larger and is positioned further away from the first position in the queue.

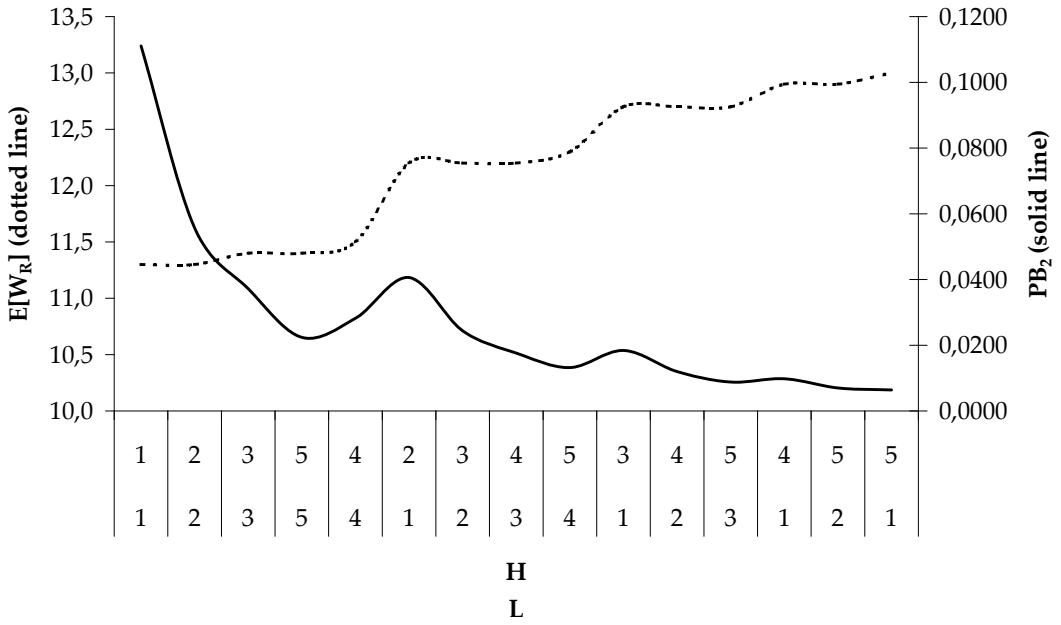
Table 5.2: Results for various positions and sizes of the appointment window

L	H	$\mathbb{E}[L_R]$	$\mathbb{E}[W_R]$	$\mathbb{P}_{B_2}$
1	1	10.1	11.3	0.1111
1	2	10.8	12.2	0.0406
1	3	11.2	12.7	0.0184
1	4	11.4	12.9	0.0098
1	5	11.5	13.0	0.0064
2	2	10.1	11.3	0.0557
2	3	10.8	12.2	0.0244
2	4	11.2	12.7	0.0120
2	5	11.4	12.9	0.0070
3	3	10.1	11.4	0.0372
3	4	10.8	12.2	0.0176
3	5	11.2	12.7	0.0088
4	4	10.2	11.5	0.0282
4	5	10.8	12.3	0.0132
5	5	10.2	11.4	0.0224

5.4.2 Comparison with the Non-Priority Queue

We compute $\mathbb{E}[L_R]$ for the same queuing system, but now the queue discipline is FCFS for both regular and priority patients (we still refer to priority patients, even though these (care pathway) patients do not have priority anymore), and there is no blocking

Figure 5.6: Waiting time for regular patients, $\mathbb{E}[W_R]$, versus blocking probability for priority patients, $\mathbb{P}_{B_2}$



of priority patients. The expected number of patients at the facility, $\mathbb{E}[L]$, is given by $\lim_{z\to 1} P'_{\Pi}(z)$, where it is easy to derive that $P_{\Pi}(z)$ in this case is given by:

$$P_{\Pi}(z) = (1-\rho)\frac{G(z)(1-z)}{G(z)-z}, \tag{5.28}$$

so that

$$\mathbb{E}[L] \;=\; (1-\rho)\frac{2G'(1)(1-G'(1))+G''(1)}{2(G'(1)-1)^2}. \tag{5.29}$$

In case of absence of the priority patients we have that $G'(1) = \rho_1$ and $G''(1) = \rho_1^2$, and thus $\mathbb{E}[L] = \mathbb{E}[L_R] = 2.7$ (note that ρ in (5.29) is equal to ρ_1). If the priority patients also arrive at the facility, $G(z)$ is the product of the two probability generating functions of the independent geometric arrival processes, and thus $G'(1) = \rho_1 + \rho_2$, and $G''(1) = 2\rho_1^2 + 2\rho_1\rho_2 + 2\rho_2^2$ (note that ρ in (5.29) is equal to $\rho_1 + \rho_2$). We obtain $\mathbb{E}[L] = 11.9$, and $\mathbb{E}[L_R] = \frac{\rho_1}{\rho_1+\rho_2}\mathbb{E}[L] = 10.5$, $\mathbb{E}[W] = 11.8$. So even though priority patients are not blocked, the mean waiting time for regular patients is shorter. Only a priority discipline where a single slot is reserved for priority patients results in a slightly shorter waiting time for regular patients (see Table 5.2).

5.5 Discussion

In this chapter we analyzed the single server queue in discrete time with two types of patients. Both patient types arrive according to a geometric arrival process and have a service requirement of 1 slot. Priority patients claim upon arrival an empty slot, h, ('appointment') in a pre-defined appointment window, and have absolute priority over regular patients. We have derived the blocking probability for priority patients and the mean waiting time for regular patients. The methodology we developed is mainly meant as a capacity planning tool, so that managers can study the effect of for instance the values of the lower and upper bound of the appointment window. In reality, a steady state situation, especially in an environment that does not offer 24/7 service such as an outpatient clinic, will maybe not be reached. However, given the managerial insights that the methodology gives, we still feel it can be very valuable in these cases.

Throughout the chapter we assumed that when h was already taken, the claim of the new arrived priority patient is advanced to slot $(h-1,\ldots,L)$, until a free slot was found. It is straightforward to analyze the queue where the claims are set back to slots $(h+1,\ldots,H)$. Also the possibility to choose any distribution for the slot pick probability, p_h, introduces a lot of flexibility. The choice for the distribution of p_h will especially influence the mean waiting time for regular patients. For example, the case where $p_H = 1$, $p_h = 0 \quad \forall \quad h \neq H$, makes maximal use of the appointment window in the case that the slots are advanced when a picked slot is already claimed, and thus $\mathbb{E}[W_R]$ will be larger than in the case that $p_H \neq 1$.

The effect of increasing H gradually reduces when H becomes larger, and will lead to computational issues. Currently, the computations for $\mathbb{E}[L_R]$ using a software program such as Matlab become already quite involved for $H \approx 10$. This is not necessarily a problem and allows for analysis of many problem instances, but deserves attention in future research. The symmetry in D might be useful to simplify the analysis and size of the solution space. Note that simulation has the same computational limitations.

Of course, the size of appointment window $(L,\ldots,H)$ has a significant influence on both the priority patient blocking probability, $\mathbb{P}_{B_2}$, and the regular patient waiting time, $\mathbb{E}[W_R]$. When the window size $H - L + 1$ is decreased, $\mathbb{P}_{B_2}$ will increase but $\mathbb{E}[W_R]$ will decrease. It is obvious that the trade-off between these two competing performance measures lies exactly here. A rule of thumb that comes into mind from the Subsection 5.4.2 and the graph in Figure 5.6, is that by reserving one slot for priority patients a few slots (3–5) from the first queue position, results in acceptable outcomes for both the waiting time for regular patients and the blocking probability for priority patients. However, a mean waiting time of over 11 slots (also in the case without priorities) is quite long, so the load of the system should be subject of study as well. In future research we plan to further investigate the exact trade-off and come to a rule of thumb. Furthermore, we plan to analyze queuing networks consisting of this type of queues.

Chapter 6

Allocating MRI Scan Capacity

6.1 Introduction

We consider an MRI scanning facility run by a Radiology department, that has to distribute MRI scan capacity among several competing hospital departments. The departments have private information regarding their future demands. For a fair allocation, Radiology depends on the information that the departments provide. How can the Radiology department motivate the users to give an honest forecast of their demands in order to ensure a fair allocation?

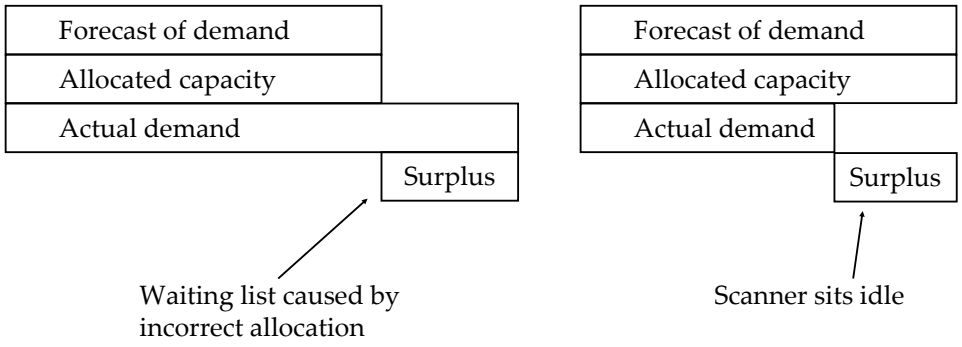
Various types of MRI scans exist, each used to inspect different parts of the body [143]. Examples are scans of the heart, breasts, nervous system, and bones. It is common practice in most hospitals to dedicate adjacent time slots (blocks) in the appointment schedule to identical MRI types. The demand for MRI scans can vary widely over time, especially in academic institutions. New treatment protocols may result in an in- or decrease of MRI requests; the same holds for the recruitment of new patient cohorts and changes in the hospital's patient mix. This asks for a periodical allocation of MRI capacity. For this it is common that hospital departments provide Radiology with a demand forecast for the next period. Overestimating demand may be tempting, since it is likely that this leads to a larger share of the scarce capacity. The quality of the MRI schedule depends on the quality of the forecast. It is therefore essential for the Radiology department that hospital departments put maximum effort into providing a reliable and honest forecast, and do not over- or underestimate their demand.

6.1.1 Problem Example

We illustrate the necessity of a reliable and honest forecast with an example of a facility with two scanning types. For the first scanning type, a forecast that is lower than the actual demand for the next period is provided. For the other scanning type, a forecast

that is higher than the actual demand for the next period is provided. Suppose that the capacity allocated by Radiology equals the forecast of demand. Then for the first scanning type, a waiting list develops because of incorrect allocation (Figure 6.2(a)). For the second scanning type, not all allocated capacity is needed and thus the scanner sits idle (Figure 6.2(b)). We see that it is very well possible that in the same period, the MRI

Figure 6.1: Example for two scanning types



(a) Scanning type 1 (b) Scanning type 2

scanners are idle during certain blocks due to less actual demand for one type of scans, while at the same time the waiting list for another scan type increases caused by a lack of capacity.

6.1.2 Approach

The problem of capacity allocation to multiple competing users, as sketched above, has several key properties. Namely (i) the users do not cooperate, (ii) the actual demands of the users are private information, and (iii) the resource wants the users to truthfully reveal their actual demand. Relevant models that capture these properties are combinatorial auction models [51], where multiple bidders can place bids on several items at the same time, and Bayesian games [87]–[89], non-cooperative games where each player has incomplete information about the characteristics of the other users. While a Bayesian game model uses only information on the user's demand, in a combinatorial auction also the price the users are willing to pay is required. This, combined with the relatively simple analysis of the Bayesian game model compared to that of the combinatorial auction model, determined our choice for the Bayesian game approach. We are interested in conditions under which the users tell the truth, that is, they provide the resource with their actual demand.

6.1.3 Literature

Bayesian Games are extensively described by Harsanyi [87]–[89]. For an introduction on this class of games we refer the reader to [71]. In the literature on Bayesian games, two types of models are often studied. In the first type of model a single resource communicates with several users. The users do not cooperate, and the resource has private information. An application of this model is given in [90]. In the second type of model a single resource communicates with a single user. Now, the user has private information. Examples can for instance be found in [168, 207]. Unlike these types of models, we consider a single resource and multiple non-cooperative users with private information. To the best of our knowledge, this has hardly been studied so far.

There is a vast body of literature on capacity allocation with truth-telling in the area of supply chain management, see for example [37, 132, 206]. The main research questions are how a supplier should allocate his capacity, and how the supplier can induce his buyers to reveal their private information. Furthermore, many papers on capacity and/or resource allocation in health care are available, such as [192], but these do not consider private information and truth-telling. This chapter contributes to the literature by studying capacity allocation under private information in a health care setting.

Several other problems in a health care context have been studied using Bayesian Games. An application area is the patient-doctor relationship, where either the patient [207] or the doctor [178] has private information. Another example is given in [193], where the authors consider the principle of kidney exchange. Patients waiting for a kidney transplant present one or more potential donors. These donors however are not a match to the patient they are related to. In order to find matching pairs, an exchange group of several patients and their donors is formed. In the paper it is demonstrated with a Bayesian Game that it is advantageous in some cases for patients not to reveal all information they possess about their donors. In [16] an economic application is given. Multiple hospitals are regulated by a central authority; hospitals do not cooperate with each other. The regulator has incomplete information on the production information hospitals possess. A Bayesian Game is used to study the effect of the information gap on the production contracts the regulator offers the hospitals. We conclude this paragraph with mentioning [139], in which the international trading and pricing of pharmaceuticals is studied. The author suggests to introduce asymmetric information with respect to the local demand function of the country the products are sold to. When the problem is modeled as a Bayesian Game, it can be shown that in equilibrium parallel imports of pharmaceuticals occur, in contrast to the complete information situation.

6.1.4 Contents of Paper

Since the approach is not limited to the MRI scan example, we use generic terminology (resources and users) in Sections 6.2–6.4. First we provide a detailed description of the

model. In the Results sections that follows we show that for two allocation mechanisms an optimal strategy for users is to provide an honest forecast of their demand, which enables the resource to make a fair allocation. We demonstrate the approach with a numerical example in Section 6.5. We conclude with the discussion and conclusions section.

6.2 Model

In this section we formulate the Bayesian Game. An overview of the notation introduced is given in Table 6.1. The allocation of capacity goes as follows. Users provide

Table 6.1: Notation introduced in Model section

Symbol	Description
C	Total amount of capacity available
F_i	Forecast of demand by user i (i.e. request to resource)
A_i	Capacity allocated to user i
D_i	Actual demand of user i
x	Reward per unit of allocated capacity
y	Penalty cost per unit of surplus capacity

their forecast F_i for the next period. The resource allocates capacity, resulting in an allocated amount A_i per user. During the period the users reveal their actual demand. This process is repeated each period. We make the following assumptions:

(i) All users make rational choices, i.e. they want to maximize benefits and minimize costs.

(ii) The total amount requested by the users exceeds the resource's capacity: $\sum_j F_j > C$.

(iii) The shared resource cannot allocate more capacity than is available: $\sum_j A_j \leq C$.

(iv) No user has an actual demand that is higher than the resource's capacity: $0 \leq D_i \leq C$.

(v) No user has any information about the private demand of any other user. Let $D_{-i} = \{D_j\}_{j \neq i}$ represent the demands of users other than user i. We model the knowledge of user i by the uniform distribution on $[0, C]^{n-1}$:

$$p_i(D_{-i}) = \begin{cases} \frac{1}{C^{n-1}}, & \text{if} \quad D_{-i} \in [0, C]^{n-1}, \\ 0, & \text{else;} \end{cases} \tag{6.1}$$

thus all demands are equally likely.

6.2.1 Utility Function

User i has a utility function V_i that measures the immediate happiness or reward [191]. The reward is the weighted difference between the allocated amount A_i and a penalty for overestimation. The weights are x per unit of allocated capacity and y per unit that is overestimated, $x, y > 0$. The utility function for user i is given by:

$$V_i \;=\; xA_i - y\max\{F_i - D_i, 0\}. \tag{6.2}$$

Each user aims to maximize its utility.

6.2.2 The Allocation Mechanism

The resource needs an allocation mechanism to distribute the capacity over the users. Desirable properties of an allocation mechanism are:

(i) Each user receives a nonnegative amount: $A_i \geq 0$.

(ii) All capacity is allocated: $\sum_j A_j = C$.

(iii) Each user receives at most the amount it requests: $A_i \leq F_i$.

(iv) If the capacity of the resource increases, then all users should obtain more (until they reach their forecast): A_i is increasing in C.

Many allocation mechanisms satisfy these properties. Three mechanisms that are used often in practice are the proportional rule, the constrained equal award rule and the constrained equal loss rule [186]. The proportional rule allocates capacity proportional to the forecasts:

$$A_i = \frac{F_i}{\sum_j F_j} C. \tag{6.3}$$

The constrained equal award rule divides the capacity equally among the users, with the constraint that a user cannot obtain more than was requested:

$$A_i = \min\{\alpha, F_i\}, \tag{6.4}$$

with α such that $\sum_j A_j = C$. Third, the constrained equal loss rule divides the shortage of capacity equally among the users such that any user receives a nonnegative amount:

$$A_i = \max\{F_i - \beta, 0\}, \tag{6.5}$$

with β such that $\sum_j A_j = C$.

6.2.3 Bayesian Game Formulation

Now we formulate the problem as a Bayesian game. Each user provides a forecast F_i, which is a function of his private actual demand D_i. We write $F_i(D_i)$ to denote this dependency. This forecast reflects the claim of user i on the available capacity. The allocated capacity A_i depends on all requests $F_j(D_j)$, $j = 1, \ldots, N$, and hence also on all the private demands. The goal of each user is to maximize his expected utility by selecting a suitable strategy. A strategy $F_i(D_i)$ of user i specifies which forecast the user should announce as a function of its private information D_i. The strategies $F^* = (F_1^*(D_1), \ldots, F_N^*(D_N))$ are a so-called Bayesian Nash equilibrium if for each user i and for any private demand D_i the requested number of units $F_i^*(D_i)$ maximizes the expected utility of the user:

$$F_i^*(D_i) = \arg\max_{F_i} \int_{[0,C]^{n-1}} V_i(F_{-i}^*, F_i; D_i) p_i(D_{-i}) \mathrm{d}D_{-i}, \tag{6.6}$$

where (F_{-i}^*, F_i) denotes the strategies F^* in which the strategy $F_i^*(D_i)$ of user i is replaced by F_i, $D_{-i} = \{D_j\}_{j \neq i}$ is the collection of private demands for users other than i, and $p_i(D_{-i})$ is the prior belief of user i about D_{-i} [146]. Hence, given the uncertainty on the private demands of the other users, it does not pay for user i to deviate from his equilibrium strategy because that will result in lower expected utility.

6.3 Results for Proportional Rule

In this section the capacity is allocated according to the proportional rule. Then the utility function of user i is:

$$V_i(F; D_i) = x \frac{F_i}{\sum_j F_j} C - y \max\{F_i - D_i, 0\}, \tag{6.7}$$

which depends on the demands $F = \{F_1, \ldots, F_N\}$ of all users, and on the user's privately known actual demand D_i. We show that when the number of users exceeds 3, it is optimal for the users to provide an honest forecast. When the number of users is equal to 2 or 3, the same result holds under weak conditions.

6.3.1 Equal Cost and Reward Parameters

To simplify calculations, we set $x = y = 1$ in the utility function, so $V_i(F; D_i) = \frac{F_i}{\sum_j F_j} C - \max\{F_i - D_i, 0\}$ (we consider other cost and reward parameters in section 6.3.2). We investigate when truth-telling, $F_i(D_i) = D_i$, is a Bayesian Nash equilibrium. Without

loss of generality we consider user $i = 1$. His expected utility, given that the other users truthfully reveal their demand, equals:

$$\begin{aligned}\mathbb{E}[V_1(F;D_1)] &= \int_0^C \cdots \int_0^C \frac{\frac{F_1}{F_1+\sum_{j=2}^N D_j} C - \max\{F_1 - D_1, 0\}}{C^{N-1}} \mathrm{d}D_2 \cdots \mathrm{d}D_N \\ &= \frac{1}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{F_1}{F_1 + \sum_{j=2}^N D_j} \mathrm{d}D_2 \cdots \mathrm{d}D_N - \max\{F_1 - D_1, 0\}.\end{aligned} \tag{6.8}$$

To analyze when truth-telling maximizes this expected utility, we calculate the derivative with respect to F_1. The values of F_1 where the derivative equals zero or does not exist, and the boundary values 0 and C are candidate values for a maximum. If the derivative equals zero for some value F_1 then we use the second derivative of the expected utility to check whether this value is indeed a maximum or minimum. We begin with stating a preliminary result on these derivatives and their properties.

Theorem 6.3.1 *Consider the situation with N users. The derivative of the expected utility equals:*

$$\frac{\partial \mathbb{E}[V_1(F;D_1)]}{\partial F_1} = \frac{1}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{\sum_{j=2}^N D_j}{(F_1 + \sum_{j=2}^N D_j)^2} dD_2 \cdots dD_N - \mathbb{1}_{\{F_1 > D_1\}}, \tag{6.9}$$

where $\mathbb{1}_E$ is the indicator function of the event E that takes the value 1 if E is true and 0 otherwise. This derivative is positive if $F_i < D_i$; the expected utility is then increasing in F_i. The second derivative of the expected utility,

$$\frac{\partial^2 \mathbb{E}[V_1(F;D_1)]}{\partial F_1^2} = \frac{1}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{-2\sum_{j=2}^N D_j}{(F_1 + \sum_{j=2}^N D_j)^3} dD_2 \cdots dD_N, \tag{6.10}$$

is always negative. So, the derivative of the expected utility is decreasing in F_i, in particular for $F_i > D_i$.

Proof Without loss of generality let $i = 1$. If $F_1 < D_1$, then the derivative (6.9) reduces to:

$$\frac{1}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{\sum_{j=2}^{N} D_j}{(F_1 + \sum_{j=2}^{N} D_j)^2} \mathrm{d}D_2 \cdots \mathrm{d}D_N, \tag{6.11}$$

which is always positive; the expected utility is increasing in F_1. It is easy to see that the second derivative (6.10) is negative. Hence, the derivative (6.9) of the expected utility is decreasing in F_1, in particular for $F_1 > D_1$. ∎

According to this theorem, the expected utility is increasing if $F_1 < D_1$. Therefore, user 1 wants to set F_1 as large as possible. Because $F_1 < D_1$, user 1 sets $F_1 = D_1$ in the limit.

Also by Theorem 6.3.1 the derivative of the expected utility is decreasing in F_1. Now if this derivative is negative for all forecasts $F_1 > D_1$, then the expected utility is decreasing in F_1. So, user 1 wants to choose F_1 as small as possible. Because $F_1 > D_1$, user 1 wants to select $F_1 = D_1$ in the limit. In this case we conclude that truth-telling is a Bayesian Nash Equilibrium; user 1 always tells the truth.

In the next subsections we investigate for several numbers of users when the derivative of the expected utility for $F_i > D_i$ is indeed negative, and under which conditions truth-telling is an equilibrium.

Truth-telling in Case of Two Users

In this section we analyze the allocation problem with two users. Then the derivative (6.9) of the expected utility for $F_1 > D_1$ equals:

$$\int_0^C \frac{D_2}{(F_1 + D_2)^2} \mathrm{d}D_2 - 1 = \ln\left(\frac{F_1 + C}{F_1}\right) + \frac{F_1}{F_1 + C} - 2, \tag{6.12}$$

We want to know for which values of F_1 this derivative is negative. If so, then the expected utility of user 1 is decreasing and this user will select $F_1 = D_1$ — the truth-telling outcome — to maximize its expected utility.

Theorem 6.3.2 *Consider the situation with two users. Truth-telling is a Bayesian Nash equilibrium if the private demand of any user is at least* 18.9% *of the total capacity.*

Proof Without loss of generality consider user $i = 1$. By Theorem 6.3.1, the derivative (6.12) is a decreasing function in F_1. This derivative is negative for all requests $F_1 \in (D_1, C]$ if it is negative for $F_1 = D_1$:

$$\ln\left(\frac{D_1 + C}{D_1}\right) + \frac{D_1}{D_1 + C} - 2 \leq 0. \tag{6.13}$$

This inequality holds if $D_1 \geq b_2 C$ with $b_2 \approx 0.189$, where b_2 is such that the derivative (6.12) is equal to zero for $F_1 = b_2 C$. ∎

In other words, the private demand of either of the two users should be larger than roughly one-fifth of the capacity of the resource. The lower bound of 18.9% on the proportion of privately known demand to the resource's capacity may be too restrictive. What happens if this lower bound is not met for user i, so $D_i < 0.189C$? According to the analysis in the proof of Theorem 6.3.2 the expected utility of this user is maximal in forecast $F_i \approx 0.189C$. This forecast is larger than the actual demand D_i; user i overestimates its private demand.

Truth-telling in Case of Three and More Users

For three to six users, the results are as follows.

Theorem 6.3.3 *Truth-telling is a Bayesian Nash equilibrium for $N = 3$ users if the private demand of any user is at least 8.0% of the total capacity. For $N = 4, 5,$ and 6 users, truth-telling is a Bayesian Nash equilibrium.*

Proof First consider $N = 3$ users. Without loss of generality focus on user 1 and on the case $F_1 > D_1$. According to (6.9), the derivative of the expected utility of user 1 equals:

$$\begin{aligned} &\frac{1}{C}\int_0^C \int_0^C \frac{D_2 + D_3}{(F_1 + D_2 + D_3)^2} \mathrm{d}D_2 \mathrm{d}D_3 - 1 \\ &= \frac{2F_1}{C} \ln\left(\frac{F_1(F_1 + 2C)}{(F_1 + C)^2}\right) + 2\ln\left(\frac{F_1 + 2C}{F_1 + C}\right) - 1. \end{aligned} \tag{6.14}$$

We know from Theorem 6.3.1 that this expression is decreasing in F_1. Hence, truth-telling is a Bayesian Nash equilibrium if this expression is non-positive for $F_1 = D_1$,

$$\frac{2D_1}{C} \ln\left(\frac{D_1(D_1 + 2C)}{(D_1 + C)^2}\right) + 2\ln\left(\frac{D_1 + 2C}{D_1 + C}\right) - 1 \leq 0. \tag{6.15}$$

Numerical evaluation by Waterloo Maple version 14, reveals that this inequality holds if $D_1 \geq b_3 C$ with $b_3 \approx 0.080$.

For the situation with more than three users, the complexity of the derivatives (6.9) increases rapidly. We once again use Theorem 6.3.1 to establish that truth-telling is a Bayesian Nash equilibrium if the first derivative is non-positive for $F_1 = D_1$. Numerical evaluation reveals that the inequality is satisfied for N users for all $D_1 \geq 0$. Hence, truth-telling is always a Bayesian Nash equilibrium for four till six users. ∎

Hence, for a Bayesian Nash equilibrium in a situation with three users, we have a lower bound on the demand per user. Note that this bound is smaller than the bound in the situation with two users. The lower bound disappears if we consider at least four users. For situations with 7 users, we were not able to perform the necessary calculations within reasonable time limits. We therefore conducted a simulation study. We tested 10 cases, for I = 7-10, 12, 15, 20, 30, 50 and 100 departments. In each case, we used a fixed capacity C equal[1] to 2500, and randomly drew from a uniform $(0, C)$ distribution the forecast and demand values for $I - 1$ departments. Then for the remaining department i we checked whether it was optimal, given the utility function (6.7), to provide a forecast that was equal to the demand. We tested each case 1000 times, and for all 10x1000 = 10,000 instances truth-telling was an optimal strategy for the department we studied. Based on these results, we conjecture the following proposition.

Proposition 6.3.4 *If there are more than 6 users, then truth-telling is a Bayesian Nash equilibrium.*

Note that truth-telling is not a unique Bayesian Nash equilibrium, since there is another (trivial) Bayesian Nash equilibrium, namely $F_i = 0$ for all i. However, this is not of any practical value considering the problem setting.

6.3.2 Different Cost and Reward Parameters

In this section we return to the general utility function $V_i(F; D_i)$ without the restriction $x = y = 1$. We analyze what happens to the lower bounds on the actual demands of the departments, as stated in the theorems 6.3.2 and 6.3.3. The expected utility for user 1 now equals:

$$\mathbb{E}[V_1(F; D_1)] = \frac{x}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{F_1}{F_1 + \sum_{j=2}^{N} D_j} \mathrm{d}D_2 \cdots \mathrm{d}D_N - y(F_1 - D_1)^+. \tag{6.16}$$

The following theorem generalizes Theorem 6.3.1, and is therefore presented without proof.

[1]The value of C = 2500 is based on the average capacity of one MRI scanner per year, given that it operates 50 weeks/year for 10 hours/working day, processing scans that on average take one hour.

Theorem 6.3.5 *Consider the situation with N users. The derivative of the expected utility:*

$$\frac{\partial \mathbb{E}[V_1(F;D_1)]}{\partial F_1} = \frac{x}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{\sum_{j=2}^{N} D_j}{(F_1 + \sum_{j=2}^{N} D_j)^2} dD_2 \cdots dD_N - y \mathbb{1}_{\{F_1 > D_1\}}, \tag{6.17}$$

is positive if $F_i < D_i$; the expected utility is then increasing in F_i. The second derivative of the expected utility:

$$\frac{\partial^2 \mathbb{E}[V_1(F;D_1)]}{\partial F_1^2} = \frac{x}{C^{N-2}} \int_0^C \cdots \int_0^C \frac{-2\sum_{j=2}^{N} D_j}{(F_1 + \sum_{j=2}^{N} D_j)^3} dD_2 \cdots dD_N, \tag{6.18}$$

is negative. The derivative of the expected utility is decreasing in F_i, in particular for $F_i > D_i$.

First, consider $N = 2$ users. According to Theorem 6.3.5, the expected utility is increasing for $F_1 < D_1$. Hence, user 1 chooses $F_1(D_1) = D_1$ in the limit in case $F_1 < D_1$. If $F_1 > D_1$ then the derivative of the expected utility equals:

$$x \left(\ln \left(\frac{F_1 + C}{F_1} \right) + \frac{F_1}{F_1 + C} - 1 \right) - y, \tag{6.19}$$

which is a generalization of (6.12). Also by Theorem 6.3.5, this derivative is decreasing in F_1. Hence, if it is non positive for $F_1 = D_1$ then it takes negative values for all $F_1 > D_1$. This happens if $D_1 \geq b_2 C$ where the lower bound b_2 is a root of expression (6.19) after substituting $F_1 = b_2 C$. Thus b_2 solves:

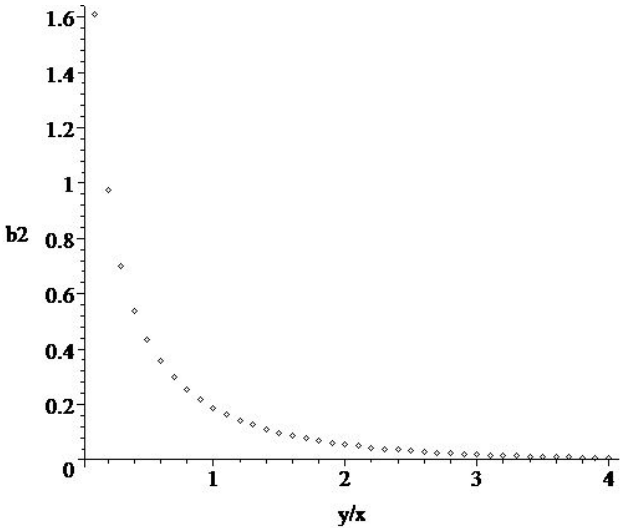
$$\ln \left(\frac{b_2 + 1}{b_2} \right) + \frac{b_2}{b_2 + 1} - 1 - y/x = 0. \tag{6.20}$$

This equation shows that lower bound b_2 is a function of y/x, the relative value of the 'cost' parameter y to the 'reward' parameter x (see Figure 6.2).

Observe that for $y/x = 1$ the lower bound b_2 agrees with the result in Theorem 6.3.2. If y/x increases then the penalty function with weight y becomes more and more important compared to the value of the allocated capacity with weight x. Since the user adds so much relative value to the penalty, truth-telling more and more easily becomes a Bayesian Nash equilibrium. The lower bound b_2 decreases, and in particular, b_2 tends to zero as y/x increases.

We perform the same analysis for situations with three and four users, see Figure 6.3. For three users, the lower bound b_3 is positive as long as $y/x \leq 1.3$. For larger values

Figure 6.2: Lower bound b_2 as a function of y/x.



of y/x there is no positive solution to (6.20). Thus, if $y/x > 1.3$ then truth-telling is always a Bayesian Nash equilibrium; there is no lower bound on the demand of the users to ensure an equilibrium. We observe the same for four users. The lower bound b_4 is positive for $y/x < 0.8$. For larger values of y/x truth-telling is always a Bayesian Nash equilibrium. The analysis for five and more users goes along the same lines, and is therefore omitted.

6.4 Results for Constrained Rules

In this section we analyze the effects of capacity allocation when using the constrained equal award rule or the constrained equal loss rule.

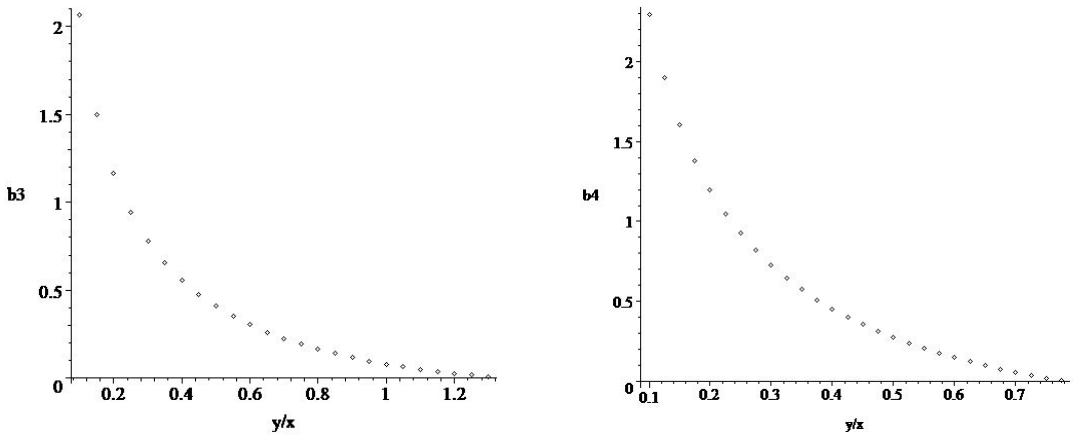
6.4.1 Results for Constrained Equal Award Rule

If the capacity is allocated according to the constrained equal award rule, then the utility function of user i is:

$$V_i(F; D_i) = x \min\{\alpha, F_i\} - y \max\{F_i - D_i, 0\}, \tag{6.21}$$

with α such that $\sum_j \min\{\alpha, F_j\} = C$. Consider the situation with two users. For simplifications we set $x = y = 1$ in the utility function. Without loss of generality consider user

Figure 6.3: The lower bounds $b_N C$ for $N = 3$ and $N = 4$ users as a function of y/x.



(a) Lower bound b_3 for 3 users. (b) Lower bound b_4 for 4 users.

$i = 1$. Assume the second user is truthful, $F_2(D_2) = D_2$. Then:

$$\alpha = \begin{cases} C/2, & F_1 \geq C/2,\ D_2 \geq C/2, \\ C - F_1, & F_1 < C/2 \leq D_2, \\ C - D_2, & D_2 < C/2 \leq F_1, \\ C/2, & F_1 < C/2, D_2 < C/2. \end{cases} \tag{6.22}$$

We investigate if and when truth-telling is a Bayesian Nash equilibrium.

Theorem 6.4.1 *Consider the constrained equal award rule and $N = 2$ users. Then truth-telling is a Bayesian Nash equilibrium.*

Proof First, if $F_1 \leq C/2$ then:

$$\begin{aligned} \mathbb{E}[V_1(F; D_1)] &= \int_0^C (\min\{\alpha, F_1\} - \max\{F_1 - D_1, 0\}) \frac{1}{C} \mathrm{d}D_2 \\ &= \frac{1}{C} \int_0^C F_1 \mathrm{d}D_2 - \max\{F_1 - D_1, 0\} \\ &= F_1 - \max\{F_1 - D_1, 0\}. \end{aligned} \tag{6.23}$$

Second, for $F_1 > C/2$:

$$\begin{aligned} \mathbb{E}[V_1(F; D_1)] = & \frac{1}{C} \int_0^{C-F_1} F_1 \mathrm{d}D_2 + \frac{1}{C} \int_{C-F_1}^{C/2} (C - D_2) \mathrm{d}D_2 \\ & + \frac{1}{C} \int_{C/2}^{C} \frac{1}{2} C \mathrm{d}D_2 - \max\{F_1 - D_1, 0\} \\ = & -\frac{F_1^2}{2C} + F_1 + \frac{1}{8} C - \max\{F_1 - D_1, 0\}. \end{aligned} \tag{6.24}$$

The derivative of the expected utility of user 1 is:

$$\frac{\partial \mathbb{E}[V_1(F;D_1)]}{\partial F_1} = \begin{cases} 1 - \mathbb{1}_{\{F_1 > D_1\}}, & F_1 < C/2, \\ -\frac{F_1}{C} + 1 - \mathbb{1}_{\{F_1 > D_1\}}, & F_1 > C/2. \end{cases} \tag{6.25}$$

First consider $D_1 \leq C/2$. If $F_1 \leq C/2$ then the maximal expected utility is D_1 for $F_1 \in [D_1, C/2]$. If $F_1 > C/2$ then the maximal utility in the limit for $F_1 \to C/2$ is also D_1. Hence, there are multiple best replies for user 1. Truth-telling is an equilibrium.

Second, consider $D_1 > C/2$. If $F_1 \leq C/2$ then the maximal expected utility is $C/2$ for $F_1 = C/2$. For $F_1 > C/2$ the maximal expected utility is $\frac{5}{8}C$ for $F_1 \in (C/2, D_1]$. Hence, given $F_2 = D_2$ the best reply of user 1 is to set F_1 such that $C/2 < F_1 \leq D_1$. Truth-telling is a mutual best reply. Therefore, truth-telling is a Bayesian Nash equilibrium. ∎

Under the constrained equal award rule, truth-telling is an equilibrium, but it is hard to determine the other equilibria. Furthermore, the analysis of the constrained equal award rule increases in complexity with the number of users. Therefore, the resource might prefer the proportional rule. For this reason, we restrict our analysis of this rule to the case of two users.

6.4.2 Results for Constrained Equal Loss Rule

When using the constrained equal loss rule, the utility function of user i is:

$$V_i(F;D_i) = x \max\{F_i - \beta, 0\} - y \max\{F_i - D_i, 0\}, \tag{6.26}$$

with β such that $\sum_j \max\{F_j - \beta, 0\} = C$. Consider a situation with two users. At first, we set $x = y = 1$ in the utility function for simplicity. Assume the second user is truthful, $F_2 = D_2$. Then:

$$\beta = \frac{1}{2}(F_1 + D_2 - C) \tag{6.27}$$

is the equal amount of loss for both users. We investigate if and when truth-telling is a Bayesian Nash equilibrium. Without loss of generality consider user 1.

Theorem 6.4.2 *Consider the constrained equal loss rule and $N = 2$ users. Then truth-telling is a Bayesian Nash equilibrium.*

Proof Since

$$\begin{aligned} \frac{1}{C} & \int_0^C \max\{F_1 - \beta, 0\} \mathrm{d}D_2 \\ &= \frac{1}{C} \int_0^{C-F_1} F_1 \mathrm{d}D_2 + \frac{1}{C} \int_{C-F_1}^{C} \frac{1}{2}(F_1 - D_2 + C) \mathrm{d}D_2 \\ &= F_1 - F_1^2/(4C), \end{aligned} \tag{6.28}$$

the expected utility of user 1 equals:

$$\mathbb{E}[V_1(F; D_1)] = F_1 - \frac{F_1^2}{4C} - \max\{F_1 - D_1, 0\}. \tag{6.29}$$

The derivative with respect to F_1 is:

$$\frac{\partial \mathbb{E}[V_1(F; D_1)]}{\partial F_1} = 1 - \frac{F_1}{2C} - \mathbb{1}_{\{F_1 > D_1\}}. \tag{6.30}$$

The expected utility is increasing for $F_1 \leq D_1$, decreasing for $F_1 > D_1$, and thus $F_i(D_i) = D_i$ is the unique best response. Truth-telling is a Bayesian Nash equilibrium. ∎

The constrained equal loss rule performs better than the proportional rule, since truth-telling is an equilibrium without a lower bound on the private demands of the users. Next, we consider situations with three users. We consider the expected utility of user 1 and assume that the users 2 and 3 tell the truth, $F_i(D_i) = D_i$, for $i = 2, 3$. To determine the value of the loss β that is equally divided, we consider several cases.

First, suppose that all users obtain a positive part of the capacity, $A_i > 0$ for all i. Then $F_1 - \beta + D_2 - \beta + D_3 - \beta = C$, or $\beta = (F_1 + D_2 + D_3 - C)/3$. Thus:

$$A_1 = F_1 - \beta = \frac{1}{3}(2F_1 + C - D_2 - D_3). \tag{6.31}$$

This amount is positive, $A_1 > 0$, if and only if:

$$D_2 + D_3 < C + 2F_1. \tag{6.32}$$

Similarly, $A_2 > 0$ if and only if:

$$-2D_2 + D_3 < C - F_1, \tag{6.33}$$

and $A_3 > 0$ if and only if

$$D_2 - 2D_3 < C - F_1. \tag{6.34}$$

Notice that at least two users should get a positive amount. If not, then one user gets all capacity, which can only happen if his request exceeds the other requests by more than the capacity C. This cannot occur since $0 \leq F_i \leq C$ for all users. Table 6.2 shows the diverse values of A_1 for the different cases that can occur. For reference, we numbered the cases from I to IV.

Theorem 6.4.3 *Consider the constrained equal loss rule and $N = 3$ users. Then truth-telling is a Bayesian Nash equilibrium.*

Table 6.2: Values of A_1 for Case I–IV

Case	True inequalities	A_1
I	(6.32), (6.33), (6.34)	$(2F_1 - D_2 - D_3 + C)/3$
II	(6.33), (6.34)	0
III	(6.32), (6.34)	$(C + F_1 - D_3)/2$
IV	(6.32), (6.33)	$(C + F_1 - D_2)/2$

Proof The expected utility for user 1 is:

$$\begin{aligned}\mathbb{E}[V_1(F; D_1)] &= \int_0^C \int_0^C (\max\{F_1 - \beta, 0\} - \max\{F_1 - D_1, 0\}) \tfrac{1}{C^2} \mathrm{d}D_2 \mathrm{d}D_3 \\ &= \tfrac{1}{C^2} \int_0^C \int_0^C \max\{F_1 - \beta, 0\} \mathrm{d}D_2 \mathrm{d}D_3 - \max\{F_1 - D_1, 0\}. \end{aligned} \tag{6.35}$$

We focus on the calculation of the first term for several values of F_1. First, if $F_1 = 0$ then the outcome of the double integral is also 0. Next, consider $0 < F_1 < C/2$. Taking into account the four cases in the table above, we obtain:

$$\begin{aligned}\frac{1}{C^2} & \int_0^C \int_0^C \max\{F_1 - \beta, 0\} \mathrm{d}D_2 \mathrm{d}D_3 \\ = & \iint_I (2F_1 - D_2 - D_3 + C)/3 \mathrm{d}D_2 \mathrm{d}D_3 + \iint_{II} 0 \mathrm{d}D_2 \mathrm{d}D_3 \\ & + \iint_{III} (C + F_1 - D_3)/2 \mathrm{d}D_2 \mathrm{d}D_3 + \iint_{IV} (C + F_1 - D_2)/2 \mathrm{d}D_2 \mathrm{d}D_3 \\ = & \frac{1}{36C^2}(2C^3 + 12C^2F_1 + 24CF_1^2 - 29F_1^3) + 0 + \frac{F_1^3}{6C^2} + \frac{F_1^3}{6C^2} \\ = & \frac{1}{36C^2}(2C^3 + 12C^2F_1 + 24CF_1^2 - 17F_1^3). \end{aligned} \tag{6.36}$$

Similarly, for $C/2 \leq F_1 \leq C$ we obtain:

$$\frac{1}{C^2} \int_0^C \int_0^C \max\{F_1 - \beta, 0\} \mathrm{d}D_2 \mathrm{d}D_3 = \frac{1}{36C^2}(24C^2F_1 - F_1^3). \tag{6.37}$$

For $0 \leq F_1 < C/2$, the derivative of this expected utility with respect to F_1 is:

$$\frac{1}{36C^2}(12C^2 + 48CF_1 - 51F_1^2) - \mathbb{1}_{\{F_1 > D_1\}}. \tag{6.38}$$

The first term is between 0 and 1 due to $F_1 \in [0, C]$. For $C/2 < F_1 \leq C$, the derivative of the expected utility with respect to F_1 is:

$$\frac{1}{12C^2}(8C^2 - F_1^2) - \mathbb{1}_{\{F_1 > D_1\}}. \tag{6.39}$$

Also here, the first term lies between 0 and 1. So, $F_1 = D_1$ is the unique maximum. In both cases, the expected utility is increasing for $F_1 \leq D_1$ and decreasing otherwise. Hence, $F_1 = D_1$ maximizes the expected utility. We conclude that truth-telling is a Bayesian Nash equilibrium. ∎

Once again, the constrained equal loss rule has truth-telling as an equilibrium. This result is better than the proportional rule, since now we have no lower bound on the private demand of the users. If there are more than three users, the complexity of the analysis grows rapidly. For N users we have to consider $2^N - N - 1$ special cases. Again, we use simulation to study these cases. In the simulation study for I = 4-10, 12, 15, 20, 30, 50 and 100 departments, truth-telling is an optimal strategy.

Different Cost and Reward Parameters

In this paragraph we analyze general utility functions with $x \neq y$. We are interested for which values of x and y truth-telling remains a Bayesian Nash equilibrium. Given the complexity of the analysis we restrict ourselves to the case with $N = 2$ users.

Theorem 6.4.4 *Consider the constrained equal loss rule, $N = 2$ users and weights $x \neq y$ in the utility function. Truth-telling is a Bayesian Nash equilibrium if $\frac{y}{x} \geq 1 - \frac{1}{2C}\min\{D_1, D_2\}$.*

Proof From the proof of the previous theorem, the expected utility of user 1 is:

$$\mathbb{E}[V_1(F; D_1)] = x\left(F_1 - \frac{F_1^2}{4C}\right) - y(F_1 - D_1)^+. \tag{6.40}$$

The derivative with respect to F_1 is:

$$x\left(1 - \frac{F_1}{2C}\right) - y\mathbb{1}_{\{F_1 > D_1\}}. \tag{6.41}$$

Hence, the expected utility is increasing for $F_1 \leq D_1$. User 1 has maximal expected utility in $F_1 = D_1$ if the expected utility is decreasing for $F_1 > D_1$. This occurs if $x(1 - D_1/(2C)) - y \leq 0$, or $y/x \geq 1 - D_1/(2C)$. The result follows since this inequality should hold for all users. ∎

6.5 Numerical Example

We illustrate the model with a numerical example, which is based on the experience of one of the authors while working as a hospital consultant. We return to the MRI scanner

example from the Introduction section, and consider four departments that each make requests for a specific scanning type, namely oncological (O), cardiovascular (C), neurological (N), and musculoskeletal (M). Capacity is distributed proportionally according to the requests, and the cost and reward parameters are both equal to 1, as in section 6.3.1. The MRI scan facility has a fixed capacity C of 1000 scans per month. In this example we chose to use the proportional allocation mechanism. Since we consider more than three departments, there is no lower bound on the demand of the users. Also, the proportional allocation rule is intuitive and easy to apply.

We start in month 1, and obtain the estimates of future demand (F_i). Recall that capacity is allocated by the Radiology department, having no knowledge on the actual demand D_i. The demand forecasts F_i and allocated capacities A_i are given in the first two columns of Table 6.3. At the end of month 1, the actual demand D_i is known. This information can be used to penalize the departments, if necessary. The other columns of Table 6.3 give the actual demand D_i, the deviation of the allocated amount A_i and forecast F_i from D_i, and the value of the utility function V_i.

Table 6.3: Month 1: forecast of demand F_i, allocated capacity A_i, actual demand D_i, deviation of D_i from F_i and A_i, and utility V_i

Department	F_i	A_i	D_i	$F_i - D_i$	% $F_i - D_i$	$A_i - D_i$	V_i
O	137	126	127	10	8%	-1	116
C	130	119	85	45	53%	34	74
N	630	578	623	7	1%	-45	571
M	193	177	238	-45	-19%	-61	177
All	1090	1000	1073	17	2%	-73	938

We see that in month 1 the waiting list increases with 107 MRI scans (scanning types oncological, neurological, and musculoskeletal), while there is unused capacity of 34 MRI scans (scanning type cardiovascular). Note that there is no penalty on the surplus demand related to the allocated capacity (i.e. $D - A$), since we only focus on the truth-telling element in the problem. In the example, it is implicitly assumed that surplus demand is lost. This lost demand could represent MRI scans that are performed at another institution, or not performed at all, because the physician decides upon another method of diagnostics.

We assume that the departments learn from the penalty given at the end of month 1 and therefore in month 2 provide an honest estimate (i.e. $F_i = D_i$ for all i). Without loss of generality, we assume that the actual demands of the departments in month 2 equal that of month 1. Capacity is again allocated proportionally to the requests. See Table 6.4 for results.

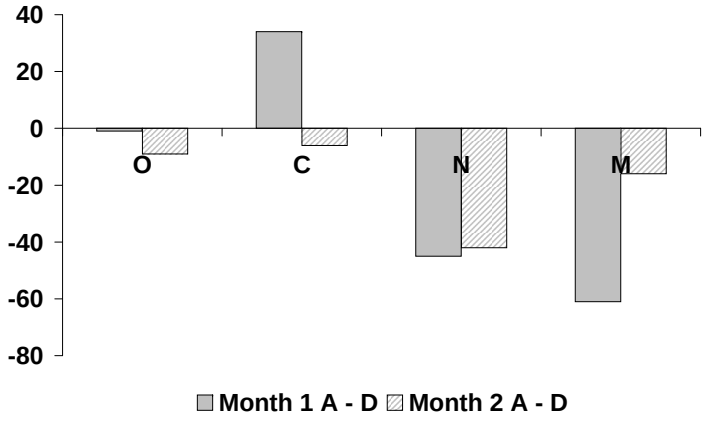
In month 2 the waiting list increases with 73 MRI scans, which equals the capacity shortage of $\sum_j D_j - C$, but there is no unused capacity. Figure 6.4 compares the difference between the allocated capacity A_i and actual demand D_i for both months. Furthermore,

Table 6.4: Month 2: forecast of demand F_i, allocated capacity A_i, actual demand D_i, deviation of D_i from F_i and A_i, and utility V_i

Department	F_i	A_i	D_i	$F_i - D_i$	% $F_i - D_i$	$A_i - D_i$	V_i
O	127	118	127	0	0%	-9	118
C	85	79	85	0	0%	-6	79
N	623	581	623	0	0%	-42	581
M	238	222	238	0	0%	-16	222
All	1073	1000	1073	0	0%	-73	1000

we see an increase in utility for all departments compared to month 1, while capacity is distributed more fairly.

Figure 6.4: Difference between allocated capacity A_i and actual demand D_i for both months.



6.6 Discussion

In this chapter we have studied a Radiology department (the resource) that has to distribute MRI scanning capacity among competing hospital departments (the users). Radiology uses forecasts of demand, provided by the hospital departments, to distribute the scanning capacity. The actual value of their demand is only known to the hospital departments. When the departments over- or underestimate the demand it can occur that the actual demand is less than the allocated capacity (i.e. the scanner sits idle) or the actual demand is larger than the allocated capacity. Both situations can arise simultaneously. In order to have a fair allocation, where all available capacity is actually used, Radiology should motivate the departments to provide an honest forecast of their demand.

We have introduced a generic approach to study the allocation of capacity to the users. Using a Bayesian game framework we show that under several capacity allocation

mechanisms it is an optimal strategy for each user to provide an honest demand forecast (the truth-telling equilibrium), and as a result the resource can fairly distribute the available capacity. When the number of users is small, certain restrictions on the relative size of the demands apply for the proportional allocation mechanism.

The penalty cost y on the surplus capacity requested by the departments will not be customary in most hospitals. However, hospitals are transforming to more professionally organized institutions. Incorporated in this transform is the usage of internal costing models in which departments reimburse each other for their services. The reimbursement also provides an incentive to make a more efficient use of available resources. Introducing a penalty cost will consequently become less involved than in the former traditional hospital organization.

Topics for further research would for instance be the reward users place on allocated capacity. Even though the three capacity allocation mechanisms are intuitively appealing, and satisfy the desired properties of an allocation mechanism as stated in section 6.2.2, other allocation mechanisms also might be of interest and may be better related to reality for some practical cases. Also, using a combinatorial auction to model the problem, as mentioned in the introduction section, could be a valuable extension. From an organizational point of view it would be appropriate to investigate the requirements for a successful implementation of the methodology.

We have shown that even with minor restrictions on the behavior of users, it is possible to attain a truth-telling equilibrium, where the shared resource is fairly allocated and all capacity is used.

Part III

Challenges Associated with Urgent Patient Flow

Chapter 7

Planning & Scheduling of Semi-Urgent Surgeries

7.1 Introduction

We consider a surgical department where elective, urgent and semi-urgent (synonym: semi-elective) patients are treated. An example of a department with such characteristics is a neurosurgery department. Urgent treatment is, among others, required for ruptured aneurysms, epidural or subdural hematomas, cauda equina syndrome, and (instable) spine fractures compromising the myelum or cauda equina. Semi-urgent pathologies include, among others, intracranial oncology, spine fractures with no or minimal neurological symptoms, drain dysfunctionalities, and disc herniations with unbearable pain or severe neurological deficits. Apart from these pathologies, the majority of neurosurgery patients do not require surgery within one or two weeks, and these are regarded as elective.

There is a definite trade-off between two major intertwined issues with respect to available surgical capacity: allocation of capacity to surgical departments and optimization of the surgical schedule within departments. On the one hand, when the target is minimal use of surgical resources, a more efficient surgical schedule may reduce the slack in the schedule, and therefore reduce the required capacity while keeping the societal costs due to patient cancellation and waiting constant. On the other hand, when the target is minimal societal costs due to patient cancellation and waiting, a more efficient surgical schedule may reduce these while keeping the allocated surgical resources constant. The trade-off is thus between societal costs and required surgical capacity. Allocating capacity to a surgical department usually is subject to additional constraints such as the restriction in the total available time, the time allocated to other departments, labor regulations (e.g., opening hours of the operating rooms), staff restrictions (e.g., available number of surgeons), and the possibility to handle exceptions (e.g., in over-time).

In this chapter we take the capacity allocated to a surgical department as a starting

point. We aim for robust patient scheduling schemes. We focus on the setting of a neurosurgery department treating urgent, semi-urgent and elective patients. Urgent patients are usually treated in a separate OR, but semi-urgent patients need to be fitted into the regular OR schedule. When a semi-urgent patient arrives, an elective patient is canceled to accommodate this (prioritized) patient. The cancellation of a surgery negatively affects the patient [171]. Medical professionals tend to feel sorry for the canceled patient and aim to reschedule the surgery as soon as possible. Thus, a canceled elective patient receives a semi-urgent status, and rescheduling this surgery possibly causes the cancellation of another elective patient. This knock-on effect results in a clear dependency between semi-urgent patient arrivals and cancellation of elective patients in subsequent weeks.

Several strategies are known from literature to cope with non-elective patients. One strategy is to reserve a small amount of time for emergency patients for whom surgery is required on the day of arrival in each elective patient OR [205], instead of dedicating one or several ORs to emergent cases [18]. Another possibility is to determine the elective patient schedule given the mean number of emergencies [72]. In all papers reviewed in [38], acute cases have to be performed at least on the day of arrival, as opposed to the semi-urgent surgeries that are studied in this chapter. In both [18] and [154] the authors distinguish between emergency surgeries (which have to be performed *now*) and urgent surgeries (which have to be performed within a day). In [72] and [118] stochastic programming is applied to support the scheduling of add-on cases, but in both papers these cases have to be completed on the day of arrival.

In [25] the authors start from a different viewpoint and determine, using a simulation model, how many elective cases can be performed in a dedicated orthopedic trauma OR. They state that when elective patients are willing to accept that their surgery might be canceled because of an incoming trauma patient, a higher throughput can be achieved. In [56] a trade-off is made between overtime and unused OR time. The paper has an operational viewpoint, by scheduling patients on an individual level. This is similar to the methodology presented in [98], where mathematical algorithms are used to schedule individual cases in available OR blocks.

The problem setting described here shows a similarity with the news vendor problem, where at the start of each decision period for that period the available capacity is matched with the required resources, and unmatched requests are discarded at the end of the period (see e.g. [56, 176, 134, 148] for news vendor problems applied to OR problems). The news vendor problem problem does not incorporate scheduling of discarded requests in subsequent periods, which is precisely the problem when elective surgeries are canceled and re-scheduled in subsequent periods. Modeling this knock-on effect is the natural domain of queuing theory. In this chapter, we therefore invoke the powerful theory of queues to analyze the cancellation rate of elective patients given a pre-specified surgical capacity, and the influence of canceling patients on the cancellation rate in the future.

For a surgical department with given capacity handling elective, urgent and semi-urgent patients, this chapter investigates reservation schemes of OR time for semi-urgent surgeries. As the arrival pattern of semi-urgent patients is unpredictable, the reserved OR may remain unused since elective patients cannot be scheduled so shortly before their surgery. We study the trade-off between cancellations of elective surgeries due to semi-urgent surgeries, and unused OR time due to excessive reservation of OR time for semi-urgent surgeries.

In the next section we first evaluate, using a queuing theory framework, the long run OR capacity needed to accommodate every incoming semi-urgent surgery. Second, we introduce another queuing model that enables a trade-off between the cancellation rate of elective surgeries and unused OR time. In Section 7.3 we develop a decision support tool, based on Markov decision theory, that assists the scheduling process of elective and semi-urgent surgeries. We demonstrate our results in Section 7.4 with actual data obtained from a department of neurosurgery, followed by the discussion and conclusion in Section 7.5.

7.2 Model and Long Term Behavior

The goal of the strategic model presented in this section is to provide an estimate for the amount of OR time that should be reserved for all semi-urgent surgeries in the long run. Therefore, we do not distinguish between the one- and two-week streams or take overtime into account. These components of the problem are discussed in the tactical model presented in Section 7.3. Obviously, dynamically adjusting the amount of reserved OR time according to the inflow of semi-urgent surgeries would result in little unused OR time. However, given hospital policy that dictates that elective patients should be planned weeks in advance, such an adaptive policy would impose canceling the elective patients that were planned in the claimed slots. In order to make the trade-off between cancellation of surgeries and unused OR capacity, a constant amount of OR time is reserved for semi-urgent surgeries. A summary of the notation used is listed in Table 7.1.

7.2.1 Assumptions and Model Parameters

The time available per OR day is divided into K slots of equal length. Surgeries can have a duration of $1, 2, .., K$ slots ($K < \infty$), and are categorized according to this duration. When a surgery has an mean duration of more than K slots, it is also included into the category of surgeries with length K slots. The total number of OR slots assigned to the department per week (m) equals the number of OR days per week multiplied by K. In order to accommodate semi-urgent patients, every week a fixed number of slots (s) is reserved ($0 \leq s \leq m$). Given the impact of the surgery on the patient and the

undesirability of performing semi-urgent surgeries in overtime, we assume, in line with medical practice (see the Introduction), that canceled elective patients become semi-urgent patients the following week. These patients need to undergo surgery within one week of their canceled surgery.

Progression of the Number of Semi-Urgent Slots

We focus on the number of semi-urgent slots waiting at the start of week n (W_n). This equals the amount of semi-urgent slots that arrived during the previous week (R_{n-1}) plus the elective slots that were canceled during the previous week in order to accommodate surplus semi-urgent slots. Elective slots are canceled if the reserved capacity for semi-urgent slots is insufficient. Recall that, in accordance with medical practice, the canceled elective slots from week n become semi-urgent slots in week $n+1$. Therefore, for our analysis of W_n, elective slots are not canceled, but instead the surplus semi-urgent slots from week n are transfered to week $n+1$. An example of the progression in the number of semi-urgent slots waiting at the start of week n is given in Figure 7.1.

Table 7.1: Notation introduced in Section 7.2

Symbol	Description
K	Number of slots available per OR day
m	Total number of slots assigned to department
s	Number of slots reserved for semi-urgent surgeries
W_n	Number of semi-urgent slots waiting for surgery at the start of week n
W	Number of semi-urgent slots waiting for surgery at the start of a week in a stationary regime
$\mathbf{q}$	Equilibrium distribution of W
$P_W(z)$	Generating function of W
λ	Arrival rate of semi-urgent surgeries
p_k	$\mathbb{P}$(Surgery is of length k slots), $k = 1, 2, ..K$
R_n	Number of semi-urgent slots that arrive during week n
$P_R(z)$	Generating function of the number of arrivals per week
N_e	Number of unused reserved semi-urgent slots per week
N_c	Number of canceled elective slots per week
C_e	Cost of one unused reserved semi-urgent slot
C_c	Cost of one canceled elective slot
C_t	Total Costs

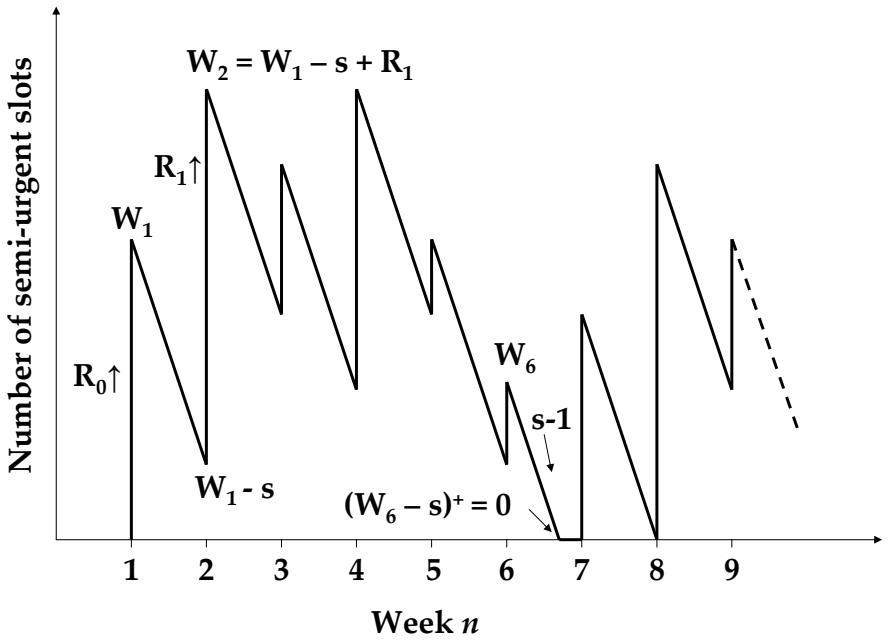
The Arrival Process

The number of arriving semi-urgent slots per week is equal to the sum of the number of slots per arriving patient. Patients arrive independently according to a Poisson process,

furthermore the number of slots per arriving patient is random. Therefore we can model the arrival process with the compound Poisson process [187]. The arrival rate of semi-urgent patients is λ. Let p_k denote the probability that an arriving semi-urgent surgery is of size k slots, $k = 1, .., K$. The generating function of the arrival process is [187]:

$$P_R(z) = \sum_{j=0}^{\infty} \mathbb{P}(R = j) z^j = e^{-\lambda(1-\sum_{k=1}^{K} p_k z^k)}, \quad \text{where} \quad \sum_{k=1}^{K} p_k = 1, \quad \text{and} \quad |z| \leq 1. \tag{7.1}$$

Figure 7.1: An example of the progression of the number of semi-urgent slots waiting at the start of the week ($s = 3$)



7.2.2 Stability of the System

From the description in subsection 7.2.1 (see also Figure 7.1) it is clear that the number of semi-urgent slots waiting at the start of week $n+1$ equals the number of semi-urgent slots that arrived during week n plus the number of surplus semi-urgent slots of week n:

$$\begin{aligned} W_{n+1} &= R_n + \{W_n - s\}^+, \quad n = 1, 2, \ldots \quad \text{and} \\ W_1 &= R_0, \end{aligned} \tag{7.2}$$

where $\{x\}^+ = 0$ if $x < 0$ and x otherwise. This is the Lindley equation for the sojourn time in a $GI/G/1$ queue [203]. The limit for $n \to \infty$ on W_{n+1} converges in distribution to W if $\mathbb{E}[R] < s$, and therefore we can conclude that as long as the mean weekly amount of semi-urgent slot arrivals, $\mathbb{E}[R]$, is strictly smaller than the number of slots allocated to semi-urgent surgeries, s, the system is stable and the capacity reserved for these slots should be sufficient on average. It follows that there is a minimum amount of capacity (s_{min}) that should be reserved for semi-urgent surgeries: $s_{min} = \lceil \mathbb{E}[R] \rceil$, where $\lceil x \rceil$ equals x rounded up to the nearest integer.

7.2.3 The Number of Semi-Urgent Slots Waiting

At the start of every week the state of the system is inspected. We represent the system by a slotted queuing model in discrete time [32]. We can distinguish between two situations: (1) more semi-urgent slots are waiting than can be completed in one week (epochs 2-6 and 9 in Figure 7.1), and (2) less (epoch 7 in Figure 7.1) or an equal amount of semi-urgent slots are waiting (epoch 8 in Figure 7.1) than can be completed. We obtain the following expressions for the transition probabilities:

$$\mathbb{P}(W_{n+1} = w_{n+1} | W_n = w_n) = \begin{cases} \mathbb{P}(R_n = w_{n+1} - w_n + s) & \text{if} \quad w_n - s > 0 \\ \mathbb{P}(R_n = w_{n+1}) & \text{otherwise.} \end{cases} \tag{7.3}$$

Define P as the matrix with transition probabilities. Let $\mathbf{q} = (q_0 \quad q_1 \quad \cdots)$ denote the equilibrium distribution of W, the number of semi-urgent slots waiting at the start of a week, where $q_i = \mathbb{P}(W = i)$. The q_i's can be computed as $\mathbf{q} = \mathbf{q}P$. An expression for the generating function of the equilibrium probabilities q_i is [32]:

$$P_W(z) \quad = \quad \frac{P_R(z) \sum_{i=0}^{s-1} q_i (z^s - z^i)}{z^s - P_R(z)}, \quad |z| \leq 1, \tag{7.4}$$

with $P_R(z)$ as given in (7.1). To obtain an exact expression for $P_W(z)$ we have to determine the s unknows $q_0, q_1, ..., q_{s-1}$. By Rouché's Theorem [110] it can be shown that the denominator of $P_W(z)$ has $s - 1$ zeros inside the unit disk [2]. Since $P_W(z)$ is a generating function and therefore bounded for all $|z| \leq 1$, the zeros of the denominator are zeros of the numerator as well [32]. Thus we obtain $s - 1$ equations for the s unknowns $q_0, q_1, ..., q_{s-1}$. To derive the last equation, we use that $P_W(1) = 1$. In order to find the $s - 1$ zeros of the denominator of $P_W(z)$, we start by solving:

$$z^s - P_R(z) \quad = \quad 0, \quad \text{which is equivalent to} \quad z^s = e^{-\lambda(1 - \sum_{k=1}^{K} p_k z^k)}. \tag{7.5}$$

We replace this equation by $s-1$ equations, where each z_j is a solution of the above equation [31]:

$$z_j = F(z_j)e^{2\pi\tilde{\imath}\frac{j}{s}}, \quad \text{with} \quad F(z) = e^{-\frac{\lambda}{s}(1-\sum_{k=1}^{K} p_k z^k)}, \quad \text{and} \quad \tilde{\imath} = \sqrt{-1}. \tag{7.6}$$

For each value of j ($j = 1, 2, ..., s-1$), we numerically solve this equation by using fixed point iteration [12]:

$$\begin{aligned} z_j^{(n+1)} &= F\left(z_j^{(n)}\right) e^{2\pi\tilde{\imath}\frac{j}{s}}, \quad n = 0, 1, ... \quad \text{and} \\ z_j^{(0)} &= 0. \end{aligned} \tag{7.7}$$

The z_j's that are found with this procedure are also zeros of the numerator of $P_W(z)$. We thus obtain $s-1$ equations for the unknowns $q_0, .., q_{s-1}$ that with the added equation $P_W(1) = 1$ define $P_W(z)$, $|z| \leq 1$.

7.2.4 Performance Measures

We are particularly interested in the mean number of canceled elective slots per work week ($\mathbb{E}[N_c]$), and the mean number of empty reserved semi-urgent slots per work week ($\mathbb{E}[N_e]$). For the latter it follows from (7.4) and $P_W(1) = P_R(1) = 1$ that:

$$\begin{aligned} \mathbb{E}[N_e] &= \sum_{i=0}^{s-1}(s-i)q_i \\ &= s - \mathbb{E}[R]. \end{aligned} \tag{7.8}$$

The mean number of elective slots that are canceled per week equals:

$$\begin{aligned} \mathbb{E}[N_c] &= \sum_{i=s}^{\infty}(i-s)q_i \\ &= \sum_{i=0}^{\infty} iq_i - s + \sum_{i=0}^{s-1}(s-i)q_i \\ &= P_W'(1) - \mathbb{E}[R]. \end{aligned} \tag{7.9}$$

Since

$$P_W'(1) = \mathbb{E}[R] + \frac{\sum_{i=0}^{s-1} q_i(s^2 - i^2 - s + i) - s^2 + s + P_R''(1)}{2(s - \mathbb{E}[R])}, \tag{7.10}$$

where

$$P_R''(1) = \lambda \sum_{k=1}^{K} k(k-1)p_k + \mathbb{E}^2[R], \tag{7.11}$$

we see that:

$$\mathbb{E}[N_c] = \frac{\sum_{i=0}^{s-1} q_i(s^2 - i^2 - s + i) - s^2 + s + P_R''(1)}{2(s - \mathbb{E}[R])}. \tag{7.12}$$

7.2.5 Cost Structure

Let C_e and C_c be the costs of one empty semi-urgent slot and one canceled elective slot. The expected total costs then equal:

$$\mathbb{E}[C_t] = \mathbb{E}[N_e]C_e + \mathbb{E}[N_c]C_c. \tag{7.13}$$

The optimal number of slots to reserve for semi-urgent surgeries (s^*) depends on the choice of C_e and C_c, and is the value of s that minimizes $\mathbb{E}[C_t]$.

7.3 Optimal Allocation of Surgery Slots

Given the stochasticity of the arrival process of semi-urgent patients, there will be weeks when the allocated capacity s^* is not sufficient. In this case the department can choose to perform the surplus semi-urgent patients this week, and cancel elective patients. On the other hand, the department can choose to postpone the semi-urgent surgeries until next week. A major drawback of this operational mode is that new semi-urgent patients arrive, who together with the postponed patients from this week, pose a huge demand on available resources. Furthermore, as mentioned in the Introduction, if the number of semi-urgent slots waiting for treatment exceeds the weekly amount of OR slots available, semi-urgent surgeries have to be performed in overtime, which is very undesirable as well. In this section we describe a Markov decision model that provides a scheduling strategy for surplus semi-urgent slots, given the parameters obtained with the queuing model. A summary of the additional notation introduced in this section is given in Table 7.2.

7.3.1 Assumptions

In this model we employ a more detailed view of the process, and consider the inflow of the two types of semi-urgent surgeries separately: the first type of semi-urgent surgeries need to be performed within one week, the second type of semi-urgent surgeries need to be performed within two weeks. Given the system status at the beginning of week n, we decide how many one- and two-week semi-urgent slots should be performed this week. Since one-week semi-urgent surgeries have to be performed *this* week, *all* incoming surgeries of this type are scheduled for *this* week. First the reserved slots $(1, 2, .., s^*)$ are used, and if additional one-week semi-urgent demand remains, elective slots are canceled. One-week semi-urgent demand that is still unaccommodated is performed in overtime. There are several options for scheduling two-week patients. A logical choice would be to first schedule all one-week slots, then schedule two-week slots in the reserved slots of this week that are still available. Subsequently, it has to be decided whether to perform the remaining two-week slots either this or next week. If the remaining two-week slots are scheduled for next week, no elective slots have to be canceled this week. On the other hand, postponed two-week semi-urgent slots have evolved into one-week semi-urgent slots the next week. The existence of these slots, together with newly arrived one-week semi-urgent slots, can result in a vast amount of semi-urgent demand that possibly has to be treated in overtime. In this section, a Markov decision model is presented that enables a trade-off between these two factors. For an overview of Markov decision theory, see [158]. In the model, we make the following assumptions:

- All one-week semi-urgent slots are planned this week.
- Two week semi-urgent slots not planned this week become one-week semi-urgent slots next week.
- Elective slots canceled this week become two-week semi-urgent slots next week.

7.3.2 The Markov Decision Model

We use a Markov decision model with infinite planning horizon to support the department in deciding how many two-week slots should be planned in a certain week (action a_n). The system state at the start of week n, $(n = 0, 1, ..., \infty)$, is given by $w_n = (w1_n, w2_n)$, where $w1_n$ and $w2_n$ are the number of one- and two-week semi-urgent slots waiting at that moment. The action chosen depends on the number of two-week slots waiting and on the part of capacity that is already allocated to one-week slots. Summarizing, the range for action a_n is determined by $(0, 1, .., \min(w2_n, (m - w1_n)^+))$.

Table 7.2: Additional notation introduced in Section 7.3

Symbol	Description
$W1_n$	Number of one-week semi-urgent slots waiting for surgery at the start of week n
$W2_n$	Number of two-week semi-urgent slots waiting for surgery at the start of week n
$w_n = (w1_n, w2_n)$	System state at start of week n
a_n	Action chosen in week n
$R1_n$	Number of one-week semi-urgent slot arrivals during week n
$R2_n$	Number of two-week semi-urgent slot arrivals during week n
λ_1	Arrival rate of one-week semi-urgent surgeries
λ_2	Arrival rate of two-week semi-urgent surgeries
p_{1k}	$\mathbb{P}$(one-week semi-urgent surgery is of length k slots), $k = 1, 2, .., K$
p_{2k}	$\mathbb{P}$(two-week semi-urgent surgery is of length k slots), $k = 1, 2, .., K$
$N_{e,n}$	Number of unused reserved semi-urgent slots during week n
$N_{c,n}$	Number of canceled elective slots during week n
$N_{o,n}$	Number of slots performed in overtime during week n
C_o	Cost of performing one slot in overtime
$C_{t,n}$	Total costs incurred in week n
α	Discount factor
δ^*	Optimal policy
δ_M	Monotone policy

Transition Probabilities

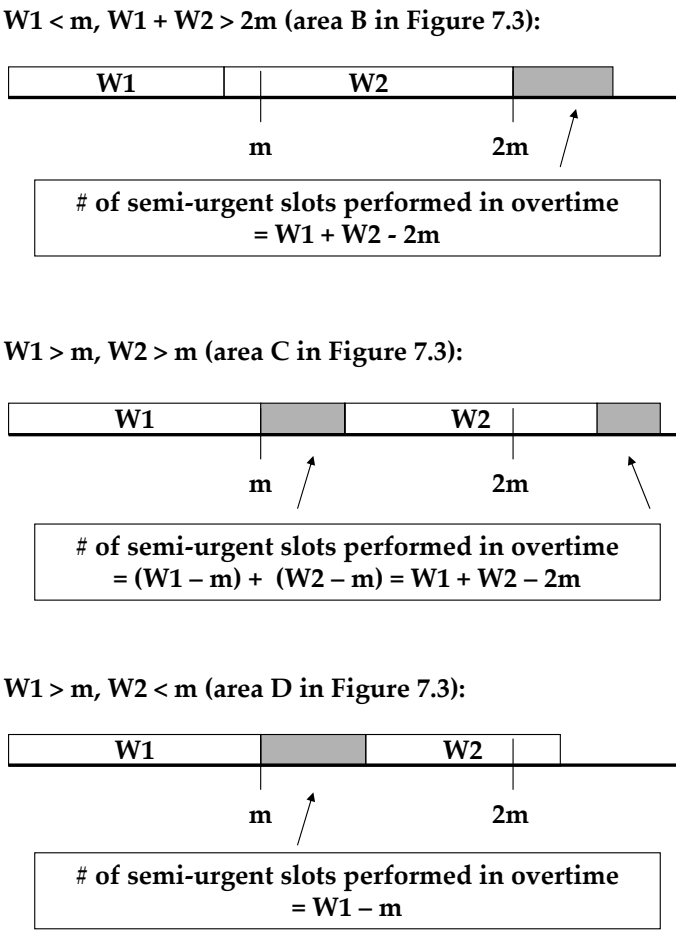
Let the random variables $R1_n$ and $R2_n$ denote the number of one- and two-week semi-urgent slot arrivals during week n, where $R1_n + R2_n = R_n$. Similarly to the queuing model presented in Section 7.2, $R1$ and $R2$ follow a compound Poisson distribution, with arrival rates λ_1 and λ_2, and p_{1k} and p_{2k} the probability that a one- and two-week surgery is of length k slots.

Recall that m slots are available each week for both elective and semi-urgent surgeries. Therefore, when the number of one-week semi-urgent slots waiting exceeds m, or when the sum of one- and two-week semi-urgent slots waiting exceeds $2m$, the surplus semi-urgent slots are performed in overtime. Figure 7.2 shows how the number of slots performed in overtime is calculated. In our model, we take into account the overtime by including (high) costs for each overtime surgery slot. However, the slots performed in overtime do not affect the system state, as they have left the system in the subsequent week. Thus, the state space A of the system is described as follows:

$$A = \{w = (w1, w2) : w1, w2 = 0, 1, \ldots; w1 \leq m; w1 + w2 \leq 2m\}. \tag{7.14}$$

The state space is depicted in Figure 7.3. The areas B, C and D and the arrows correspond to the three different cases of handling the overtime slots (see also Figure 7.2).

Figure 7.2: Number of semi-urgent slots performed in overtime: three different cases



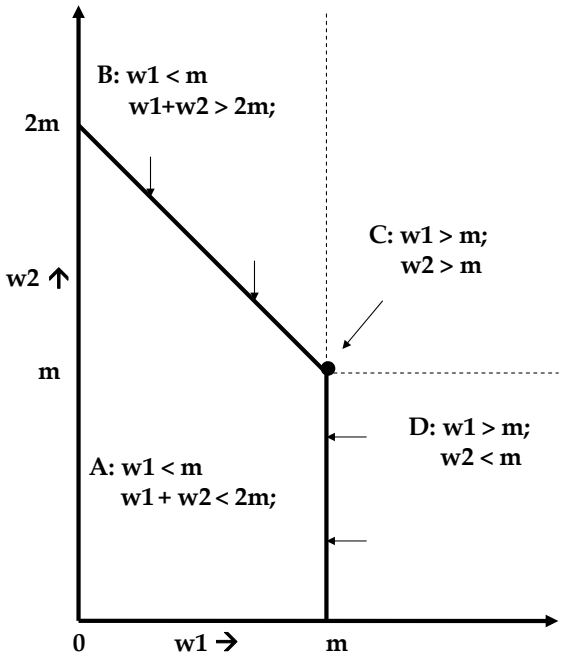
For notational purposes, let

$$\begin{aligned}\mathbb{P}(w_n|w_{n-1},a) = \quad & \mathbb{P}(W1_n = w1_n, W2_n = w2_n|W1_{n-1} = w1_{n-1};\\ & W2_{n-1} = w2_{n-1}; a_{n-1} = a).\end{aligned} \tag{7.15}$$

Now we define these transition probabilities for each $w_n \in A$.
If $w1_n < m$ and $w1_n + w2_n < 2m$ then no slots are performed in overtime in week n and thus we have:

$$\begin{aligned}\mathbb{P}(w_n|w_{n-1},a) = \quad & \mathbb{P}(R1_{n-1} = w1_n - w2_{n-1} + a)\\ & \times\mathbb{P}(R2_{n-1} = w2_n - (w1_{n-1} - s + a)^+),\\ & w1_n < m, w1_n + w2_n < 2m,\end{aligned} \tag{7.16}$$

Figure 7.3: State space of the system



and $(w1_{n-1} - s + a)^+$ is the number of canceled elective slots.
Now, assume that at the start of week n we have $w1$ one-week semi-urgent slots and $w2$ two-week semi-urgent slots waiting. If $w = (w1, w2) \in B$ then some slots have to be performed in overtime as explained above. Thus, according to the overtime policy depicted in Figure 7.2, the next state is given by $w1_n = w1$, $w2_n = 2m - w1_n$, a point on the boundary between A and B, as pointed out with arrows in Figure 7.3. Including this into transition probabilities, we derive:

$$\begin{aligned} \mathbb{P}(w_n|w_{n-1}, a) = \quad & \mathbb{P}(R1_{n-1} = w1_n - w2_{n-1} + a) \\ & \times \mathbb{P}(R2_{n-1} \geq w2_n - (w1_{n-1} - s + a)^+), \\ & w1_n < m, w1_n + w2_n = 2m. \end{aligned} \tag{7.17}$$

Analogously, if at the start of week n the number of waiting semi-urgent slots is described by $w \in C$, then the next state is $w_n = (m, m)$, and thus the transition probabilities for this state are given by:

$$\begin{aligned} \mathbb{P}(w_n|w_{n-1}, a) = \quad & \mathbb{P}(R1_{n-1} \geq w1_n - w2_{n-1} + a) \\ & \times \mathbb{P}(R2_{n-1} \geq w2_n - (w1_{n-1} - s + a)^+), \\ & w_n = (m, m). \end{aligned} \tag{7.18}$$

Finally, $w \in D$ will result in the state with $w1_n = m$, and we obtain:

$$\begin{aligned}\mathbb{P}(w_n|w_{n-1}, a) = \quad & \mathbb{P}(R1_{n-1} \geq w1_n - w2_{n-1} + a) \\ & \times \mathbb{P}(R2_{n-1} = w2_n - (w1_{n-1} - s + a)^+), \\ & w1_n = m, w2_n < m. \end{aligned} \tag{7.19}$$

Note that $\mathbb{P}(R1_n \leq x) = \mathbb{P}(R2_n \leq x) = 0$ if $x < 0$.

Performance Measures

The performance measures that were introduced for the queuing model are calculated on a weekly basis. Given the state $w_n = (w1_n, w2_n)$ and action a, the number of unused reserved semi-urgent slots and the number of canceled electives can be established as follows:

$$\begin{aligned} N_{e,n} &= (s - w1_n - a)^+, \quad \text{and} \\ N_{c,n} &= (w1_n - s + a)^+. \end{aligned} \tag{7.20}$$

Besides, we introduce a new performance measure, $\mathbb{E}[N_o]$ the mean number of semi-urgent slots that have to be performed in overtime next week as a consequence of the chosen action of this week. In week n, this amount depends on the number of slots at the start of week n, as described in Figure 7.2. Computing $E[N_{o,n+1}|w_n, a]$ works as follows:

$$\begin{aligned}\mathbb{E}[N_{o,n+1}|w_n, a] = \quad & \sum_{\substack{w1<m \\ w1+w2>2m}} (w1 + w2 - 2m) \\ & \times \mathbb{P}(R1_n = w1 - w2_n + a)\mathbb{P}(R2_n = w2 - (w1_n - s + a)^+) \\ & + \sum_{\substack{w1>m \\ w1+w2>2m}} (w1 + w2 - 2m) \times \mathbb{P}(R1_n = w1 - w2_n + a) \\ & \times \mathbb{P}(R2_n = w2 - (w1_n - s + a)^+) \\ & + \sum_{\substack{w1>m \\ w2<m}} (w1 - m) \\ & \times \mathbb{P}(R1_n = w1 - w2_n + a)\mathbb{P}(R2_{n-1} = w2 - (w1_n - s + a)^+). \end{aligned} \tag{7.21}$$

Cost Structure

The costs incurred for unused semi-urgent slots (C_e) and canceled elective slots (C_c) are equivalent with those introduced in Section 7.2. An extra cost, C_o, for performing one-

and two-week slots in overtime is introduced. The expected total costs incurred in week n equal:

$$\mathbb{E}[C_{t,n}] = \mathbb{E}[N_{e,n}]C_e + \mathbb{E}[N_{c,n}]C_c + \mathbb{E}[N_{o,n+1}]C_o. \tag{7.22}$$

7.3.3 Determining the Optimal Policy

In the process of coming to an optimal policy δ^* that defines an optimal action for each state w_n, we want to take into account the costs incurred today and in the future. However, we consider the costs experienced today as being more important than those experienced in the future. Therefore we use discount factor α, $\alpha \in (0, 1)$, in order to recalculate future costs to the cost level of today. Define $V_\delta(w_0)$ as the expected discounted costs over an infinite horizon, given initial state w_0:

$$V_\delta(w_0) = \mathbb{E}_\delta \left[\sum_{n=0}^{\infty} \alpha^n C_{t,n} \left(W_n, a_n \right) | w_0 \right]. \tag{7.23}$$

Let $V(w_0)$ denote the minimal value of $V_\delta(w_0)$:

$$V(w_0) = \min_{\delta} \quad V_\delta(w_0). \tag{7.24}$$

For each initial state w_0 and every action a, in an optimal policy it should hold that:

$$V(w_0) \leq \quad C_{t,0} \left(w_0, a_0 \right) + \alpha \sum_{w_1} \mathbb{P}(w_1 | w_0, a) V(w_1). \tag{7.25}$$

This gives us the optimality equation:

$$V(w_0) = \quad \min_{a \in \delta} \quad \{ C_{t,0} \left(w_0, a_0 \right) + \alpha \sum_{w_1} \mathbb{P}(w_1 | w_0, a) V(w_1) \}. \tag{7.26}$$

The optimal policy δ^* consists of the values of a that solve the optimality equation for each state. In order to find an optimal policy δ^*, we use the policy iteration algorithm [167]. Since the state and action space are finite, the policy iteration algorithm converges in a finite number of steps. Note that it is never optimal to perform two-week slots in overtime, since even if they are postponed and then cannot be treated in regular time, they can be treated in overtime next week as well.

7.4 Planning & Scheduling at a Neurosurgery Department

In this section we illustrate our modeling and optimization approach by considering a neurosurgery department situated in an academic hospital in the Netherlands. Department staff feared that dedicating scarce OR time to the uncertain stream of semi-urgent patients would lead to an excessive amount of unused OR capacity, and therefore decided to plan almost only elective patients in the available OR time. As a consequence, in daily operation, a large portion of elective surgeries was canceled in order to accommodate semi-urgent surgeries. Furthermore, many ad hoc decisions were needed to ensure that all patients would receive the care they needed. Supported by our models, we show possibilities for improvement.

All surgeries performed by the department can be characterized by the estimated OR time as follows: a) one third of an OR day, b) two thirds of an OR day, c) one OR day, and d) more than one OR day. With this in mind the OR day is divided into three slots of equal length ($K = 3$). Type 1 surgeries have an estimated duration of one slot, type 2 of two slots, and type 3 surgeries an estimated duration of three or more slots. Therefore, it is either possible to perform in one OR day i) three type 1 surgeries, ii) one type 1 and one type 2 surgery, or iii) one type 3 surgery. The department is assigned 8 OR days each week. With each day consisting of 3 slots, the department has 24 slots per week at its disposal (i.e. $m = 24$).

7.4.1 Data

The data needed for the model, semi-urgent patient arrivals, their mean surgery duration and semi-urgent state (i.e. surgery within one or two weeks) were recorded for a consecutive period of ten weeks. The characteristics of the arrival process are in line with the compound Poisson arrival process as outlined in [187]. Furthermore, the variance to mean ratio (vmr), defined as $\frac{\sigma^2}{\mu}$, which equals 1 for the Poisson distribution, shows that modeling the patient arrival process at this department with a Poisson process gives a conservative estimate for the aggregated semi-urgent patient stream (vmr = 0.25, so the variance is lower than would be expected from the Poisson distribution), while it provides a good estimate for the one-week semi-urgent patient flows (vmr = 1.03) and a slight conservative estimate for the two-week semi-urgent patient flow (vmr = 0.76). Therefore we feel confident that the compound Poisson process is an appropriate choice for modeling the arrival process of semi-urgent surgeries at this department. Table 7.3 gives the parameter values derived from the data, used in the queuing model. Since in the Markov decision model a distinction is made between one- and two-week semi-urgent surgeries, different parameter values for the compound Poisson process apply (Table 7.4). The cost parameters as defined in Section 7.2 and 7.3 should be determined by the department, and depend on the emphasis the department wants to put on either canceling patients or having an empty OR. For example, when $C_e = 10$

Table 7.3: Parameter values for queuing model (Section 7.2)

Parameter	Value
λ	$\frac{11}{2}$
p_1	$\frac{29}{55}$
p_2	$\frac{11}{55}$
p_3	$\frac{15}{55}$

Table 7.4: Parameter values for Markov decision model (Section 7.3)

Parameter	Value
λ_1	$\frac{31}{10}$
λ_2	$\frac{12}{5}$
p_{11}	$\frac{20}{31}$
p_{12}	$\frac{5}{31}$
p_{13}	$\frac{6}{31}$
p_{21}	$\frac{9}{24}$
p_{22}	$\frac{6}{24}$
p_{23}	$\frac{9}{24}$

and $C_c = 1$, having an empty semi-urgent slot is considered ten times worse than canceling one elective slot. Since the department considers performing semi-urgent slots in overtime as very undesirable, we emphasize on this by fixing C_o on 100. We consider three combinations for C_e and C_c (Table 7.5). For the department under consideration, CC_1 is a reasonable cost configuration. To demonstrate our methodology we also use two other cost configurations.

Table 7.5: Cost combinations

Name	C_e	C_c	C_o
CC_1	1	1	100
CC_2	10	1	100
CC_3	1	10	100

7.4.2 Determining the Required Number of Semi-Urgent Slots

We start by calculating the minimal amount of semi-urgent slots required (s_{min}), which is equal to $\lceil \mathbb{E}[R] \rceil$ (see Section 7.2.2). Since

$$\mathbb{E}[R] = \lambda \sum_{k=1}^{K} k p_k, \tag{7.27}$$

we have that $s_{min} = \lceil 9.6 \rceil = 10$. The department estimated that approximately 40% of surgeries performed during regular OR days is of the semi-urgent type, which is supported by the data ($\frac{9.6}{24} = 40\%$). Given that s may vary from s_{min} to m, we obtain the results from Table 7.6. The optimal value of $\mathbb{E}[C_t]$ for each cost combination is given in bold. Note the vast amount of canceled elective slots for $s = 10$. This shows that focusing on the average behavior of a system can result in unsatisfactory (and maybe unexpected) system outcomes. In Figure 7.4 $\mathbb{E}[N_e]$ and $\mathbb{E}[N_c]$ are compared graphically.

Table 7.6: Queuing model outcomes

s	$\mathbb{E}[N_e]$	$\mathbb{E}[N_c]$	$\mathbb{E}[C_t(CC_1)]$	$\mathbb{E}[C_t(CC_2)]$	$\mathbb{E}[C_t(CC_3)]$
10	0.40	23.81	24.21	27.81	238.54
11	1.40	5.42	6.82	**19.42**	55.64
12	2.40	2.50	4.90	26.50	27.36
13	3.40	1.37	**4.77**	35.37	17.14
14	4.40	0.82	5.22	44.82	12.58
15	5.40	0.51	5.91	54.51	10.47
16	6.40	0.32	6.72	64.32	9.61
17	7.40	0.21	7.61	74.21	**9.45**
18	8.40	0.13	8.53	84.13	9.72
19	9.40	0.08	9.48	94.08	10.25
20	10.40	0.05	10.45	104.05	10.94
21	11.40	0.03	11.43	114.03	11.74
22	12.40	0.02	12.42	124.02	12.62
23	13.40	0.01	13.41	134.01	13.54
24	14.40	0.01	14.41	144.01	14.48

We see in Table 7.6 that for CC_1 the optimal value of s^* equals 13 ($4\frac{1}{3}$ days), for CC_2, s^* equals 11 ($3\frac{2}{3}$ days), and for CC_3, s^* equals 17 ($5\frac{2}{3}$ days).

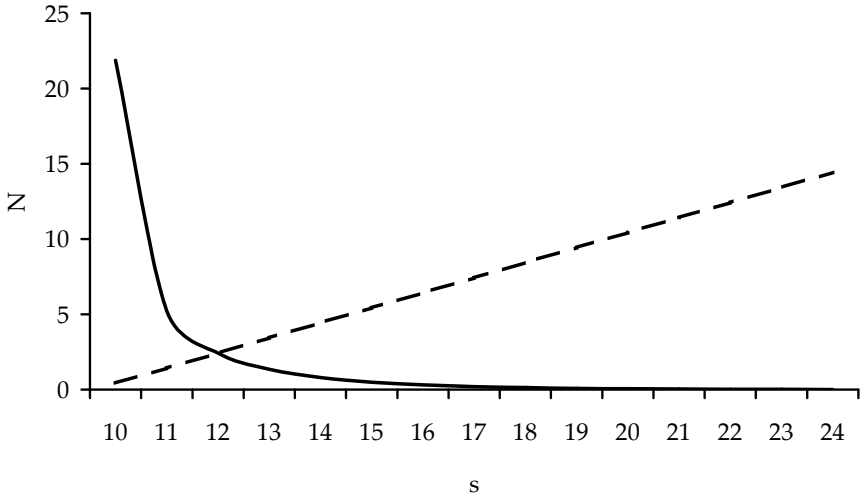
7.4.3 Allocation of Two-Week Semi-Urgent Slots

We now use the Markov decision model to schedule the one- and two-week semi-urgent slots. Our goal is to find an optimal policy that prescribes the number of two-week semi-urgent slots to plan, given any possible system state.

Monotone Policy

It is possible that in the optimal policy action a is not monotone increasing in $w2_n$. Although this form of the optimal policy is not uncommon in literature [114], it may be hard for medical professionals to implement. Therefore we proceed as follows. We determine an optimal policy δ^*, as described in Section 7.3.3. We then check whether a is

Figure 7.4: $\mathbb{E}[N_e]$ (interrupted line) and $\mathbb{E}[N_c]$ for $s = (\lceil s_{min} \rceil, .., m)$



monotone increasing in $w2_n$. If this is the case, we maintain this optimal policy. Otherwise, we create a monotone policy, δ_M, based on the optimal policy, where the number of two-week slots to plan (the chosen action) is not allowed to decrease. Such a monotone policy is not necessarily optimal, even in the class of monotone policies.

Obtained Policies

The cost combinations CC_1, CC_2, and CC_3 are used to obtain three policies from the Markov decision model. For cost combinations CC_1 and CC_3 we find monotone increasing optimal policies, given in Figures 7.5 and 7.7. For cost combination CC_2 a monotone policy was created, given in Figure 7.6. A discount factor of α = 0.95 is used in all cases. We find that $\mathbb{E}[C_t] = 4.01$ for CC_1, $\mathbb{E}[C_t] = 20.21$ for CC_2, and $\mathbb{E}[C_t] = 7.48$ for CC_3. The horizontal axes in the figures show the possible values of w_1 and w_2. When these are combined the system state is obtained. On the vertical axis the action that is chosen for each state is given. The set of actions for all possible states forms the policy δ. Recall that the action chosen consists only of the number of two-week semi-urgent slots to plan this week, since one-week semi-urgent slots are completed this week. While δ^* for CC_1 and CC_3 is straightforward - plan two-week slots up to s^* and postpone the remaining two-week slots until next week, the policy obtained for CC_2 is quite different. In several states it occurs that even when the number of one-week slots exceeds s^*, elective slots are canceled in order to accommodate two-week slots. This action is chosen to avoid overtime, a result of s^* being close to s_{min}. Similar to the queuing model outcomes, this shows that maintaining a cost structure similar to CC_2, which results in choosing an s^* which is close to $E[R]$, leads to the cancellation of elective slots.

Figure 7.5: δ^* for CC_1 ($s^* = 13$)

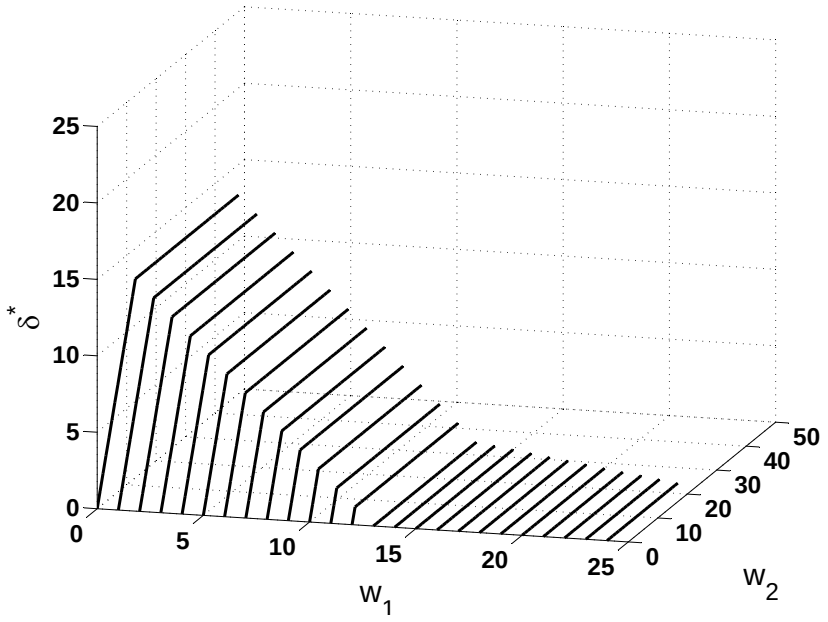


Figure 7.6: δ_M for CC_2 ($s^* = 11$)

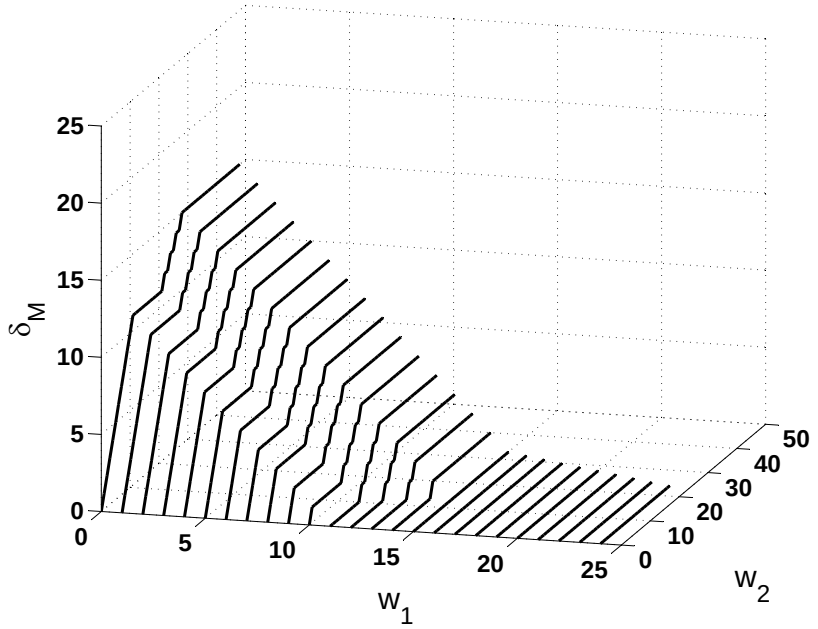
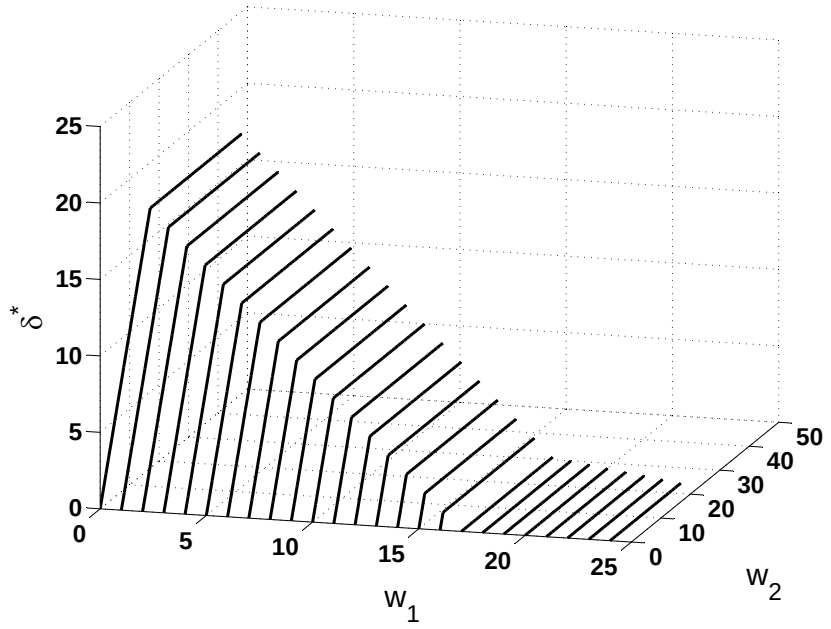


Figure 7.7: δ^* for CC_3 ($s^* = 17$)



7.5 Discussion

In this chapter we have developed a methodology to handle the semi-urgent patient flow at a surgical department. On a strategic level, we have determined the OR capacity needed to accommodate all semi-urgent patients on the long run, and we have described a queuing model that allows for a trade-off between the number of elective patients canceled and the amount of unused OR time. Given the amount of slots dedicated to semi-urgent patients, the distribution of the number of elective slots canceled, and the distribution of the number of unused semi-urgent slots can be derived with the queuing model, as is shown in Section 7.4. An insight that follows from these results is that focusing on only the average behavior of a system can result in undesired system outcomes, in this case the cancellation of many elective patients. Since semi-urgent patient arrivals and elective cancellations are dependent, even over consecutive periods, a natural modeling approach lies in the area of queuing theory.

On a tactical level, we have outlined a Markov decision model that supports the allocation of one- and two-week semi-urgent surgeries. This model provides a guideline for the weekly scheduling of semi-urgent patients. The policies obtained with the model can be transfered to a spreadsheet program and with little effort developed into a tool that is easy to use. The added value of the Markov decision model is that it simplifies the

scheduling task substantially. Note that all models can be used for arbitrary parameter values.

In the methodology presented, both models involve the planning and scheduling of individual slots. It is not taken into account that when a surgery takes more than one slot, all slots must be scheduled adjacently in the same OR on the same day. To quantify this effect, we calculated the mean number of semi-urgent slots treated for the example in the case study where $s^* = 13$. When considering all possible states, consisting of the number of one-, two- and three-slot semi-urgent surgeries waiting, the mean equals 8.86 when taking into account the adjacency requirement (i.e. in the situation where we have 4 full OR days of 3 slots and a single slot on another OR day). Note that in these calculations we assumed that a rational planner would aim to maximize the number of semi-urgent slots treated in the available time. Given that we consider an instance of the problem where s^* is relatively small, so there is little freedom to fill the OR days, the deviation of 7.7% from the value of 9.60 slots (calculated with the queuing model) will be smaller in most other (larger) instances of the problem. However, the adjacency requirement results in a slightly higher demand for semi-urgent slots.

A topic for further research would be to extend the presented methodology with an operational model that schedules individual surgeries. We consider the total OR time allocated to a surgical department by OR management as given. Of course, it is possible to establish the optimal amount of allocated OR time, and doing so first could result in a better performance. One of our other aims is to carry out an extensive data analysis to support an implementation of our methodology at the neurosurgery department discussed in Section 7.4.

Chapter 8

Implementation Study: Neurosurgery Planning

8.1 Introduction

Planning difficulties at the Neurosurgery department of LUMC in the winter of 2007/2008 initiated the study presented in Chapter 7. A large portion (≈40%) of the patient flow at this department was considered semi-urgent. Semi-urgent surgeries, to be performed soon but not necessarily today, pose an uncertain demand on available hospital resources, and interfere with the planning of elective surgeries. For a highly utilized OR, reservation of OR time for semi-urgent surgeries avoids excessive cancellations of elective surgeries, but may also result in unused OR time, since arrivals of semi-urgent patients are unpredictable. The queuing model we presented in Chapter 7 allowed for a trade-off between cancellations of elective surgeries due to semi-urgent surgeries, and unused OR time due to excessive reservation of OR time for semi-urgent surgeries. After discussing the model outcomes of a case study that represented their situation (see Section 7.4), the Neurosurgery department decided to implement the methodology, starting with the allocation of OR slots to semi-urgent surgeries.

In this chapter we study how the department implemented the methodology. We monitored the surgical planning process for a period of 25 weeks. We analyze the results in terms of canceled elective surgeries and the occupation of the OR. Next to that we identify factors that still complicate the planning process, and present general observations following from discussing the planning process with the department staff.

8.2 Methods

As follows from Section 7.4, at least 10 slots ($3\frac{1}{3}$ days) of OR time should be allocated to semi-urgent surgeries in order to avoid excessive growth of the waiting list. During the

monitoring of the planning process, it turned out that the amount of OR time allocated to semi-urgent surgeries varied per week and depended on the availability of OR time (which was variable as well). We therefore took the available OR time and amount of time allocated to semi-urgent surgeries as outcomes of this study.

The surgeries were planned as follows. During week n, the semi-urgent slots of week $n+1$ and $n+2$ were gradually filled with patients. Usually a temporary planning existed that changed frequently. The planning of week n would be finalized in week $n-1$, so that patients received notice about a week before. If an elective patient had to be canceled, the patient received the semi-urgent status. In the remainder of this chapter, when we refer to semi-urgent patients we do not include elective patients who received the semi-urgent status after a cancellation.

8.2.1 Data Collection and Analysis

The planning process was monitored for a period of 25 weeks. For all elective and semi-urgent surgeries performed during this period the following information was recorded:

- Name and hospital ID of the patient.
- Date of the surgery.
- Operating room the surgery was performed.
- Access time, which is in this case defined as time spent on the waiting list.
- Duration of the surgery.
- Elective or semi-urgent status of the patient.
- Number of times the surgery was canceled before admission.
- Number of times the surgery was canceled after admission.
- Any relevant additional information.

The available OR time was also recorded for each week. To determine the OR occupation for the Neurosurgery department, data on urgent surgeries performed in these ORs during regular OR hours was also collected.

The time between the first patient-doctor contact and the day of surgery was usually longer than the access time, since patients entered the waiting list when all preparational activities were completed.

8.2.2 Additional Measures

Next to the allocation of OR slots to semi-urgent surgeries, the department implemented additional measures to improve the planning process, namely:

- A part of OR time was divided in blocks and the blocks were dedicated to specific surgery types (beginning of 2010).
- A semi-urgent surgery of a specific type had also to be carried out during these blocks.
- A data manager was hired who did the OR planning, supervised by a neurosurgeon (October 2010).
- The department started collaborations with several hospitals in the neighborhood, to enable the exchange of patients, surgeries and surgeons between the hospitals and thus increase flexibility (November 2010).
- An effort was made to shorten the waiting list and make a treatment plan for each patient (mid 2011).

8.3 Results

We first present the results on the patient level, and then zoom in to the slot level. In the 25-week period, 265 elective and semi-urgent surgeries were performed. If the same patient had multiple surgeries on several occasions, these were counted separately. Of the 265 patients, 98 were semi-urgent (36.98%), and 167 were elective (63.02%). The department had a total of 161 OR days at its disposal. Initially, there were 166 OR days available, but due to public holidays (3 days) and a Neurosurgical conference (2 days), 5 days were canceled. Furthermore, the OR department also canceled one day but provided an extra day in another week as well. The available OR capacity fluctuated per week. Per week on average 6.44 days (SD: 1.29) were available, instead of the initial value of 8 days from the case study in Section 7.4. During week 9 – 17 the OR was partly closed because of the summer holiday and so the available capacity was less: 6.11 days (SD: 0.78, 95%CI: 5.57;6.65). Outside the summer holiday the available capacity was 6.63 days (SD: 1.50, 95%CI: 5.92;7.34).

8.3.1 Elective Patient Cancellations

A total of 31 (18.56%) elective patients included in the study were canceled one or several times for surgery. An additional 3 patients were canceled during the monitoring

period and subsequently the indication for the surgery was not present anymore. If patients were canceled, their surgery was on average rescheduled within 12.29 days (SD: 13.71). The high standard deviation shows that for many patients the time between the cancellation and the day of surgery was longer than the average of two weeks. Specific requirements on the availability of surgeons who needed to perform the surgery usually made the rescheduling more complicated. Table 8.1 shows for the 31 patients how often and how they were canceled. A cancellation after admission usually resulted in the patient being discharged and subsequently re-admitted a day prior to the new surgery date.

Table 8.1: Cancellation mode and occurrence

Cancellation mode	Number of patients	Percentage
Before admission, single cancellation	9	29.03%
Before admission, multiple cancellations	2	6.45%
After admission, single cancellation	9	29.03%
Combinations		
Before admission, single cancellation & After admission, single cancellation	5	16.13%
Before admission, multiple cancellations & After admission, single cancellation	1	3.23%
Unknown	5	16.13%

All cancellations were related to the planning of additional semi-urgent patients. The OR department also keeps track of surgery cancellations; usually this is also presented as a quality indicator for the hospital. However, in this registration system a cancellation is only recorded when it is 24 hours or less before the intended day of surgery. In this study we view cancellations from a patient perspective and incorporate the cancellation of all surgeries the patient had received notion of, regardless of the time between the cancellation and the intended surgery date.

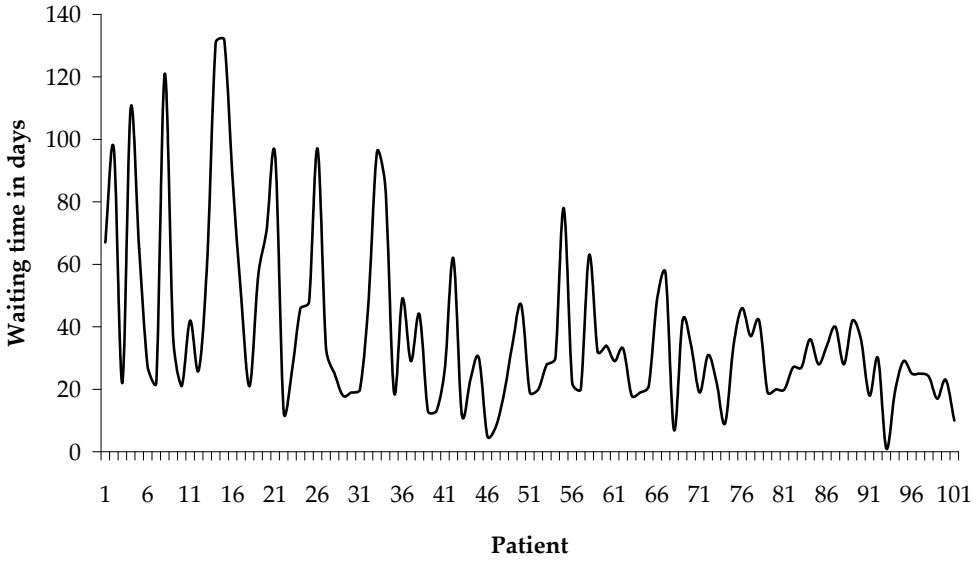
8.3.2 Access Time

To provide a complete overview of the planning process we also present results on the access time. For 12 patients (3 semi-urgent and 9 elective) the access time could not be recorded. The mean access time was 8.56 days (SD: 9.91) for semi-urgent patients. In the case of 8 semi-urgent patients, the access time was prolonged due to patient related factors: for example, the condition of an elective patient deteriorated and consequently the patient obtained a semi-urgent status. Some patients used medication that could affect the outcome of surgery adversely and their surgery had to be postponed, or the patient was too ill for surgery. If these patients are excluded, the mean access time for semi-urgent patients decreases to 6.77 days (SD: 6.92 days).

For the elective patients the mean access time was 86.60 days (SD: 135.32), which is approximately 12.5 weeks. It is difficult to estimate the expected access time for an ar-

bitrary elective patient who is still on the list, since this can only be calculated retrospectively for patients who already received surgery. The department put a lot of effort in shortening the waiting list and therefore 8 patients, who were on the waiting list for over a year, were called in again at the outpatient clinic and subsequently had surgery (if still necessary). If these patients are excluded, the mean access time for elective patients reduces to 60.07 days (SD: 52.60), which is approximately 8.5 weeks. The department's focus on the waiting list is also apparent from Figure 8.1 which shows that for the patients who arrived and had surgery during the monitoring period, the access time gradually decreased during the monitoring period.

Figure 8.1: Access time for new patients in chronological order



8.3.3 Elective Patient Cancellations and Unused OR Time on a Slot Level

We now zoom in to the slot level. Table 8.2 gives the results per week in terms of the number of canceled elective slots and unused OR time (also given in slots). Recall that one OR day consists of 3 slots.

For 4 elective and 5 semi-urgent patients the surgery duration was not recorded. We assumed that the duration of their surgery was equal to the mean duration of surgeries of the patients for whom the duration was recorded (2.10 slots for elective patients and 2.09 slots for semi-urgent patients). The weeks for which we had to make this assumption are depicted with $*$ in the Table. We see that the cancellations of elective slots are relatively low (16.43%), and the number of unused OR slots is minimal (1.26%). The

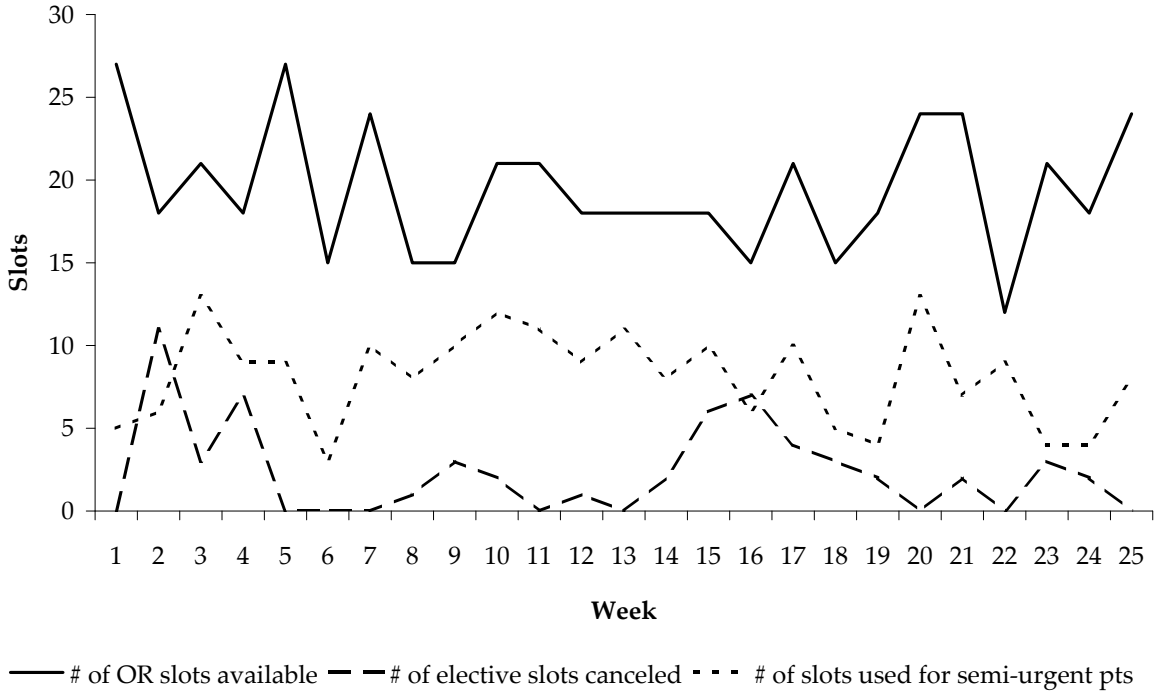
Table 8.2: Results for week 1 – 25 in slots (slots of urgent surgeries that were performed in the regular OR hours are also given. All values in the table were calculated using the actual duration of the surgeries. In case of the canceled elective slots we used the duration of the (first canceled) surgery once it was performed)

Week	OR capacity	Completed elective	Completed semi-urgent	Completed urgent	Overtime	Canceled elective	Unused OR capacity
1	27	24	5	0	2	0	0
2	18	13	6	0	1	11	0
3	21	11	13	0	3	3	0
4	18	12	9	0	3	7	0
5	27	23	9	2	7	0	0
6	15	12	3	3	3	0	0
7	24	16	10	0	2	0	1
8	15	10	8	0	3	1	0
9	15	6	10	0	1	3	0
10	21	12	12	0	3	2	1
11	21	13*	11	0	3	0	1
12	18	16	9	0	7	1	0
13	18*	11	11	0	4	0	1
14	18	11	8	0	1	2	0
15	18	14	10	2	8	6	1
16	15	10	6	3	4	7	0
17	21	14	10	0	3	4	0
18	15	14	5	0	4	3	0
19	18	15	4	0	1	2	0
20	24	9*	13	3	1	0	0
21	24*	25*	7	0	8	2	0
22	12	6	9	0	3	0	0
23	21	21	4	0	4	3	0
24	18*	18*	4	0	4	2	0
25	24*	23*	8	0	7	0	1
Total	486	359	204	13	90	59	6
Mean	19.44	14.36	8.16	0.52	3.60	2.36	0.24
SD	3.98	5.32	2.91	1.08	2.20	2.81	0.44

cancelation percentage is slightly lower than on a patient level; this is caused by dividing the surgeries into slots. On the other hand, the mean overtime per week is 3.60 slots, which adds up to more than 1 OR day, so that part of the low cancellations were realized at the cost of overtime. On average there are 8.16 semi-urgent slots completed and 2.36 elective slots canceled which receive the semi-urgent status as well. This adds up to a total of 10.52 slots per week that cannot be used to for elective surgeries, on average approximately half of the capacity.

Figure 8.2 gives a graphical representation of the results per week.

Figure 8.2: Graphical representation of results per week



8.3.4 General Observations

During the monitoring period we also made a number of general observations. Patients who were on the waiting list for over a year were difficult to plan, for example because the information on their health status was not up to date anymore. Since there are usually enough other patients to plan, these patients are postponed time after time. The department had to make a real effort to make a treatment plan for these patients so that they could ultimately be removed from the waiting list. Patients have a lot of influence themselves on the date of their surgery. For patients frequently calling about their pain or their waiting time, an effort would be made to plan their surgery earlier. Some patients postpone their surgery themselves; for example they first want to go on a vacation, have work obligations or have doubts whether they want the surgery or not. The latter issue is also related to the moment the patient receives notice of the surgery. This is usually quite late (a week or less prior to the surgery date) and for some patients then the time to (mentally or physically) prepare themselves is too short. On the other hand, the late moment of informing the patient gives a lot of flexibility in the planning

process. It allows the planner to reschedule elective patients on numerous occasions, without them being aware of it. The continuous arrivals of semi-urgent patients, which requires frequent rescheduling of elective patients, demands that the planner is on top of the process. Since the planning was standardized to some extent, it could be performed by an administrative employee instead of a neurosurgeon.

8.4 Discussion

In this chapter we studied how the Neurosurgery department of LUMC implemented the methodology presented in Chapter 7. The aim of the methodology was to determine the amount of OR slots to dedicate to semi-urgent surgeries, by making a trade-off between cancellations of elective surgeries due to semi-urgent surgeries, and unused OR time due to excessive reservation of OR time for semi-urgent surgeries. Next to the queuing model that enabled this trade-off, the methodology consisted of a Markov model that allowed to distinguish between semi-urgent surgeries that had to be performed within one or two weeks. The department started with the implementation of the queuing model results and will later on decide on implementation of the policies obtained with the Markov model (see Subsection 7.4.3).

Implementation studies are always challenging. The real world is not a model, and many factors that influence the outcomes of an implementation somehow cannot be incorporated. However, from the results we can conclude that the department, supported by a few additional measures, succeeded in implementation of the methodology. Even though the OR capacity turned out to be less than expected, the number of canceled elective slots was kept to a minimum while the available OR capacity was highly utilized (98.74%). It should be noted that many elective and semi-urgent surgeries were performed in overtime, which likely also contributed to these results. The continous devotion of the administrative employee to improve the OR planning was essential. It might be worthwile to consider sharing the planning task among two employees, to guarantee continuity in case of holidays, illness and resignation.

By maintaining a temporary schedule to the last possible moment, the number of elective patient cancellations could be kept to a minimum. When the time between informing the patient on the surgery date and the actual surgery is long, the patient has more time to prepare for the surgery. However, this reduces the flexibility for the planner to move the patient around the OR schedule and will increase the number of elective patient cancellations. The Neurosurgery department considered being canceled as worse for the patient than being called in late. The focus in this study was on the cancellation of elective patients. Some semi-urgent patients were canceled as well, but rescheduled quickly. Given the urgent character of the procedure, most patients understand the planning their surgery is more involved and the date might be subject to change.

The shortening of the waiting list made the pool of patients to choose from smaller. This

reduced flexibility and made the planning process more difficult. We see that, especially in a smaller department, there is yet another trade-off: a long waiting list versus an involved planning and possibly unused OR time. The introduction of dedicated blocks for specific surgery types lead to additional planning challenges. When there are many patients to choose from, it is not hard to find a patient to fill a block. When the number of patients with a specific surgery type on the waiting list decrease, it might become a problem. An option is to rethink the size, number and composition of the blocks.

When we started this study, we were not sure whether to incorporate the summer holiday. It turned out that the summer holiday was, in terms of available capacity, not significantly different from the weeks in the monitoring period before and after. The only difference was that the capacity during the summer holiday was more constant. It is always difficult in these kind of studies to find a period that represents the normal situation and is long enough for a reliable data analysis. These results suggest that it might be worthwhile to just take an arbitrary period and consider that as normal.

Chapter 9

The Emergency Observation and Assessment Ward

9.1 Introduction

Over the last period, the ED has become more and more crowded, resulting in among others an increased LOS and prolonged waiting times for patients. Also, ED crowding may result in increased mortality rates and lower quality of care [95]. These problems are not only caused by an aging population [166], a higher demand for acute care [137], and the inability to transfer patients to inpatient beds [68, 137], but also by hospital restructuring leading to fewer inpatient beds and more ambulatory care [172].

There are several measures hospitals can take in order to improve ED patient flow [138]. A recent development is the creation of an Emergency Observation and Assessment Ward (EOA Ward). The definition and purpose of such wards varies across hospitals. Also, in literature a consistent definition seems to be lacking. The review papers [50] and [173] provide a comprehensive overview of definitions and concepts for EOA Wards. Patient types that can be admitted vary, for example sometimes only medical patients are considered [173]. Patients that need intensive care are usually excluded [50, 173]. At an EOA Ward, patients are temporarily (less than 24 hours, 24-48 hours) hospitalized until a bed at an inpatient ward becomes available. ED patients who have to wait for test results or require observation for a short period of time can also be admitted. Given the close monitoring, a staffed bed at this location is usually more expensive than a bed at a regular inpatient ward.

The success of an EOA Ward depends on the overlying organizational structures, together with clear agreements upon transfers to regular inpatient wards, a well-defined chain of command, and access to specialist consultations [50, 173]. Currently, little evidence is available that operating an EOA Ward reduces ED crowding (see [50, 95, 138, 173] and the references therein). This is not only related to the ambiguity in the terminology and definitions of the EOA Ward used in practice, but could also be caused by

a lack of management information. This makes it very hard to measure the effects of opening an EOA Ward on the ED patient flow. Furthermore, there may be a publication bias since it is common to report only positive experiences [173]. This chapter aims at filling this gap in literature by providing a clear model taking into account all relevant patient flows, which allows for a quantitative analysis of the benefits of an EOA Ward.

The patient flow between the ED and inpatient wards is in most hospitals organized as follows. Patients arrive at the ED, receive treatment and are then either discharged, admitted, or might die. Alternatively, patients can be admitted at another hospital. While some transfers to other hospitals are necessary, for example because the other hospital is specialized in the type of care the patient needs, other transfers are inevitable since there are no inpatient beds available at the current location. The process of finding a bed and subsequently waiting for transport can easily take several hours. During this time the patient usually occupies an ED room, resulting in fewer ED capacity and substantial delay for patients in the waiting room. Additionally, ED treatment is expensive compared to inpatient care. It is therefore also financially attractive for hospitals to continue the care process at one of their own inpatient wards, instead of transferring the patient to another hospital after ED treatment.

Since the inpatient wards admit elective patients as well, it might be difficult to set aside inpatient beds for urgent patients whose arrival is uncertain. At an EOA Ward this is avoided since only urgent and observation patients from the ED are admitted. The maximum LOS at the EOA Ward is usually short, with regular transfers to the inpatient wards. Transfer moments can be fixed (for example twice a day) or patients are transferred immediately when a bed becomes available. In addition to the elective and urgent patient flows, the inpatient wards also receive patients from the ICU.

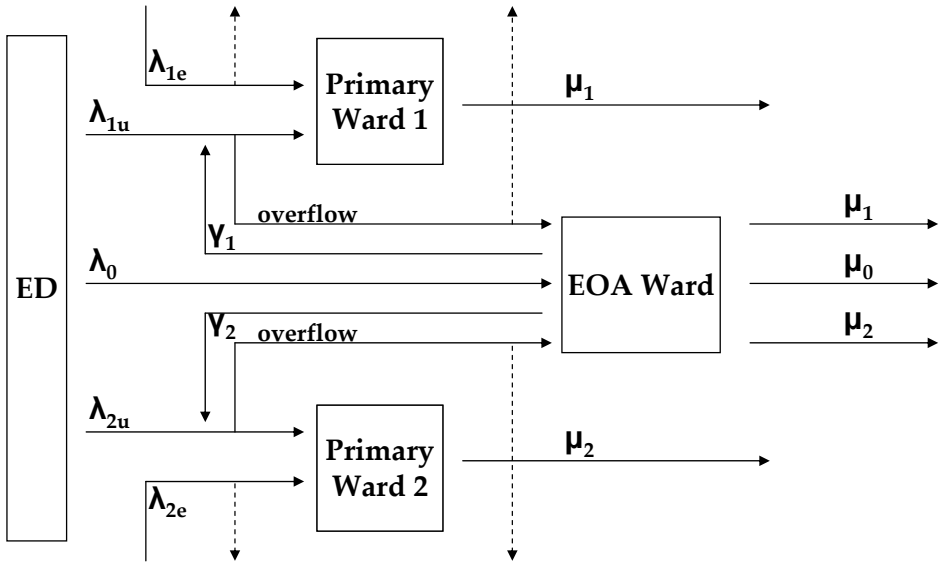
In this chapter we present a queuing model (Section 9.2) that can be used to evaluate the effect of employing an EOA Ward. We analyze a small example in Section 9.3 and calculate performance measures such as the number of elective and urgent inpatient admissions. Many work in improving ED patient flow has been done using simulation techniques (see e.g., [13, 44, 61, 96]). Fewer examples use queuing theory [47, 83, 133, 162]. Even though the EOA Ward has been subject of research quite often in the last decade, we were not able to find an analytical evaluation of its effect in terms of inpatient admissions as we present here.

9.2 Model

In this section we describe our mathematical model, which is based on Wilkinson's Equivalent Random Method (ERM) [200], a methodology developed to analyze overflow systems. The inpatient wards and EOA Ward can also be modeled as an overflow system, where the inpatient wards are the primary wards (i.e., the wards that generate the overflow of urgent patients) and the EOA Ward is the overflow ward where urgent

patients are routed if the inpatient ward is full. We have I primary wards, with capacity $c_i, i = 1, \ldots, I$. We assume that the LOS at ward i is exponentially distributed with mean μ_i^{-1}, where the LOS for elective and urgent patients at ward i is the same, but the LOS per ward may be different, so that $\mu_i \neq \mu_j$ for $i, j \in I$, $i \neq j$. Urgent patients arrive at primary ward i with rate λ_{iu}. If all beds at the primary ward are occupied, the urgent patient is routed to the EOA Ward. If the EOA Ward, which has a capacity of c_0 staffed beds, is fully occupied as well, the patient is blocked (again) and leaves.

Figure 9.1: ED – primary ward – EOA Ward patient flow; example with two primary wards



All urgent patients at the EOA Ward have the same exponential distributed LOS with rate μ_{over}. Urgent patients who originated from primary ward i are transferred from the EOA Ward to primary ward i with rate γ_i. The LOS at primary ward i is for these patients again exponential with mean μ_i^{-1}. Patients directly routed from the ED to the EOA Ward arrive with Poisson rate λ_0 and have an exponentially distributed LOS with mean μ_0^{-1}. Elective patients are blocked when the primary ward is full. The elective patient demand at the primary wards, which also incorporates patients from the ICU, is modeled with a Poisson process with rate λ_{ie}. Although elective arrivals are scheduled, random fluctuations in the number of scheduled arrivals make the Poisson assumption plausible [28]. Figure 9.1 summarizes our overflow system.

9.2.1 Global Balance Equation

We denote the number of elective and urgent patients present at primary ward i with n_{ie} and n_{iu} resp. The number of urgent patients from primary ward i present at the EOA Ward is given by n_{0i}, and the number of patients present directly routed to the EOA

Ward is denoted by n_{00}. The state space for the overflow system in Figure 9.1 is given by:

$$S: \quad \{\mathbf{n} = (n_{00}, n_{01}, \ldots, n_{0I}, n_{1e}, \ldots, n_{Ie}, n_{1u}, \ldots, n_{Iu}); \quad n_{ie} + n_{iu} \leq c_i \quad \forall i; \\ \sum_{i=0}^{I} n_{i0} \leq c_0; \quad n_{ie}, n_{iu}, n_{0i}, n_{00} \geq 0 \quad \forall i\}. \tag{9.1}$$

Denote $\pi(\mathbf{n})$ as the equilibrium probability that $\mathbf{n}$ patients are present in the system. We obtain the following global balance equation:

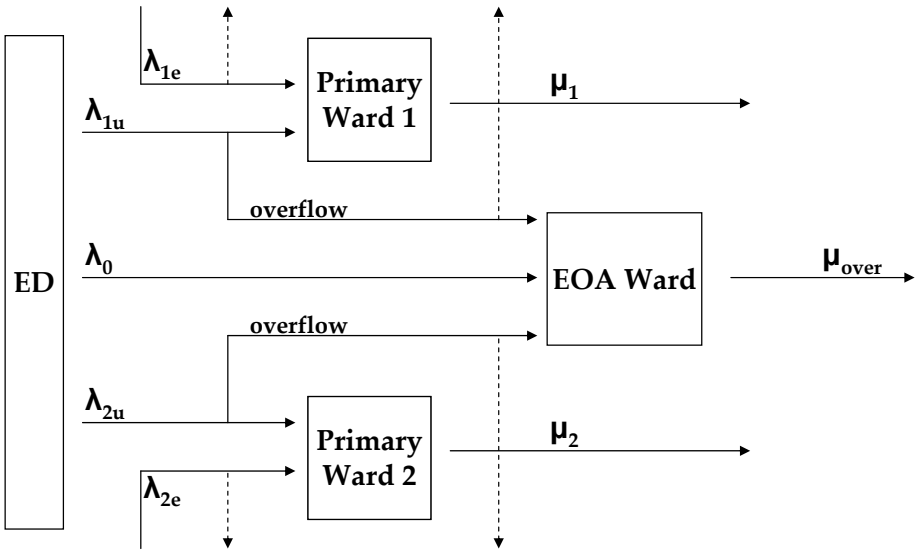
$$\begin{aligned}
\pi(\mathbf{n}) \quad & \left[\sum_{i=1}^{I} \lambda_{i,e} \mathbb{1}_{n_{ie}+n_{iu}<c_i} + \sum_{i=1}^{I} \lambda_{iu} \left(\mathbb{1}_{n_{ie}+n_{iu}<c_i} + \mathbb{1}_{n_{ie}+n_{iu}=c_i, \sum_{i=0}^{I} n_{0i}<c_0}\right)\right. \\
& + \lambda_0 \mathbb{1}_{\sum_{i=0}^{I} n_{0i}<c_0} + \sum_{i=1}^{I} (n_{ie} + n_{iu})\mu_i + \sum_{i=1}^{I} n_{0i}\mu_{over} \\
& \left. + n_{00}\mu_0 + \sum_{i=1}^{I} n_{0i}\gamma_i \mathbb{1}_{n_{ie}+n_{iu}<c_i}\right] \\
= \quad & \sum_{i=1}^{I} \lambda_{ie} \pi(\mathbf{n} - e_{ie}) \mathbb{1}_{n_{ie}>0} \\
& + \sum_{i=1}^{I} \lambda_{iu} \left(\pi(\mathbf{n} - e_{iu}) \mathbb{1}_{n_{iu}>0} + \pi(\mathbf{n} - e_{0i}) \mathbb{1}_{n_{0i}>0, n_{ie}+n_{iu}=c_i}\right) \\
& + \lambda_0 \pi(\mathbf{n} - e_{00}) \mathbb{1}_{n_{00}>0} + \sum_{i=1}^{I} (n_{ie} + 1)\mu_i \pi(\mathbf{n} + e_{ie}) \mathbb{1}_{n_{ie}+1+n_{iu}\leq c_i} \\
& + \sum_{i=1}^{I} (n_{iu} + 1)\mu_i \pi(\mathbf{n} + e_{iu}) \mathbb{1}_{n_{ie}+n_{iu}+1\leq c_i}] \\
& + \sum_{i=1}^{I} (n_{0i} + 1)\mu_{over} \pi(\mathbf{n} + e_{0i}) \mathbb{1}_{n_{0i}+1\leq c_0} + (n_{00} + 1)\mu_0 \pi(\mathbf{n} + e_{00}) \mathbb{1}_{n_{00}+1\leq c_0} \\
& + \sum_{i=1}^{I} (n_{0i} + 1)\gamma_i \pi(\mathbf{n} + e_{0i} - e_{iu}) \mathbb{1}_{n_{ie}+n_{iu}+1\leq c_i} \, .
\end{aligned} \tag{9.2}$$

This equation can be solved explicitly only for specific values of the system parameters [21]. We therefore apply the method as discussed in [128], which requires that $\mu_i = \mu_{over}$. We adapt it such that $\mu_i \neq \mu_{over}$. Our model shows strong similarities with the model of Schehrer [170], but we add an extra flow to represent the elective patients arriving at the primary ward. We first analyze the model without transfers from the EOA Ward to the primary wards. Subsequently we introduce transfers and apply the approach presented in [21] to determine the number of patients present at each ward.

9.2.2 No Transfers from the EOA Ward to the Primary Wards

We first analyze the situation where patients at the EOA Ward are not transferred to the primary wards (see Figure 9.2).

Figure 9.2: No transfers to the primary wards; example with two primary wards



We do so by adapting the global balance equation (9.2) by setting $\gamma_i = 0 \, \forall i$. Now we can solve the global balance equations explicitly. In line with [128] we use a probability generating function approach to determine the mean, $\mathbb{E}_i$, and variance, $\mathbb{V}_i$, of the overflow of urgent patients from primary ward i at the EOA Ward in case of infinite EOA Ward capacity. Since $c_0 = \infty$, and since the overflow processes from the primary wards are independent, $\mathbb{E}_i$ and $\mathbb{V}_i$ can be determined for each primary ward i in isolation. We only want to calculate the blocking probability at the overflow, and thus it is not required to know whether a patient residing at primary ward i is of the urgent or elective type. Let $n_i = n_{ie} + n_{iu}$ denote the number of patients at primary ward i, and let $\lambda_i = \lambda_{ie} + \lambda_{iu}$ denote the total arrival rate at primary ward i. The global balance equation simplifies to:

$$
\begin{aligned}
&\pi(n_{0i}, n_i)\,(\lambda_i + n_i\mu_i + n_{0i}\mu_{over}) \\
&\quad = \lambda_i \pi(n_{0i}, n_i - 1) + (n_{0i} + 1)\mu_{over}\pi(n_{0i} + 1, n_i) + (n_i + 1)\mu_i \pi(n_{0i}, n_i + 1) \\
&\quad \text{for} \quad n_i < c_i, \\
&\pi(n_{0i}, n_i)\,(\lambda_{iu} + n_i\mu_i + n_{0i}\mu_{over}) \\
&\quad = \lambda_i \pi(n_{0i}, n_i - 1) + (n_{0i} + 1)\mu_{over}\pi(n_{0i} + 1, n_i) + \lambda_{iu}\pi(n_{0i} - 1, n_i) \\
&\quad \text{for} \quad n_i = c_i.
\end{aligned} \tag{9.3}
$$

We define the probability generating function of the number of urgent patients from ward i present at the EOA Ward, $G_{i,n_i}(z)$, as:

$$G_{i,n_i}(z) = \sum_{n_{0i}=0}^{\infty} \pi(n_{0i}, n_i) z^{n_{0i}}, \tag{9.4}$$

where $|z| \leq 1$. Multiplication of (9.3) with $z^{n_{0i}}$ and the summation of the result over $n_{0i} = 0, \ldots, \infty$ yields

$$\begin{aligned}
&[\lambda_i + n_i\mu_i]\, G_{i,n_i}(z) + \mu_{over}(z-1)\frac{d}{dz}G_{i,n_i}(z) \\
&\qquad = \lambda_i G_{i,n_i-1}(z) + (n_i+1)\mu_i G_{i,n_i+1}(z) \quad \text{for} \quad 0 \leq n_i < c_i, \\
&[\lambda_{iu}(1-z) + n_i\mu_i]\, G_{i,n_i}(z) + \mu_{over}(z-1)\frac{d}{dz}G_{i,n_i}(z) \\
&\qquad = \lambda_i G_{i,n_i-1}(z) \quad \text{for} \quad n_i = c_i.
\end{aligned} \tag{9.5}$$

Now $\mathbb{E}_i$ and $\mathbb{V}_i$ can be derived from:

$$\begin{aligned}
\mathbb{E}_i &= \sum_{n_i=0}^{c_i} \frac{d}{dz} G_{i,n_i}(z)|_{z=1} \\
\mathbb{V}_i &= \sum_{n_i=0}^{c_i} \frac{d^2}{dz^2} G_{i,n_i}(z)|_{z=1} + \mathbb{E}_i - (\mathbb{E}_i)^2.
\end{aligned} \tag{9.6}$$

Taking the first derivative of (9.5) and evaluating at $z = 1$, gives:

$$\begin{aligned}
&(\lambda_i + n_i\mu_i + \mu_{over})\, g_i[n_i] \\
&\qquad = \lambda_i g_i[n_i - 1] + (n_i+1)\mu_i g_i[n_i+1] \quad \text{for} \quad 0 \leq n_i < c_i, \\
&(\mu_{over} + n_i\mu_i)\, g_i[n_i] - \lambda_{iu}\mathbb{P}_i(c_i) \\
&\qquad = \lambda_i g_i[n_i - 1] \quad \text{for} \quad n_i = c_i,
\end{aligned} \tag{9.7}$$

where $g_i[n_i] = \frac{d}{dz}G_{i,n_i}(z)|_{z=1}$ and $\mathbb{P}_i(c_i) = Erl\left(\frac{\lambda_i}{\mu_i}, c_i\right)$. Then $\mathbb{E}_i$ is obtained by the summation of (9.7) for $n_i = 0, \ldots, c_i$:

$$\mathbb{E}_i = \frac{\lambda_{iu}}{\mu_{over}}\mathbb{P}_i(c_i). \tag{9.8}$$

The variance can be calculated accordingly by taking the second derivative of (9.5) and evaluating at $z = 1$:

$$\begin{aligned}
&(\lambda_i + n_i\mu_i + 2\mu_{over})\, h_i[n_i] \\
&\qquad = \lambda_i h_i[n_i - 1] + (n_i+1)\mu_i h_i[n_i+1] \quad \text{for} \quad 0 \leq n_i < c_i, \\
&(n_i\mu_i + 2\mu_{over})\, h_i[n_i] - 2\lambda_{iu} g_i[n_i] \\
&\qquad = \lambda_i h_i[n_i - 1] \quad \text{for} \quad n_i = c_i,
\end{aligned} \tag{9.9}$$

where $h_i[n_i] = \frac{d^2}{dz^2} G_{i,n_i}(z)|_{z=1}$. The summation of the result for $n_i = 0, \ldots, c_i$ yields:

$$\mathbb{V}_i = \frac{\lambda_{iu}}{\mu_{over}} g_i[c_i] + \mathbb{E}_i - (\mathbb{E}_i)^2, \tag{9.10}$$

where $g_i[c_i]$ can be determined recursively from (9.7) with $g_i[-1] = 0$. The direct patient flow arriving at the EOA Ward can be represented by an $M/M/\infty$ queue. The mean $\mathbb{E}_0$ and variance $\mathbb{V}_0$ of this stream is therefore given by:

$$\begin{aligned} \mathbb{E}_0 &= \frac{\lambda_0}{\mu_0} \\ \mathbb{V}_0 &= \frac{\lambda_0}{\mu_0}. \end{aligned} \tag{9.11}$$

Using the information obtained in this section, we are now able to define an (artificial) equivalent primary ward with service rate μ_{over} which generates the same traffic as the i overflow (urgent) and direct streams together. Since only urgent patients are routed to the EOA Ward, the elective patients who do not cause overflow do not need to be incorporated in the equivalent primary ward. The equivalent primary ward has artificial load a and capacity C such that [200]:

$$\begin{aligned} aErl(a, C) &= \mathbb{E} \\ \mathbb{E}\left(1 - \mathbb{E} + \frac{a}{C + 1 + \mathbb{E} - a}\right) &= \mathbb{V}, \end{aligned} \tag{9.12}$$

where, since the overflow processes from the primary wards and the direct arrival process are independent, the expectation and variation of the aggregated overflow, $\mathbb{E}$ and $\mathbb{V}$, are given by:

$$\begin{aligned} \mathbb{E} &= \sum_{i=0}^{I} \mathbb{E}_i \\ \mathbb{V} &= \sum_{i=0}^{I} \mathbb{V}_i. \end{aligned} \tag{9.13}$$

The blocking probability for patients from ward i and the direct arriving patients ($i = 0$) is given by the Katz approximation [203], which takes the peakedness, $\zeta_i = \frac{\mathbb{V}_i}{\mathbb{E}_i}$, of the separate flows into account:

$$K_i = \frac{aErl(a, C + c_0)}{aErl(a, C)} \left(v(C, c_0)^{-1} + \frac{\zeta_i - 1}{\zeta - 1} \left(1 - v(C, c_0)^{-1}\right) \right), \tag{9.14}$$

with $\zeta = \frac{\mathbb{V}}{\mathbb{E}}$ and $v(C, c_0)$ can be determined recursively from:

$$\begin{aligned} v(C,j) &= \frac{aj}{aErl(a,C)\left(C + j - a - aErl(a,C)v(C,j-1)\right)}, \quad j = 1, 2, \ldots \\ v(C,0) &= 1. \end{aligned} \tag{9.15}$$

Since the direct arrival stream is Poisson, we have that $\zeta_0 = 1$. It follows from the consistency requirements $\sum_{i=0}^{I} \mathbb{E}_i K_i = \mathbb{E}K$ and $\sum_{i=0}^{I} \mathbb{E}_i \zeta_i = \mathbb{E}\zeta$, where K denotes the overall blocking probability, that the Katz approximation is exact in this case [203].

The mean number of urgent patients from primary ward i present at the EOA Ward, $\mathbb{E}[N_{0i}]$, is given by:

$$\mathbb{E}[N_{0i}] = \frac{\lambda_{iu}}{\mu_{over}} Erl\left(\frac{\lambda_i}{\mu_i}, c_i\right)(1 - K_i). \tag{9.16}$$

The mean number of patients who directly arrived at the EOA Ward, $\mathbb{E}[N_{00}]$, equals:

$$\mathbb{E}[N_{00}] = \frac{\lambda_0}{\mu_0}(1 - K_0). \tag{9.17}$$

9.2.3 Transfers from the EOA Ward to the Primary Wards

Finally, we allow patient transfers from the EOA Ward back to the primary wards. It was assumed that μ_{over} was the same for all patients, and therefore we define γ_i such that $\mu_{over} = \mu_i + \gamma_i \ \forall i$. In this scenario we obviously have that $\mu_{over} \geq \mu_i$, and this is exactly the model as depicted in Figure 9.1. We approximate the arrival rate at primary ward i, ν_i, by the sum of the arrivals of elective and urgent patients (λ_{ie} and λ_{iu} resp.) and the patients transferred from the EOA Ward, $\gamma_i \mathbb{E}[N_{0i}]$, [21]:

$$\nu_i = \lambda_{ie} + \lambda_{iu} + \gamma_i \mathbb{E}[N_{0i}]. \tag{9.18}$$

A fraction of this stream, κ_i, is routed to the EOA Ward when all beds at primary ward i are occupied:

$$\kappa_i = \lambda_{iu} + \gamma_i \mathbb{E}[N_{0i}]\ . \tag{9.19}$$

To analyze the model for $\gamma_i > 0$, we replace λ_{iu} by κ_i and λ_i by ν_i in (9.7 – 9.10). We then obtain a system of equations, which can be solved for $\mathbb{E}[N_{0i}]$ using fixed point iteration with initial value $\mathbb{E}[N_{0i}] = 0$ [21].

The mean of the total number of patients present at the EOA Ward, $\mathbb{E}[N_0]$, and the mean of the total number of patients present at primary ward i, $\mathbb{E}[N_i]$, are given by:

$$\begin{aligned} \mathbb{E}[N_0] &= \sum_{i=0}^{I} \mathbb{E}[N_{0i}] \\ \mathbb{E}[N_i] &= \frac{\nu_i}{\mu_i}\left(1 - Erl\left(\frac{\nu_i}{\mu_i}, c_i\right)\right). \end{aligned} \tag{9.20}$$

The mean occupation, ρ_i, of the EOA Ward and the primary wards, is given by:

$$\rho_i = \frac{E[N_i]}{c_i}. \tag{9.21}$$

The mean number of patients blocked or admitted at the wards can be calculated accordingly.

9.3 Results

We now use the model from Subsection 9.2.3 to analyze a simple example for a hospital with two primary wards. Primary ward 1 has a capacity of $c_1 = 200$ beds and admits only medical patients, whose mean LOS is five days (so $\mu_1 = \frac{1}{5}$). The elective patient arrival rate λ_{1e} equals 26 patients per day, and the urgent patient arrival rate λ_{1u} is 14 patients per day, so that the total patient arrival rate at ward 1, $\lambda_1 = 40$. Primary ward 2 admits only surgical patients, with $c_2 = 200$, a mean LOS of four days ($\mu_2 = \frac{1}{4}$), $\lambda_{2e} = 37$, $\lambda_{2u} = 13$, and $\lambda_2 = 50$. Adding capacity by creating so-called overbeds is not allowed.

Patients arrive directly at the EOA Ward with rate $\lambda_0 = 2$ and have a service rate of $\mu_0 = 4$. With this flow we represent patients who only require observation for a short period of time (on average six hours in this case). The mean LOS for urgent patients at the EOA Ward is set on 36 hours, so that $\mu_{over} = \frac{2}{3}$. Consequently, $\gamma_1 = \mu_{over} - \mu_1 = \frac{7}{15}$ and $\gamma_2 = \mu_{over} - \mu_2 = \frac{5}{12}$. This implies that patients with a longer LOS should be transferred back sooner to the primary ward than patients with a shorter LOS, in order to keep the LOS at the EOA Ward the same for all urgent patients. Table 9.1 summarizes the parameter values.

9.3.1 Opening the EOA Ward

Suppose the hospital considers opening an EOA Ward. We first analyze the situation where only urgent patients are admitted at this ward (so for now, we set $\lambda_0 = 0$). In Table 9.2 for $c_0 = 0$ (no EOA Ward, i.e., the old situation), and $c_0 = 4, 6, 8, 12$ the blocking probabilities for elective, $\mathbb{P}(B_{ie})$, and urgent patients, $\mathbb{P}(B_{iu})$, are given. The number

Table 9.1: Parameter values for EOA Ward example

Parameter	Value	Parameter	Value
c_1	200	μ_0	4
c_2	200	μ_{over}	$\frac{2}{3}$
λ_{1e}	26	μ_1	$\frac{1}{5}$
λ_{1u}	14	μ_2	$\frac{1}{4}$
λ_{2e}	37	γ_1	$\frac{7}{15}$
λ_{2u}	13	γ_2	$\frac{5}{12}$

of rejected elective, B_{ie}, and urgent patients, B_{iu}, is given per ward per day, and the number of admitted elective, EP/y, and admitted urgent, UP/y, patients per year are given.

Table 9.2: Results for opening an EOA Ward

c_0	$\mathbb{P}(B_{ie})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	EP/y	UP/y
0	5.4%	1.4144	5.4%	0.7616	5.4%	2.0128	5.4%	0.7072	21,753	9,323
4	6.0%	1.5547	2.4%	0.3313	5.8%	2.1469	2.1%	0.2733	21,642	9,633
6	6.2%	1.5998	1.5%	0.2024	5.9%	2.1885	1.1%	0.1468	21,610	9,726
8	6.3%	1.6352	0.1%	0.1035	6.0%	2.2110	0.1%	0.0796	21,587	9,845
12	6.4%	1.6652	0.2%	0.0215	6.0%	2.2322	0.1%	0.0169	21,577	9,840

What we see is that the blocking probability for urgent patients decreases, which was expected since we added capacity for these patients. However, since the hospital is now able to admit more urgent patients, ultimately there is less capacity available at the primary wards for the elective patients which results in repression of elective patients. An EOA Ward with four beds results in a total of 310 more (9,633 vs. 9,323) urgent patients admitted per year, but at the same time 111 less elective patients are admitted per year (21,642 vs. 21,753). It follows that the opening of the EOA Ward also affects the elective patient flow.

9.3.2 Admitting Observation Patients

Following the opening of the EOA Ward, the hospital decides that patients from the ED requiring observation should also be admitted at the EOA Ward (thus we set $\lambda_0 = 2$). It is obvious that more beds are required to maintain the decreased blocking probabilities for urgent patients (Table 9.3), but the blocking probabilities for elective patients remain about the same.

9.3.3 Increasing Urgent Admissions

As mentioned in the Introduction, one of the reasons to open an EOA Ward is to increase the number of urgent patient admissions through the ED. Table 9.4 shows for various

Table 9.3: Results for admitting observation patients

c_0	$\mathbb{P}(B_{ie})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	$\mathbb{P}(B_0)$	B_0
4	5.9%	1.5409	2.7%	0.3716	5.8%	2.1377	2.3%	0.3019	2.2%	0.4452
12	6.4%	1.6647	0.2%	0.0228	6.0%	2.2319	0.1%	0.0176	1.1%	0.0218

rates of increase, f_u, in the arrival rate of urgent patients, λ_{iu}, the required size of the EOA Ward for which $\mathbb{P}(B_{iu}) \approx 1\%$. Note that $\lambda_0 = 0$.

Table 9.4: Results for increasing urgent admissions

f_u	c_0	$\mathbb{P}(B_{ie})$	B_{1e}	$\mathbb{P}(B_{1u})$	B_{1u}	$\mathbb{P}(B_{2e})$	B_{2e}	$\mathbb{P}(B_{2u})$	B_{2u}	EP/y	UP/y
5%	8	7.6%	1.9716	1.2%	0.1758	6.9%	2.5465	0.8%	0.1176	21,342	10,262
10%	10	9.1%	2.3664	1.0%	0.1601	7.9%	2.9208	0.7%	0.0954	21,065	10,748
20%	12	12.2%	3.1753	1.3%	0.2112	9.9%	3.6731	1.0%	0.1502	20,500	11,768
50%	22	22.2%	5.7827	1.3%	0.2816	16.6%	6.1434	1.0%	0.1955	18,646	14,612

We see that an increase in the number of urgent patient admissions has a tremendous effect on the number of elective patient admissions. For example in the case of a 10% increase, the number of elective patient admissions decreases from 21,753 to 21,065 per year.

9.3.4 Maintaining the Number of Elective Patient Admissions

If the hospital wishes to maintain the number of elective patient admissions, it has two options: increase the number of beds at the primary wards or stop transferring patients from the EOA Ward back to the primary wards. The latter option would transform the EOA Ward to a long stay ward for urgent patients, and therefore we only analyze the first option. Table 9.5 gives for each value of f_u the required number of beds at the primary wards, c_1 and c_2, and the extra number of beds required at the primary wards, c_+, compared to the initial situation where $c_1 = c_2 = 200$, such that the elective patient blocking probability is maintained at its initial value of $\approx 5\%$. Again, note that $\lambda_0 = 0$.

Table 9.5: Results for maintaining the number of elective patient admissions

f_u	c_0	c_1	c_2	c_+
0%	6	203	202	5
5%	6	207	205	12
10%	6	210	208	18
20%	6	215	215	30
50%	8	238	229	67

Since the number of inpatient beds increases, more urgent patients can be admitted directly at the primary wards and thus less EOA Ward capacity is required to keep $\mathbb{P}(B_{iu}) \approx 1\%$.

9.4 Discussion

Inpatient wards designed to improve the urgent patient flow have gained increased popularity during the last decade. In this chapter we have developed a queuing model, based on Wilkinson's Equivalent Random Method [200], that allows for a quantification of the effects of these EOA Wards in terms of elective and urgent patient admissions. In addition to the extensions to the ERM made in [128] and [170], our model enables the analysis of an overflow system for which the service rate at the primary wards (or cells in the ERM terminology) is not equal to the service rate at the overflow ward (or cell). Furthermore we have added an extra arrival stream to the primary wards which is blocked when all beds are occupied. For analytical tractability, we assumed the LOS at the primary wards and EOA Ward is exponentially distributed. In [128] the authors show for a similar system that the outcomes are insensitive to the service time distribution, as is the case for many loss models.

With a simple example of a hospital with two primary wards, we have shown that opening an EOA Ward results in an increase of urgent patient admissions, but at the same time in a decrease in the number of elective patient admissions. The elective patients are repressed by the extra admitted urgent patients who return to the primary ward from the EOA Ward. We assumed that the urgent patient flow remained constant over time. In reality, the added capacity may attract extra urgent patients, which will in turn result in even less capacity for elective patients. To overcome this effect, next to the EOA Ward capacity created, additional inpatient beds should be added. This in turn results in a decrease of the number of EOA Ward beds required, which makes the EOA Ward a small ward that is possibly difficult to staff. In the example we incorporated only two, very large inpatient wards. In case of more wards with smaller capacities than in our example, the blocking probabilities at these wards are more sensitive to an increase in patient arrivals and thus the repression effect will remain and possibly even worsen. Another important factor to consider is the maximum LOS at the EOA Ward. In our example we used a maximum LOS of 36 hours. When this is shortened to 24 hours, the number of urgent admissions will increase and thus the repression effect will get worse. When determining the maximum LOS this phenomenon should be taken into account as well. The model allows for an easy evaluation of the repression effect in case of changes in the EOA Ward LOS.

EOA Wards definitely have advantages, such as the fact that the admission of urgent patients is centrally organized, or that admissions during the night for regular inpatient wards are avoided. Given the results presented in this chapter, we can also perceive the EOA Ward as an instrument to control the elective/urgent patient ratio. However, the effect on elective patient admissions should not be neglected, but rather, studied before the decision to open an EOA Ward is made. Other possibilities to improve urgent patient flow are likely to have less adverse effects and should also be taken into consideration.

Epilogue

We discussed the difficult decisions that have to be made on the distribution of healthcare resources in Chapter 1. In this concluding chapter we briefly review the key results obtained and how they can support decision making processes in a healthcare setting.

We started in Chapter 2 with a review of queuing theory and networks of queues in particular. We showed that these methodologies are very well suited to model and analyze healthcare networks, even though there are still a couple of mathematical hurdles to take. Modeling hospital departments as queuing networks, thus capturing their interdependency, will become increasingly important in order to make the transfer from a single to a multi-departmental modeling approach.

In Part II four different capacity distribution problems related to outpatient clinics and diagnostic facilities were outlined. The re-distribution of tasks and responsibilities among different groups of healthcare professionals resulted in better clinic performance in Chapter 3. Although the implementation was successful, only the problem at hand was solved, as is the case with many improvement studies. A challenge for the future would be to develop methodologies that ensure continuous monitoring and improvement, so that when the circumstances or the clinic environment changes, potential problems will be recognized and solved before they even appear.

The scheduling of appointment patients during periods of low walk-in demand can lead to acceptable access and waiting times for resp. appointment and walk-in patients in Chapter 4. Introducing a reservation policy for priority jobs, for example patients in a care pathway, allowed for a trade-off between accessibility and waiting time for resp. priority and non-priority jobs (Chapter 5). The latter two models show that performance targets in terms of access and waiting times can be achieved for several patient groups at the same time.

With the game-theoretic model in Chapter 6 we showed that hospital departments can be stimulated to provide a reliable estimate of future MRI demand, so that a sensible distribution of MRI capacity is ensured. Implementation of the models of Chapters 4–6 would definitely be challenging, but also worthwhile given the potential efficiency gain.

Part III also contained three chapters, this time on topics related to urgent patient flow. Chapter 7 and 8 outlined a methodology and consequently its implementation, designed to determine the amount of OR time that should be allocated to semi-urgent

patients. The goal was to balance the cancellation of elective patients and unused OR time, caused by variations in semi-urgent patient arrivals. The LUMC Neurosurgery department was successfull in implementing the method, but the process required a vast amount of time and effort. It would be interesting to study whether the implementation of this methodology in other hospitals would involve the same complications the LUMC encountered.

In Chapter 9 we studied the distribution of staffed beds among regular inpatient wards and the EOA Ward. The latter type of ward is designed to increase the admission rates of urgent patients through the ED. We showed that an increase of urgent admissions through a new ward, results in a repression of elective patients at the inpatient wards. To overcome this effect, additional inpatient beds should be added. These results clearly demonstrate the potential danger of employing a single department view.

All models presented in parts II and III allow for quantitative analysis of resource distribution problems in a healthcare setting. In the process of defining a mathematical model, thorough discussion on the process and problem is required, which is an advantage. Additionally, the 'clean' model outcomes shift the attention from specific employees or departments to steps in the process that might be improved or removed. We can conclude that mathematical modeling contributes to higher quality, more sound decision making in healthcare. Of course, we are not there yet. As Chapter 8 shows, the development of a mathematical model is only the first, and perhaps even the *easiest* step. The next challenge lies in the implementation.

Bibliography

[1] Aaby K, Herrmann JW, Jordan CS, Treadwell M, Wood K (2006) *Montgomery county's public health service uses operations research to plan emergency mass dispensing and vaccination clinics.* Interfaces 36(6):569-579

[2] Adan IJBF, van Leeuwaarden JSH, Winands EMM (2006) *On the application of Rouché's theorem in queueing theory.* Operations Research Letters 34(3):355-360

[3] Adan IJBF, van der Wal J (2011) *Mean Values Techniques.* In: Boucherie RJ, van Dijk NM (eds) *Queueing networks: a fundamental approach.* Springer, New York, NY, USA

[4] Albin SL, Barrett J, Ito D, Mueller JE (1990) *A queueing network analysis of a health center.* Queueing Systems 7:51-61

[5] Allen D (2009) *From boundary concept to boundary object: the practice and politics of care pathway development*. Social Science & Medicine 69(3):354-361

[6] Allen D, Rixson L (2008) *How has the impact of 'care pathway technologies' on service integration in stroke care been measured and what is the strength of the evidence to support their effectiveness in this respect?* International Journal of Evidence-Based Healthcare 6(1):78-110

[7] American Society of Anesthesiologists, ASA Score, retrieved April 20, 2011, from *www.asahq.org/For-Members/Clinical-Information/ASA-Physical-Status-Classification-System.aspx*

[8] Asaduzzaman Md, Chaussalet TJ, Adeyemi S, Chahed S, Hawdon J, Wood D, Robertson NJ (2010) *Towards effective capacity planning in a perinatal network centre.* Archives of Disease in Childhood Fetal Neonatal 95:F283-287

[9] Asaduzzaman Md, Chaussalet TJ, Robertson NJ (2010) *A loss network model with overflow for capacity planning of a neonatal unit.* Annals of Operations Research 178(1):67-76

[10] Ashton R, Hague L, Brandreth M, Worthington D, Cropper S (2004) *A simulation-based study of a NHS Walk-in Centre*. Journal of the Operational Research Society 56(2):153-161

[11] Asmussen S (2003) *Applied Probability and Queues*. 2nd ed. Springer, New York, NY, USA

[12] Atkinson KE (1978) *An introduction to numerical analysis.* 2nd ed. John Wiley & Sons, New York, NY, USA

[13] Bagust A, Place M, Posnett JW (1999) *Dynamics of bed use in accommodating emergency admissions: stochastic simulation model.* British Medical Journal 319(7203): 155-158

[14] Bailey NTJ (1952) *A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times.* Journal of the Royal Statistical Society, Series B-Statistical Methodology 14(2):185-199

[15] Baskett F, Chandy KM, Muntz RR, Palacios FG (1975) *Open, closed, and mixed networks of queues with different classes of customers.* Journal of the Association for Computing Machinery 22(2):248-260

[16] Beitia A (2003) *Hospital quality choice and market structure in a regulated duopoly.* Journal of Health Economics 22(6):1011-1036

[17] Berezner SA, Kriel CF, Krzesinski AE (1995) *Quasi-reversible multiclass queues with order independent departure rates*. Queueing Systems 19(4):345-359

[18] Bhattacharyya T, Vrahas MS, Morrison SM, Kim E, Wiklund RA, Smith RM, Rubash HE (2006) *The value of the dedicated orthopaedic trauma operating room.* The Journal of Trauma: Injury, Infection, and Critical Care 60(6):1336-1341

[19] Bitran GR, Morabito R (1996) *Survey open queueing networks: optimization and performance evaluation models for discrete manufacturing systems.* Production and Operations Management 5(2):163-193

[20] Blair EL, Lawrence CE (1981) *A queueing network approach to health care planning with an application to burn care in New York state.* Socio-Economic Planning Sciences 15(5):207-216

[21] Borst S, Boucherie RJ, Boxma OJ (1999) *ERMR: a generalised Equivalent Random Method for overflow systems with repacking*. Proceedings of the 16th International Teletraffic Congress

[22] vanden Bosch PMV, Dietz DC (2000) *Minimizing expected waiting in a medical appointment system*. IIE Transactions 32(9):841-848

[23] vanden Bosch PMV, Dietz DC, Simeoni JR (1999) *Scheduling customer arrivals to a stochastic service system*. Naval Research Logistics 46(5):549-559

[24] Boucherie RJ, van Dijk NM (2011) *Queueing networks: a fundamental approach.* Springer, New York, NY, USA

[25] Bowers J, Mould G (2004) *Managing uncertainty in orthopaedic trauma theatres.* European Journal of Operational Research 154(3):599-608

[26] Brailsford SC, Harper PR, Patel B, Pitt M (2009) *An analysis of the academic literature on simulation and modelling in health care*. Journal of Simulation 3:130-140

[27] Brockmeyer E, Halström HL, Jensen A (1948) *The life and works of A.K. Erlang.* Translations of the Danish Academy of Technical Sciences 2

[28] de Bruin AM, Bekker R, van Zanten L, Koole GM (2010) *Dimensioning hospital wards using the Erlang loss model.* Annals of Operations Research 178(1):23-43

[29] de Bruin AM, van Rossum AC, Visser MC, Koole GM (2007) *Modeling the emergency cardiac in-patient flow: an application of queuing theory*. Health Care Management Science 10(2):125-137

[30] Bruneel H (1993) *Performance of discrete-time queueing systems*. Computers & Operations Research 20(3):303-320

[31] Bruneel H, Kim BG (1993) *Discrete-time models for communication systems including ATM.* Kluwer Academic Publishers, Norwell, MA, USA

[32] Bruneel H, Wuyts I (1994) *Analysis of discrete-time multiserver queueing models with constant service times.* Operations Research Letters 15(5):231-236

[33] Burke EK, de Causmaecker P, vanden Berghe G, van Landeghme H (2004) *The state of the art of nurse rostering*. Journal of Scheduling 7(6):441-499

[34] Burke PJ (1956) *The output of a queuing system.* Operations Research 4(6):699-704

[35] Buzacott JA, Shanthikumar JG (1993) *Stochastic models of manufacturing systems.* Prentice Hall, Englewood Cliffs, NJ, USA

[36] Buzen JP (1973) *Computational algorithms for closed queueing networks with exponential servers.* Communications of the ACM 16(9):527-531

[37] Cachon GP, Lariviere MA (1999) *Capacity choice and allocation: strategic behavior and supply chain.* Management Science 45(8):1091-1108

[38] Cardoen B, Demeulemeester E, Beliën J (2010) *Operating room planning and scheduling: a literature review*. European Journal of Operational Research 201(3):921-932

[39] Carter MW (2002) *Diagnosis: mismanagement of resources*.
OR/MS Today 29(2):26-32

[40] Cayirli T, Veral E (2003) *Outpatient scheduling in health care: a review of literature*. Production and Operations Management 12(4):519-549

[41] Cayirli T, Veral E, Rosen H (2006) *Designing appointment scheduling systems for ambulatory care services*. Health Care Management Science 9(1):47-58

[42] Cayirli T, Veral E, Rosen H (2008) *Assessment of patient classification in appointment system design*. Production and Operations Management 17(3):338-353

[43] Chaussalet TJ, Xie H, Millard P (2006) *A closed queueing network approach to the analysis of patient flow in health care systems*. Methods of Information in Medicine 45(5):492-497

[44] Coats TJ, Michalis S (2001) *Mathematical modelling of patients flow through an accident and emergency department.* Emergency Medicine Journal 18(3): 190-192

[45] Cochran JK, Bharti A (2006) *A multi-stage stochastic methodology for whole hospital bed planning under peak loading.* International Journal of Industrial and Systems Engineering 1(1):8-36

[46] Cochran JK, Bharti A (2006) *Stochastic bed balancing of an obstetrics hospital.* Health Care Management Science 9(1):31-45

[47] Cochran JK, Roche KT (2009) *A multi-class queuing network analysis methodology for improving hospital emergency department performance*. Computers & Operations Research 36(5):1497-1512

[48] Cohen JW (1982) *The single server queue.* 8th ed. North-Holland Publishing Company, Amsterdam, the Netherlands

[49] Conway JB, Goldberg J, Chung F (1992) *Preadmission anaesthesia consultation clinic*. Canadian Journal of Anesthesia 39(10):1051-1057

[50] Cooke MW, Higgins J, Kidd P (2003) *Use of emergency observation and assessment wards: a systematic literature review.* Emergency Medicine Journal 20(2):138-142

[51] Cramton P, Shoham Y, Steinberg R (2006) *Combinatorial auctions*. The MIT Press, Cambridge, MA, USA

[52] Creemers S (2009) *Appointment-driven queueing systems*. PhD thesis, Katholieke Universiteit Leuven

[53] Creemers S, Lambrecht M (2010) *Queueing models for appointment-driven systems*. Annals of Operations Research 178(1):155-172

[54] Creemers S, Lambrecht M (2011) *Modeling a hospital queueing network.* In: Boucherie RJ, van Dijk NM (eds) *Queueing networks: a fundamental approach.* Springer, New York, NY, USA

[55] Dexter F (1999) *Design of appointment systems for preanesthesia evaluation clinics to minimize patient waiting times: a review of computer simulation and patient survey studies*. Anesthesia & Analgesia 89(4):925-931

[56] Dexter F, Macario A, O'Neill L (2000) *Scheduling surgical cases into overflow block time - computer simulation of the effects of scheduling strategies on operating room labor costs.* Anesthesia & Analgesia 90(4):980-988

[57] Dickson D, Ford RC, Laval B (2005) *Managing real and virutal waits in hospitality and service organizations*. Corneel Hotel and Restaurant Administration Quarterly 46(1):52-68

[58] van Dijk NM, Kortbeek N (2009) *Erlang loss bounds for OT–ICU systems.* Queueing Systems 63(1):253-280

[59] Disney's Fastpass, Wikipedia the free encyclopedia, retrieved from *http://en.wikipedia.org/w/index.php?title=Disney%27s_Fastpass&oldid=462223439* on August 30, 2011

[60] Dobson G, Hasija S, Pinker EJ (2011) *Reserving capacity for urgent patients in primary care*. Production and Operations Management 20(3):456-473

[61] Duguay C, Chetouane F (2007) *Modeling and improving emergency department systems using discrete event simulation.* Simulation 83(4):311-320

[62] Dunnill MG, Pounder RE (2004) *Medical outpatients: changes that can benefit patients*. Clinical Medicine 4(1):45-49

[63] Edward GM, de Haes JC, Oort FJ, Lemaire LC, Hollmann MW, Preckel B (2008) *Setting priorities for improving the preoperative assessment clinic: the patients' and the professionals' perspective*. British Journal of Anaesthesia 100(3):322-326

[64] Edward GM, Razzaq S, de Roode A, Boer F, Hollmann MW, Dzoljic M, Lemaire LC (2008) *Patient flow in the preoperative assessment clinic*. European Journal of Anaesthesiology 25(4):280-286

[65] Elkhuizen SG, Limburg M, Bakker PJM, Klazinga NS (2006) *Evidence-based reengineering: re-engineering the evidence: a systematic review of the literature on business process redesign (BPR) in hospital care*. International Journal of Health Care Quality Assurance 19(6):477-499

[66] Fackrell M (2009) *Modelling healthcare systems with phase-type distributions.* Health Care Management Science 12(1):11-26

[67] Ferschl MB, Tung A, Sweitzer B, Huo D, Glick DB (2005) *Preoperative clinic visits reduce operating room cancellations and delays.* Anesthesiology 103(4):855-859

[68] Forster AJ, Stiell I, Wells G, Lee AJ, van Walraven C (2003) *The effect of hospital occupancy on emergency department length of stay and patient disposition.* Academic Emergency Medicine 10(2):127-133

[69] Foster EM, Hosking MR, Ziya S (2010) *A spoonful of math helps the medicine go down: an illustration of how healthcare can benefit from mathematical modeling and analysis*. BMC Medical Research Methodology 10(1):60-70

[70] Free University, Department of Mathematics, Erlang-C Calculator, retrieved December 9, 2011, from *www.few.vu.nl/~koole/ccmath/ErlangC/*

[71] Fudenberg D, Tirole J (1993) *Game theory*. 3rd ed. The MIT Press, Cambridge, MA, USA

[72] Gerchak Y, Gupta D, Henig M (1996) *Reservation planning for elective surgery under uncertain demand for emergency surgery.* Management Science 42(3):321-334

[73] Gibby GL, Schwab WK (1998) *Availability of records in an outpatient preanesthetic evaluation clinic*. Journal of Clinical Monitoring and Computing 14(6):385-391

[74] Glouberman S, Mintzberg H (2001) *Managing the care of health and the cure of disease - part I: differentiation*. Health Care Management Review 26(1):56-69

[75] Glouberman S, Mintzberg H (2001) *Managing the care of health and the cure of disease - part II: integration*. Health Care Management Review 26(1):70-84

[76] Goddard J, Tavakoli M (2008) *Efficiency and welfare implications of managed public sector hospital waiting lists*. European Journal of Operational Research 184(2):778-792

[77] Gordon WJ, Newell GF (1967) *Closed queuing systems with exponential servers.* Operations Research 15(2):254-265

[78] Green LV (2006) *Queueing analysis in healthcare.* In: Hall (ed) *Patient flow: reducing delay in healthcare delivery.* Springer, New York, NY, USA

[79] Green LV, Kolesar PJ, Soares J (2001) *Improving the SIPP approach for staffing service systems that have cyclic demands.* Operations Research 49(4):549-564

[80] Green LV, Savin S (2008) *Reducing delays for medical appointments: a queueing approach*. Operations Research 56(6):1526-1538

[81] Green LV, Savin S, Wang B (2006) *Managing patient service in a diagnostic medical facility*. Operations Research 54(1):11-25

[82] Green LV, Soares J (2007) *Computing time-dependent waiting time probabilities in* $M(t)/M/s(t)$ *queueing systems*. Manufacturing & Service Operations Management 9(1):54-61

[83] Green LV, Soares J, Giglio JF, Green RA (2006) *Using queueing theory to increase the effectiveness of emergency department provider staffing.* Academic Emergency Medicine 13(1):61-68

[84] Gross D, Shortle JF, Harris CM (2008) *Fundamentals of queueing theory.* 4th ed. John Wiley & Sons, Hoboken, NJ, USA

[85] Gupta D, Denton B (2008) *Appointment scheduling in health care: challenges and opportunities*. IIE Transactions 40(9):800-819

[86] Harper PR, Gamlin HM (2003) *Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach*. OR Spectrum 25(2):207-222

[87] Harsanyi JC (1967) *Games with incomplete information played by "Bayesian" players, I-III. Part I. The basic model.* Management Science 14(3):159-182

[88] Harsanyi JC (1968) *Games with incomplete information played by "Bayesian" players, I-III. Part II. Bayesian equilibrium points.* Management Science 14(5):320-334

[89] Harsanyi JC (1968) *Games with incomplete information played by "Bayesian" players, I-III. Part III. The basic probability distribution of the game.* Management Science 14(7):486-502

[90] Hart OD (1983) *Optimal labour contracts under asymmetric information: an introduction.* The Review of Economic Studies 50(1):3-35

[91] Hassin R, Mendel S (2008) *Scheduling arrivals to queues: a single-server model with no-shows*. Management Science 54(3):565-572

[92] Hepner DL, Bader AM, Hurwitz S, Gustafson M, Tsen LC (2004) *Patient satisfaction with preoperative assessment in a preoperative assessment testing clinic*. Anesthesia & Analgesia 98(4):1099-1105

[93] Ho CJ, Lau HS (1992) *Minimizing total cost in scheduling outpatient appointments*. Management Science 38(12):1750-1764

[94] Hollingsworth B, Dawson PJ, Maniadakis N (1999) *Efficiency measurement of health care: a review of non-parametric methods and applications*. Health Care Management Science 2(3):161-172

[95] Hoot NR, Aronsky D (2008) *Systematic review of emergency department crowding: causes, effects, and solutions*. Annals of Emergency Medicine 52(2):126-136

[96] Hoot NR, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, Aronsky D (2008) *Forecasting emergency department crowding: a discrete event simulation.* Annals of Emergency Medicine 52(2):116-125

[97] Horn RA, Johnson CR (1985) *Matrix Analysis*. Cambridge University Press, Cambridge, UK

[98] van Houdenhoven M, van Oostrum JM, Hans EW, Wullink G, Kazemier G (2007) *Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling.* Anesthesia & Analgesia 105(3):707-714

[99] Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW (2011) *Taxonomic classification of planning decisions in health care: a review of the state of the art in OR/MS.* Memorandum 1944, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands

[100] Jackson JR (1957) *Networks of waiting lines.* Operations Research 5(4):518-521

[101] Jiang L, Giachetti RE (2008) *A queueing network model to analyze the impact of parallelization of care on patient cycle time.* Health Care Management Science 11(3):248-261

[102] Jun JB, Jacobson SH, Swisher JR (1999) *Application of discrete-event simulation in healthcare: a survey*. Journal of the Operational Research Society 50(2):109-123

[103] Kaandorp GC, Koole G (2007) *Optimal outpatient appointment scheduling*. Health Care Management Science 10(3):217-229

[104] Kelly FP (1979) *Reversibility and stochastic networks.* Available online from *www.statslab.cam.ac.uk/ frank/rsn.html*

[105] Kelly FP (1991) *Loss networks.* The Annals of Applied Probability 1(3):319-378

[106] Kemeny JG, Snell JL (1976) *Finite Markov Chains*. 2nd ed. Springer, New York, NY, USA

[107] Klassen KJ, Rohleder TR (1996) *Scheduling outpatient appointments in a dynamic environment*. Journal of Operations Management 14(2):83-101

[108] Klassen KJ, Rohleder TR (2004) *Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment*. International Journal of Service Industry Management 15(2):167-186

[109] van Klei WA, Moons KG, Rutten CL, Schuurhuis A, Knape JT, Kalkman CJ, Grobbee DE (2002) *The effect of outpatient preoperative evaluation of hospital inpatients on cancellation of surgery and length of hospital stay*. Anesthesia & Analgesia 94(3):644-649

[110] Kleinrock L (1967) *Queuing systems: theory, vol. 1.* John Wiley & Sons, New York, NY, USA

[111] Kleinrock L (1976) *Queueing systems: computer applications, vol. 2.* John Wiley & Sons, New York, NY, USA

[112] Koizumi N, Kuno E, Smith TE (2005) *Modeling patient flows using a queueing network with blocking.* Health Care Management Science 8(1):49-60

[113] Kolisch R, Sickinger S (2008) *Providing radiology health care services to stochastic demand of different customer classes*. OR Spectrum 30(2):375-395

[114] Koole G (1997) *Assigning a single server to inhomogeneous queues with switching costs.* Theoretical Computer Science 182(1-2):203-216

[115] Kopach R, DeLaurentis PC, Lawley M, Muthuraman K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D (2007) *Effects of clinical characteristics on successful open access scheduling*. Health Care Management Science 10(2):111-124

[116] Kostami V, Ward AR (2009) *Managing service systems with an offline waiting option and customer abandonment*. Manufacturing & Service Operations Management 11(4):644-656

[117] LaGanga LR, Lawrence SR (2007) *Clinic overbooking to improve patient access and increase provicer productivity**. Decision Sciences 38(2):251-276

[118] Lamiri M, Xie X, Dolgui A, Grimaud F (2008) *A stochastic model for operating room planning with elective and emergency demand for surgery.* European Journal of Operetional Research 185(3):1026-1037

[119] Latouche G, Ramaswami V (1999) *Introduction to matrix analytic methods in stochastic modeling*. ASA/SIAM Series on Statistics and Applied Probability

[120] Latouche G, Taylor P (2002) *Matrix-analytic methods: theory and applications.* Proceedings of the 4th International Conference on Matrix-Analytic Methods in Stochastic Models

[121] Law AM, Kelton WD (1991) *Simulation modeling and analysis.* McGraw-Hill, New York, NY, USA

[122] Lee DKK, Zenios SA (2009) *Optimal capacity overbooking for the regular treatment of chronic conditions.* Operations Research 57(4):852-865

[123] Lee JA (1949) *The anaesthetic out-patient clinic.* Anaesthesia 4(4):169-74

[124] Lehaney B, Clarke SA, Paul RJ (1999) *A case of an intervention in an outpatients department*. Journal of the Operational Research Society 50(9):877-891

[125] Lew E, Pavlin DJ, Amundsen L (2004) *Outpatient preanaesthesia evaluation clinics*. Singapore Medical Journal 45(11):509-516

[126] Liao CJ, Pegden CD, Rosenshine M (1993) *Planning timely arrivals to a stochastic production or service system*. IIE Transactions 25(5):63-73

[127] Little JDC (1961) *A proof for the queuing formula* $L = \lambda W$. Operations Research 9(3):383-387

[128] Litvak N, van Rijsbergen M, Boucherie RJ, van Houdenhoven M (2008) *Managing the overflow of intensive care patients*. European Journal of Operational Research 185(3):998-1010

[129] Liu J, Tao L, Xiao B (2011) *Discovering the impact of preceding units' characteristics on the wait time of cardiac surgery unit from statistic data*. PLoS ONE 6(7):e21959

[130] Liu L, Liu X (1998) *Dynamic and static job allocation for multi-server systems*. IIE Transactions 30(9):845-854

[131] Liu N, Ziya S, Kulkarni VG (2010) *Dynamic scheduling of outpatient appointments under patient no-shows and cancellations*. Manufacturing & Service Operations Management 12(2):347-364

[132] Mallik S (2007) *Contracting over multiple parameters: capacity allocation in semiconductor manufacturing.* European Journal of Operational Research 182(1):174-193

[133] Mayhew L, Smith D (2008) *Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments.* Health Care Management Science 11(1):11-21

[134] McIntosh C, Dexter F, Epstein RH (2006) *The impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: a tutorial using data from an Australian hospital.* Anesthesia & Analgesia 103(6):1499-1516

[135] McManus ML, Long MC, Cooper A, Litvak E (2004) *Queuing theory accurately models the need for critical care resources*. Anesthesiology 100(5):1271-1276

[136] Morton A, Cornwell J (2009) *What's the difference between a hospital and a bottling factory?* British Medical Journal 339(2727):428-430

[137] Moskop JC, Sklar DP, Geiderman JM, Schears RM, Bookman KJ (2009) *Emergency department crowding, part 1 - concept, causes, and moral consequences.* Annals of Emergency Medicine 53(5):605-611

[138] Moskop JC, Sklar DP, Geiderman JM, Schears RM, Bookman KJ (2009) *Emergency department crowding, part 2 - barriers to reform and strategies to overcome them.* Annals of Emergency Medicine 53(5):612-617

[139] Müller-Langer F (2007) *A game theoretic analysis of parallel trade and the pricing of pharmaceutical products.* German Working Papers in Law and Economics 6:1-42

[140] Murray M, Berwick DM (2003) *Advanced access: reducing waiting and delays in primary care*. Journal of the American Medical Association 289(8):1035-1040

[141] Nelson RD (1995) *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modelling.* Springer, New York, NY, USA

[142] Neuts MF (1989) *Structured stochastic matrices of $M/G/1$ type and their applications*. Marcel Dekker New York, NY, USA

[143] Novelline RA, (2004) *Squire's fundamentals of radiology*. 6th ed. Harvard University Press, Cambridge, MA, USA

[144] OECD Health Data 2011, retrieved December 9, 2011, from *www.oecd.org/health/healthdata*

[145] ORchestra Bibliography, available online at *www.utwente.nl/choir/en/orchestra/*

[146] Osborne MJ, Rubinstein A (1994) *A course in game theory.* The MIT Press, Cambridge, MA, USA

[147] Osorio C, Bierlaire M (2009) *An analytic finite capacity queueing network model capturing the propagation of congestion and blocking.* European Journal of Operational Research 196(3):996-1007

[148] Pandit JJ, Dexter F (2009) *Lack of sensitivity of staffing for 8-hour sessions to standard deviation in daily actual hours of operating room time used for surgeons with long queues.* Anesthesia & Analgesia 108(6):1910-1915

[149] Parker BM, Tetzlaff JE, Litaker DL, Maurer WG (2000) *Redefining the preoperative evaluation process and the role of the anesthesiologist*. Journal of Clinical Anesthesia 12(5):350-356

[150] Patient Flow Improvement Center Amsterdam, Erlang-B calculator, retrieved December 9, 2011, from *www.vumc.nl/afdelingen/pica/Software/erlang_b/*

[151] Patrick J, Puterman ML (2007) *Improving resource utilization for diagnostic services through flexible inpatient scheduling: a method for improving resource utilization*. Journal of the Operational Research Society 58(2):235-245

[152] Patrick J, Puterman ML, Queyranne M (2008) *Dynamic multi-priority patient scheduling for a diagnostic resource*. Operations Research 56(6):1507-1525

[153] Pegden CD, Rosenshine M (1990) *Scheduling arrivals to queues*. Computers & Operations Research 17(4):343-348

[154] Pham DN, Klinkert A (2008) *Surgical case scheduling as a generalized job shop scheduling problem.* European Journal of Operational Research 185(3):1011-1025

[155] Pollard JB (2002) *Economic aspects of an anesthesia preoperative evaluation clinic*. Current Opinion in Anaesthesiology 15(2):257-261

[156] Preater J (2002) *Queues in health*. Health Care Management Science 5(4):283

[157] Pullman M, Rodgers S (2010) *Capacity management for hospitality and tourism: a review of current approaches*. International Journal of Hospitality Management 29(1):177-187

[158] Puterman ML (1994) *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, New York, NY, USA

[159] QtsPlus Software, available at *ftp://ftp.wiley.com/public/sci_tech_med/queueing_theory/*

[160] Qu X, Rardin RL, Williams JAS, Willis DR (2007) *Matching daily healthcare provider capacity to demand in advanced access scheduling systems*. European Journal of Operational Research 183(2):812-826

[161] Qu X, Shi J (2009) *Effect of two-level provider capacities on the performance of open access clinics*. Health Care Management Science 12(1):99-114

[162] Ramakrishnan M, Sier D, Taylor PG (2005) *A two-time-scale model for hospital patient flow*. IMA Journal of Management Mathematics 16(3):197-215

[163] Reilly TA, Marathe VP, Fries BE (1978) *A delay-scheduling model for patients using a walk-in clinic*. Journal of Medical Systems 2(4):303-313

[164] Rising EJ, Baron R, Averill B (1973) *A systems analysis of a university-health-service outpatient clinic*. Operations Research 21(5):1030-1047

[165] Robert P (2003) *Stochastic networks and queues*. Springer, New York, NY, USA

[166] Roberts DC, McKay MP, Shaffer A (2008) *Increasing rates of emergency department visits for elderly patients in the United States, 1993 to 2003*. Annals of Emergency Medicine 51(6):769-774

[167] Ross SM (1982) *Introduction to stochastic dynamic programming.* Academic Press, New York, NY, USA

[168] Sappington D (1983) *Limited liability contracts between principal and agent.* Journal of Economic Theory 29(1):1-21

[169] Schäfer W, Kroneman M, Boerma W, van den Berg M, Westert G, Devillé W, van Ginneken E (2010) *The Netherlands: health system review*. Health Systems in Transition 12(1):1-229

[170] Schehrer RG (1997) *A two moment method for overflow systems with different mean holding times*. Proceedings of the 15th International Teletraffic Congress

[171] Schofield WN, Rubin GL, Piza M, Yin Lai Y, Sindhusake D, Fearnside MR, Klineberg PL (2005) *Cancellation of operations on the day of intended surgery at a major Australian referral hospital.* Medical Journal of Australia 182(12):612-615

[172] Schull MJ, Szalai J, Schwartz B, Redelmeier DA (2001) *Emergency department overcrowding following systematic hospital restructuring: trends at twenty hospitals of ten years.* Academic Emergency Medicine 8(11):1037-1043

[173] Scott I, Vaughan L, Bell D (2009) *Effectiveness of acute medical units in hospitals: a systematic review.* International Journal for Quality in Health Care 21(6):397-407

[174] Sickinger S, Kolisch R (2009) *The performance of a generalized Bailey-Welch rule for outpatient appointment scheduling under inpatient and emergency demand*. Health Care Management Science 12(4):408-419

[175] Stoer J, Bulirsch R (2002) *Introduction to numerical analysis*. 3rd ed. Springer New York, NY, USA

[176] Strum DP, Vargas LG, May JH, Bashein G (1997) *Surgical suite utilization and capacity planning: a minimal cost analysis model.* Journal of Medical Systems 21(5):309-322

[177] Su S, Shih CL (2003) *Managing a mixed-registration-type appointment system in outpatient clinics*. International Journal of Medical Informatics 70(1):31-40

[178] Su-sheng W, Zhao-kun K, Charlene X, Jing X (2008) *Bayesian Nash equilibrium analysis of medical market under asymmetry information.* Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing

[179] Swisher JR, Jacobson SH, Jun JB, Balci O (2001) *Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation*. Computers & Operations Research 28(2):105-125

[180] Taha HA (1997) *Operations research: an introduction.* Prentice Hall, Englewood Cliffs, NJ, USA

[181] Takagi H (1988) *Queuing analysis of polling models*. ACM Computing Surveys (CSUR) 20(1):5-28

[182] Takagi H (2000) *Analysis and application of polling models.* In: Haring G, Lindemann C, Reiser M (eds) *Performance evaluation: origins and directions.* Lecture Notes in Computer Science 1769, Springer Verlag, Berlin, Germany

[183] Takahashi Y, Hashida O (1991) *Delay analysis of discrete-time priority queue with structured inputs*. Queueing Systems 8(1):149-164

[184] Taylor HM, Karlin S (1998) *An introduction to stochastic modeling*. 3rd ed. Academic Press, San Diego, CA, USA

[185] Thomas SJ, Williams MV, Burnet NG, Baker CR (2001) *How much surplus capacity is required to maintain low waiting times?* Clinical Oncology 13(1):24-28

[186] Thomson W (2003) *Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: a survey.* Mathematical Social Sciences 45(3):249-297

[187] Tijms HC (2003) *A first course in stochastic models.* John Wiley & Sons, Chichester, UK

[188] Tucker JB, Barone JE, Cecere J, Blabey RG, Rha CK (1999) *Using queueing theory to determine operating room staffing needs*. Journal of Trauma 46(1):71-79

[189] Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, Litvak N (2010) *A survey of health care models that encompass multiple departments*. International Journal of Health Management and Information 1(1):37-69

[190] Vandaele N, van Nieuwenhuyse I, Cupers S (2003) *Optimal grouping for a nuclear magnetic resonance scanner by means of an open queueing model*. European Journal of Operational Research 151(1):181-192

[191] Varian HR (1992) *Microeconomic analysis*. 3rd ed. W.W. Norton & Company, New York, NY, USA

[192] Vermeulen IB, Bothe SM, Elkhuizen SG, Lameris H, Bakker PJM, Poutré HL (2009) *Adaptive resource allocation for efficient patient scheduling.* Artificial Intelligence in Medicine 46(1):67-80

[193] Villa S, & Patrone F (2009) *Incentive compatibility in kidney exchange problems.* Health Care Management Science 12(4):351-362

[194] de Vuyst S, Wittevrongel S, Bruneel H (2005) *Delay differentiation by reserving space in queue*. Electronics Letters 41(9):564

[195] Walraevens J, Steyaert B, Bruneel H (2002) *Delay characteristics in discrete-time $GI/G/1$ queues with non-preemptive priority queueing discipline*. Performance Evaluation 50(1):53-75

[196] Wang PP (1999) *Sequencing and scheduling N customers for a stochastic server*. European Journal of Operational Research 119(3):729-738

[197] Westbay Online Traffic Calculators, Erlang-B Calculator, retrieved December 9, 2011, from *www.erlang.com/calculator*

[198] Westert GP, Burgers JS, Verkleij H (2009) *The Netherlands: regulated competition behind the dykes?* Brithish Medical Journal 339(7725):839-842

[199] Whitt W (1983) *The queueing network analyzer.* The Bell System Technical Journal 62(9):2779-2815

[200] Wilkinson RI (1956) *Theories for toll traffic engineering in the U.S.A.* The Bell System Technical Journal 35:421-514

[201] Williams P, Tai G, Lei Y (2010) *Simulation based analysis of patient arrival to health care systems and evaluation of an operations improvement scheme*. Annals of Operations Research 178(1):263-279

[202] Winston WL (1994) *Operations research: applications and algorithms.* 3th ed. Duxbury Press, Belmont, CA, USA

[203] Wolff RW (1989) *Stochastic modeling and the theory of queues.* Prentice Hall, Englewood Cliffs, NJ, USA

[204] Worthington DJ (1987) *Queueing models for hospital waiting lists*. Journal of the Operational Research Society 38(5):413-422

[205] Wullink G, van Houdenhoven M, Hans EW, van Oostrum JM, van der Lans M, Kazemier G (2007) *Closing emergency rooms improves efficiency.* Journal of Medical Systems 31(6):543-546

[206] Xiao T, Yang D (2009) *Risk sharing and information revelation mechanism of a one-manufacturer and one-retailer supply chain facing an integrated competitor.* European Journal of Operational Research 196(3):1076-1085

[207] Xie B, Dilts DM, Shor M (2006) *The physician-patient relationship: the impact of patient-obtained medical information.* Health Economics 15(8):813-833

[208] Xie H, Chaussalet T, Rees M (2007) *A semi-open queueing network approach to the analysis of patient flow in healthcare systems.* Proceedings of the 20th IEEE International Symposium on Computer-Based Medical Systems

[209] Young T, Brailsford S, Connell C, Davies R, Harper P, Klein JH (2004) *Using industrial processes to improve patient care*. British Medical Journal 328(7432):162-164

[210] Zachary S, Ziedins I (2011) *Loss networks.* In: Boucherie RJ, van Dijk NM (eds) *Queueing networks: a fundamental approach.* Springer, New York, NY, USA

[211] Zeng G (2003) *Two common properties of the Erlang-B function, Erlang-C function, and Engset blocking function.* Mathematical and Computer Modelling 37(12-13):1287-1296

[212] Zonderland ME, Boer F, Boucherie RJ, de Roode A, van Kleef JW (2009) *Redesign of a university hospital preanesthesia evaluation clinic: a queuing theory approach*. Anesthesia & Analgesia 109(5):1612-1621

[213] Zonderland ME, Boucherie RJ, Litvak N, Vleggeert-Lankamp CLAM (2010) *Planning and scheduling of semi-urgent surgeries*. Health Care Management Science 13(3):256-267.

Acronyms

CI	Confidence Interval
CT	Computed Tomography (Scanner)
ED	Emergency Department
EOA	Emergency Observation and Assessment (Ward)
ERM	Equivalent Random Method
FCFS	First Come First Serve
ICU	Intensive Care Unit
LCFS	Last Come First Serve
LUMC	Leiden University Medical Center
LOS	Length of Stay
MRI	Magnetic Resonance Imaging (Scanner)
OR	Operating Rooms
PAC	Preanesthesia Evaluation Clinic
PS	Processor Sharing
QNA	Queuing Network Analyzer
scv	Squared coefficient of variation

Summary

In this dissertation we study several problems related to the management of healthcare and the cure of disease. In each chapter a hospital capacity distribution problem is analyzed using techniques from operations research, also known as mathematical decision theory. The problems considered are inspired by logistical challenges faced by Leiden University Medical Center (LUMC). Several of the solutions we present in this dissertation have been implemented at LUMC.

Considering our aging population, shrinking workforce and the current hospital efficiency levels it will be difficult, if not impossible, to provide an appropriate level of care for the sick and the elderly in the coming decades. Given what is currently at stake, it is hard to understand that it is quite common in hospitals to avoid explicit decisions on resource allocation and capacity distribution and just anticipate on ad-hoc basis on problems that occur. This is sometimes accompanied with very undesirable system outcomes such as patient cancellations and extremely long access (the time the patient spends on the waiting list) or waiting (the time the patient spends in the hospital waiting) times. The models we present allow for a quantification of consequences of capacity distribution decisions. The item that is distributed can either be time, or another kind of resource such as staffed beds. With the models a clear and succinct understanding of the problem, its possible solutions, and implications of these solutions can be obtained.

This dissertation consists of three parts. The first part serves as an introduction and contains the Introduction Chapter, 1, which discusses recent developments in the healthcare sector and the role of Operations Research therein, and Chapter 2, which provides an introduction to queues, networks of queues, and their applications in healthcare.

In the second part of the dissertation we focus on challenges for outpatient clinics and diagnostic facilities. In Chapter 3 we study the reorganization of an outpatient clinic. We demonstrate how the involvement of essential employees combined with a queuing network model designed to support the decision making process results in a successful intervention. Key points in the intervention are the rescheduling of appointments and the reallocation of tasks.

Chapter 4 presents a methodology to develop appointment schedules for outpatient clinics with unscheduled (walk-in) and scheduled (appointment) patients. The goal is an appointment schedule that keeps waiting time at the facility for unscheduled patients

below an acceptable level, while controlling the access time for scheduled patients. A cyclic queuing and a Markov decision model are combined with an algorithm to determine an appointment schedule that satisfies all requirements.

Chapter 5 is motivated by the increasing popularity of care pathways in outpatient clinics. Hospitals aim to optimize the flow of patients in a care pathway by prioritizing them in the appointment planning process. As a result, regular patients who are not in a care pathway may experience increased waiting times. We develop a queuing model with a reservation scheme that allows for a trade-off between the accessibility for patients from the care pathway and waiting time for regular patients at an outpatient clinic.

In Chapter 6 we consider an MRI scanning facility run by a Radiology department. Several medical departments compete for capacity and have private information regarding their demand for scans. The fairness of the capacity allocation by the Radiology department depends on the quality of the information provided by the medical departments. We employ a game-theoretic approach that stimulates the disclosure of true demand, so that capacity can be allocated fairly.

In the last part we study challenges that evolve when urgent and elective patient flow meet. Chapter 7 studies the trade-off between cancellations of elective surgeries due to semi-urgent surgeries, and unused operating room (OR) time due to excessive reservation of OR time for semi-urgent surgeries. Semi-urgent surgeries, to be performed soon but not necessarily today, pose an uncertain demand on available hospital resources, and interfere with the planning of elective patients. For a highly utilized OR, reservation of OR time for semi-urgent surgeries avoids excessive cancellations of elective surgeries, but may also result in unused OR time, since arrivals of semi-urgent patients are unpredictable. We use a discrete-time queuing and a Markov decision model to smooth the planning process.

Using the methodology presented in Chapter 7, part of the OR capacity of the Neurosurgery department at LUMC was allocated to semi-urgent surgeries. In Chapter 8 we study the implementation process and the effect of dedicating OR slots to semi-urgent surgeries on elective patient cancellations and OR utilization.

A recent development to reduce Emergency Department crowding and increase urgent patient admissions is the opening of an Emergency Observation and Assessment Ward (EOA Ward). At these wards urgent patients are temporarily hospitalized until they can be transferred to an inpatient bed. In Chapter 9 we present an overflow model to evaluate the effect of employing an EOA Ward on elective and urgent patient admissions.

All models presented in parts II and III allow for a quantitative analysis of resource distribution problems in healthcare. We can conclude that mathematical modeling contributes to higher quality, more sound decision making in healthcare.

Samenvatting

In dit proefschrift bestuderen we verschillende problemen gerelateerd aan de organisatie van de gezondheidszorg en de behandeling van patiënten. Gebruikmakend van technieken uit de Operationele Research, ook wel bekend als Mathematische Besliskunde, wordt in elk hoofdstuk een probleem geanalyseerd dat betrekking heeft op de verdeling van (een gedeelte van) de ziekenhuiscapaciteit. Logistieke uitdagingen, geïnspireerd op de dagelijkse praktijk van het Leids Universitair Medisch Centrum (LUMC), vormen de basis van de bestudeerde materie. Verschillende oplossingen die in dit proefschrift gepresenteerd worden, zijn ondertussen geïmplementeerd in het LUMC.

De vergrijzing, de kleiner wordende beroepsbevolking en het huidige efficiëntieniveau in de meeste ziekenhuizen, maken het moeilijk, zo niet onmogelijk, om voldoende zorg en verzorging voor de ouderen en zieken in onze samenleving te garanderen. Gezien wat er op het spel staat, is het moeilijk te begrijpen dat het in ziekenhuizen eerder regel dan uitzondering is om expliciete beslissingen over de verdeling van schaarse goederen en capaciteit te vermijden. Vaak wordt ad hoc geanticipeerd op problemen die spontaan lijken te ontstaan en vergezeld worden door ongewenste bij-effecten zoals het afzeggen van patiënten voor behandelingen en zeer lange toegangstijden (de tijd die de patiënt op de wachtlijst staat) en wachttijden (de tijd die de patiënt wacht in het ziekenhuis). De modellen die we in dit proefschrift presenteren maken de gevolgen van capaciteitsbeslissingen inzichtelijk. Wat voor soort capaciteit verdeeld wordt verschilt; voorbeelden zijn tijd (bij de arts, op de MRI scanner) of een bed op de verpleegafdeling. Met de modellen wordt een diepgaand inzicht verkregen van het probleem, de mogelijke oplossingen en de gevolgen van de gekozen oplossing.

Dit proefschrift bestaat uit drie delen. In het eerste deel wordt een inleiding op recente ontwikkelingen in de gezondheidszorg en de rol van Operationele Research hierin (Hoofdstuk 1), gevolgd door een inleiding in de wachtrijtheorie, met speciale aandacht voor netwerken van wachtrijen en toepassingen van wachtrijtheorie in de gezondheidszorg (Hoofdstuk 2).

In het tweede gedeelte van dit proefschrift leggen we de focus op uitdagingen voor poliklinieken en diagnostische afdelingen. In Hoofdstuk 3 bestuderen we de reorganisatie van een polikliniek. We demonstreren hoe een netwerk van wachtrijen, ontwikkeld om het besluitsvormingsproces te ondersteunen, het personeel van de polikliniek faciliteert bij het uitvoeren van een succesvolle interventie. De belangrijkste veranderingen zijn

het herverdelen van taken en het op een ander moment plannen van afspraken.

Hoofdstuk 4 presenteert een methodologie waarmee afspraakschema's voor polikliniekken met zowel inloop- als afspraakpatiënten ontworpen kunnen worden. Het doel is een afspraakschema waarbij de wachttijd voor inlooppatiënten acceptabel is, terwijl de toegangstijd voor afspraakpatiënten beperkt blijft. Een cyclisch wachtrijmodel en een Markov beslissingsmodel worden gecombineerd met een algoritme om een afspraakschema te ontwerpen dat aan alle eisen voldoet.

Hoofdstuk 5 is gemotiveerd door de toenemende populariteit van zorgpaden. Ziekenhuizen proberen de patiëntenstroom in een zorgpad zoveel mogelijk te stroomlijnen door deze patiënten voorrang te geven bij het plannen van afspraken. Een direct gevolg is dat patiënten die niet in een zorgpad zijn opgenomen, hinder kunnen ondervinden in de vorm van langere wacht- en toegangstijden. We ontwikkelen een wachtrijmodel met een reserveringsschema, waarbij we voor een polikliniek een afweging maken tussen de beschikbaarheid van afspraken voor zorgpadpatiënten enerzijds, en de wachttijd voor reguliere patiënten anderzijds.

In Hoofdstuk 6 beschouwen we een MRI afdeling. Verschillende medische afdelingen proberen zoveel mogelijk tijd op de scanner te verwerven, en geven daarbij soms een onjuiste inschatting van het verwachte aantal scans voor de komende periode. De schaarse MRI capaciteit kan alleen eerlijk verdeeld worden wanneer alle medische afdelingen een goede inschatting geven. We laten zien, door een speltheoretisch model toe te passen, dat het mogelijk is om de afdelingen zo te stimuleren dat ze dit ook doen.

In het laatste gedeelte van dit proefschrift bestuderen we uitdagingen die ontstaan wanneer spoed en electieve (planbare) patiëntenstromen elkaar kruisen. Hoofdstuk 7 behandelt de afweging tussen het afzeggen van electieve operaties door de tussenkomst van semi-spoed patiënten, en ongebruikte operatietijd door het reserveren van teveel tijd voor deze patiënten. Door de onvoorspelbaarheid van het precieze aantal semi-spoed patiënten per week is de benodigde hoeveelheid operatietijd onzeker en komt de planning van electieve patiënten in het gedrang. Wanneer de operatiekamers (OK) veel gebruikt worden, biedt het reserveren van tijd voor semi-spoed patiënten uitkomst, aangezien zo minder electieve patiënten afgezegd worden. Echter, door de eerder genoemde onvoorspelbaarheid kan het ook zijn dat de OK leeg staat. We presenteren een wachtrijmodel en een Markov beslissingsmodel om het planningsproces te reguleren.

Naar aanleiding van de resultaten uit Hoofdstuk 7 werd door de afdeling Neurochirurgie van het LUMC een gedeelte van de OK capaciteit toegewezen aan semi-urgente operaties. In Hoofdstuk 8 presenteren we een praktijkstudie, waarbij het implementatieproces en het effect van de reservering op het aantal afgezegde electieve operaties en de OK bezetting bestudeerd worden.

De acute opnameafdeling is een recente ontwikkeling, bedoeld om beddentekorten op de reguliere verpleegafdelingen te omzeilen en zo het aantal spoedopnames via de eerste hulp te verhogen. We analyseren in Hoofdstuk 9, met behulp van een overflow model, de invloed van de acute opnameafdeling op het aantal opgenomen spoed- en electieve patiënten.

Alle modellen gepresenteerd in deel II en III maken een kwantitatieve analyse van capaciteitsverdeelproblemen mogelijk. We kunnen concluderen dat wiskundig modelleren in positieve zin bijdraagt aan besluitvormingsprocessen in de gezondheidszorg.

About the Author

Maartje Zonderland was born in Warnsveld (the Netherlands) on January 24, 1982. She obtained her VWO diploma, with an International Baccalaureate Certificate in English, at the Lorentz College in Arnhem in 1999. In 2007, Maartje completed her studies at the University of Twente having received B.Sc. degrees in Industrial Engineering & Management and Applied Mathematics, and an M.Sc. degree in Applied Mathematics. She carried out her final master's project at Leiden University Medical Center, which lead to an appointment as a staff consultant at this hospital. Additionally, Maartje started with her Ph.D. studies at the University of Twente, supervised by prof. dr. Richard Boucherie and dr. Fred Boer. In the autumn of 2010, Maartje spend two months at the Centre for Research in Healthcare Engineering, University of Toronto in Canada to work with prof. Michael Carter. She returned to Canada in the summer of 2011, this time to work with prof. David Stanford at the University of Western Ontario, London. Her Ph.D. research culminates with this dissertation.

Publications

1. Zonderland ME, Boer F, Boucherie RJ, de Roode A, van Kleef JW (2009) *Redesign of a University Hospital Preanesthesia Evaluation Clinic: a Queuing Theory Approach*. Anesthesia & Analgesia 109(5):1612-1621 (basis for Chapter 3).

2. Zonderland ME, Boucherie RJ, Litvak N, Vleggeert-Lankamp CLAM (2010) *Planning and Scheduling of Semi-Urgent Surgeries*. Health Care Management Science 13(3):256-267 (basis for Chapter 7).

3. Hulshof PJH, Boucherie RJ, van Essen JT, Hans EW, Hurink JL, Kortbeek N, Litvak N, Vanberkel PT, van der Veen E, Veltman B, Vliegen IMH, Zonderland ME (2011) *ORchestra: an online reference database of OR/MS literature in health care*. Health Care Management Science 14(4):383-384.

4. Zonderland ME, Boucherie RJ (2011) *Queuing Networks in Healthcare Systems*. In: Randolph W. Hall (eds.) *Handbook of Healthcare System Scheduling*. Springer, New York, NY, USA (basis for Chapter 2).

5. Zonderland ME, Boucherie RJ, Al Hanbali A (2011) *Appointments for Care Pathway Patients*. Memorandum 1961, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands (basis for Chapter 5).

6. Zonderland ME, Boucherie RJ, Carter MW, Stanford DA (2011) *The Emergency Observation and Assessment Ward*. Memorandum 1967, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands (basis for Chapter 9).

7. Kortbeek N, Zonderland ME, Boucherie RJ, Litvak N, Hans EW (2011) *Designing Cyclic Appointment Schedules for Systems with Scheduled and Unscheduled Arrivals*. Memorandum 1968, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands (basis for Chapter 4).

8. Zonderland ME, Timmer JB (2011) *Optimal Allocation of MRI Scan Capacity among Competing Hospital Departments*. To appear in: European Journal of Operational Research, special issue on Operations Research in Healthcare (basis for Chapter 6).