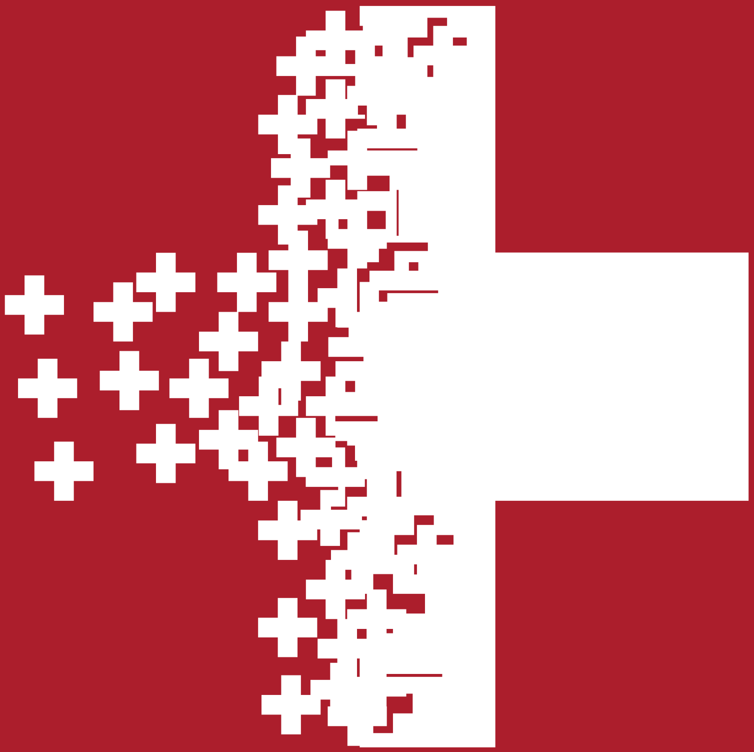
Integrated Decision Making in Healthcare

An Operations Research and Management Science Perspective



Peter J.H. Hulshof

Integrated Decision Making in Healthcare

An Operations Research and Management Science Perspective

Peter J.H. Hulshof

Graduation committee:

Chairman and secretary:	Prof. dr. K.I. van Oudenhoven - van der Zee
Promotors:	Prof. dr. ir. E.W. Hans Prof. dr. R.J. Boucherie
Members:	Prof. dr. P.R. Harper Prof. dr. G. Kazemier Dr. ir. M.R.K. Mes Prof. dr. M. Uetz Prof. dr. B. Werners Prof. dr. W.H.M. Zijm

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Center for Telematics and Information Technology (No. 13-269, ISSN 1381-3617)
Beta Research School for Operations Management and Logistics (No. D-172)
Industrial Engineering and Business Information Systems

Center for Healthcare Operations Improvement and Research
University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands

This research was financially supported by the Dutch Technology Foundation STW by means of the project 'Logistical Design for Optimal Care' (No. 08140)

Publisher: Ipskamp Drukkerijen, Enschede, the Netherlands
Cover design: Bart de Wit van Weperen

ISBN 978-90-365-0858-2
DOI 10.3990/1.9789036508582

INTEGRATED DECISION MAKING IN HEALTHCARE

AN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE PERSPECTIVE

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 donderdag 21 november 2013 om 16.45 uur

door

Peter Jan Hendrik Hulshof

geboren op 19 juni 1983
te Groenlo, Nederland

This thesis is approved by promotors:

Prof. dr. ir. Erwin W. Hans
Prof. dr. Richard J. Boucherie

Contents

CHAPTER 1

Introduction, motivation and outline

The pressure on healthcare systems rises as both demand for healthcare and expenditures are increasing steadily. As a result, healthcare professionals face the challenging task to design and organize the healthcare delivery process more effectively and efficiently. Designing and organizing processes is known as planning and control. Healthcare planning and control lags behind manufacturing and control for various reasons. One of the main causes is the fragmented nature of healthcare planning and control. Healthcare organizations such as hospitals are typically organized as a cluster of autonomous departments, where planning and control is also often functionally dispersed. As the clinical course of patients traverses multiple, thus interdependent, departments, an integrated approach to healthcare planning and control is likely to bring improvements.

This thesis aims to contribute to integrated decision making in healthcare in two ways. First, we develop a framework, taxonomy, and extensive literature review to support healthcare professionals in structuring and positioning planning and control decisions in healthcare. The framework and taxonomy can be used to break down planning functions, determine managerial responsibilities and deficiencies, and to identify adjacent planning decisions influencing each other. Second, we propose planning approaches to develop resources allocation and patient admission plans for multiple departments, multiple resources and multiple patient types, thereby integrating decision making for a chain of healthcare resources. The planning approaches are developed with techniques from Operations Research and Management Science (OR/MS).

This introductory chapter is organized as follows. In Section 1.1, we discuss planning and control and investigate the reasons why healthcare planning and control is lagging behind manufacturing planning and control. In Section 1.2, we discuss integrated decision making in healthcare. Section 1.3 introduces the field of OR/MS, and Section 1.4 explains the outline of this thesis.

1.1 Healthcare planning and control lags behind

In recent decades, developed countries have achieved major improvements in population health. The average life expectancy at birth went from 74.6 in 1980 to 81.2 in 2010 for several developed countries [378]. Infant mortality rates have dropped in the same period from 10.7 to 3.7 per 1000 live births. The significant improvement in population health comes with rapidly and steadily increasing healthcare costs. Figure 1.1 illustrates that the absolute healthcare costs per capita have increased dramatically in various developed countries between 1980 and 2010. In this period, the average percentage of GDP spent on healthcare in these countries has increased from 7.2% to 11.2% [378], and this is forecasted to grow further. For example, in the Netherlands it may increase to 22-31% in 2040 [92].

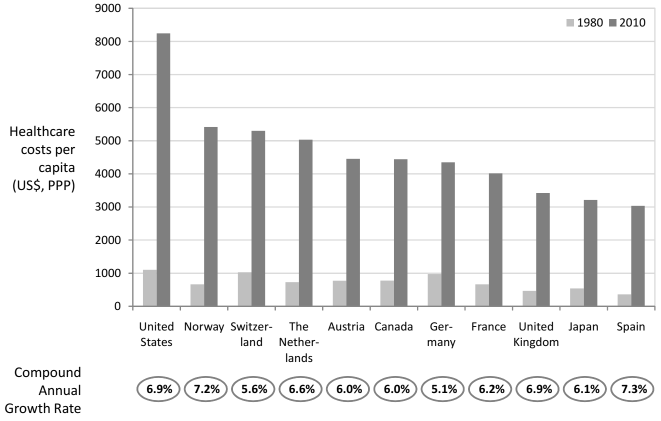


Figure 1.1: Healthcare costs per capita in US$ for several OECD countries. The Compound Annual Growth Rate depicts the average year-over-year growth rate of healthcare costs in the period 1980 to 2010. Source: OECD Health Data 2013 [378].

Growing healthcare costs increase the pressure on healthcare systems to investigate and implement innovative methods to contain these costs. In 2006, the Dutch government decided to introduce a controlled form of competition in healthcare, to spur a decrease in healthcare costs and continue to provide equitable access to good quality healthcare [516]. This has led to significant competition and consolidation in the Dutch health insurers market, where currently the largest four insurers have over 90% market share. By merging, these healthcare insurers have increased their negotiating power, which enables them to demand lower prices from healthcare providers. (In addition, competition between healthcare providers has increased, partly due to the entry of a significant number of freestanding clinics [435]. As a result of these developments, healthcare providers face the challenging task to organize their processes more effectively and efficiently. Designing and organizing processes is known as planning and control, which comprises integrated coordination of resources (staff, equipment and materials) and product flows, in such a way that the organization's objectives are realized [9].

Planning and control has a rich tradition in manufacturing [234]. Due to the increase in demand and expenditure for healthcare [384], planning and control in healthcare has received a growing amount of attention over the last ten years, both in practice and in the literature. It is therefore not surprising that the Operations Research & Management Science (OR/MS) research community's interest in healthcare applications is rapidly increasing as well [59]. In fact, the attendance of the conference of the EURO Working Group on Operational Research Applied to Health Services (ORAHS) [382] has increased from around 50 in 2002 to 140 in 2013, and involves an increasing number of countries. Within these research efforts, planning and control is a key focal area - the subject of more than 35% of the ORAHS publications [61]. INFORMS organized the first conference on healthcare in 2009, presenting over 400 abstracts in a variety of OR healthcare topics [234].

The growing attention for resource capacity planning and control in healthcare contributes to further closing the gap with planning and control in manufacturing. Common reasons stated in the literature for the fact that this gap exists include [234]:

1. Healthcare organizations are professional organizations that often lack cooperation between, or commitment from, involved parties (doctors, administrators, etc.). These groups have their own, sometimes conflicting, objectives, as is aptly illustrated by Glouberman and Mintzberg in their "four faces of healthcare" framework [200, 201].

2. Due to the state of information systems in healthcare, crucial information required for planning and control is often not available [87]. Although Diagnosis Related Groups (DRGs) and electronic health record systems have spurred the need for financial and clinical information management systems, these systems tend to be poorly integrated with operational information systems. This lack of integration is impeding the advance of integrated planning and control in healthcare, both organization-wide and between organizations. This was recognized already in 1995 [425], but developments until now have been slow [290].

3. Since large healthcare providers, such as hospitals, generally consist of autonomously managed departments, managers tend not to look beyond the border of their department, and planning and control is fragmented [401, 425].

4. The Hippocratic Oath taken by doctors forces them to focus on the patient at hand, whereas planning and control addresses the entire patient population, both within and beyond the scope of an individual doctor [354, 353].

5. While healthcare managers are dedicated to provide the best possible service, they lack the knowledge and training to make the best use of the available resources [87].

6. As healthcare managers often feel that investing in better administration diverts funds from direct patient care [87], managerial functions are often ill-defined, overlooked, poorly addressed, or functionally dispersed.

Four of these six reasons mention the fragmented nature of healthcare delivery and the lack of an integrated perspective on decision making in practice as the main cause why healthcare planning and control lags behind. In the next section, we discuss an integrated healthcare planning and control perspective in detail.

1.2 Integrated planning and control

The healthcare delivery process from the patient's perspective generally is a composition of several care services, and a patient's pathway typically includes several care stages performed by various healthcare services and professionals. The effectiveness and efficiency of healthcare delivery is a result of planning and control decisions made for the care services involved in each care stage. The quality of decisions in each care service depends on the information available from and the restrictions imposed by other care services [263]. As such, patient care pathways connect multiple departments and resources together as an integrated network. Fluctuations in both patient arrivals (e.g., seasonality) and resource availability (e.g., holidays) at one department may impact the entire care pathway. For patients, this may result in varying waiting times for each separate stage in a care process, and from a hospital's perspective, this results in varying resource utilizations and service levels. Suboptimization is a threat when at each stage in the patient's care pathway, resource capacity decisions are taken in isolation. Therefore, integrated decision making on all involved resources, taking into account a care chain perspective [80, 231, 401], seems necessary.

Due to the segmented organizational structure of healthcare delivery, also the OR/MS literature has initially focused predominantly on autonomous, isolated decision making. Such models ignore the inherent complex interactions between decisions for different services in different organizations and departments. Although the benefits of an integrated approach are often recognized [80, 231, 401], relatively few articles integrate decision making for a chain of resources or departments along the patient's care process. In 1999, the survey [282] identified this void in OR/MS literature, and the level of complexity of such models is depicted as the main barrier. In 2010, the survey [491], reviewing OR/MS models that encompass patient flows across multiple departments, addressed the question whether this void has since been filled. The authors conclude that the lack of models for complete healthcare processes still existed. Although a body of literature focusing on two-departmental interactions was identified, very few contributions were found on complete hospital interactions, let alone on complete healthcare system interactions [263].

Part of why this void existed is the lack of planning frameworks for integrated planning and control in healthcare. Chapter 2 discusses the various frameworks available in healthcare, and illustrates that these frameworks are not suitable for multiple managerial areas, multiple departments or multiple healthcare organizations. In manufacturing planning and control such integrated frameworks, encompassing multiple managerial areas, are more common. To fill this void, we develop a healthcare planning and control framework and taxonomy in Chapters 2 and 3, which can be used to structure and identify planning and control decisions, such that relations between planning decisions, resouces and departments can be identified. The framework in Chapter 2 can be used to identify planning decisions for multiple managerial areas: financial planning, materials planning, medical planning and resource capacity planning. They are defined as follows. *Medical planning* comprises decision making by clinicians regarding medical protocols, treatments, diagnoses and triage. *Financial planning* addresses how an organization should manage its costs and revenues to achieve its objectives under current and future organizational and economic circumstances. *Materials planning* addresses the acquisition, storage, distribution and retrieval of all consumable resources/materials, such as suture materials, blood, bandages, food, etc. *Resource capacity planning* addresses the dimensioning, planning, scheduling, monitoring, and control of renewable resources. The taxonomy presented in Chapter 3 is a further specification of the framework developed in Chapter 2, within the managerial area of resource capacity planning. The taxonomy distinguishes between different services in healthcare, such as ambulatory care, inpatient care, and residential care.

The framework and taxonomy distinguish between several hierarchical decision making levels that reflect the disaggregation of decision making as time progresses and more information gradually becomes available [534]. We build upon the "classical" hierarchical decomposition often used in manufacturing planning and control, which discerns *strategic*, *tactical*, and *operational* levels of control [9]. We extend this decomposition by discerning between *offline* and *online* on the operational level. This distinction reflects the difference between "in advance" decision making and "reactive" decision making (monitoring and control). We will discuss these levels in more detail in Chapters 2 and 3.

Overlooked or poorly addressed managerial functions can be encountered on all levels of control [87], but are most often found on the tactical level of control [425]. In fact, to many, tactical planning is less tangible than operational planning and even strategic planning. Inundated with operational problems, managers are inclined to solve problems at hand (i.e., on the operational level). We refer to this phenomenon as the "real-time hype" of managers. A claim for "more capacity" is the universal panacea for many healthcare managers. It is, however, often overlooked that instead of such drastic strategic measures, tactically allocating and organizing the available resources may be more effective and cheaper. Consider for example a "master schedule" or "block plan", which is the tactical allocation of blocks of resource time (e.g., operating theatres, or CT-scanners) to specialties and/or patient categories during a week. Such a block plan should be periodically revised to react to variations in supply and demand. However, in practice, it is more often a result of "historical development" than of analytical considerations [501].

We investigate the voids in the literature with regards to models for integrated decision making and models that cover the tactical planning level with an extensive and structured literature review in Chapter 3. Our literature review confirms that both voids still exist in the literature. To fill these voids, we propose methods from OR/MS for tactical planning in Chapters 4 to 6. Chapters 4 and 5 propose approaches to allocate resources and develop a patient admission plan for multiple departments, multiple resources and multiple patient types, thereby integrating decision making for a chain of hospital resources. Chapter 6 discusses the tactical problem of patient flow through an outpatient clinic.

1.3 Operations Research and Management Science

Operations Research and Management Science (OR/MS) is an interdisciplinary branch of applied mathematics, engineering and sciences that uses various scientific research-based principles, strategies, and analytical methods including mathematical modeling, statistics and algorithms to improve an organization's ability to enact rational and meaningful management decisions [265]. OR/MS has been applied widely to resource capacity planning and control in manufacturing. Since the 1950s, the application of OR/MS to healthcare also accomplishes essential efficiency gains in healthcare delivery.

In OR/MS in healthcare, many different topics have been addressed in the literature, such as operating room planning, appointment scheduling and. Reference databases aim to provide a selective overview of the extensive set of articles published in healthcare. For example, the comprehensive bibliography on operating room management of Dexter [134]. The Center for Healthcare Operations Improvement and Research (CHOIR) of the University of Twente has introduced and maintains the online literature database 'ORchestra' [262, 383], in which references in the field of OR/MS in healthcare are categorized by medical and mathematical subject. All the articles mentioned in Chapter 3 are categorized in ORchestra [263].

The broad field of OR/MS contains various methods and techniques that can be applied to planning and control problems. We have aimed to categorize these in five broad categories, which are described in more detail in Chapter 3. In Chapters 4 to 6, we have tapped into the rich field of OR/MS by applying a wide variety of its methods and techniques. In Chapter 4, we apply *mathematical programming*, *heuristics* and *computer simulation*. In Chapter 5, we adopt techniques from *Markov processes*, *mathematical programming* and *computer simulation*. In Chapter 6, we use *queueing theory* and *computer simulation* to develop our conclusions.

1.4 Outline of this thesis

Chapter 1 introduces the main topics in this thesis.

Chapter 2 proposes a modern framework for healthcare planning and control. The framework integrates all managerial areas involved in healthcare delivery operations and all hierarchical levels of control, to ensure completeness and coherence of responsibilities for every managerial area.It serves as a foundation for the taxonomy which is presented in Chapter 3.

Published as: E.W. Hans, M. van Houdenhoven, P.J.H. Hulshof. A framework for healthcare planning and control. In: Handbook of Healthcare System Scheduling, Randolph Hall (editor), International Series in Operations Research & Management Science 168:303-320, 2012.

Chapter 3 provides a taxonomy to identify, structure and position planning and control decisions within the area of resource capacity planning and control in healthcare. Following the taxonomy, we provide a comprehensive overview of the typical decisions to be made in resource capacity planning and control in healthcare, and a structured review of relevant OR/MS articles for each planning decision.

Published as: P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, and P.J.M. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. Health Systems 1(2):129-175, 2012.

Chapter 4 proposes a method to develop a tactical resource allocation and elective patient admission plan. These tactical plans allocate available resources to various care processes and determine the selection of patients to be served that are at a particular stage of their care process. Our method is developed in a Mixed Integer Linear Programming (MILP) framework and copes with multiple resources, multiple time periods and multiple patient groups with various uncertain treatment paths through the hospital, thereby integrating decision making for a chain of hospital resources.

Published as: P.J.H. Hulshof, R.J. Boucherie, E.W. Hans, and J.L. Hurink. Tactical resource allocation and elective patient admission planning in care processes. Health Care Management Science 16(2):152-166, 2013.

Chapter 5 proposes a method to develop a tactical resource allocation and elective patient admission plan taking stochastic elements into consideration, thereby potentially providing more robust tactical plans. We develop our solution approach within the Approximate Dynamic Programming (ADP) framework, and it can also cope with with multiple resources, multiple time periods and multiple patient groups with various uncertain treatment paths through the hospital.

Submitted as: P.J.H. Hulshof, M.R.K. Mes, R.J. Boucherie, and E.W. Hans. Tactical planning in healthcare using Approximate Dynamic Programming. Memorandum

2014, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2013.

Chapter 6 evaluates two policies for patient flow through an outpatient clinic. This tactical planning problem is investigated with a queueing theoretic and a discrete-event simulation approach to evaluate the performance of the two policies for different parameter settings. These models can be used by managers of outpatient clinics to compare the two policies and choose a particular policy when redesigning the patient process, and to calculate the required number of consultation rooms each policy.

Published as: P.J.H. Hulshof, P.T. Vanberkel, R.J. Boucherie, E.W. Hans, M. van Houdenhoven, and J.C.W. van Ommeren. Analytical models to determine room requirements in outpatient clinics. OR Spectrum 34(2):391-405, 2012.

This thesis concludes with an **epilogue**, in which the main findings in this thesis are discussed.

Chapters 1 to 6 in this thesis are self-contained, so that they can also be read in isolation. Therefore, minor passages may overlap in these chapters.

CHAPTER 2

Framework for healthcare planning and control

2.1 Introduction

Chapter 1 introduced various problems that cause healthcare planning and control to lag behind manufacturing planning and control. To help overcome these problems, we propose and demonstrate a hierarchical framework for healthcare planning and control in this chapter. This framework serves as a tool to structure and break down all functions of healthcare planning and control. In addition, it can be used to identify planning and control problems and to demarcate the scope of organization interventions. It is applicable broadly, from an individual hospital department to an entire hospital, or to a complete supply chain of care providers.The framework facilitates a dialogue between clinical staff and managersto design the planning and control mechanisms. These mechanisms are necessary to translate the organization's objectives into effective and efficient healthcare delivery processes [126]. It covers all managerial areas involved in healthcare delivery operations and all levels of control, to ensure completeness and coherence of responsibilities for every managerial area.

We will argue in Section 2.2 that while frameworks for planning and control do exist in the literature, they mostly focus on one managerial area - in particular resource capacity planning or materials planning - and mostly only focus on hospitals. The contribution of our framework is that it encompasses all managerial areas, including those typically overlooked by others. In particular, medical planning (i.e., decision making by clinicians) and financial planning should not be overlooked when healthcare delivery processes are to be redesigned or optimized. Another contribution of the framework is its hierarchical decomposition of managerial levels, which is an extension of the classical strategic-tactical-operational breakdown [9], often used in manufacturing. Finally, while most frameworks focus on hospitals, our framework can be applied to any type of healthcare delivery organization.

This chapter is organized as follows. Section 2.2 outlines the literature on frameworks for planning and control. Section 2.3 presents the generic framework for healthcare planning and control. Section 2.4 describes how to identify managerial problems with the framework, and demonstrates its application.

Section 2.5 presents our conclusions.

2.2 Literature

In this section we give an overview of the state-of-the art in the literature of both manufacturing planning and control and healthcare planning and control. We also discuss the strengths and weaknesses of the existing frameworks.

Almost all well-known frameworks for manufacturing planning and control (MPC) organize planning and control functions hierarchically. It reflects the natural process of increasing disaggregation in decision making as time progresses, and more information becomes available [534]. It also reflects the hierarchical (department) structure of most organizations [78]. Many MPC frameworks use the hierarchical decomposition into a strategic, tactical, and operational level, as first done by Anthony in 1965 [9].

The classical MPC frameworks have a specific orientation on either *production planning* (e.g., hierarchical production planning [248]), or *technological (or process) planning* (e.g., computer aided process planning [342]), or *material planning* (e.g., Material Requirements Planning (MRP) [386]). As argued in [534], this myopic orientation to one managerial area is the main cause that these MPC frameworks are inadequate in practice. Modern MPC frameworks integrate these orientations: the frameworks in [233, 534] are designed for integrated MPC in highly complex organizations, such as engineer-to-order manufacturers.

Various researchers have proposed frameworks for (hierarchical) planning and control in healthcare. In the remainder of this section, we give an overview of existing frameworks for healthcare planning and control.

First introduced in [414], and later expanded on in [425], two papers propose a hierarchical framework that is based on application of the Manufacturing Resource Planning (MRP-II) concept. This framework considers both resource capacity planning and material planning, and focuses specifically on hospitals. It relies on Diagnostic Related Groups (DRGs), which serve as the "bill of materials" in MRP-II, to derive the resource and material requirements of patient groups. In [425] it is proposed to use DRGs to facilitate integrated hospitalwide planning and control. This framework is criticized in [503], in which is argued that although DRGs are an excellent tool to market and finance hospitals, they are not a good basis for logistical control and managing day-to-day operations.

In [500], a framework is proposed for production control in hospitals based on the design requirements discussed in [125]. The approach assumes the common situation that a hospital is organized in relatively independent business units. It is limited to resource capacity planning, for which it distinguishes five hierarchical levels: *strategic planning, patient volumes planning and control, resources planning and control, patient group planning*, and *patient planning and control*. These levels address "offline" (in advance) decision making. "Online" (reactive) operational control functions such as reactive planning (e.g., add-on scheduling upon arrival of an emergency case) and monitoring are not considered in their framework.

In [76], the authors emphasize that due to the differing complexity and information requirements of the various decisions, organizational planning processes are commonly hierarchical in nature. The first step, on a strategic level, involves strategy formation, process layout design, and long-term capacity dimensioning. Subsequent steps relate increasingly to operational concerns, with a decreasing planning horizon and increasing information availability. The hierarchical levels of control are linked: for example long-term capacity dimensioning decisions shape the capacity restrictions for subsequent operational decision making. The performance, which is measured at an operational level, is the result of how well the various hierarchical planning activities are integrated. In [78], the authors indicate that the literature neglects cooperation between different managerial areas at the strategic level of hospital planning and control. They argue that to attain exceptional operational performance, it is important that the hospital's strategy consistently and coherently integrates operations issues from areas like *Finance, Marketing, Operations*, and *Human Resources*.

In [49], the authors focus on an operating theatre setting, for which they propose a hierarchical framework for resource planning and appointment scheduling with three hierarchical levels: *strategic*, *administrative (tactical)*, and *operational* planning.

We conclude that all existing frameworks for healthcare planning and control focus on hospitals, and are hierarchical in nature. However, like many MPC frameworks they also focus on just one managerial area - mostly resource capacity planning. Integration of managerial areas is neglected, as well as the reactive decision functions, which are important given the inherently stochastic nature of healthcare processes. Modern MPC frameworks [233, 534], however, address multiple managerial areas as well as the three well-known hierarchical levels of control. These frameworks were designed for engineer-to-order or manufacture-to-order environments, where uniquely specified products are produced on demand. In this aspect, these environments resemble healthcare delivery. Therefore, these MPC frameworks offer a sound basis for our framework for healthcare planning and control. However, for application in healthcare, they require significant modification. In the following section, we introduce our generic framework.

2.3 Framework

We propose a four-by-four generic framework for healthcare planning and control thats pans four hierarchical levels of control, and four managerial areas. We first discuss the managerial areas (2.3.1), and then the hierarchical decomposition (2.3.2). We then combine these two dimensions to form the framework for healthcare planning and control (2.3.3). Finally, we discuss the context of the framework and how it affects the content (2.3.4).

2.3.1 Managerial areas

As outlined in Section 2.2, most existing frameworks in the literature focus on one managerial area. We propose to include multiple managerial areas for healthcare planning and control, specifically: *medical planning, resource capacity planning, materials planning*, and *financial planning*. We describe these areas in more detail below.

Medical planning

The role of engineers/process planners in manufacturing is performed by clinicians in healthcare. We refer to healthcare's version of "technological planning" as *medical planning*. Medical planning comprises decision making by clinicians regarding for example medical protocols, treatments, diagnoses, and triage. It also comprises development of new medical treatments by clinicians. The more complex and unpredictable the healthcare processes, the more autonomy is required for clinicians. For example, activities in acute care are necessarily planned by clinicians, whereas in elective care (e.g., ambulatory surgery), standardized and predictable activities can be planned centrally by management.

Resource capacity planning

Resource capacity planning addresses the dimensioning, planning, scheduling, monitoring, and control of *renewable* resources. These include equipment and facilities (e.g., MRIs, physical therapy equipment, bed linen, sterile instruments, operating theatres, rehabilitation rooms), as well as staff.

Materials planning

Materials planning addresses the acquisition, storage, distribution and retrieval of all *consumable* resources/materials, such as suture materials, prostheses, blood, bandages, food, etc. Materials planning typically encompasses functions like warehouse design, inventory management and purchasing.

Financial planning

Financial planning addresses how an organization should manage its costs and revenues to achieve its objectives under current and future organizational and economic circumstances. Since healthcare spending has been increasing steadily [119], market mechanisms are being introduced in many countries as an incentive to encourage cost-efficient healthcare delivery (e.g., [516]). An example is the introduction of DRGs, which enables the comparison of care products and their prices. As healthcare systems differ per country, so does financial planning in healthcare organizations. As financial planning heavily influences the way the processes are organized and managed, we include this managerial area in our framework. For example, the authors of [506] argue that in the US, the tactical allocation of temporary expansions in operating theatre capacity should be based on the contribution margin of the involved surgical (sub)specialties. This criterion is not likely to be used in countries with a non-competitive healthcare system, such as the UK or the Netherlands. Financial planning in healthcare concerns functions such as investment planning, contracting (e.g., with healthcare insurers), budget and cost allocation, accounting, cost price calculation, and billing.

We have selected medical planning, resource capacity planning, materials planning and financial planning as the four managerial areas, as we consider them relevant in all our research projects that revolve around optimization of healthcare operations [101].

2.3.2 Hierarchical decomposition

As argued in Section 2.2, decision making disaggregates as time progresses and information gradually becomes available. We build upon the "classical" hierarchical decomposition often used in manufacturing planning and control, which discerns *strategic*, *tactical*, and *operational* levels of control [9]. We extend this decomposition by discerning between *offline* and *online* on the operational level. This distinction reflects the difference between "in advance" decision making and "reactive" decision making. We explain the resulting four hierarchical levels below, where the tactical level is explained last. The tactical level is often considered less tangible than the strategic and operational levels, as we will further explain in Section 2.4. Therefore, we explain the more tangible levels first, before addressing the tactical level.

We do not explicitly state a decision horizon length for any of the hierarchical planning levels, since these depend on the specific characteristics of the application. An emergency department for example inherently has shorter planning horizons than a long-stay ward in a nursing home.

Strategic level

Strategic planning addresses structural decision making. These decisions are the bricks and mortar of an organization [326]. It involves defining the organization's mission (i.e., "strategy" or "direction"), and the decision making to translate this mission into the design, dimensioning, and development of the healthcare delivery process. Inherently, strategic planning has a long planning horizon and is based on highly aggregated information and forecasts. Examples of strategic planning are resource capacity expansions (e.g., acquisition of MRI machines), developing and/or implementing new medical protocols, forming a purchasing consortium, a merger of nursing homes, and contracting with health insurers.

Offline operational level

Operational planning (both “offline” and “online”) involves the short-term decision making related to the execution of the healthcare delivery process. There is low flexibility on this planning level, since many decisions on higher levels have demarcated the scope for the operational level decision making. The adjective “offline” reflects that this planning level concerns the in advance planning of operations. It comprises the detailed coordination of the activities regarding current (elective) demand. Examples of offline operational planning are: treatment selection, appointment scheduling, nurse rostering, inventory replenishment ordering, and billing.

Online operational level

The stochastic nature of healthcare processes demands reactive decision making. “Online” operational planning involves control mechanisms that deal with monitoring the process and reacting to unforeseen or unanticipated events. Examples of online planning functions are: triaging, add-on scheduling of emergencies, replenishing depleted inventories, rush ordering surgery instrument sterilization, handling billing complications.

Tactical level

In between the strategic level, which sets the stage (e.g., regarding location and size), and the operational level, which addresses the execution of the processes, lies the tactical planning level.We explain tactical planning in relation to strategic and operational planning.

While strategic planning addresses structural decision making, tactical planning addresses the organization of the operations / execution of the healthcare delivery process (i.e., the “what, where, how, when and who”). In this way, it is similar to operational planning; however, decisions are made on a longer planning horizon. The length of this intermediate planning horizon lies somewhere between the strategic planning horizon and operational planning horizon. Following the concept of hierarchical planning, intermediate tactical planning has more flexibility than operational planning, is less detailed, and has less demand certainty. Conversely, the opposite is true when compared to strategic planning.

For example, while capacity is fixed in operational planning, temporary capacity expansions like overtime or hiring staff are possible in tactical planning. Also, while demand is largely known in operational planning, it has to be (partly) forecasted for tactical planning, based on (seasonal) demand, waiting list information, and the ‘downstream’ demand in care pathways of patients currently under treatment. Due to this demand uncertainty, tactical planning is less detailed than operational planning. Examples of tactical functions are admission planning, block planning, treatment selection, supplier selection and budget allocation.

	Medical planning	***Resource capacity planning***	***Materials planning***	***Financial planning***
Strategic	Research, development of medical protocols	Case mix planning, capacity dimensioning, workforce planning	Supply chain and warehouse design	Investment plans, contracting with insurance companies
Tactical	Treatment selection, protocol selection	Block planning, staffing, admission planning	Supplier selection, tendering	Budget and cost allocation
Offline operational	Diagnosis and planning of an individual treatment	Patient-to-appointment assignment, workforce scheduling	Materials purchasing, determining order sizes	DRG billing, cash flow analysis
Online operational	Triage, diagnosing emergencies and complications	Monitoring, emergency coordination	Rush ordering, inventory replenishing	Billing complications and changes

← managerial areas →

← hierarchical decomposition →

Figure 2.1: Example application of the framework for healthcare planning and control to a general hospital.

2.3.3 Framework

Integrating the four managerial areas and the four hierarchical levels of control shapes a four-by-four positioning framework for healthcare planning and control. While the dimensions of the framework are generic, the content depends on the application at hand. The framework can be applied anywhere from the department level (for example to an operating theatre department) to organization-wide, or to a complete supply chain of care providers. Depending on the context, the content of the framework may be very different. Figure 2.1 shows the content of the framework when applied to a general hospital as a whole. The inserted planning and control functions are examples, and not exclusive.

2.3.4 Context of the framework

As argued in the previous section, the content of the framework should be accommodated to the context of the application. Regarding the context we discern the internal and external environment characteristics.

The *internal* environment characteristics are scoped by the boundaries of the organization. This involves all characteristics that affect planning and control, regarding for example patient demand (e.g., variability, complexity, arrival intensity, medical urgency, recurrence), organizational culture and structure.

The way healthcare organizations are organized is perhaps most influenced by their *external* environment. For example, a "STEEPLED" analysis (an extension of "PESTEL", see e.g., [278]) can be done to identify external factors that influence healthcare planning and control, now or in the future. "STEEPLED" is an abbreviation for the following external environment factors:

- Social factors (e.g., education, social mobility, religious attitudes)

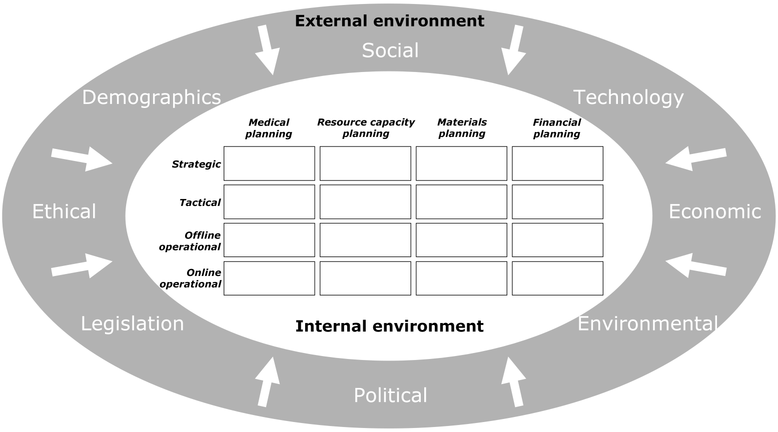


Figure 2.2: The framework and the organization's external environment.

- Technology (e.g., medical innovation, transport infrastructure)
- Economic factors (e.g., change in health finance system)
- Environmental factors (e.g., ecological, recycling)
- Political factors (e.g., change of government policy, privatization)
- Legislation / Legal(e.g., business regulations, quality regulations)
- Ethical factors (e.g., business ethics, confidentiality, safety)
- Demographics (e.g., graying population, life expectancy, obesity)

These factors largely explain the differences amongst countries in the management approach of healthcare organizations. Figure 2.2 illustrates how the framework can be observed in light of the organization's external environment.

2.4 Application

The primary objective of the framework is to structure the various planning and control functions. In this section, we give examples of how the framework can be applied. Section 2.4.1 discusses how the framework can be used to identify managerial deficiencies. Section 2.4.2 gives an example of an application of the framework to an integrated model for primary care outside office hours.

2.4.1 Identification of managerial deficiencies

Once the content of the framework has been established for a given application, further analysis of this content may identify managerial problems. In the remainder of this section, we discuss examples of four kinds of typical problems:

1. Deficient or lacking planning functions
2. Inappropriate planning approaches
3. Lack of coherence between planning functions
4. Planning functions that have conflicting objectives

Deficient or lacking planning functions

Overlooked or poorly addressed managerial functions can be encountered on all levels of control [87], but are most often found on the tactical level of control [425]. In fact, to many, tactical planning is less tangible than operational planning and even strategic planning. Inundated with operational problems, managers are inclined to solve problems at hand (i.e., on the operational level). We refer to this phenomenon as the "real-time hype" of managers. A claim for "more capacity" is the universal panacea for many healthcare managers. It is, however, often overlooked that instead of such drastic strategic measures, tactically allocating and organizing the available resources may be more effective and cheaper. Consider for example a "master schedule" or "block plan", which is the tactical allocation of blocks of resource time (e.g., operating theatres, or CT-scanners) to specialties and/or patient categories during a week. Such a block plan should be periodically revised to react to variations in supply and demand. However, in practice, it is more often a result of "historical development" than of analytical considerations [501].

An example of a deficient planning function is when autonomy is given to or assumed by the wrong staff member. We illustrate this with two examples: (1) Spurred by the Oath of Hippocrates, clinicians may try to 'cheat' the system to advance a patient. A clinician may for example admit an outpatient to a hospital bed to shorten access time for diagnostics (which is lower for inpatients). The resulting bed occupation may lead to operating room blocking. Although this may appear suboptimal from a central management point of view, it may be necessary from a medical point of view. The crux is to put the autonomy where it is actually needed. This depends on the application at hand. As argued earlier, the more complex and unpredictable the healthcare processes, the more autonomy is required for clinicians. Standardized and predictable activities can however be planned centrally by management, which is advantageous from an economies-of-scale viewpoint. (2) Medical equipment shared by different departments is hoarded to ensure immediate availability [117]. This leads to excessive inventory (costs), which may be significantly reduced by centralizing equipment management and storage. A typical example is the hoarding of intravenous drips by wards.

Inappropriate planning approaches

There are many logistical paradigms, such as Just-In-Time (JIT), Kanban, Lean, Total Quality Management (TQM), and Six Sigma, all of which have reported success stories. As these paradigms are mostly developed for industry, they generally cannot be simply copied to healthcare without loss in fidelity. “The tendency to uncritically embrace a solution concept, developed for a rather specific manufacturing environment, as the panacea for a variety of other problems in totally different environments has led to many disappointments” [534]. The structure provided by the framework helps to identify whether a planning approach is suitable for a planning function in a particular organizational environment. Planning approaches are only suitable if they fit the internal and external characteristics of the involved application. They have to be adapted to/designed for the characteristics that are unique for healthcare delivery, such as: (1) patient participation in the service process; (2) simultaneity of production and consumption; (3) perishable capacity; (4) intangibility of healthcare outputs; and (5) heterogeneity [387].

Lack of coherence between planning functions

The effectiveness and efficiency of healthcare delivery is not only determined by how the various planning functions are addressed; this is also determined by how they interact. As healthcare providers such as hospitals are typically formed as a cluster of autonomous departments, planning is also often functionally dispersed. The framework structures planning functions, and provides insight in their horizontal (cross-management) and vertical (hierarchical) interactions. *Horizontal interaction* between managerial areas in the framework provides that required medical information and protocols, and all involved resources and materials, are brought together to enable both effective and efficient healthcare delivery. *Downward vertical interaction* concerns concretizing higher level objectives and decisions on a shorter planning horizon. For example, capacity dimensioning decisions on a strategic level (e.g., number of CT scanners) impose hard restrictions on tactical and operational planning and scheduling. *Upward vertical interaction* concerns feedback about the realization of higher level objectives. For example the capacity of MRI machines is determined on the strategic level to attain a certain service level (e.g., access time). Feedback from the tactical and operational level is then needed to observe whether this objective is actually attained, and to advise to what extent the capacity is sufficient.

Planning functions that have conflicting objectives

As argued, the framework structures planning functions and their horizontal and vertical interactions. The framework can thus identify conflicting objectives between planning functions. For example, minimally-invasive surgery generally results in significantly reduced length of stay in wards and improved quality of care, but results in higher costs and increased capacity consumption for the operating theatre department. These departments are often managed autonomously and independently, which leads to sub-optimal decision making from both the patient's and the hospital's point of view.

Conflicting objectives also occur between two care providers in an inter-organizational care chain. For example, a nursing home's efforts to maximize occupancy may lead to bed blocking in hospitals. Aligning planning functions between healthcare organizations may identify and solve such problems.

2.4.2 Primary care outside office hours

In this section we give an example application of the framework. First we introduce the context: the concept of an integrated organization that provides primary care outside office hours. We then demonstrate how the framework can facilitate the discussion regarding the design of such an organization.

Introduction

The organization of primary care outside office hours, which involves telephone triage, urgent consultations and house calls, has received increasing attention in many countries [223]. In parts of Europe, general practitioners (GPs) are required by law to provide this type of care, and in some countries, GPs cooperate in primary care cooperatives (PCCs) to jointly provide primary care outside office hours. Within a PCC, the GPs can alternate who is responsible outside office hours. As a result, these GPs do not have to be available outside office hours at all times. As an alternative to the PCC, patients requiring primary care outside office hours can visit the emergency department (ED) of a hospital. Although EDs are intended for complex emergent care, they deal with a relatively large group of patients that could have been served by a GP. For example a study at King's College Hospital in the United Kingdom reports that 41% of patients visiting the ED could have been treated by a GP [116]. It is more costly to serve these so-called 'self-referrals' at the ED. Therefore, methods are proposed to ensure these patients are served by GPs and do not visit an ED.

One of the proposed methods is an integrated model, where the PCC is located in close proximity to the ED, with a joint triage system. Integrated models are effective in the UK [316], and are also favored by the Netherlands as the appropriate system for emergency care [489]. A survey showed that the integrated model significantly decreases the number of self-referrals in the ED, since these patients can be referred to the PCC [489]. The integration is thus cost effective from a societal point of view [116, 489]. It is, however, under debate whether the integration is cost effective for the EDs and PCCs [489]. For EDs, the integration decreases the number of patient visits, possibly around 50% [223]. This reduces turnover, and all kinds of economies-of-scale advantages. In the Netherlands, the hourly rate for primary care outside office hours for GPs (set by government and paid by health insurers) is considered low and not profitable. Hence, GPs do not welcome the increased workload.

Application of the framework

To successfully implement an integrated ED/PCC, the involved parties must address the aforementioned problems, and discuss how to manage the new organization's planning and control. To facilitate this discussion in a structured way, the framework can be instrumental. We mention some of the key issues per managerial area:

- *Medical planning*: How does the case of joint triage affect the role and responsibilities of the GPs, who before were considered the 'gatekeepers' of healthcare delivery?
- *Resource capacity planning*: What are the "24/7" resource capacity requirements? Is collaboration of ED and PCC staff possible despite the fact that they work for two independent cost centers - if so, to what extent should they collaborate?
- *Materials planning*: Should the ED and PCC jointly purchase materials? Where should inventories be kept, and who has ownership?
- *Financial planning*: Is an integration of ED and PCC cost effective for hospitals, GPs, insurance companies, society? Is it profitable for the ED to employ general practitioners for self-referrals instead of integrating with a PCC? Should hospitals, insurance companies, or the government compensate GPs for the increased workload? Should the ED and PCC be integrated into one cost center?

Based on the outcomes of the discussion around the aforementioned issues, the framework can be used further to design appropriate planning and control on all hierarchical levels and in all managerial areas.

2.5 Conclusion

In this chapter we propose a reference framework for healthcare planning and control, which hierarchically structures planning and control functions in multiple managerial areas. It offers a common language for all involved decision makers: clinical staff, managers, and experts on planning and control. This allows coherent formulation and realization of objectives on all levels and in all managerial areas [126]. The framework is widely applicable to any type of healthcare provider or to specific departments within a healthcare organization. The contents of the framework depend on the application at hand, for example an organizational intervention, a decision making process or a healthcare delivery process.

While existing management and control approaches use either an "organizational unit"/vertical perspective or a "business process"/horizontal perspective, our framework accommodates both approaches. The framework facilitates a structural analysis of the planning and control functions and their interaction. Moreover, it helps to identify managerial problems regarding, for example, planning functions that are deficient or inappropriate, that lack coherence, or have conflicting objectives.

When managerial deficiencies have been identified, the framework can be used to demarcate the scope of organization interventions. In general, focusing on problems on lower hierarchical levels reduces uncertainty, as inherently the planning horizon is shorter and more information is available. However, flexibility (e.g., regarding resource expansion) is also lower. Focusing on problems on higher hierarchical levels increases the potential impact (e.g., cost savings, waiting time reduction, quality of care); however, required investments are usually also higher, and effects of interventions are felt on a longer term.

Regardless of the focal point of organization interventions, the framework emphasizes the implications from and for adjacent managerial functions. It can thus be prevented that stake holding decision makers are not involved, and that interventions like “more capacity” (the universal panacea) are not made without considering the possible effects for all underlying and related planning functions. As a result, interventions will have a higher chance of success.

As argued in Chapter 1, the literature regarding the application of OR/MS in healthcare is expanding rapidly. This framework can also be instrumental in the design of taxonomies for, for example, literature on outpatient department (appointment) planning, operating theatre planning and scheduling, and inventory management of medical supplies. Scientific papers can be positioned in the framework to illustrate the managerial area(s) they focus on, and the hierarchical level of decision making in the considered problem(s). Similarly, also algorithmic developments can be classified and positioned in the framework.

The framework can easily be extended to include other managerial areas or hierarchical levels. In particular information management is a managerial area that should go hand in hand with development of innovative organization-wide planning approaches. “Business-IT Alignment” addresses how companies can apply information technology to formulate and achieve their goals on the various hierarchical levels [317]. Another relevant managerial area that can be included is quality and safety management, which is involved in almost all care delivery processes, and can be decomposed hierarchically. The framework can also be expanded in the hierarchical decomposition. There may be different functions on a single hierarchical level within a managerial area, which by themselves have a natural hierarchy. For example decisions regarding the construction of a new building are of a higher level than decisions regarding the expansion of a ward, while both are strategic decisions.

CHAPTER 3

Taxonomic classification and structured review of planning decisions

3.1 Introduction

In this chapter, we aim to guide healthcare professionals and Operations Research and Management Science (OR/MS) researchers through the broad field of OR/MS in healthcare. We provide a structured overview of the typical decisions to be made in resource capacity planning and control in healthcare, and provide a review of relevant OR/MS articles for each planning decision.

The contribution of this chapter is twofold. First, to position the planning decisions, we present a taxonomy. This taxonomy provides healthcare managers and OR/MS researchers with a method to identify, break down and classify planning and control decisions. The taxonomy contains two axes. The vertical axis reflects the hierarchical nature of decision making in resource capacity planning and control, and the horizontal axis the various healthcare services. The vertical axis is strongly connected, because higher-level decisions demarcate the scope of and impose restrictions on lower-level decisions. Although healthcare delivery is generally organized in autonomous organizations and departments, the horizontal axis is also strongly interrelated as a patient pathway often consists of several healthcare services from multiple organizations or departments.

Second, following the vertical axis of the taxonomy, and for each healthcare service on the horizontal axis, we provide a comprehensive specification of planning and control decisions in resource capacity planning and control. For each planning and control decision, we structurally review the key OR/MS articles and the OR/MS methods and techniques that are applied in the literature to support decision making. No structured review exists of this nature, as existing reviews are typically exhaustive within a confined scope, such as simulation modeling in healthcare [282] or outpatient appointment scheduling [89], or are more general to the extent that they do not focus on the concrete specific decisions.

This chapter is organized as follows. Section 3.2 presents our taxonomy. Section 3.3 explains the objectives, scope, and search method for our structured review. Sections 3.4 to 3.9 identify, classify and discuss the planning and control decisions. Section 3.10 concludes this chapter with a discussion of our findings.

3.2 Taxonomy

Taxonomy is the practice and science of classification. It originates from biology where it refers to a hierarchical classification of organisms. The National Biological Information Infrastructure [370] provides the following definition of taxonomy: "Taxonomy is the science of classification according to a pre-determined system, with the resulting catalog used to provide a conceptual framework for discussion, analysis, or information retrieval; ...a good taxonomy should be simple, easy to remember, and easy to use." With exactly these objectives, we present a taxonomy for resource capacity planning and control in healthcare.

Planning and control decisions are made by healthcare organizations to design and operate the healthcare delivery process. It requires coordinated long-term, medium-term and short-term decision making in multiple managerial areas. In Capter 2, a framework is presented to subdivide these decisions in four hierarchical, or temporal, levels and four managerial areas. These hierarchical levels and the managerial area of resource capacity planning and control form the basis for our taxonomy. For the hierarchical levels, [234] applies the well-known breakdown of *strategic*, *tactical* and *operational* [9]. In addition, the operational level is subdivided in *offline* and *online* decision making, where *offline* reflects the in advance decision making and *online* the real-time reactive decision making in response to events that cannot be planned in advance. The four managerial areas are: medical planning, financial planning, materials planning and resource capacity planning. They are defined as follows. *Medical planning* comprises decision making by clinicians regarding medical protocols, treatments, diagnoses and triage. *Financial planning* addresses how an organization should manage its costs and revenues to achieve its objectives under current and future organizational and economic circumstances. *Materials planning* addresses the acquisition, storage, distribution and retrieval of all consumable resources/materials, such as suture materials, blood, bandages, food, etc. *Resource capacity planning* addresses the dimensioning, planning, scheduling, monitoring, and control of renewable resources. Our taxonomy is a further specification of the healthcare planning and control framework of Chapter 2, in the managerial area of resource capacity planning.

The taxonomy contains two axes. The vertical axis reflects the hierarchical nature of decision making in resource capacity planning and control, and is derived from Chapter 2 [234]. On the horizontal axis of our taxonomy we position different services in healthcare. We identify *ambulatory care services*, *emergency care services*, *surgical care services*, *inpatient care services*, *home care services*, and *residential care services*. The taxonomy is displayed in Figure 3.1. We elaborate on both axes in more detail below.

Vertical axis

Our taxonomy is intended for planning and control decisions within the boundaries of a healthcare delivery organization. Every healthcare organization operates in a particular external environment. Therefore, all planning and control decisions are made in the context of this external environment. The external environment is characterized by factors such as legislation, technology and social factors.

The nature of planning and control decision making is such that decisions disaggregate as time progresses and more information becomes available [534]. Aggregate decisions are made in an early stage, while more detailed information supports decision making with a finer granularity in later stages. Because of this disaggregating nature, most well-known taxonomies and frameworks for planning and control are organized hierarchically [234, 534]. As the impact of decisions decreases when the level of detail increases, such a hierarchy also reflects the top-down management structure of most organizations [42].

For completeness we explicitly state the definitions of the four hierarchical levels as given in Chapter 2 and [234], which we position on the vertical axis of our taxonomy. The definitions are adapted to specifically fit the managerial area of resource capacity planning and control.

- *Strategic planning* addresses structural decision making. It involves defining the organization's mission (i.e., 'strategy' or 'direction'), and the decision making to translate this mission into the design, dimensioning, and development of the healthcare delivery process. Inherently, strategic planning has a long planning horizon and is based on highly aggregated information and forecasts. Examples of strategic planning are determining the facility's location, dimensioning resource capacities (e.g., acquisition of an MRI scanner, staffing) and deciding on the service and case mix.

- *Tactical planning* translates strategic planning decisions to guidelines which facilitate operational planning decisions. While strategic planning addresses structural decision making, tactical planning addresses the organization of the operations/execution of the healthcare delivery process (i.e., the 'what, where, how, when and who'). As a first step in tactical planning, patient groups are characterized based on disease type/diagnose, urgency and resource requirements. As a second step, the available resource capacities, settled at the strategic level, are divided among these patient groups. In addition to the allocation in time quantities, more specific timing information can already be added, such as dates or time slots. In this way, blueprints for the operational planning are created that allocate resources to different tasks, specialties and patient groups. Temporary capacity expansions like overtime or hiring staff are also part of tactical planning. Demand has to be (partly) forecasted, based on (seasonal) demand, waiting list information, and the 'downstream' demand in care pathways of patients currently under treatment. Examples of tactical planning are staff-shift scheduling and the (cyclic) surgical block schedule that allocates operating time capacity to patient groups.

- *Operational planning* (both 'offline' and 'online') involves the short-term decision making related to the execution of the healthcare delivery process. Following the tactical blueprints, execution plans are designed at the individual patient level and the individual resource level. In operational planning, elective demand is entirely known and only emergency demand has to be forecasted. In general, the capacity planning flexibility is low on this level, since decisions on higher levels have demarcated the scope for the operational level decision making.

- *Offline operational planning* reflects the in advance planning of operations. It comprises the detailed coordination of the activities regarding current (elective) demand. Examples of offline operational planning are patient-to-appointment assignment, staff-to-shift assignment and surgical case scheduling.
- *Online operational planning* reflects the control mechanisms that deal with monitoring the process and reacting to unplanned events. This is required due to the inherent uncertain nature of healthcare processes. An example of online operational planning is the real-time dynamic (re)scheduling of elective patients when an emergency patient requires immediate attention.

Note that the decision horizon lengths are not explicitly given for any of the hierarchical planning levels, since these depend on the specific characteristics of the application. For example, an emergency department inherently has shorter planning horizons than a long-stay ward in a nursing home. Furthermore, there is a strong interrelation between hierarchical levels. Top-down interaction exists as higher-level decisions demarcate the scope of and impose restrictions on lower-level decisions. Conversely, bottom-up interaction exists as feedback about the healthcare delivery realization supports decision making in higher levels.

Horizontal axis

On the horizontal axis of our taxonomy we position the different services in healthcare. The complete spectrum of healthcare delivery is a composition of many different services provided by many different organizations. From the perspective of resource capacity planning and control, different services may face similar questions. To capture this similarity, we distinguish six clusters of healthcare services. The definitions of the six care services are obtained from the corresponding MeSH terms provided by PubMed [362]. For each care service we offer several examples of facilities that provide this service.

- *Ambulatory care services* provide care to patients without offering a room, a bed and board, and they may be free-standing or part of a hospital. In ambulatory care services, we position primary care services and community services as well as hospital-based services such as the outpatient clinic, since these services face similar questions from a resource capacity planning perspective.

Examples of ambulatory care facilities are outpatient clinics, primary care services and the hospital departments of endoscopy, radiology and radiotherapy.

- *Emergency care services* are concerned with the evaluation and initial treatment of urgent and emergent medical problems, such as those caused by accidents, trauma, sudden illness, poisoning, or disasters. Emergency medical care can be provided at the hospital or at sites outside the medical facility. Examples of emergency care facilities are hospital emergency departments, ambulances and trauma centers.
- *Surgical care services* provide operative procedures (surgeries) for the correction of deformities and defects, repair of injuries, and diagnosis and cure of certain diseases. Examples of surgical care facilities are the hospital's operating theater, surgical daycare centers and anesthesia facilities.
- *Inpatient care services* provide care to hospitalized patients by offering a room, a bed and board. Examples are intensive care units, general nursing wards, and neonatal care units.
- *Home care services* are community health and nursing services that provide multiple, coordinated services to a patient at the patient's home. Home care services are provided by a visiting nurse, home health agencies, hospitals, or organized community groups using professional staff for healthcare delivery. Examples are medical care at home, housekeeping support and personal hygiene assistance.
- *Residential care services* provide supervision and assistance in activities of daily living with medical and nursing services when required. Examples are nursing homes, psychiatric hospitals, rehabilitation clinics with overnight stay, homes for the aged, and hospices.

Note that the horizontal subdivision is not based on healthcare organizations, but on the provided care services. Therefore, it is possible that a single healthcare organization offers services in multiple clusters. It may be that a particular facility is used by multiple care services, for example a diagnostics department that is used in both ambulatory and emergency care services. In addition, a patient's treatment often comprises of consecutive care stages offered by multiple care services. The healthcare delivery realization within one care service is impacted by decisions in other services, as inflow and throughput strongly depend on these other services. Therefore, resource capacity planning and control decisions are always made in the context of decisions made for other care services. Hence, like the interrelation in the vertical levels, a strong interrelation exists between the horizontal clusters.

This taxonomy provides a method to identify, break down and classify planning and control decisions in healthcare. This is a starting point for a complete specification of planning decisions and helps to gain understanding of the interrelations between various planning decisions. Hence, healthcare professionals

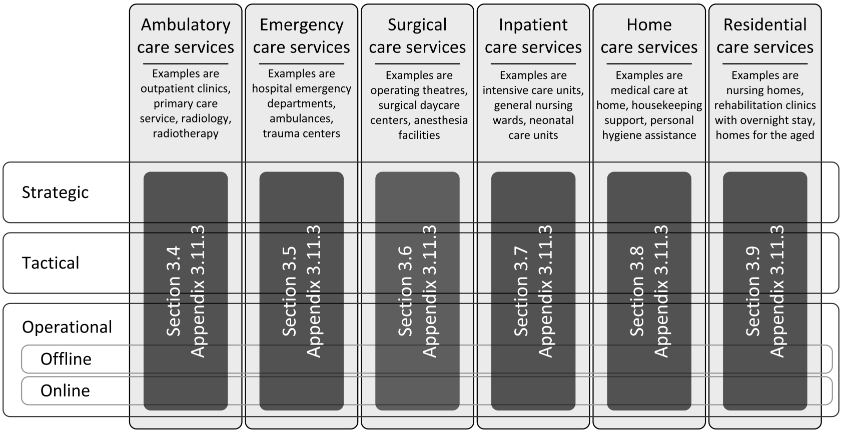

Ambulatory care services
Examples are outpatient clinics, primary care service, radiology, radiotherapy
Emergency care services
Examples are hospital emergency departments, ambulances, trauma centers
Surgical care services
Examples are operating theatres, surgical daycare centers, anesthesia facilities
Inpatient care services
Examples are intensive care units, general nursing wards, neonatal care units
Home care services
Examples are medical care at home, housekeeping support, personal hygiene assistance
Residential care services
Examples are nursing homes, rehabilitation clinics with overnight stay, homes for the aged
Strategic
Tactical
Operational
Offline
Online
Section 3.4 Appendix 3.11.3
Section 3.5 Appendix 3.11.3
Section 3.6 Appendix 3.11.3
Section 3.7 Appendix 3.11.3
Section 3.8 Appendix 3.11.3
Section 3.9 Appendix 3.11.3

Figure 3.1: The taxonomy for resource capacity planning and control decisions in healthcare.

can identify lacking, insufficiently defined and incoherent planning decisions within their department or organization. It also gives the opportunity to identify planning decisions that are not yet addressed often in the OR/MS literature. Therefore, in Section 3.3, with our taxonomy as the foundation, we provide an exhaustive specification of planning decisions for each care service, combined with a review of key OR/MS literature.

3.3 Objectives, scope, and search method

In this section, we identify the resource capacity planning and control decisions for each of the six care services in our taxonomy. The decisions are classified according to the vertical hierarchical structure of our taxonomy. For each identified planning decision we will discuss the following in our overview:

- What is the concrete *decision*?
- Which *performance measures* are considered?
- What are the *key trade-offs*?
- What are *main insights and results* from the literature?
- What are *general conclusions*?
- Which *OR/MS methods* are applied to support decision making?

The identified planning decisions are in the first place obtained from available books and articles on healthcare planning and control. Our literature search method will be explained in more detail below. In addition, to be as complete as possible, expert opinions from healthcare professionals and OR/MS specialists are obtained to identify decisions that are not yet well-addressed in the literature and for this reason cannot be obtained from the literature. In this introduction, we first discuss the scope of the identified planning decisions and the applied OR/MS methods, and next we present the applied literature search method.

Scope. Numerous processes are involved in healthcare delivery. We focus on the resource capacity planning and control decisions to be made regarding the *primary process* of healthcare delivery. In the management literature, the primary process is defined as the set of activities that are directly concerned with the creation or delivery of a product or service [400]. Thus, we do not focus on *supporting activities*, such as procurement, information technology, human resource management, laboratory services, blood services and instrument sterilization.

We focus on OR/MS methods that quantitatively support and rationalize decision making in resource capacity planning and control. Based on forecasting of demand for care (see [387] for forecasting techniques), these methods provide optimization techniques for the design of the healthcare delivery process. Outside our scope is statistical comparison of performance of healthcare organizations, so-called benchmarking, of which Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA) are well-known examples [107]. Quantitative decision making requires measurable performance indicators by which the quality of healthcare delivery can be expressed. A comprehensive survey of applied performance measures in healthcare organizations is provided in [325]. Next, practical implementation of OR/MS methods may require the development of Information Communications Technology (ICT) solutions (that are possibly integrated in healthcare organizations' database systems); this is also outside the scope of this chapter.

The spectrum of different OR/MS methods is wide (see for example [253, 460, 473, 520] for introductory books). In this review, we distinguish the following OR/MS methods: computer simulation [319], heuristics [1], Markov processes (which includes Markov reward and decision processes) [473], mathematical programming [388, 434], queueing theory [424]. For a short description of each of these OR/MS methods, the reader is referred to Appendix 3.11.1.

Literature search method. As the body of literature on resource capacity planning and control in healthcare is extensive, we used a structured search method and we restricted to articles published in ISI-listed journals to ensure that we found and filter key and state-of-the-art contributions. Table 3.1 displays our search method. To identify the search terms as listed in Appendix 3.11.2 and to create the basic structure of the planning decision hierarchy for each care service, we consulted available literature reviews [49, 59, 61, 74, 84, 89, 99, 166, 182, 183, 224, 228, 229, 271, 282, 287, 289, 304, 336, 351, 363, 368, 393, 399, 404, 410, 413, 451, 452, 479, 491] and books [62, 231, 312, 361, 387, 503]. Additional search terms were obtained from the index of *Medical Subject Headings* (MeSH) [362] and available synonyms. With these search terms, we performed a search on the database of Web of Science (WoS) [510]. WoS was chosen as it contains articles from all ISI-listed journals. It is particularly useful as it provides the possibility to select *Operations Research and Management Science* as a specific subject category and to sort references on the number of citations.

We identify a base set containing the ten most-cited articles in the predefined subject category of *Operations Research and Management Science*. Starting from this base set, we include all articles from ISI-listed journals that are referred by or refer to one of the articles in the base set and deal with resource capacity planning and control decisions. As such, we ensure that we also review recent work that may not have been cited often yet. In addition, we include articles published in Health Care Management Science (HCMS), which is particularly relevant for OR/MS in healthcare and obtained an ISI listing in 2010. To be sure that by restricting to WoS and HCMS, we do not neglect essential references, we also performed a search with our search terms on the databases of Business Source Elite [159], PubMed [405] and Scopus [437]. This search did not result in significant additions to the already found set of articles. The literature search was updated up to May 10, 2012.

Step 1:	Identify search terms from reviews, books and MeSH
Step 2:	Search the OR/MS subject category in WoS with the search terms
Step 3:	Select a base set: the ten most-cited articles relevant for our review
Step 4:	Perform a backward and forward search on the base set articles
Step 5:	Search relevant articles from HCMS

Table 3.1: The search method applied to each care service.

In the following sections, we provide the structured reviews per care service that is in the taxonomy's horizontal axis. Section 3.4 is devoted to ambulatory care services, Section 3.5 to emergency care services, Section 3.6 to surgical care services, Section 3.7 to inpatient care services, Section 3.8 to home care services and Section 3.9 to residential care services. For each care service, the review is subdivided in strategic, tactical, offline operational and online operational planning. In Appendix 3.11.3, tables are included in which the identified planning decisions are listed for each care service, together with applied OR/MS methods and literature references per planning decision. When for different care services a similar planning decision is involved, we use the same term. Our intention is that Sections 3.4-3.9 are self-contained, so that they can be read in isolation. Therefore, minor passages are overlapping. When in the description of a planning decision an article is cited, while it does not appear in the 'methods'-list, it means that this article contains a relevant statement about this planning decision, but the particular planning decision is not the main focus of the article.

3.4 Ambulatory care services

Ambulatory care services provide medical interventions without overnight stay, i.e., the patient arrives at the facility and leaves the facility on the same day. These medical interventions comprise for example diagnostic services (e.g, CT scan, MRI scan), doctor consultations (e.g., general practitioner, hospital specialist), radiotherapy treatments or minor surgical interventions. Demand for ambulatory care services is growing in most western countries since 2000 [377]. The existing literature has mainly focused on the offline operational planning decision of appointment scheduling.

Strategic planning

Regional coverage. Ambulatory care planning on a regional level aims to create the infrastructure to provide healthcare to the population in its catchment area. This regional coverage decision involves determining the number, size and location of facilities in a certain region to find a balanced distribution of facilities with respect to the geographical location of demand [153]. The main trade-off in this decision is between patient accessibility and efficiency. Patient accessibility is represented by access time and travel distance indicators. Efficiency is represented by utilization and productivity indicators [153, 451]. Common regional planning models incorporate the dependency of demand on the regional demographic and socioeconomic characteristics [2].

Methods: computer simulation [348, 420, 454, 471], heuristics [2, 153], literature review [451].

Service mix. An organization decides the particular services that the ambulatory care facility provides. The service mix stipulates which patient types can be consulted. In general, the service mix decision is not made at an ambulatory care service level, but at the regional or hospital level, as it integrally impacts the ambulatory, emergency, surgical and inpatient care services. This is also expressed by [501] in which for example inpatient resources, such as beds and nursing staff, are indicated as 'following' resources. This may be the reason that we have not found any references focusing on service mix decisions for ambulatory care services in specific.

Methods: no articles found.

Case mix. Every ambulatory care facility decides on a particular case mix, which is the volume and composition of patient groups that the facility serves. The settled service mix restricts the decisions to serve particular patient groups. Patient groups can be classified based on disease type, age, acuteness, home address, etc. The case mix influences almost all other planning decisions, such as a facility's location, capacity dimensions and layout. Also, demand for different patient groups in the case mix may vary, which influences required staffing levels significantly [450, 458]. However, case mix decision making has not received much attention in the OR/MS literature. In the literature, the case mix is often treated as given.

Methods: computer simulation [458], mathematical programming [450].

Panel size. The panel size is the number of potential patients of an ambulatory care facility [218]. Since only a fraction of these potential patients, also called calling population, actually demands healthcare, the panel size can be larger than the number of patients a facility can serve. The panel size is particularly important for general practitioners, as they need an accurate approximation of how many patients they can subscribe or admit to their practice. A panel size should be large enough to have enough demand to be profitable and to benefit from economies of scale, as a facility's costs per patient decrease when the panel size increases [454]. On the other hand, when the panel size is too large, access times may grow exponentially [218].

Methods: computer simulation [454], queueing theory [218].

Capacity dimensioning. Ambulatory care facilities dimension their resources, such as staff, equipment and space, with the objective to (simultaneously) maximize clinic profit, patient satisfaction, and staff satisfaction [458]. To this end, provider capacity must be matched with patient demand, such that performance measures such as costs, access time and waiting time are controlled. Capacity is dimensioned for the following resource types:

- *Consultation rooms.* The number of consultation rooms that balances patient waiting times and doctor idle time with costs for consultation rooms [264, 451, 457, 458].

- *Staff.* Staff in the ambulatory care services concern for example doctors, nurses and assistants [35, 282, 348, 421, 450, 451, 454, 457, 458, 508, 515].

- *Consultation time capacity.* The total consultation time that is available, for example for an MRI scanner or a doctor [115, 160, 162].

- *Equipment.* Some ambulatory care services require equipment for particular consultations, for example MRI scanners, CT scanners and radiotherapy machines [187, 348, 471].

- *Waiting room.* The waiting room is dimensioned such that patients and their companions waiting for consultation can be accommodated [458].

When capacity is dimensioned to cover average demand, variations in demand may cause long access and waiting times [471]. Basic rules from queueing theory demonstrate the necessity of excess capacity to cope with uncertain demand [213]. Capacity dimensioning is a key decision, as it influences how well a facility can meet demand and manage access and waiting times.

Methods: computer simulation [160, 162, 187, 264, 348, 421, 454, 457, 458, 471, 515], Markov processes [508], mathematical programming [450], queueing theory [35, 115, 162, 264], literature review [282, 451].

Facility layout. The facility layout concerns the positioning and organization of various physical areas in a facility. A typical ambulatory care facility consists of a reception area, a waiting area, and consultation rooms [190]. The facility layout is a potentially cost-saving decision in ambulatory care facilities [190, 387], but we found no articles that used an OR/MS approach to study the layout of an ambulatory care facility. Yet, the handbook [387] discusses heuristics for facility layout problems in healthcare.

Methods: heuristics [387].

Tactical planning

Patient routing. Ambulatory care typically consists of multiple stages. We denote the composition and sequence of these stages as the route of a patient. An effective and efficient patient route should match medical requirements, capacity requirements and restrictions, and the facility's layout. For a single facility, identifying different patient types and designing customized patient routes for each type prevents superfluous stages and delays [348]. For example, instead of two visits to a doctor and a medical test in between, some patient types may undergo a medical test before visiting the doctor, which saves valuable doctor time. Parallel processing of patients may increase utilization of scarce resources (e.g., a doctor or a CT scanner) [187, 264]. When parallel processing is applied, idle time of the scarce resource is reduced by preparing patients for consultation during the consultation time of other patients. Performance is typically measured by total visit time, waiting time, and queue length.

Methods: computer simulation [97, 187, 264, 348, 454], queueing theory [264, 535].

Capacity allocation. On the tactical level, resource capacities settled on the strategic level are subdivided over all patient groups. To do so, patient groups are first assigned to resource types.

- *Assign patient groups to resource types.* The assignment of patient groups to available resources requires knowledge about the capabilities of for example clinical staff, support staff or medical equipment, and the medical characteristics of patients. The objective is to maximize the number of patients served, by calculating the optimal assignment of patient groups to appropriately skilled members of clinical staff [450]. Efficiency gains are possible when certain tasks can be substituted between clinical staff, either horizontally (equally skilled staff) or vertically (lower skilled staff) [451].

- *Time subdivision.* The available resource capacities, such as staff and equipment, are subdivided over patient groups. For example, general practitioners divide their time between consulting patients and performing prevention activities for patients [227]. When patient demand changes over time (e.g., seasonality), a dynamic subdivision of capacity, updated based on current waiting lists, already planned appointments and expected requests for appointments, performs better than a long-term, static subdivision of resource capacity [498].

Methods: computer simulation [498], mathematical programming [227, 450], literature review [503].

Temporary capacity change. The balance between access times and resource utilization may be improved when resource capacities can temporarily be increased or decreased, to cope with fluctuations in patient demand [498]. For example, changing a CT scanner's opening hours [498] or changing doctor consultation time [162].

Methods: computer simulation [162, 498].

Access policy. In appointment-driven facilities, the access policy concerns the waiting list management that deals with prioritizing waiting lists so that access time is equitably distributed over patient groups. In the traditional approach, there is one queue for each doctor, but when patient queues are pooled into one joint queue, patients can be treated by the first available doctor, which reduces access times [496]. Another policy is to treat patients without a scheduled appointment, also called 'walk-in' service. In between scheduled and walk-in service is 'advanced access' (also called 'open access', or 'same-day scheduling'). With advanced access, a facility leaves a fraction of the appointment slots vacant for patients that request an appointment on the same day or within a couple of days. The logistical difficulty of both walk-in service and advanced access is a greater risk of resource idle time, since patient arrivals are more uncertain. However, implementation of walk-in/advanced access can provide significant benefits to patient access time, doctor idle time and doctor overtime, when the probability of patients not showing up is relatively large [390, 419]. A proper balance between traditional appointment planning and walk-in/advanced access further decreases access times and increases utilization [417, 535]. The specification of such a balanced design will be discussed below.

Methods: computer simulation [12, 178, 333, 390, 417, 496], heuristics [333], Markov processes [390], queueing theory [419, 535].

Admission control. Given the access policy decisions, admission control involves the rules according to which patients are selected to be admitted from the waiting lists. Factors that are taken into account are for example resource availability, current waiting lists and expected demand. Clearly, this makes admission control and capacity allocation mutually dependent. This is for example the case in [498], where the capacity subdivision for a CT scanner is settled by determining the number of patients to admit of each patient group. Access times can be controlled by adequate admission control [195, 203, 280, 498]. Admission control plays a significant role in advanced access or walk-in policies. Successful implementation of these policies requires a balance between the reserved and demanded number of slots for advanced access or walk-in patients. Too many reserved slots results in resource idle time, and too little reserved slots results in increased access time [408, 409].

Methods: computer simulation [498], heuristics [203], Markov processes [195, 203], mathematical programming [280, 408, 409].

Appointment scheduling. Appointment schedules are blueprints that can be used to provide a specific time and date for patient consultation (e.g., an MRI scan or a doctor visit). Appointment scheduling comprises the design of such appointment schedules. Typical objectives of this design are to minimize patient waiting time, maximize resource utilization or minimize resource overtime. A key trade-off in appointment scheduling is the balance between patient waiting time and resource idle time [89, 254, 288]. Appointment scheduling is comprehensively reviewed in [89, 229]. In an early article [514], the Bailey-Welch appointment scheduling rule is presented, which is a robust and well-performing rule in many settings [254, 283, 297]. References differ in the extent in which various aspects are incorporated in the applied models. Frequently modeled aspects that influence the performance of an appointment schedule are patient punctuality [178, 323, 518], patients not showing up ('no-shows') [178, 179, 255, 283], walk-in patients or urgent patients [12, 178, 417, 535], doctor lateness at the start of a consultation session [178, 179, 332, 421], doctor interruptions (e.g., by comfort breaks or administration) [179, 323], and the variance of consultation duration [254]. These factors can be taken into account when modeling the following key decisions that together design an appointment schedule.

- *Number of patients per consultation session.* The number of patients per consultation session is chosen to control patient access times and patient waiting times. When the number of patients is increased, access times may decrease, but patient waiting times and provider overtime tend to increase [85, 178, 254].

- *Patient overbooking.* Patients not showing up, also called 'no-shows', cause unexpected gaps, and thus increase resource idleness [254]. Overbooking of patients, i.e., booking more patients into a consultation session than the number of planned slots, is suggested to compensate no-shows in [299, 303, 320, 369, 453]. Overbooking can significantly improve patient access times and provider productivity, but it may also increase patient waiting time and staff overtime [299, 303]. Overbooking particularly provides benefits for large facilities with high no-show rates [299].

- *Length of the appointment interval.* The decision for the length of the planned appointment interval or slot affects resource utilization and patient waiting times. When the slot length is decreased, resource idle time decreases, but patient waiting time increases [179]. For some distributions of consultation time, patient waiting times and resource idle time are balanced when the slot length equals the expected length of a consultation [89]. The slot length can be chosen equal for all patients [179, 254, 514], but using different, appropriate slot lengths for each patient group may decrease patient waiting time and resource idle time when expected consultation times differ between patient groups [160].

- *Number of patients per appointment slot.* Around 1960, it was common to schedule all patients in the first appointment slot of a consultation session [184]. This minimizes resource idle time, but has a negative effect on patient waiting times [399, 421]. Later, it became common to distribute patients evenly over the consultation session to balance resource idle time and patient waiting time. In [184] various approaches in between these two extremes are evaluated, such as two patients in one time slot and zero in the next.

- *Sequence of appointments.* When different patient groups are involved, the sequence of appointments influences waiting times and resource utilization. Appointments can be sequenced based on patient group or expected variance of the appointment duration. In [297] various rules for patient sequencing are compared. Alternatively, when differences between patients exist with respect to the variation of consultation duration, sequencing patients by increasing variance (i.e., lowest variance first) may minimize patient waiting time and resource idle time [89].

- *Queue discipline in the waiting room.* The queue discipline in the waiting room affects patient waiting time, and the higher a patient's priority, the lower the patient's waiting time. The queue discipline in the waiting room is often assumed to be first-come-first-serve (FCFS), but when emergency patients and walk-in patients are involved, the highest priority is typically given to emergency patients and the lowest priority to walk-in patients [89]. Priority can also be given to the patient that has to visit the most facilities on the same day [348].

- *Anticipation for unscheduled patients.* Facilities that also serve unscheduled patients, such as walk-in and urgent patients, require an appointment scheduling approach that anticipates these unscheduled patients by reserving slack capacity. This can be achieved by leaving certain appointment slots vacant [151], or by increasing the length of the appointment interval [89]. Reserving too little capacity for unscheduled patients results in an overcrowded facility, while reserving too many may result in resource idle time. Often, unscheduled patients arrive in varying volumes during the day and during the week. When an appropriate number of slots is reserved for unscheduled patients, and appointments are scheduled at moments that the expected unscheduled demand is low, patient waiting times decrease and resource utilization increases [380, 417, 535]. In the online operational level of this section, we discuss referring unscheduled patients to a future appointment slot when the facility is overcrowded.

Methods: computer simulation [14, 85, 90, 133, 160, 178, 179, 239, 254, 255, 288, 303, 323, 332, 333, 348, 380, 417, 458, 503, 504, 514, 518], heuristics [85, 283, 333], Markov processes [184, 219, 283, 297, 329, 369, 453], mathematical programming [27, 85, 129, 418], queueing theory [58, 115, 151, 299, 320, 418, 503, 535], literature review [89, 229, 282, 451].

Staff-shift scheduling. Shifts are duties with a start and end time [74]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [166]. More attractive schedules promote job satisfaction, increase productivity, and reduce turnover. While staff dimensioning on the strategic level has received much attention, shift scheduling in ambulatory care facilities seems underexposed in the literature. In [71], shift schedules are developed for physicians, who often have disproportionate leverage to negotiate employment terms, because of their specialized skills. Hence, physicians often have individual arrangements that vary by region, governing authority, seniority, specialty and training. Although these individual arrangements impose requirements to the shift schedules, there is often flexibility for shifts of different lengths and different starting times to cope with varying demand during the day or during a week. In this context, the handbook [387] discusses staggered shift scheduling and flexible shift scheduling. In the first alternative, employees have varying start and end times of a shift, but always work a fixed number of hours per week. In the latter, cheaper alternative, a core level of staff is augmented with daily adjustments to meet patient demand.

Methods: computer simulation [395], mathematical programming [71], literature review [74, 166, 231, 387].

Offline operational planning

Patient-to-appointment assignment. Based on the appointment scheduling blueprint developed on the tactical level, patient scheduling comprises scheduling of an appointment in a particular time slot for a particular patient. A patient may require multiple appointments on one or more days. Therefore, we distinguish scheduling a *single appointment*, *combination appointments* and *appointment series*.

- *Single appointment.* Patients requiring an appointment often have a preference for certain slots. When information is known about expected future appointment requests and the expected preferences of these requests, a slot can be planned for this patient to accommodate the current patient, but also to have sufficient slots available for future requests from other patients. This can for example be necessary to ensure that a sufficient number of slots is available for advanced access patients [230, 509], or to achieve equitable access for all patient groups to a diagnostic facility [391].
- *Combination appointments.* Combination appointments imply that multiple ap-

pointments for a single patient are planned on the same day, so that a patient requires fewer hospital visits. This is the case when a patient has to undergo various radiotherapy operations on different machines within one day [397].

- *Appointment series.* For some patients, a treatment consisting of multiple (recurring) appointments may span a period of several weeks or months. The treatment is planned in an appointment series, in which appointments may have precedence relations and certain requirements for the time intervals in between. In addition, the involvement of multiple resources may further complicate the planning of the appointment series. The appointment series have to fit in the existing appointment schedules, which are partly filled with already scheduled appointments. Examples of patients that require appointment series are radiotherapy patients [109, 110, 111] and rehabilitation patients [100].

Methods: heuristics [100, 397, 509], Markov processes [230, 391, 509], mathematical programming [109, 110, 111].

Staff-to-shift assignment. On the tactical level, staff shift scheduling results in shifts that have to be worked. In staff-to-shift assignment on the offline operational level, a date and time are given to staff members to perform particular shifts. For example, a consultation session is scheduled for a doctor on a particular day and time, and with a certain duration. For an endoscopy unit, the authors of [280] develop a model to schedule available doctors to endoscopy unit shifts.

Methods: mathematical programming [280], literature review [231].

Online operational planning

Dynamic patient (re)assignment. After patients are assigned to slots in the appointment schedule, the appointments are carried out on their planned day. During such a day, unplanned events, such as emergency or walk-in patients, extended consultation times, and equipment breakdown, may disturb the planned appointment schedule. In such cases, real-time dynamic (re)scheduling of patients is required to improve patient waiting times and resource utilization in response to acute events. For example, to cope with an overcrowded facility walk-in patients can be rescheduled to a future appointment slot to improve the balance of resource utilization over time [411]. Dynamic patient (re)assignment can also be used to decide which patient group to serve in the next time slot in the appointment schedule [219], for example based on the patient groups' queue lengths. When inpatients are involved in such decisions, they are often subject to rescheduling [89], since it is assumed that they are less harmed by a rescheduled appointment as they are already in the hospital. However, longer waiting times of inpatients may be more costly, since it may mean they have to be hospitalized longer [120].

Methods: computer simulation [411], Markov processes [120, 219, 329], mathematical programming [120].

Staff rescheduling. At the start of a shift, the staff schedule is reconsidered. Before and during the shift, the staff capacities may be adjusted to unpredicted demand fluctuations and staff absenteeism by using part-time, on-call nurses, staff overtime, and voluntary absenteeism [222, 399].

Methods: no articles found.

3.5 Emergency care services

Emergency care services have the goal to reduce morbidity and mortality resulting from acute illness and trauma [412, 526]. To attain this goal, rapid response of an ambulance and transportation to an emergency care center (e.g., an emergency department in a hospital, or an emergency location near a disaster) are required [40, 412]. Patients arrive to the emergency department (ED) of a hospital as a self-referral, through ambulatory care services or by ambulance [60]. A frequently reported and studied problem in emergency care is that of long ED waiting times. One of the causes of long ED waiting time is treatment of a high number of self-referrals that could also be treated in ambulatory care services (e.g., by general practitioners). To cope with this problem, EDs increasingly cooperate with ambulatory care services, for example by combining the ED with a service that provides primary care outside office hours, or by opening an ambulatory walk-in center to which these patients can be referred [315]. The body of OR/MS literature directed to emergency care services is large. The existing literature mainly focuses on the strategic decisions regional coverage and capacity dimensioning for ambulances, and the tactical decision staff-shift scheduling.

Strategic planning

Regional coverage. To be able to provide rapid response to an acute illness or trauma, emergency care services need to be geographically close to their customer base, where emergencies can potentially occur [399]. Given a geographical region with a certain spatial distribution of service requests (i.e. emergency demand), the locations, types and number of emergency care facilities have to be decided. The objective is to find a balanced distribution of facilities to guarantee a desired level of service [39, 328]. This level of service can for example be measured by the maximum time it takes for a patient to travel to the closest ED, or the maximum response time that an ambulance requires to reach a specified region. The main trade-off in the decision where to locate emergency care centers and ambulances is between the level of service for emergency patients and costs [39, 45, 79, 237, 276]. Below, we will elaborate on this trade-off for both emergency care centers and ambulances.

- *Emergency care centers.* The decision where to locate emergency care centers, such as an ED in a hospital, is determined such that locations where emergencies may occur have at least one emergency care center within a target travel time or distance [413]. When a large-scale emergency or disaster occurs, an emergency care center may be unable to provide emergency care services (e.g., the facility is destroyed). In [257], this possibility is incorporated in regional coverage models that can be used to determine good locations for (temporary) emergency care centers after a disaster.

- *Ambulances (e.g., vans, motorcycles, helicopters, airplanes).* The decision where to locate ambulances is determined such that a specified region can be reached within a target response time by one or more ambulances, or that the average or maximum response time to a potential emergency is minimized [45, 79, 158, 165, 276]. The response time of an ambulance concerns the time elapsed from notification of an emergency until an ambulance arrives at the emergency location [180]. Other factors to take into account in planning the locations of ambulances are the likelihood of timing and location of an emergency, staff availability, location constraints (e.g., a place where staff can rest), and the emergency care center where patients are potentially transported to [36, 62, 79, 158, 180, 205, 267, 276, 399, 412, 526].

Methods: computer simulation [62, 165, 180, 186, 204, 237, 267, 412, 429, 459, 526], heuristics [22, 36, 38, 164, 197, 266], Markov processes [21, 257], mathematical programming [18, 36, 38, 39, 40, 45, 79, 118, 158, 165, 186, 205, 207, 237, 257, 276, 412, 413, 446, 459, 475], queueing theory [36, 197, 266, 314, 339, 446], literature review [67, 214, 276, 328, 399, 413].

Service mix. An organization decides the particular services that the emergency care facility provides. Facilities may provide services for particular types of emergency patients, which are possibly classified by severity of trauma. For example, a first-aid center may provide services that are adequate for minor emergencies, while an academic medical center is equipped to treat the most complex and severe traumas. In this case, treating minor emergencies at the first-aid center may alleviate the use of expensive resources in the academic medical center, and may be more cost-effective from a societal viewpoint. In order to balance provided emergency care and the cost of emergency care resources within a region or country, the service mix decision may be governed by societal influences and governmental regulations.

In general, the service mix decision is not made at an emergency care service level, but at the hospital level, as it integrally impacts the ambulatory, emergency, surgical and inpatient care services. The decided service mix dictates the case mix of emergency patients that can be served by the emergency care facility. Emergency care facilities in general do not decide a particular case mix, as they are often obliged to serve arriving emergency patients with any type of injury or disease [91, 393].

Methods: no articles found.

Ambulance districting. A covered region may be subdivided into several districts to which available ambulances are assigned. In subdividing a region and assigning ambulances to districts, it is the objective to minimize response times, while balancing the workload [36, 86, 314]. When an emergency occurs in a district, one of the available ambulances within that district is dispatched to the emergency [36, 314]. When none are available (e.g., when all ambulances are responding to a call), an ambulance from a different district may be dispatched to the emergency [314]. Such *interdistrict* dispatching decreases average response times, especially for relatively smaller districts. This is the effect of so-called pooling of resources [429]. Due to this possibility of *interdistrict* dispatching, only predicting the workload generated within the assigned district may not lead to a well-balanced ambulance districting decision. In the analysis of the districting problem, overlapping districts, mobile locations, and interdistrict dispatching should be included for an accurate prediction of workload balance [314].

Methods: computer simulation [204, 429], heuristics [36], mathematical programming [36], queueing theory [36, 86, 314].

Capacity dimensioning. Emergency care facilities dimension their resources with the objective to attain a reliable level of service while minimizing costs [39, 40, 376]. Often, this level of service is represented by a response target, for example $x\%$ of the emergency patients should be reached (ambulance) or seen (ED) within y minutes. An imbalance in supply and demand can lead to congestion or overcrowding in the ED [62]. ED overcrowding results in long waiting times, patients who leave the ED without being seen, and ambulance diversions [91, 220, 393]. This leads to patient dissatisfaction, medical errors, and decreased staff satisfaction [393]. A typical cause of congestion in the ED is the delay in admitting emergency patients to an inpatient bed due to congested medical care units and ICUs [16, 91, 106, 375, 376, 491]. Congestion may also be caused by insufficient available resources in the surgical care services (e.g., operating time capacity) and ambulatory care serivces (e.g., diagnostic equipment). Moreover, coordinated decision making for resource capacity dimensioning both within, and in services relating to emergency care services, reduces delays for emergency patients [16, 60, 91, 106, 311, 491]. The following resources are dimensioned:

- *Ambulances.* Ambulances exist in different transport modalities (helicopters, vans, cars) carrying different types of equipment and staff [38, 39, 40, 165, 186, 267, 412, 429, 446, 468, 526]. Ambulances collect emergency patients, but also perform less urgent transfers of patients between care facilities [468]. The number of ambulances should be chosen to include buffer capacity, to cope with fluctuations in demand and ambulance availability. Fluctuations in demand may be caused by expected demand peaks, such as large events, or unexpected demand peaks, such as large-scale accidents [526].

- *Waiting room.* The waiting rooms is possibly separated for patients awaiting results and patients awaiting initiation of service [106, 393].

- *Treatment rooms.* Treatment rooms comprise treatment beds or treatment chairs [91, 106, 311, 393]. Occupation of treatment rooms can be alleviated by letting patients await their lab test results in the waiting room and not in the treatment room [91, 106].

- *Emergency wards.* These are observation wards for a temporary stay, possibly before admission to the general wards [16, 106, 375, 376], also called Acute Admission Unit (AAU). Capacity is generally given in the number of beds.

- *Equipment.* Equipment may be required for emergency procedures, including treatment beds, treatment chairs and diagnostic equipment [91, 106, 393]. In general, diagnostic testing is considered outside the control of the emergency care services [181]. Emergency patients may require an X-ray or other diagnostic testing, and may have to compete with inpatients and outpatients for diagnostic resource capacity. Ineffective management of the diagnostic department causes delays in the emergency care services [181]. Installing diagnostic equipment in the ED may decrease the waiting time for diagnostic results, and therewith the overall length of stay of a patient in the ED [393].

- *Staff.* Staff in emergency care services is composed of different skill and responsibility levels, for example doctors, emergency nurses and support staff [60, 181, 220, 282, 311, 375, 376, 393, 532]. Required staffing dimensions, and thereby staff costs, may be reduced by passing on non-critical patients from doctors to lower-qualified, less-costly staff, releasing doctors to work on the critical cases [60]. Moreover, flexibility in staffing can be used to cost-effectively match uncertain emergency demand with resource capacity. For example, staff members may be 'on call' while working elsewhere or being off-duty, and they are called upon when additional staff is required in the ED to cope with unexpected demand peaks [376].

Methods: computer simulation [16, 40, 60, 91, 165, 181, 186, 267, 311, 315, 375, 412, 429, 526, 532], heuristics [38], mathematical programming [38, 39, 165, 375, 376, 412], queueing theory [106, 220, 446, 468], literature review [62, 282, 393].

Facility layout. The facility layout concerns the positioning and organization of different physical areas in a facility. Hospital managers aim to find the layout of the emergency care facility that maximizes the number of emergency patients that can be examined, given the budgetary and building constraints. Letting patients wait for their lab results in a waiting area instead of the treatment room enables the treatment rooms to be used more effectively, which can decrease patient waiting time [106]. Moreover, integration of the facility layout decision and the *patient routing* decision may decrease costs.

Methods: computer simulation [532], heuristics [387], literature review [393].

Tactical planning

Patient routing. An emergency patient process consists of multiple stages. We denote the composition and sequence of these stages as the route of a patient. Patient routes are designed to minimize patient waiting time, maximize patient throughput and increase staff utilization [60, 282]. A typical patient process is as follows. Patients arrive to the hospital as a self-referral, through ambulatory care services or by ambulance [60]. Generally, upon arrival at the ED, patients see a 'triage-nurse', who prioritizes these patients into urgency categories [91]. After triage and possibly a wait in the waiting room, patients see a medical staff member that aims to establish a diagnosis of the patient's condition timely and cost-effectively. In this phase, diagnostic tests (e.g., laboratory, X-ray) are typically required. Although more expensive, it may be decided to directly administer multiple diagnostic tests, to reduce the time to establish a diagnosis and patient waiting time [282]. When a diagnosis is determined, possibly a treatment is carried out at the ED. This treatment may be continued in the operating room or a medical care unit in the hospital. If (further) treatment is not required, the patient is discharged, possibly with a referral to an ambulatory care clinic [310].

To minimize patient waiting time, maximize patient throughput and increase staff utilization, alternative patient routing systems within the emergency care services may be developed. For example, a 'fast-track system' in the ED separates the patients with minor injuries and illnesses from the more severe traumas [106, 282, 349]. It reduces waiting time for patients with minor injuries and illnesses [349], but may lead to increased waiting time for the other patient groups, since less resources are available for these groups [181]. This may be acceptable, when the effect is not too large [349] and the increased waiting times are still within the set targets for each patient group [60, 282]. As relatively many steps in the emergency care process depend on effective and efficient processing in other care services (e.g., diagnostic services, surgical care services, and inpatient care services), coordinated decision making between the services involved in the emergency care process, reduces delays for emergency patients [60, 91, 106, 181].

Methods: computer simulation [60, 91, 181, 310, 349, 494], queueing theory [106, 352], literature review [282, 393].

Admission control. Admission control involves the rules according to which patients are selected to be served. The admission control rules first prescribe that the highest priority (life-threatened) patients are seen immediately, and that other patients can be deferred to the waiting room until they can be seen by a clinician [60]. Secondly, they define the order in which waiting patients are selected to be served. In the triage process, mentioned earlier in the *patient routing* decision, emergency patients are classified into 'triage categories' (often five) during an assessment by a qualified medical practitioner [91]. Typically, waiting time targets are set for each triage category, as a particular waiting time has a different impact on the health status of two patients in different urgency groups [352]. In general, patients are served in the order of triage category of decreasing urgency. However, applying more dynamic rules that take into account the number of waiting patients per triage category can enhance the compliance to the waiting time targets per category [60].

Methods: computer simulation [60, 91], queueing theory [352].

Staff-shift scheduling. Shifts are hospital duties with a start and end time [74]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [166]. The objective of shift scheduling is to generate shifts that minimize the number of staff hours required to cover the desired staffing levels [404]. The required staffing levels are determined by calculating how much staff is required to reach a given service level target, for example $x\%$ of the patients should be seen in y minutes.

For an ED, patient demand varies significantly throughout the week and throughout the day. Therefore, identical staffing schedules each day and each hour may seem convenient and practical, but they are likely suboptimal [216]. Implementing different staffing levels based on patient arrival rates for different moments within the day and week may decrease patient waiting times and reduce the number of patients that leave an ED without being seen [220]. To calculate the required staffing levels on each moment of the day, the working day is typically divided into planning intervals [215]. The required staffing level in each interval is dependent on the patient arrivals in that interval, but also by delayed congestion effects from prior intervals [215, 220]. Therefore, it can be beneficial to let a change in staffing level follow a change in patient arrival rate after a certain delay in time [215].

When staffing levels are determined, a set of shifts can be developed to meet those staffing levels as close as possible. Staggered shift scheduling is when shifts do not have to start and finish at the same time. This results in more flexibility to accommodate shifts to the required staffing levels at specific intervals, leading to improved utilization of resources [447]. Shift schedules are impacted by the preferences of staff and by laws prescribing emergency staff is only available for a limited number of hours [164, 220].

Methods: computer simulation [267, 447, 448, 532], heuristics [447, 448], queueing theory [215, 216, 220], literature review [231, 282, 393].

Offline operational planning

Staff-to-shift assignment. In staff-to-shift assignment, a date and time are given to a staff member to perform a particular shift. The objective is to attain the tactically settled staffing levels for each shift while minimizing costs, such as overtime by regular staff or staff hired temporarily from an agency [20]. Staff-to-shift assignments can be noncyclic and cyclic, where in the latter a staff member constantly repeats the same shift pattern [88]. In staff-to-shift assignment, labor laws, staff availability, and staff satisfaction have to be taken into account [20, 25, 124]. In [164], the staff-to-shift assignment for ambulance staff is coordinated with the *regional coverage* decision for ambulances, to maximize the provided service level for patients in a region.

Methods: heuristics [88], mathematical programming [20, 25, 88, 124, 164].

Online operational planning

Ambulance dispatching. Ambulance dispatching concerns deciding which ambulance to send to an emergency patient [8]. When calls to report an emergency event come in, a physician, nurse or paramedic assesses whether the reported emergency requires an ambulance. If so, the call is transferred to the dispatcher, who decides which ambulance will respond [446]. Many dispatching rules exist, and a commonly used rule is to send the ambulance closest to the emergency [330]. However, when predictions on future emergency calls are incorporated, sending the closest ambulance is not always optimal, as dispatching an ambulance makes it temporarily unavailable to respond to other calls. Shorter overall response times can be achieved when future demand is also incorporated in the dispatching decision [321]. When multiple calls come in, prioritizing calls and dispatching accordingly may balance ambulance workload [204]. Prioritizing and dispatching based on urgency improves response rates for the high-urgent calls [330]. After the dispatching decision has been made and an ambulance is traveling to the emergency, a request for emergency care may be canceled, leading to resource idle time [237].

Methods: computer simulation [8, 321, 330, 526], heuristics [321], mathematical programming [330], queueing theory [468].

Facility selection. When an ambulance has collected a patient, the emergency facility to which to bring a collected emergency patient has to be decided [526]. It is the aim to select the facility that minimizes the patient's travel time and is 'adequate' to serve the patient. The prospective emergency facility may for example be a local health clinic, a first-aid center or a hospital ED, and its appropriateness depends on the match between the facility's services, resources and bed availability, and the services required for the medical condition of the patient [429]. Delivering the emergency patient to the closest appropriate ED also leads to higher ambulance availability, as ambulance travel time is minimized [429].

Methods: computer simulation [429].

Ambulance routing. When an ambulance is dispatched to a particular emergency, the fastest route between an ambulance's location and the emergency location needs to be determined with the aim to minimize ambulance response times. Information with respect to distance, traffic, road work, accessibility can be taken into account. No specific contributions have been found for ambulance routing with our search method, but note that the a wide range of contributions in the general problem of vehicle routing exists [476].

Methods: no articles found.

Ambulance relocation. When ambulances are unavailable, for example because they are dispatched to emergency cases or they are in repair, they may leave a significant fraction of population without ambulance coverage [194]. In this case, to maximize regional coverage and to decrease response times, ambulances may be relocated [8, 67, 194, 350]. Relocation improves flexibility to respond to fluctuating patient demand [526] and dynamic traffic conditions [431]. When relocating ambulances dynamically, one aims to control the number of relocations to avoid successively relocating the same set of ambulances, long travel times between the initial and final location, and repeated round trips between the same two locations [67, 194]. However, the increased movements of ambulances caused by dynamic relocation, may also pose advantages. There is a higher chance of receiving a call while on the road, which may result in a decrease in response times caused by shorter turn-out times, i.e., the time for a crew to get ready before they can drive to an emergency when they are dispatched [350]. Real-time dynamic relocation is increasingly implementable, due to the increased availability of location information and the decreasing price of computing power [350].

Methods: computer simulation [8, 194, 526], Markov processes [350, 431], mathematical programming [194], literature review [67].

Treatment planning and prioritization. Each patient may follow a tailored set of stages through the ED, for example patients may receive different diagnostic tests and visit different types of doctors [91]. It is the objective to dynamically plan these stages, such that resource utilization is maximized and waiting time between stages is minimized. Planning the sequence of these stages and the selection of which task for which patient is performed at each point in time, includes various factors, such as urgency, medical requirements, resource availability, and patient waiting time. This planning decision is highly interrelated with the strategic and tactical level decisions *facility layout*, *patient routing*, and *admission control*. As these decisions shape the process flow for a particular patient and set the priority rules applying to the patient. Furthermore, there is a significant interdependence between *medical decision making* and *resource capacity planning* in this planning decision.

Methods: computer simulation [91, 181].

Staff rescheduling. When emergency demand for ambulances or in the emergency care facility is significantly higher than predicted or when staff is lower than expected (e.g., absent due to illness), additional staff may be required. Especially when senior doctors, who are required for key decisions such as discharge and particular treatments of a patient, are unexpectedly unavailable, it is recommended to call in an additional senior doctor [375].

Methods: computer simulation [532], mathematical programming [375].

3.6 Surgical care services

Surgeries are physical interventions on tissues, generally involving cutting of a patient's tissues or closure of a previously sustained wound, to investigate or treat a patient's pathological condition. Surgical care services have a large impact on the operations of the hospital as a whole [31, 50, 84], and they are the hospital's largest revenue center [84, 131]. Surgical care services include ambulatory surgical wards, where patients wait and stay before and after being operated. We do not classify such wards as inpatient care services, since patients served on ambulatory basis do not require an overnight stay. The proportion of ambulatory surgeries, which are typically shorter, less complex and less variable [398], is increasing in many hospitals [351]. There is a vast amount of literature on OR/MS in surgical care services, comprehensively surveyed in [49, 84, 134, 224, 228, 229, 336, 351, 404, 452, 506]. These surveys are used to create the *taxonomic* overview of the planning decisions.

Strategic planning

Regional coverage. At a regional level, the number, types and locations of surgical care facilities have to be determined to find a balanced distribution of facilities with respect to the geographical location of demand [153]. The main trade-off in this decision is between patient accessibility and facility efficiency. Coordination of activities between hospitals in one region, can provide significant cost reductions at surgical care facilities and downstream facilities [56, 428].

Methods: computer simulation [56], mathematical programming [428].

Service mix. An organization selects the particular services that the surgical care facility provides. The service mix stipulates which surgery types can be performed, and therefore impacts the net contribution of a facility [256]. Specific examples of services are medical devices to perform noninvasive surgeries and robotic services for assisting in specialized surgery [130]. In general, the service mix decision is not made at a surgical care service level, but at the regional or hospital level, as it integrally impacts the ambulatory, emergency, surgical and inpatient care services.

Methods: no articles found.

Case mix. The case mix involves the number and types of surgical cases that are performed at the facility. Often, diagnosis-related groups (DRGs), which classify patient groups by relating common characteristics such as diagnosis, treatment and age to resource requirements, are used to identify the patient types included in the case mix [260]. The case mix is chosen with the objective to optimize net contribution while considering several internal and external factors [224, 260]. Internal factors include the limited resource capacity, the settled service mix, research focus, and medical staff preferences and skills [50, 224, 281]. External factors include societal preferences, the disease processes affecting the population in the facility's catchment area [50], the case mix of competing hospitals [149], and the restricted budgets and service agreements in government funded systems [50]. High profit patient types may be used to cross-subsidize the unprofitable ones, possibly included for research or societal reasons [50].

Methods: computer simulation [281], mathematical programming [50, 260], literature review [224].

Capacity dimensioning. Surgical care facilities dimension their resources with the objective to optimize hospital profit, idle time costs, surgery delays, access times, and staff overtime [334, 433]. Therefore, provider capacity must be matched with patient demand [433] for all surgical resource types. The capacity dimensioning decisions for different resource types are highly interrelated and performance is improved when these decisions are coordinated both within the surgical care facility and with capacity dimension decisions in services outside the surgical care facility, such as medical care units and the Intensive Care Unit (ICU) [62, 432, 490, 491]. The following resources are dimensioned:

- *Operating rooms*. Operating rooms can be specified by the type of procedures that can be performed [23, 228, 282, 433].
- *Operating time capacity*. This concerns the number of hours per time period the surgical care services are provided [281, 351, 432, 469, 490]. Operating time capacity is determined by the number of operating rooms and their opening hours [334].
- *Presurgical rooms*. These rooms are used for preoperative activities, for example induction rooms for anesthetic purposes [351].
- *Recovery wards*. At these wards, patients recover from surgery [300, 301, 302, 432, 433]. The recovery ward is also called Post Anesthesia Care Unit (PACU) [224].
- *Ambulatory surgical ward*. At this ward, outpatients stay before and after surgery.
- *Equipment*. Equipment may be required to perform particular surgeries. Examples are imaging equipment [229] or robotic equipment [130]. Equipment may be transferable between rooms, which increases scheduling flexibility.
- *Staff*. Staff in surgical care services include surgeons, anesthesiologists, surgical assistants and nurse anesthetists [5, 72, 130, 256]. Staffing costs are a large portion of costs in surgical care services [10, 130]. Significant cost savings can be achieved by increasing staffing flexibility [130], for example by (i) cross-training surgical assistants for multiple types of surgeries [228], (ii) augmenting nursing staff with short-term contract nurses [130], and (iii) drawing nurses from less critical parts in the hospital during demand surges [130].

Methods: computer simulation [281, 300, 301, 302, 334, 432, 433, 490], heuristics [72, 130, 256], mathematical programming [23, 72, 130, 469], queueing theory [334], literature review [282, 351].

Facility layout. The facility layout concerns the positioning and organization of different physical areas in a facility. The aim is to determine the layout of the surgical care facility which maximizes the number of surgeries that can take place, given the budgetary and building constraints. A proper integration of the facility layout decision and the patient routing decision decreases costs and increases the number of patients operated [340]. For example, when patients are not anesthetized in the operating room, but in an adjacent induction room, patients can be operated with shorter switching times in between. In [351], contributions that model a facility layout decision for surgical care services are reviewed.

Methods: computer simulation [340], heuristics [387], literature review [351].

Tactical planning

Patient routing. A surgical process consists of multiple stages. We denote the composition and sequence of these stages as the route of a patient. The surgical process consists of a preoperative, perioperative and postoperative stage [224, 228, 398]. The preoperative stage involves waiting and anesthetic interventions, which can take place in induction rooms [340] or in the operating room [351]. The perioperative stage involves surgery in the operating room, and the postoperative stage involves recovery at a recovery ward [224]. Recovery can also take place in the operating room when a recovery bed is not immediately available [13]. Surgical patients requiring a bed are admitted to a (inpatient or outpatient) medical care unit before the start of the surgical process, where they return after the surgical process [274]. Efficient patient routes are designed with the objective to increase resource utilization [340].

Methods: computer simulation [340], heuristics [13], mathematical programming [13, 398], literature review [224, 351].

Capacity allocation. On the tactical level, resource capacities settled on the strategic level are subdivided over patient groups. The objectives of capacity allocation are to trade off patient access time and the utilization of surgical and postsurgical resources [49, 146, 224, 336, 469], to maximize contribution margin per hour of surgical time [84], to maximize the number of patients operated, and to minimize staff overtime [235]. Capacity allocation is a means to achieve an equitable distribution of access times [469]. Hospitals commonly allocate capacity through *block* scheduling [177, 224, 503]. Block scheduling involves the subdivision of operating time capacity in blocks that are assigned to patient groups [224, 228]. Capacity is allocated in three consecutive steps. First, patient groups are identified. Second, resource capacities, often in the form of operating time capacity, are subdivided over the identified patient groups. Third, blocks of assigned capacity are scheduled to a specified date and time.

- *Patient group identification.* In general, patient groups are classified by (sub)specialty, medical urgency, diagnosis or resource requirements. Identification by medical urgency distinguishes elective, urgent and emergent cases [84, 170, 224, 228]. Elective cases can be planned in advance, urgent cases require surgery urgently, but can incur a short waiting period, and emergency patients require surgery immediately [57, 84]. Examples of patient grouping by resource requirements are inpatients, day-surgery patients [228] and grouping patients by the equipment that is required for the surgery [130].

- *Time subdivision.* With the earlier mentioned objectives, operating time is subdivided over the identified patient groups based on expected surgery demand. This is often a politically charged and challenging task, since various surgical specialties compete for a profitable and scarce resource. What makes it even more complex is that hospital management and surgical specialties may have conflicting objectives [52]. When allocating operating time capacity to elective cases, a portion of total operating time capacity is reserved for emergency cases, which arrive randomly [196]. Staff overtime is the result when the reserved capacity is insufficient to serve all arriving emergency patients, but resource idle time increases when too much capacity is reserved, causing growth in elective waiting lists [57, 307, 308, 396, 536]. Capacity can be reserved by dedicating one or more operating rooms to emergency cases, or by reserving capacity in elective operating rooms [84, 306, 444].

- *Block scheduling.* In the last step of capacity allocation, a date and time are assigned to blocks of allocated capacity [31]. Several factors have to be considered in developing a block schedule. For example, (seasonal) variation in surgery demand, the number of available operating rooms, staff capacities, surgeon preferences, and material and equipment requirements [31, 428]. Block schedules are often developed to be cyclic, meaning the block schedule is repeated periodically. A (cyclic) block schedule is also termed a Master Surgical Schedule (MSS) [488]. Cyclic block schedules may not be suitable for rare elective procedures [224, 488]. For these procedures, capacity can be reserved in the cyclic block schedule [506], or non-cyclical plans may provide an outcome. When compared to cyclic plans, non-cyclic [130, 143, 144], or variable plans [282], increase flexibility, decrease staffing costs [130] and decrease patient access time [231, 282]. However, cyclic block schedules have the advantage that they make demand more predictable for surgical and downstream resources, such as the ICU and general wards, so that these resources can increase their utilization by anticipating demand more structurally [488].

In addition to block scheduling, the literature also discusses *open* scheduling and *modified block* scheduling. Open scheduling involves directly scheduling all patient groups in the total available operating time capacity, without subdividing this capacity first. Although open scheduling is more flexible than block scheduling, open scheduling is rarely adopted in practice [52, 224], because it is not practical with regards to doctor schedules and increases competition for operating time capacity [336, 404]. Modified block scheduling is when only a fraction of operating time capacity is allocated by means of block scheduling [140, 224]. Remaining capacity is allocated and scheduled in a later stage, which increases flexibility to adapt the capacity allocation decision based on the latest information about fluctuating patient demand [224].

Capacity allocation decisions in surgical care services impact the performance of downstream inpatient care services [31, 33, 49, 84, 130, 336, 403, 491, 492, 493]. Variability in bed utilization and staff requirements can be decreased by incorporating information about the required inpatient beds for surgical cases in allocating surgical capacity [4, 31, 33, 217, 428, 487, 488]. In contributions that model downstream services, it is often the objective to level the bed occupancy in the wards or the ICU, to decrease the number of elective surgery cancellations [31, 84, 403, 428, 462, 469, 487, 488], or to minimize delays for inpatients waiting for surgery [533].

Methods: computer simulation [57, 140, 143, 144, 307, 396, 533], heuristics [31, 32, 33, 462, 501], Markov processes [196, 492, 493, 536], mathematical programming [31, 32, 33, 51, 52, 95, 130, 231, 307, 403, 428, 462, 469, 470, 487, 488, 533], queueing theory [536], literature review [49, 84, 224, 228, 282, 336, 387, 491, 503, 506].

Temporary capacity change. Available resource capacity could be temporarily changed in response to fluctuations in demand [334]. When additional operating time capacity is available, it can be allocated to a particular patient group, for example based on contribution margin [228, 506] or access times [469], or it can be proportionally subdivided between all patient groups [52, 469].

Methods: computer simulation [140], mathematical programming [52, 130, 469], literature review [228, 231, 506].

Unused capacity (re)allocation. Some time periods before the date of carrying out a settled block schedule, capacity allocation decisions may be reconsidered in order to reallocate capacity that remains unused [148, 228, 250] and to allocate capacity not allocated before (for example in *modified block* scheduling, discussed in *capacity allocation*). When unused capacity is released sufficiently early before the surgery time is planned, better quality reallocations are possible than when the unused capacity is released on the same day it is available [250]. Unused capacity is (re)allocated with the same objectives as the *capacity allocation* decision.

Methods: computer simulation [140, 148], heuristics [148], Markov processes [250], literature review [228].

Admission control. Admission control involves the rules according to which patients from different patient groups are selected to undergo surgery in the available operating time capacity. There is a strong reciprocal relation between admission control decisions and capacity allocation decisions: capacity allocation decisions demarcate the available operating time capacity for surgeries, and admission control decisions influence the required operating time capacity. Admission control has the objective to balance patient service, resource utilization and staff satisfaction [49]. It is established by developing an admission plan that prescribes how many surgeries of each patient group to perform on each day, taking the block schedule into account [4]. Balancing the number of scheduled surgical cases throughout the week prevents high variance in utilization of involved surgical resources, such as operating rooms and recovery beds, and downstream inpatient care resources, such as ICU and general ward beds [3, 4, 31, 291, 336, 478]. Resource utilization can be improved by using call-in patients [49] and overbooking [57]. Call-in patients are given a time frame in which they can be called in for surgery when there is sufficient space available in the surgical schedule. Overbooking of patients involves planning more surgical cases than available operating time capacity to anticipate for no-shows [31]. Most patients requiring surgical care enter the hospital through the ambulatory care services. Although this makes admission control and capacity allocation policies for both ambulatory and surgical care services interdependent, not much literature is available on the interaction between ambulatory and surgical care services [491].

Methods: computer simulation [57, 138, 291, 478], Markov processes [365], mathematical programming [3, 4], literature review [49, 224].

Staff-shift scheduling. Shifts are hospital duties with a start and end time [74]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [166]. The objective of shift scheduling is to generate shifts that minimize the number of staff hours required to cover the desired staffing levels [404]. The desired staffing levels are impacted by the capacity allocation decisions. Hence, integrated decision making for capacity allocation and staff-shift scheduling minimizes required staff [32]. Flexible shifts can improve performance [52, 139]. One example is staggered shift scheduling, which implies that employees have varying start and end times of shifts [387]. It can be used to plan varying, but adequate staffing levels during the day, and to decrease overtime [52, 139].

Methods: heuristics [135], mathematical programming [32, 71, 150], literature review [231, 404].

Offline operational planning

Staff-to-shift assignment. In staff-to-shift assignment, a date and time are given to a staff member to perform a particular shift. The literature on shift scheduling and assignment in healthcare mainly concerns inpatient care services [166], which we address in Section 3.7.

Methods: no articles found.

Surgical case scheduling. Surgical case scheduling is concerned with assigning a date and time to a specific surgical case. Availability of the patient, a surgeon, an anesthetist, nursing and support staff, and an operating room is a precondition [49]. Surgical case scheduling is an offline operational planning decision, since it results in an assignment of individual patients to planned resources and not in blueprints for assigning surgical cases to particular slots. The objectives of surgical case scheduling are numerous: to achieve a high utilization of surgical and postsurgical resources, to achieve high staff and patient satisfaction, and to achieve low patient deferrals, patient cancellations, patient waiting time, and staff overtime [84, 130, 176, 274, 338, 398, 422, 444, 511]. The execution of a surgical case schedule is affected by various uncertainties in the preoperative stage duration, surgical procedure duration, switching time, postsurgical recovery duration, emergency patient interruption, staff availability, and the starting time of a surgeon [224, 398]. These uncertainty factors should be taken into account in surgical case scheduling.

Although surgical case scheduling can be done integrally in one step [13, 142, 143, 177, 336, 398, 422, 455], it is often decomposed in several steps. In the latter case, first, the planned length of a surgical case is decided. Second, a date and an operating room are assigned to a surgical case on the waiting list (also termed the 'advance scheduling' [336]). Third, the sequence of surgical cases on a specific day is determined [225, 336] (also termed the 'allocation scheduling' [336]). Fourth, starting times for each surgical case are determined. Below, we explain these four steps in more detail.

- *Planned length of a surgical case.* The planned length of a surgical case is the reserved operating time capacity in the surgical schedule for the surgical case duration, switching time and slack time. Surgical case duration, which is often estimated for each patient individually [381], is impacted by factors as the involved surgeon's experience, and the acuteness, sex, and age of the patient [136, 381]. Switching time between surgical cases includes cleaning the operating room, performing anesthetic procedures, or changing the surgical team [147]. Slack capacity is reserved as a buffer to deal with longer actual surgery durations than expected in advance [235]. When too little time is reserved, staff overtime and patient waiting time occur, and when too much time is reserved, resources incur idle time [147, 381, 511].

- *Assigning dates and operating rooms to surgical cases.* Dates and operating rooms are assigned to the elective cases on the surgical waiting list, following the settled admission control decisions [23, 175, 176, 235, 274, 341, 416]. The available blocks of operating time capacity are filled with elective cases. When too few cases are planned, utilization decreases, leading to longer waiting lists. Conversely, when too many cases are planned, costs increase due to staff overtime [57, 416]. Assigning dates and operating rooms to surgical cases can be done by assigning an individual surgical case, or by jointly assigning multiple cases to various possible dates and times. The latter is more efficient as more assignment possibilities can be considered [143].

- *Sequencing of surgical cases.* When the set of surgical cases for a day or for a block is known, the sequence in which they are performed still has to be determined. Factors to consider in the sequencing decision are doctor preference [224], medical or safety reasons [81, 274], patient convenience [81, 82], and resource restrictions [83]. Various rules for sequencing surgical cases are known [23, 81, 82, 226, 398, 416, 444]. In general, the traditional first-come-first-serve (FCFS) rule is outperformed by a longest-processing-time-first (LPTF) rule [49, 300, 302, 387]. When the variation of surgical case duration is known, sequencing surgical cases in the order of increasing case duration variation (i.e., lowest-variance-first) may yield further improvements [131, 511].

- *Assigning starting times to surgical cases.* The planned start time of each surgical case is decided [226]. This provides a target time for planning the presurgical and postsurgical resources, and for planning the doctor schedules [511]. The actual start time of a surgical case is impacted by the planned and actual duration of all preceding surgical cases [23, 511] and the completion time of the preoperative stage [145].

Emergency cases may play a significant role during the execution of the surgical case schedule [224]. Hence, incorporating knowledge about emergency cases, for example predicted demand, in surgical case scheduling decreases staff overtime and patient waiting time [57, 196, 306, 307, 308]. Often, surgical case scheduling is done in isolation. However, efficiency gains may be achieved by also considering decisions in other care services [81, 84, 102, 274, 398]. For example, without coordination with the ICU, a scheduled case may be rejected on its day of surgery due to a full ICU [398]. The contributions [13, 81, 102, 177, 256, 337, 366, 398, 433] do incorporate other care services, such as the patient wards and ICUs.

Methods: computer simulation [10, 57, 102, 137, 140, 142, 143, 146, 170, 226, 300, 302, 306, 307, 433, 470, 511], heuristics [10, 13, 83, 131, 136, 175, 177, 225, 226, 256, 306, 308, 341, 416, 422, 455, 486], Markov processes [196, 228, 365, 381], mathematical programming [13, 23, 81, 82, 83, 95, 102, 129, 130, 131, 175, 176, 177, 225, 274, 306, 307, 308, 338, 341, 396, 398, 416, 422, 444, 469], queueing theory [511], literature review [49, 84, 229, 336, 351, 387, 451].

Online operational planning

Emergency case scheduling. Emergency cases requiring immediate surgery are assigned to reserved capacity or to capacity obtained by canceling or delaying elective procedures [488]. It is the objective to operate emergency cases as soon as possible, but also to minimize disturbance of the surgical case schedule [229]. When emergency cases cannot be operated immediately, prioritizing of emergency cases is required to accommodate medical priorities or to minimize average waiting time of emergency cases [141, 398].

Methods: mathematical programming [141, 398], literature review [229].

Surgical case rescheduling. When the schedule is carried out, unplanned events, such as emergency patients, extended surgery duration and equipment breakdown may disturb the surgical case schedule [3, 338]. Hence, the surgical case schedule often has to be reconsidered during the day to mitigate increasing staff overtime, patient waiting time and resource idle time. Rescheduling may involve moving scheduled surgeries from one operating room to another and delaying, canceling or rescheduling surgeries [338].

Methods: mathematical programming [3, 338], literature review [228, 229].

Staff rescheduling. At the start of a shift, the staff schedule is reconsidered. Before and during the shift, the staff capacities may be adjusted to unpredicted demand fluctuations and staff absenteeism by using part-time, on-call nurses, staff overtime, and voluntary absenteeism [222, 399].

Methods: no articles found.

3.7 Inpatient care services

Inpatient care refers to care for a patient who is formally admitted (or 'hospitalized') for treatment and/or care and stays for a minimum of one night in the hospital [379]. Due to progress in medicine inpatient stays have been shortened, with many admissions replaced by more cost-effective outpatient procedures [377, 387]. Resource capacity planning has received much attention in the OR/MS literature, with capacity dimensioning being the most prominently studied decision.

Strategic planning

Regional coverage. At a regional planning level, the number, types and locations of inpatient care facilities have to be decided. To meet inpatient service demand, the available budget needs to be spent such that the population of each geographical area has access to a sufficient supply of inpatient facilities of appropriate nature and within acceptable distance [62]. Coordinated regional coverage planning between various geographical areas supports the realization of equity of access to care [47, 426]. To achieve this, local and regional bed occupancies need to be balanced with the local and regional probability of admission refusals resulting from a full census. The potential pitfall of deterministic approaches as used in [426] is that resource requirements are underestimated and thus false assurances are provided about the expected service level to patients [243].

Methods: computer simulation [243], mathematical programming [62, 426], queueing theory [47].

Service mix. The service mix is the set of services that healthcare facilities offer. healthcare facilities that offer inpatient care services can provide a more complex mix of services and can accommodate patient groups with more complex diagnoses [451]. In general, the inpatient care service mix decision is not made at an inpatient care service level, but at the regional or hospital level, as it integrally impacts the ambulatory care facilities, the operating theater and the wards. This may be the reason that we have not found any references focusing on service mix decisions for inpatient care services in specific.

Methods: no articles found.

Case mix. Given the service mix decision, the types and volumes of patients that the facility serves need to be decided. The settled service mix decision restricts the decisions to serve particular patient groups. Patient groups can be classified based on disease type, demographic information, and resource requirements [221]. In addition, whether patient admissions are elective or not is an influential characteristic on the variability of the operations of inpatient care services [482]. The case mix decision influences almost all other decisions, in particular the care unit partitioning and capacity dimensioning decisions [26].

Methods: computer simulation [221], heuristics [26, 482].

Care unit partitioning. Given the service and case mix decisions, the hospital management has to decide upon the medical care units in which the inpatient care facility is divided. We denote this decision as care unit partitioning. It addresses both the question which units to create and the question which patient groups to consolidate in such care units. Each care unit has its designated staff, equipment and beds (in one or more wards). The objective is to guarantee care from appropriately skilled nurses and required equipment to patients with specific diagnoses, while making efficient use of scarce resources [26, 156, 157, 206, 243, 251, 439, 499].

First, the desirability of opening shared higher-level care units like Intensive Care Units (ICU) or Medium Care Units (MCU) should be considered [480]. Second, the general wards need to be specified. Although care unit partitioning is traditionally done by establishing a care unit for each specialty, or sometimes even more diagnosis specific [451], specialty-based categorization is not necessarily optimal. Increasingly, the possibilities and implications of consolidating inpatient services for care related groups is investigated to gain from the economies-of-scale effect, so-called 'pooling' [522]. For example, many hospitals merge the cardiac and thoracic surgery unit [217], or allow gynecologic patients in an obstetric unit during periods of low occupancy [355]. In such cases, the overflow rules need to be specified on the tactical level. For geriatric departments, it has to be decided whether to separate or consolidate assessment, rehabilitation and long-stay care [359, 360]. Also, multi-specialty wards can be created for patients of similar length of stay, such as day-care, short-, week- and long-stay units [439, 499], or for acute patients [259, 482]. Concentrating emergency activities in one area (a Medical Assessment Unit) can improve efficiency and minimize disruption to other hospital services [376]. One should be cautious when pooling beds for patient groups with diverging service level [217] or nursing requirements [309]. A combined unit would require the highest service and nurse staffing level for all patient groups. As a result, acceptable utilization may be lower than with separate units. Also, pooling gains should be weighed against possible extra costs for installing extra equipment on each bed [309]. To conclude, the question whether to consolidate or separate clinical services from a logistical point of view is one that should be answered for each specific situation, considering demand characteristics but also performance preferences and requirements [243]. Obviously, the care unit partitioning decision is highly interrelated with the capacity dimensioning decisions, to be discussed next.

Methods: computer simulation [156, 157, 206, 243, 259, 439], heuristics [26, 309, 482], mathematical programming [376], queueing theory [217, 251, 355, 359, 360, 480, 522].

Capacity dimensioning. In conjunction with the care unit partitioning, the size of each care unit needs to be determined. Care unit size is generally expressed in the number of staffed beds, as this number is often taken as a guideline for dimensioning decisions for other resources such as equipment and staff.

- *Beds.* The common objective is to dimension the number of beds of a single medical care unit such that occupancy of beds is maximized while a predefined performance norm is satisfied [208, 371, 373, 415, 497, 519]. The typical performance measure is the percentage of patients that have to be rejected for admission due to lack of bed capacity: the admission refusal rate. Several other consequences of congested wards can be identified, all being a threat to the provided quality of care. First, patients might have to be transferred to another hospital in case of an emergency [108, 292, 347, 524]. Second, patients may (temporarily) be placed in less appropriate units, so-called misplacements [112, 156, 157, 217, 242, 247, 524]. Third, backlogs may be created in emergency rooms or surgical recovery units [104, 206, 217, 375, 376]. Fourth, elective admissions or surgeries may have to be postponed, by which surgical waiting lists may increase [7, 112, 208, 523, 524], which negatively impacts the health condition of (possibly critical) patients [478, 484]. Finally, to accommodate a new admission in critical care units, one may predischarge a less critical patient to a general ward [152, 517].

  The number of occupied beds is a stochastic process, because of the randomness in the number of arrivals and lengths of stay [295]. Therefore, slack capacity is required and thus care units cannot operate at 100% utilization [123, 217]. Often, inpatient care facilities adopt simple deterministic spreadsheet calculations, leading to an underestimation of the required number of beds [104, 112, 123, 238, 243]. Hospitals commonly apply a fixed target occupancy level (often 85%), by which the required number of beds is calculated. Such a policy may result in excessive delays or rejections [19, 217, 243, 295, 371]. The desirable occupancy level should be calculated as a complex function of the service mix, the number of beds and the length of stay distribution [242, 243]. This non-linear relationship between number of beds, mean occupancy level and the number of patients that have to be rejected for admission due to lack of bed capacity is often emphasized [7, 123, 242, 247, 295, 371, 372, 415]. In determining the appropriate average utilization, the effect of economies-of-scale due to the so-called portfolio effect plays a role: larger facilities can operate under a higher occupancy level than smaller ones in trying to achieve a given patient service level [217, 243, 244, 295], since randomness balances out. However, possible economies-of-scope due to more effective treatment or use of resources should not be neglected [217]. Units with a substantial fraction of scheduled patients can in general operate under a higher average utilization [217]. The effect of variability in length of stay on care unit size requirements is shown to be less pressing than often thought by hospital managers [217, 484]. Reducing the average length of stay shows far more potential. For care units that have a demand profile with a clear time-dependent pattern, these effects are preferably explicitly taken into account in modeling and decision making, to capture the seasonal [247, 335], day-of-week [152, 188, 247, 258] and even hour-of-day effects [29, 68, 104, 242]. This especially holds for units with a high fraction of emergencies admissions [451].

Capacity decisions regarding the size of a specific care unit can affect the operations of other units. Therefore, the number of beds needs to be balanced among interdependent inpatient care units [7, 66, 104, 105, 221, 243, 251, 259, 327, 347, 451]. Models that consider only a single unit neglect the possibility of admitting patients in a less appropriate care unit and thus the interaction between patient flows and the interrelationship between care units. Next to estimating utilization and the probability of admission rejections or delays, models that do incorporate multiple care units, also focus on the percentage of time that patients are placed in a care unit of a lower level or less appropriate care unit, or in a higher level care unit [11, 108, 202, 217, 327, 439]. The first situation negatively impacts quality of care as it can lead to increased morbidity and mortality [478] and the second negatively impacts both quality of care, as it may block admission of another patient, and efficient resource use [217, 439]. Some multi-unit models explicitly take the patient's progress through multiple treatment or recovery stages into account and try to dimension the care units such that patients can in each stage be placed in the care units that are most suitable regarding their physical condition [104, 108, 123, 172, 192, 199, 240, 245, 246, 259, 347, 439, 480].

- *Equipment.* In [499] it is stated that pooling equipment among care units can be highly beneficial. However, no references have been found explicitly focusing on this planning decision. This might be explained by the fact that the care unit partitioning and size decisions are generally assumed to be translatable to equipment capacity requirements. Therefore, many of the references mentioned under these decisions are useful for the capacity dimensioning of equipment.

- *Staff.* The highest level of personnel planning is the long-term workforce capacity dimensioning decision. This decision concerns both the number of employees that have to be employed, often expressed in the number of full time equivalents, and the mix in terms of skill categories [241, 375]. For inpatient care services it mainly concerns nursing staff. To deliver high-quality care, the workforce capacity needs to be such that an appropriate level of staff can be provided in the different care units in the hospital [166, 202]. In addition, holiday periods, training, illness and further education need to be addressed [74].

  Workforce flexibility is indicated as a powerful concept in reducing the required size of workforce [74, 127, 202, 451]. To adequately respond to patient demand variability and seasonal influences, it pays off to have substitution possibilities of different employee types, to use overtime, and to use part-time employees and temporary agency employees [451]. Just as with pooling bed capacity, economies-of-scale can be gained when pooling nursing staff among multiple care units. Nurses cross-trained to work in more than one unit can be placed in a so called floating nurse pool [74, 202, 309, 451]. Note that flexible staff can be significantly more expensive [222]. Also, [318] indicates that to maintain the desired staff capacity, it is necessary to determine the long-term human resource planning strategies with respect to recruiting, promotion and training. To conclude, integrating the staff capacity dimensioning decision with the care unit size decision yields a significant efficiency gain [202].

Methods: computer simulation [7, 19, 104, 108, 112, 156, 157, 206, 221, 222, 238, 241, 242, 243, 244, 259, 292, 295, 347, 371, 372, 373, 375, 415, 439, 478, 497, 517, 519, 523, 524], heuristics [309], Markov processes [7, 68, 172, 192, 199, 245, 246, 247, 335], mathematical programming [127, 202, 241, 318, 327, 375, 376], queueing theory [11, 29, 66, 104, 105, 123, 152, 188, 208, 217, 240, 251, 258, 292, 327, 415, 480, 484], literature review [74, 166, 399, 451].

Facility layout. The facility layout concerns the positioning and organization of different physical areas in a facility. To determine the inpatient care facility layout, it needs to be specified which care units should be next to each other [387] and which care units should be close to other services like the surgical, emergency and ambulatory care facilities [77]. Ideally, the optimal physical layout of an inpatient care facility is determined given the decisions on service mix, case mix, care unit partitioning and care unit size. However, in practice, it often happens vice versa: physical characteristics of a facility constraint service mix, care unit partitioning and care unit size decisions [77, 499]. Newly-built hospitals are preferably designed such that they support resource pooling and have modular spaces so that they are as flexible as possible with respect to care unit partitioning and dimensioning [499].

Methods: computer simulation [77], heuristics [387], mathematical programming [77].

Tactical planning

Bed reallocation. Given the strategic decision making, tactical resource allocation needs to ensure that the fixed capacities are employed such that inpatient care is provided to the right patient groups at the right time, while maximizing resource utilization. Bed reallocation is the first step in tactical inpatient care service planning. Medium-term demand forecasts may expose that the care unit partitioning and size decisions fixed at the strategic level are not optimal. If the ward layout is sufficiently flexible, a reallocation of beds to units or specialties based on more specific demand forecast can be beneficial [26, 242, 501]. In addition, demand forecasts can be exploited to realize continuous reallocation of beds in anticipation for seasonality in demand [284]. To this end, hospital bed capacity models should incorporate monthly, daily and hourly demand profiles and meaningful statistical distributions that capture the inherent variability in demand and length of stay [238]. When reallocating beds, the implications for personnel planning, and involved costs for changing bed capacity, should not be overlooked [6].

Methods: computer simulation [242, 284], heuristics [26, 501], mathematical programming [6], queueing theory [284].

Temporary bed capacity change. To prevent superfluous staffing of beds, beds can temporarily be closed by reducing staff levels [217]. This may for instance be in response to predicted seasonal or weekend demand effects [238, 244]. The impact of such closings on the waiting lists at referring outpatient clinics and the operating room is studied by [522, 523]. Temporary bed closings may also be unavoidable as a result of staff shortages [347]. In such cases hospitals can act pro-actively, to prevent bed closings during peak demand periods [26].

Methods: computer simulation [238, 244, 347, 523], heuristics [26], queueing theory [217, 522].

Admission control. To provide timely access for each different patient group, admission control prescribes the rules according to which various patients with different access time requirements are admitted to nursing wards. At this level, patients are often categorized in elective, urgent and emergency patients. Admission control policies have the objective to match demand and supply such that access times, rejections, surgical care cancellations and misplacements are minimized while bed occupancy is maximized. The challenge is to cope with variability in patient arrivals and length of stay. Smoothing patient inflow, and thus workload at nursing wards, prevents large differences between peak and non-peak periods, and so realizes a more efficient use of resources [4, 238, 502].

Patient resource requirements are another source of variability in the process of admission control. Most references only focus on maximizing utilization of bed resources. This may lead to extreme variations in the utilization of other resources like diagnostic equipment and nursing staff [451]. Also, as with temporarily closing of beds, possible effects of admission control policies on the waiting lists at referring outpatient clinics and the operating room should not be neglected [443]. Admission control policies can be both static (following fixed rules) and dynamic (changing rules responding to the actual situation).

- *Static bed reservation.* To anticipate for the estimated inflow of other patient groups, two types of static bed reservation can be distinguished. The first is refusing admissions of a certain patient type when the bed census exceeds a threshold. For example, to prevent the rejection of emergent admission requests, an inpatient care unit may decide to suspend admissions of elective patients when the number of occupied beds reaches a threshold [167, 189, 270, 285, 347, 355, 415, 443]. As such, a certain number of beds is reserved for emergency patients. This reservation concept is also known as 'earmarking'. Conversely, [293, 478] indicate that earmarking beds for elective postoperative patients can minimize operating room cancellations. In the second static level the number of reserved beds varies, for example per weekday. Examples of such a policy are provided in [46, 481] where for each work day a maximum reservation level for elective patients is determined.

- *Dynamic bed reservation.* Dynamic bed reservation schemes take into account the actual 'state' of a ward, expressed in the bed census per patient type. Together with a prediction of demand, the reservation levels may be determined for a given planning horizon [296] or it may be decided to release reserved beds when demand is low. Examples of the latter are found in [293], where bed reservations for elective surgery are released during weekend days, and [30], where admission quota are proposed per weekday. In [249], an extension to dynamic reservation is proposed which concerns calling in semi-urgent patients from an additional waiting list on which patients are placed who needs admission within 1-3 days.

- *Overflow rules.* In addition to the bed reservation rules, overflow rules prescribe what happens in the case that all reserved beds for a certain patient type are occupied. In such cases, specific overflow rules prescribe which patient types to place in which units [243]. Generally, patients are reassigned to the correct treatment area as soon as circumstances permit [451]. By allowing overflow and setting appropriate rules, the benefits of bed capacity pooling are utilized (see *capacity dimensioning: care unit size*), while the alignment of patients with their preferred bed types is maximized [347]. Various references focus on predicting the impact of specific overflow rules [210, 243, 251, 347, 439].

- *Influence surgical schedule.* For many inpatient care services the authority on admission control is limited due to the high dependency on the operating room schedule (see *surgical care services*). By adjusting the surgical schedule, extremely busy and slack periods can be avoided [4, 26, 152, 156, 170, 210, 217, 238, 469, 492, 493, 501, 502, 524] and cancellation of elective surgeries can be avoided [291]. In practice, the operating room planning is generally done under the assumption that a free bed is available for postoperative care [293], which may result in surgery cancellations. Therefore, both for inpatient and surgical care services coordinated planning is beneficial [3, 238].

Methods: computer simulation [3, 156, 170, 210, 238, 243, 291, 293, 347, 415, 439, 469, 478, 502, 524], heuristics [26, 501], Markov processes [46, 167, 249, 251, 296, 492, 493], mathematical programming [3, 4, 30, 469], queueing theory [30, 152, 189, 217, 270, 285, 355, 443, 481]

Staff-shift scheduling Shifts are hospital duties with a start and end time [74]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [166, 289]. For inpatient care services, it generally concerns the specification of 24-hours-a-day-staffing levels divided in a day, evening and night shift, during which demand varies considerably [74, 166]. Typically, this is done for a period of one or two months [399]. Staffing levels need to be set both for each care unit's dedicated nurses and for flexible staff in floating pools [309]. Also, [127, 222] investigate the potential of on-call nurses who are planned to be available during certain shifts and only work when required.

The first step in staff shift scheduling is to determine staffing requirements with a demand model [166, 231, 289, 463], based on which the bed occupancy levels [451] and medical needs are forecasted [309]. The second step is to translate the forecasted demand in workable shifts and in the number of nurses to plan per shift, taking into account the staff resources made available at the strategic decision level [507]. Often, nurse-to-patient ratios are applied in this step [222], which are assumed to imply acceptable levels of patient care and nurse workload [525]. To improve the alignment of care demand and supply, shift scheduling is preferably coordinated with scheduled admissions and surgeries [399], which also helps avoiding high variation in nurse workload pressure [32].

Methods: computer simulation [222], heuristics [309], mathematical programming [32, 127, 507, 525], queueing theory [463], literature review [74, 166, 231, 289, 399, 451].

Offline operational planning

Admission scheduling. Governing the rules set by tactical admission control policies, on the operational decision level the admission scheduling determines for a specific elective patient the time and date of admission. We found one reference on this decision: [111] presents a scheduling approach to schedule admissions for a short-stay inpatient facility that only operates during working days, which takes into account various resource availabilities such as beds and diagnostic resources. We suggest two reasons for the lack of contributions on this decision. First, when admission control policies are thoroughly formulated, admission scheduling is fairly straightforward. Second, as described before, for postoperative inpatient care authority of admission planning is generally at the surgical care services [501].

Methods: mathematical programming [111].

Patient-to-bed assignment. Together with the admission scheduling decision, an elective patient needs to be assigned to a specific bed in a specific ward. Typically, this assignment is carried out a few days before the effective admission of the patient. The objective is to match the patient with a bed, such that both personal preferences (for example a single or twin room) and medical needs are satisfied [94, 128]. An additional objective may be to balance bed occupancy over different wards.

Methods: heuristics [94, 128], mathematical programming [94, 128].

Discharge planning. Discharge planning is the development of an individualized discharge plan for a patient prior to leaving the hospital. It should ensure that patients are discharged from the hospital at an appropriate time in their care and that, with adequate notice, the provision of other care services is timely organized. The aim of discharge planning is to reduce hospital length of stay and unplanned readmission, and improve the coordination of services following discharge from the hospital [441]. As such, discharge planning is highly dependent on availability downstream care services, such as rehabilitation, residential or home care. Therefore, a need is identified for integrated coherent planning across services of different healthcare organizations [495, 513]. Patients whose medical treatment is complete but cannot leave the hospital are often referred to as 'alternative level of care patients' or 'bed blockers' [491, 513]. Also in discharge planning it is worthwhile to anticipate for seasonality effects.

Methods: computer simulation [495], queueing theory [513], literature review [441].

Staff-to-shift assignment. Staff-to-shift assignment deals with the allocation of staff members to shifts over a period of several weeks [166]. The term 'nurse rostering' is also often used for this step in inpatient care services personnel planning [74, 99]. The objective is to meet the required shift staffing levels set on the tactical level, while satisfying a complex set of restrictions involving work regulations and employee preferences [44, 74, 99, 273, 289, 483]. Night and weekend shifts, days off and leaves have to be distributed fairly [399, 451, 525] and as much as possible according to individual preferences [44, 166]. In most cases, to compose a roster for each individual, first sensible combinations or patterns of shifts are generated (cyclic or non-cyclic), called 'lines-of-work', after which individuals are assigned to these lines-of-work [166]. Sometimes, staff-to-shift assignment is integrated with staff-shift scheduling [74, 525]. 'Self-scheduling' is an increasingly popular concept aimed at increased staff satisfaction which allows staff members to first propose individual schedules, which are taken as starting point to create a workable schedule that satisfies the staffing level requirement set on the tactical level [423].

Methods: heuristics [44, 483], mathematical programming [44, 273, 423, 483, 525], literature review [74, 99, 166, 289, 399, 451].

Online operational planning

Elective admission rescheduling. Based on the current status of both the patient and the inpatient care facility, it has to be decided whether a scheduled admission can proceed as planned. Circumstances may require postponing or canceling the admission, to reschedule it to another care unit, or to change the bed assignment. Various factors will be taken into consideration such as severity of illness, age, expected length of stay, the probable treatment outcome, the (estimated) bed availability, and the conditions of other patients (in view of the possibility of predischarging an other patient) [292, 331, 442]. This decision is generally made on the planned day of admission or a few days in advance. Rescheduling admissions can have a major impact on the operations at the surgical theater [292].

Methods: computer simulation [292], heuristics [331], queueing theory [292, 442].

Acute admission handling. For an acute admission request it has to be decided whether to admit the emergency patient and if so to which care unit, which bed, and on what notice. The tactical admission control rules act as guideline. As with rescheduling elective admissions, the status of both the patient and the inpatient care facility are taken into account [292, 442]. In [292], it is calculated how long the waiting will be if the patient is placed on 'the admission list' and [442] proposes and evaluates an admission policy to maximize the expected incremental number of lives saved from selecting the best patients for admission to an ICU.

Methods: computer simulation [292], queueing theory [292, 442].

Staff rescheduling. At the start of a shift, the staff schedule is reconsidered. Based on severity of need in each care unit, the float nurses and other flexible employees are assigned to a specific unit and a reassignment of dedicated nurses may also take place [74, 451]. In addition, before and during the shift, the staff capacities among units may be adjusted to unpredicted demand fluctuations and staff absenteeism by using float, part-time, on-call nurses overtime, and voluntary absenteeism [222, 399].

Methods: computer simulation [222], mathematical programming [406], literature review [74, 399, 451].

Nurse-to-patient assignment. At the beginning of each shift, each nurse is assigned to a group of patients to take care for. This assignment is done with the objective to provide each patient with an appropriate level of care and to balance workloads [367, 456]. Distributing work fairly among nurses improves the quality of care [367]. Generally, the assignment has to satisfy specified nurse-to-patient ratios [406]. Additionally, when patient conditions within one care unit can differ considerably, for each specific patient an estimate of the severity of the condition (and thereby expected workload) is made, in most cases on the basis of a certain severity scoring system [367]. In [406], it is explicitly taken into account that patient conditions, and therefore care needs, can vary during a shift. They state that it is preferred to also decide at the beginning of each shift to which nurse(s) unanticipated patients will be assigned.

Methods: computer simulation [456], heuristics [367], mathematical programming [367, 406].

Transfer scheduling. Throughout the inpatients' stay, the transfer scheduling is done to the appropriate inpatient care unit or to other areas within the hospital for treatments or diagnoses [399]. Transfer scheduling includes the planning of transportation. Transfer scheduling is often postponed until the time an already occupied bed is requested by a new patient. However, in [472] it is concluded that when relocation of patients is done proactively, admission delays for other patients can significantly be reduced, which has a positive effect on both quality and efficiency.

Methods: Markov processes [472].

3.8 Home care services

Home care includes medical, paramedical and social services delivered to patients at their homes [313]. It represents an alternative for hospitalization or placement in a residential care facility [37]. Home care services are a growing sector in the healthcare domain [37, 96], which might be because it is in general less costly [313] and it has a positive effect on a patient's quality of life [169]. Their development is accelerated by factors such as the ageing of the population, the increase of chronic diseases, the introduction of innovative technologies, and the continuous pressure of governments to contain healthcare costs [37, 96, 169]. Home care is provided in multi-disciplinary teams, since patients typically have a mixture of social, physical, psychological needs, and home care professionals may carry out several patient visits during a day. This diversity, multi-disciplinarity and the fact that the patient's home has to be integrated in the care supply chain makes the resource planning of home care delivery complex [37, 96, 313]. Coordination between the various disciplines is required to ensure continuity of care and to prevent overlap of care [37].

The body of OR/MS literature focusing on home care resource capacity planning in healthcare is not extensive compared to other care services. A single review was available [37], which has been a valuable starting point for the taxonomic overview in this section. It has been noted in the literature that due to the nature of home care services, intelligent portable electronic devices have a high potential in supporting home care organization and they are more and more used [28, 169]; recall that ICT solutions are not the focus of our review.

Strategic planning

Placement policy. The placement policy decision prescribes which patient types are eligible for home care services, and which are preferably admitted to an inpatient or residential care facility. The aim is to provide patients with the right treatment at the right time in the most cost-effective manner [530]. Defining placement policies requires classification systems by which health status and type of care requirements can be assessed [313, 530]. Often, for a single patient there are multiple alternatives for what type of care facility is best suitable. Optimal placement involves the consideration whether or not to treat a patient in a hospital bed, and at which point during recovery a patient is transferred from hospital care to home care [96, 530]. This makes coordination of inpatient, residential and home care resource capacities desirable [37].

Methods: heuristics [530], Markov processes [313], mathematical programming [96], literature review [37].

Regional coverage. At a regional planning level, the number, types and locations of home care agencies are decided. Unlike hospitals, which cater to a population not constrained to a specific area, home care agencies are generally responsible for the population in a given area, possibly assigned by the government [48, 305]. To meet home care service demand, the available budget needs to be spent such that the population in the area has access to a sufficient supply of home care services. Since care is delivered at a patient's home, the distance between agencies is only a provider efficiency issue, and does not affect patient accessibility [37]. No specific contributions to regional coverage planning in home care have been found.

Methods: literature review [37].

Service mix. A home care organization has to decide which services to offer. With respect to home care services, [169] distinguishes home care and home *health* care. Home care involves helping patients with everyday activities, such as bathing, dressing, eating, cooking, cleaning, and monitoring the daily medication regime. Home healthcare involves helping patients recover from an illness or injury. Therefore, home healthcare is often provided by registered nurses, therapists, and home health assistants. Another service is that of social and emotional support to patients and their family [37]. Home care services solely involving medication or meal distribution are outside the scope of our review, as these are secondary services. All found contributions treat the service mix decision as given in their models.

Methods: literature review [37].

Case mix. Aligned with the service mix, an organization needs to determine the types and volumes of patients it will serve. Patient types can be grouped according to pathology or to required type of care [37, 169]. Based on duration and content of care, [37] distinguishes four types of care: punctual care, continuous care, palliative care, rehabilitation care. For home care, the variety of required care in type, frequency and time is substantial [37, 121]. Therefore, almost all other planning decisions are geared to the case mix. However, as with the service mix decision, all found contributions treat this decision as given.

Methods: literature review [37].

Panel size. The panel size, also called calling population, is the number of potential patients of a home care facility. Since only a fraction of these potential patients actually demands home care services, the panel size can be larger than the number of patients a facility can serve. The goal is to set the panel size such that a minimum standard of service is ensured, while making efficient use of available resources [121], where service is typically measured in access time and efficiency in staff utilization. To achieve this, future care needs originating from the potential patients have to be forecasted. The panel size decision is closely connected to the *regional coverage* and *districting* decisions.

Methods: mathematical programming [121].

Districting. Districting involves the partitioning into districts of the area in which an organization is responsible for the logistics of home care visits. Typically, each district falls under the responsibility of one multi-disciplinary care team [37, 48]. Districting is done to limit the travel distances and times of care providers between the homes of patients, to improve coordination between different care providers treating the same patient and to encourage long-term relationships between providers and patients [37, 169]. Although these reasons plead for small districts, the districts should not be too small, to avoid inefficient operations [48]. Also, the objective of balancing workload among the districts in an area can be taken into account [48].

Methods: heuristics [48], literature review [37].

Capacity dimensioning. Home care organizations dimension their resources, to spend the available budget such that a satisfactory quality of care is realized and access times are minimized [37, 75]. To this end, provider capacity must be matched with patient demand. Since individual home care is in general a long-term process, the capacity dimensioning decision also requires long-term demand forecast models based on demographic information [236, 485]. In [305], it is indicated that true care needs are hard to estimate as care demands stored in historical data tend to be biased by the realized (un)availability of services. Capacity is dimensioned for the following resource types:

- *Staff.* Many healthcare professionals are involved in home care delivery, including nurses, occupational therapists, physiotherapists, speech therapists, nutritionists, home support workers, social workers, physicians and pharmacists [75, 305]. For each skill category it is determined whether to employ staff or to (temporarily) hire staff from an external agency when required [37]. Where usually professionals are dedicated to a fixed district, flexibility to respond to fluctuating demand can be achieved by allowing care providers to work in more than one district [305].

- *Equipment.* Medical and paramedical equipment and information technology equipment can be involved in providing home care [37]. Sharing resources among multiple districts may induce cost savings [389].

- *Fleet vehicles.* Means of transport, rented or bought, are required for visiting patients [37].

When capacity is dimensioned to cover average demand, variations in demand may cause long access times. Basic rules from queueing theory demonstrate the necessity of excess capacity to cope with uncertain demand [75]. Variation arises not only from uncertainty in the arrival process of care requests but also from the different levels of care required per patient [37]. The multi-disciplinary nature of home care causes a diversity of resources to be involved. Therefore, the capacity dimensioning decisions for different resource types are highly interrelated and performance is improved when these decisions are aligned. If dimensions are not properly balanced, some resources may become bottlenecks, while at the same time others are underutilized [37, 75].

Methods: computer simulation [389, 485], Markov processes [236], queueing theory [75], literature review [37].

Tactical planning

Capacity allocation. On the tactical level, resource capacities settled on the strategic level are subdivided over districts and patient groups. The objective is to equitably allocate resources: workload, access times, and quality of care (for example measured in number of visits per patient) should be balanced over districts and patient groups [75, 121]. Capacity allocation requires two steps:

- *Patient group identification.* Patients are classified by medical urgency or resource requirements [37, 75].

- *Time subdivision.* Capacity subdivision is provided in the number of care hours available per discipline per district [121, 305]. Time subdivision can be done based on number of inhabitants per district. However, [53] proposes to use a detailed demand estimation per patient group, taking into account demographical information including age and gender distributions.

To accurately respond to demand fluctuations, a dynamic subdivision of capacity, updated based on current waiting lists, already planned visits and expected requests for appointments, performs better than a static one [121, 305]. Finally, a close cooperation with other health organizations such as residential and in-patient care services may yield better future demand predictions [37].

Methods: heuristics [53, 305], mathematical programming [121], queueing theory [75], literature review [37].

Admission control. Admission control involves the rules according to which patients are selected to be admitted to home care services from the waiting lists. Admission control policies have the objective to match demand and supply such that access times are minimized while resource utilization is maximized, taking into account resource availability, current waiting lists and expected demand. Clearly, admission control and capacity allocation are interrelated. Patient needs and available resources must be balanced to prevent poor service levels or staff overutilization [121]. The challenge in admission control is to cope with various sources of variability such as variation in patient arrivals, patient home locations, urgency, number of visits per week per discipline required, patient health conditions, and treatment durations [121].

Multiple waiting lists are created based on geographical area and patient groups [75]. To provide timely access for each urgency class, patients are typically categorized in several priority groups within a waiting list [75]. A possible admission policy is to always take the patient with highest priority into service whenever capacity becomes available [75]. Another option is to develop an admission plan that prescribes how many patients of each patient group are taken into service during a certain planning horizon. The latter has the potential to simultaneously decrease access times and increase resource utilizations on the long run [121]. Incorporating forecasts of future care pathways of patients already in service and of patients on the waiting list can also improve performance [121, 313]. Finally, [252, 305] signifies that when case load is high in one district while another district is (temporarily) underutilized, it is beneficial to allow the flexibility of dynamic admissions over district borders.

Methods: Markov processes [313], mathematical programming [121, 252], queueing theory [75].

Staff-shift scheduling Shifts are duties with a start and end time [74]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [166]. Staffing levels per discipline and per district need to be such that feasible operational plans can be generated [37]. Shifts for various disciplines need to be synchronized to accommodate simultaneous visits [96] and to facilitate interdisciplinary team meetings [37]. By staffing a surplus team of which the members are able to work in whichever district that is required, the home care organization is able to respond to temporary demand fluctuations, and unplanned staff absence due to sickness [305].

Methods: heuristics [305], literature review [37].

Offline operational planning

Assessment and intake. Upon a home care request, first an assessment and intake process takes place. This process consists of assessing the patient's eligibility for home care, determining the care requirements and assigning a reference care provider. The eligibility is determined based on the strategic placement policy, together with specific personal characteristics, among which the social situation of the patient. The latter is also an important factor in the estimation of the patient's needs, since for example family assistance can reduce demands for professional support [37, 449]. The patient's health status and social situation are very specific, hence customized care programs are required [96]. Determining care requirements is in this phase done at an aggregate level, for example in hours per care type per week [169]. This is not only important from a patient's point of view, but also for the home care organization, as it dictates resource requirements on the short term [313]. Estimating a patient's care pathway and possible variation herein facilitates forecasting resource requirements on the medium term [313]. The reference care provider, also called case manager [96], is responsible for coordination of the entire multidisciplinary treatment [37]. Based on the resource requirements estimation, the reference care provider assignment can be done such that case load is balanced among home care employees [37, 252]. Inter-organizational coordination in the assessment and intake process is crucial to know about a patient arrival in an early phase [530]. This promotes continuity of care between discharge at inpatient and residential care facilities and admission to home care services [37, 96].

Methods: heuristics [449, 530], Markov processes [313], mathematical programming [96, 169, 252], literature review [37].

Staff-to-shift assignment. Staff-to-shift assignment deals with the allocation of staff members to shifts over a period of several weeks [166]. The objective is to meet the required shift staffing levels set on the tactical level, while satisfying a complex set of restrictions involving work regulations and employee preferences [74]. Weekend shifts, days off and leaves have to be distributed fairly [399, 451] and as much as possible according to individual preferences [166], which include working times, preferential days, vacation and performing particular care activities [37]. The decision is often integrated with *visit scheduling* [41, 169], the decision that is discussed next.

Methods: heuristics [41, 169], mathematical programming [41, 169], literature review [37].

Visit scheduling. Visit scheduling determines which visit will be performed, on which day and time, and by which staff member. It consists of two components: creating detailed care plans per patient, and the appointment scheduling. This visit scheduling is complex, since all patients have to be treated individually at their own home. Therefore, all tasks have to be planned in advance and synchronization of all human and material resources is required [37, 96]. Visit scheduling consists of three components:

- *Short-term care plan.* For each patient it has to be determined when, which visits by which care discipline are (medically) necessary [28, 37, 168, 169, 252, 313].
- *Staff-to-visit assignment.* Each visit has to be assigned to a certain staff member [28, 37, 41, 65, 168, 169, 252, 313].
- *Route creation.* For each care provider individual routes are constructed that determine at which day and what time each visit will be done [28, 37, 41, 65, 63, 64, 96, 168, 169].

Since the three components are highly interdependent, an integrated approach is required to determine the complete visit schedule all at once [28, 41, 96, 168, 169]. It may even be necessary to integrate the staff-to-shift assignment decision [41, 168, 169]. This integration of different planning and scheduling decisions makes home care operational planning also mathematically difficult [28, 41, 168, 169]. Typically, a visit base plan is made a few weeks in advance for a planning horizon of several weeks, which assigns specific visits to specific weekly time buckets. Then, around a week in advance the detailed visit schedule is established [28, 168, 169]. A wide set of constraints needs to be satisfied, like provider skill qualifications, working hours, geographical coherence between the district of patient and staff member, and allowed time windows for each visit [28, 37, 41, 63, 64, 168, 169]. In addition, precedence and synchronization requirements need to be satisfied, since some patients need simultaneous or sequential tasks requiring multiple resources [28, 65]. Also, uncertainty of travel times and visit durations need to be taken into account [37]. The goal is to design visit schedules that are efficient in terms of minimizing labor costs, travel time and distance [28, 63, 64, 65, 96] and such that preferences of both patient and providers are considered. Preferences of patients include preferential days, preferred staff, minimize unplanned visits, and continuity of care reflected in same day, same time and same staff [28, 41, 65, 168, 169]. Staff preference mainly concerns equity of workload, expressed in visit load and travel load [28, 168, 169, 252]. Workload imbalance can be reduced by allowing temporary deployment of staff outside their own district [28, 252].

Methods: heuristics [28, 41, 63, 64, 168, 169], Markov processes [313], mathematical programming [28, 41, 63, 64, 65, 96, 168, 169, 252], literature review [37].

Online operational planning

Visit rescheduling. The visit schedule is updated a few days in advance, the day in advance, and on the day of execution itself. Rescheduling is required to respond to unplanned events such as unplanned staff absenteeism, changed visit requirements due to changed patient health conditions, and incoming urgent care requests [28, 168, 169, 477]. It involves integrally rescheduling care plans, staff rescheduling, and rerouting [28, 168, 169]. One has to decide whether to arrange staff replacements (internally or externally) or to fit additional tasks into the routes of the available staff, and whether to postpone some less time-critical tasks to a later day. Weather conditions such as snowfalls, floods or storms, can be a source for the necessity of visit rescheduling [477].

Methods: heuristics [28, 168, 169, 477], mathematical programming [28, 168, 169, 477].

3.9 Residential care services

Residential care services cover a range of healthcare services for patients, often elderly, who have acute, chronic, palliative or rehabilitative healthcare needs that do not allow them to stay at home, but who do not strictly require a hospital stay [236]. Making residential care available to such patients avoids long-term hospital admissions, which are in general more costly [24]. The body of OR/MS literature directed to residential care services is limited. The literature has mainly focused on predicting patients' health progress, to support the strategic decisions placement policy and capacity dimensioning. The dynamics of residential care services, although on a slower time scale, are similar to that of inpatient care services. Therefore, most planning decisions and insights described under inpatient care services also apply to residential care services. This fact and the low variety in addressed planning decisions in the literature are the reasons that we choose for residential care services, as opposed to the other care services, to only present planning decisions for which we found references.

Strategic planning

Placement policy. The placement policy decision prescribes which patient types are eligible for which type of residential care services, and which are preferably admitted to inpatient or home care services. The aim is to provide patients with the right treatment at the right time using means which are most cost-efficient [24]. Defining a placement policy requires classification systems by which the health status and care requirements of a patient can be assessed [24, 530]. Often, for a single patient there are multiple alternatives for what type of care facility is most suitable. This especially holds for elderly patients, since they often suffer from multiple pathologies [358]. The placement policy involves the consideration whether to treat a patient in a hospital bed, and at which point during recovery a patient is transferred from the hospital to residential care [346, 358, 440]. This makes coordination between inpatient and residential care resource capacity planning desirable. Although hospital beds are in general more costly, a relatively short hospital stay may prevent a long stay in residential care, which may therefore be less expensive in the long run [212, 364, 438].

To derive optimal placement policies, many contributions model the movement of patients through the healthcare system including both hospital and residential care [358, 440, 467, 505, 529], the progress of patients through different health states [103, 173, 174, 191, 268, 356, 358, 359, 360, 364, 440, 464, 465, 466, 467, 529], and part of them include an estimation of related cost [343, 344, 346, 360, 394, 528]; other contributions model the relation between gender, age and clinical patient characteristics to length-of-stay and resource consumption in each stage [174, 343, 344, 345, 357, 364, 440, 485, 528, 529]. Various cost evaluations include analysis of demographics and individual life-expectations [103, 236, 394, 528, 529].

Methods: computer simulation [356, 485], heuristics [24, 530], Markov processes [103, 173, 174, 191, 236, 268, 343, 344, 345, 346, 356, 357, 358, 359, 360, 364, 394, 438, 440, 464, 465, 466, 467, 505, 528, 529], queueing theory [212].

Regional coverage. At a regional planning level, the number, types and locations of residential care facilities have to be decided. To meet residential care service demand, the available budget needs to be spent such that the population of each geographical area has access to a sufficient supply of facilities of appropriate nature [62, 114]. In general, the primary criteria for the locations are not so much closeness to customer bases but costs of site acquisition and construction, cost of operation, and speed of access to acute care facilities [399]. However, for rehabilitation care, where patients stay for a relatively short period, distance between the facility and the patients home and family is of importance to stimulate reintegration into their communities [114]. When the locations of facilities are well-spread over a region and load is balanced between facilities, equity of access to care is maximized, since the situation is avoided that some facilities have long waiting lists while other facilities have idle beds [154].

Methods: mathematical programming [62, 114, 154], literature review [399].

Case mix. Aligned with the service mix, an organization needs to determine the types and volumes of patients it will serve. Patient types can be grouped according to pathology, required type of care and resource requirements. For example, in [185], resource-utilization groups (RUGs) are presented which classifies patient groups by relating diagnosis, mental condition, and mobility, to resource requirements. The case mix decision is an influential factor with respect to almost all other planning decisions, especially to staff related decisions [185], since the length-of-stay of different patient types can be significantly different (i.e., a rehabilitation short-stay versus a geriatric long-stay).

Methods: heuristics [185].

Capacity dimensioning. Residential care organizations dimension their resources, to spend the available budget such that a satisfactory quality of care is realized, while access time is minimized and resource utilization is maximized [161]. To this end, provider capacity must be matched with patient demand. To estimate patient demand (in number and length-of stay), the earlier mentioned models for the movement of patients through health states and through the health system are applicable [173, 174, 199, 268, 343, 344, 345, 346, 356, 357, 359, 360, 364, 440, 464, 465, 466, 467, 505, 528, 529]. Due to the long-term character of residential care, long-term demand forecast models that include demographic information and survival analysis as presented in [103, 132, 236, 394, 485] have additional value. To anticipate for the uncertainty of long-term demand developments such as population ageing, scenario analysis can be applied to answer what-if questions [465, 466]. Capacity is dimensioned for the following resource types:

- *Beds*. The size of residential care facilities is generally expressed in the number of beds. This number can be taken as a guideline for dimensioning decisions for other resources such as equipment and staff. The common objective is to dimension the number of beds of facilities such that occupancy of beds is maximized while admission rejection and delay is minimized [209]. Delay in admission of patients to appropriate care facilities negatively affects therapeutic effectiveness [209]. To be able to realize quick turn-over for short-stay patients, strict separation within a facility of short-stay and long-stay patients might be preferred [161]. However, allowing overflow between longs-stay and short-stay beds potentially increases bed utilization. In that case, an appropriate balance between short-stay and long-stay beds is required [98, 161, 191, 209], to avoid short-stay bed blocking by long-stay patients. A relatively small decrease in the number of long-stay beds, with a corresponding rise in the number of short-stay beds, has a dramatic effect on the number of patients that can be treated [98, 212, 438]. In addition, balancing capacities between facilities is required, since for instance hospital discharges are highly dependent on availability of downstream care services [394]. When patients find the facilities to which they are referred to be full, they are forced to wait at their current, often unnecessarily intensive (and generally more expensive) care facilities. These patients unnecessarily block the current beds while waiting, preventing utilization by potential patients who require care at these facilities [294]. Hence, inappropriate bed dimensioning for residential services causes both degradation of quality of care and financial losses due to these 'alternative level of care' patients [98, 392]. Again, a need is identified for integrated coherent planning across services of different healthcare organizations [513].

- *Staff*. In view of the increasing residential care demand and a declining labor force, changes in staff skill mix are worthwhile to consider. It might be able to identify subtasks for which can be carried out by less qualified staff [132].

Methods: computer simulation [132, 161, 294, 356, 392, 485], Markov processes [103, 173, 174, 191, 199, 236, 268, 343, 344, 345, 346, 356, 357, 359, 360, 364, 392, 394, 438, 440, 464, 465, 466, 467, 505, 528, 529], queueing theory [98, 209, 212, 294, 513].

Tactical planning

Admission control. Admission control involves the rules according to which patients are selected to be admitted to residential care services from the waiting lists. Taking into account resource availability, current waiting lists and expected demand, admission control policies have the objective to match demand and supply such that access times are minimized, while resource utilization is maximized. In addition, access times should be equitable among patient groups [322]. Admission control requires patient group identification, which is done by clustering patients with similar pathologies and similar resource requirements [192, 322]. For each of these groups, waiting lists are created. To provide timely access for different urgency classes, patients are typically categorized in several priority groups based on medical urgency and current accommodation [392]. A patient's current accommodation plays a role, since waiting at upstream facilities leads to bed blocking, while waiting at home might lead to added stress on families. Estimating the future transitions between patient groups and urgency classes for both patients already in service and patient on the waiting lists, can support the design of good admission control policies [192, 322].

A possible admission policy is to always take the patient with highest priority into service whenever capacity becomes available. Another option is to develop an admission plan that prescribes how many patients of each patient group are taken into service during a certain planning horizon. The latter has the potential to simultaneously decrease access times and increase resource utilizations on the long run [192, 322]. In [392], a dynamic admission rule is proposed which, under the assumption that the total bed capacity is sufficient, maintains 'alternative level of care' census at hospitals below a certain threshold and maintains access times from home below a certain access time target. For such dynamic rules, a close cooperation with upstream health organizations is required, which might be challenging since reducing hospital bed blocking may not be the primary interest of residential care organizations.

Methods: computer simulation [392], mathematical programming [192, 322, 392].

Offline operational planning

Treatment scheduling. For rehabilitation patients the therapeutic process generally takes several weeks during which multiple treatments with clinicians from different disciplines have to take place. Usually, the treatment requirements are known in advance, at least for a number of weeks, so that the appointments can be scheduled in advance. The treatment is planned in an appointment series, in which appointments may have precedence relations and certain guidelines for the time intervals in between. The goal is to provide treatments at the right time and in the right sequence, while resource utilization is maximized. The quality of the schedules is highly important for the medical effectiveness and the economic efficiency of rehabilitation centers [430]. Since the amount of variables is tremendous, treatment scheduling for a complete rehabilitation center is highly complex. In [430], it is claimed that if scheduling is done by hand, it is generally done on a patient-by-patient or even appointment-for-appointment basis. Therefore, decision support tools based on OR/MS are considered as indispensable to achieve high-quality treatment scheduling.

Methods: mathematical programming [430].

3.10 Discussion

This chapter has introduced a taxonomy to identify, break down and classify decisions to be made in the managerial field of healthcare resource capacity planning and control. It has provided a structured overview of the planning decisions in six identified care services and the corresponding state of the art in OR/MS literature. Having done this, we aim for an impact that is threefold. First, we aim to support healthcare professionals in improved decision making. Second, we aim to inspire OR/MS researchers in formulating future research objectives and to position their research in a hierarchical framework. Third, we aim for interconnecting healthcare professionals and OR/MS researchers so that the potential of OR/MS can be discovered and exploited in improving healthcare delivery performance.

The vertical axis in our taxonomy represents the hierarchical nature of decision making in healthcare organizations. Aggregate decisions are made in an early stage, and finer granularity is added in later stages when more detailed information becomes available. The observed literature explicitly substantiates the relations between planning decisions both within and between hierarchical levels. Planning decisions on higher levels shape decision making on lower levels by imposing restrictions. Performance on lower levels concerns feedback about the realization of higher level objectives, thereby potentially impacting decision making on higher levels. We have seen many examples of these interactions in our review. Incorporating flexibility in planning reduces restrictions imposed by decisions settled in higher levels on lower level decision making. Increased planning flexibility involves specifying and adjusting planning decisions closer to the time of actual healthcare delivery, thereby giving the opportunity to incorporate more detailed and accurate information in decision making. The observed contributions that incorporate planning flexibility provide opportunities to improve the response to fluctuations in patient demand and thus to improve performance.

Although organized by different organizations, the healthcare delivery process from the patient's perspective generally is a composition of several care services. A patient's pathway typically includes several care stages performed by various healthcare services. The effectiveness and efficiency of healthcare delivery is a result of planning and control decisions made for the care services involved in each care stage. The quality of decisions in each care service depends on the information available from and the restrictions imposed by other care services. Therefore, in the perspective of the presented taxonomy, in addition to the vertical interaction, a strong horizontal interaction can be recognized. Suboptimization is a threat when these decisions are taken in isolation. At various points in our overview, we have observed that taking an integrated approach in decision making is beneficial. Such an integration is not straightforward as it also emerged that different care services may have conflicting objectives. Our categorization of planning decisions in Section 3.3 based on the taxonomy presented in Section 3.2, enables identification of interactions between different care services, allows detection of conflicting objectives, and helps to discover opportunities for coordinated decision making.

Due to the segmented organizational structure of healthcare delivery, also the OR/MS literature has initially focused predominantly on autonomous, isolated decision making. Such models ignore the inherent complex interactions between decisions for different services in different organizations and departments. In 1999, the survey [282] identified a void in OR/MS literature focusing on integrated healthcare systems. The level of complexity of such models is depicted as main barrier. In 2010, the survey [491], reviewing OR/MS models that encompass patient flows across multiple departments, addressed the question whether this void has since been filled. The authors conclude that the lack of models for complete healthcare processes still existed. Although a body of literature focusing on two-departmental interactions was identified, very few contributions were found on complete hospital interactions, let alone on complete healthcare system interactions. The present review of the literature confirms these observations.

To conclude, the specification of planning decisions in our taxonomy allows for identifying relations within and between hierarchical levels. Recognizing and incorporating these relationships in decision making improves healthcare delivery performance. Creating more planning flexibility in decision making demonstrates great potential. By specifying and adjusting planning decisions closer to the time of actual healthcare delivery, more detailed and accurate information can be incorporated, providing opportunities to adjust planning decisions to better match care supply and demand. Furthermore, integrated decision making for multiple care services along a care chain shows great potential. With the growing awareness of the potential benefit of such integrated decision making, an increase in the number of publications in which integrated models are presented is noticeable [84, 491]. However, it remains a challenge for OR/MS researchers to develop integral models that on the one hand provide an extensive healthcare system scope, while on the other hand incorporating a satisfactory level of detail and insight. Summarizing, for the sake of patient centeredness and cost reductions required by societal voices and pressures, we claim that both healthcare professionals and OR/MS researchers should address coordinated and integrated decision making for interrelated planning decisions, should explore the opportunities of increased flexibility, and should take an integral care chain perspective.

3.11 Appendix

3.11.1 Descriptions of the OR/MS methods

OR/MS method	*Description*
Computer simulation	Technique to imitate the operation of a real-world system as it evolves over time by developing a "simulation model". A simulation model usually takes the form of a set of assumptions about the operation of the system, expressed as mathematical or logical relations between the objects of interest in the system [319, 520].
Heuristics	Systematic methods to optimize a problem by creating and/or iteratively improving a candidate solution. Heuristics are used when exact approaches take too much computation time. They do not guarantee an optimal solution is found [1, 520].
Markov processes	Mathematical models for the random evolution of a system satisfying the so-called Markov property: given the present (state of stochastic process), the future (evolution of the process) is independent of the past (evolution of the process) [473, 521].
Mathematical programming	Optimization models consisting of an objective function, representing a reward to be maximized or a (penalty) cost to be minimized, and a set of constraints that circumscribe the decision variables [275, 388, 434].
Queueing theory	Mathematical methods to model and analyze congestion and delays at service facilities, by specifying the arrival and departure processes for each of the queues of a system [424, 521].

3.11.2 Search terms to identify the literature base set

Care service	*Search terms*
Ambulatory care services	“outpatient clinic$” OR “outpatient facilit*” OR “outpatient care” OR “ambulatory care” OR “ambulatory health center$” OR “diagnostic service$” OR “diagnostic facilit*” OR “radiology” OR “primary care” OR “general practi*” OR “community service$”
Emergency care services	“emergenc*” OR “acute” OR “accident” OR “ambulance” AND “health”
Surgical care services	“operating room$” OR “operating theat*” OR “surgery scheduling” OR “operating suite” OR “surgical” OR “surger*”
Inpatient care services	“bed$” OR “intensive care” OR “ward$” AND “hospital”
Residential care services	“nursing home$” OR “mental care” OR “rehabilitation cent*” OR “rehabilitation care” OR “long-term care" OR (“retirement” OR “geriatric” OR “elderly” AND "health")
Home care services	“home care” OR “home health care” OR “home-care” OR “home-health-care” OR “home-health care” OR “home healthcare”

A search engine can replace ‘$’ by any one character, but can also leave it empty. A search engine can replace ‘*’ by any one or multiple characters, but can also leave it empty.

3.11.3 Overview tables of the identified planning decisions

This appendix displays the overview tables of the identified planning decisions and the applied OR/MS methods for each of the six care services: ambulatory care, emergency care, surgical care, inpatient care, home care, and residential care.

In the overview tables, the following acronyms are used when referring to the methods:

Method	*Abbreviation*
Computer simulation	CS
Heuristics	HE
Markov processes	MV
Mathematical programming	MP
Queueing theory	QT
Literature review	LR

Ambulatory care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage	[348, 420, 454, 471]	[2, 153]				[451]
	Service mix						
	Case mix	[458]			[450]		
	Panel size	[454]				[218]	
	Capacity dimensioning:						
	– consultation rooms	[264, 457, 458]				[264]	[282, 451]
	– staff	[348, 421, 454, 457, 458, 515]		[508]	[450]	[35]	[282, 451]
	– consultation time capacity	[160, 162]				[115, 162]	[282, 451]
	– equipment	[187, 348, 471]					[282, 451]
	– waiting room	[458]					[282, 451]
	Facility layout		[387]				
Tactical	Patient routing	[97, 187, 264, 348, 454]				[264, 535]	
	Capacity allocation	[498]		[227, 450]			[503]
	Temporary capacity change	[162, 498]					
	Access policy	[12, 178, 333, 390, 417, 496]	[333]	[390]		[419, 535]	
	Admission control	[498]	[203]	[195, 203]	[280, 408, 409]		
	Appointment scheduling	[14, 85, 90, 133, 160, 178, 179, 239, 254, 255, 288, 303, 323, 332, 333, 348, 380, 417, 458, 503, 504, 514, 518]	[85, 283, 333]	[184, 219, 283, 297, 329, 369, 453]	[27, 85, 129, 418]	[58, 115, 151, 299, 320, 418, 503, 535]	[89, 229, 282, 451]
	Staff-shift scheduling	[395]			[71]		[74, 166, 231, 387]
Offline operational	Patient-to-appointment assignment:						
	– single appointment		[100, 509]	[230, 391, 509]			
	– combination appointments		[397]				
	– appointment series				[109, 110, 111]		
	Staff-to-shift assignment				[280]		[231]
Online operational	Dynamic patient (re)assignment	[411]		[120, 219, 329]	[120]		
	Staff rescheduling						

Emergency care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage						
	– emergency care centers	[62]	[36]	[21, 257]	[79, 207, 257, 276, 413, 475]	[36]	[214, 276, 328, 399, 413]
	– ambulances	[62, 165, 180, 186, 204, 237, 267, 412, 429, 459, 526]	[22, 36, 38, 164, 197, 266]		[18, 36, 38, 39, 40, 45, 118, 158, 165, 186, 205, 237, 266, 276, 412, 413, 446, 459]	[36, 197, 266, 314, 339, 446]	[67, 214, 276, 328, 399, 413]
	Service mix						
	Ambulance districting	[204, 429]	[36]		[36]	[36, 86, 314]	
	Capacity dimensioning:						
	– ambulances	[40, 165, 186, 267, 412, 429, 526]	[38]		[38, 39, 165, 412]	[446, 468]	
	– waiting room					[106]	[393]
	– treatment rooms	[91, 311]				[106]	[393]
	– emergency wards	[16, 315, 375]			[375, 376]	[106]	
	– equipment	[91]				[106]	[393]
	– staff	[60, 181, 311, 375, 532]			[375, 376]	[220]	[62, 282, 393]
	Facility layout	[532]	[387]				[393]
Tactical	Patient routing	[60, 91, 181, 310, 349, 494]				[106, 352]	[282, 393]
	Admission control	[60, 91]				[352]	
	Staff-shift scheduling	[267, 447, 448, 532]	[447, 448]			[215, 216, 220]	[231, 282, 393]
Offline operational	Staff-to-shift assignment		[88]		[20, 25, 88, 124, 164]		
Online operational	Ambulance dispatching	[8, 321, 330, 526]	[321]		[330]	[468]	
	Facility selection	[429]					
	Ambulance routing						
	Ambulance relocation	[8, 194, 526]		[350, 431]	[194]		[67]
	Treatment planning and prioritization	[91, 181]					
	Staff rescheduling	[532]			[375]		

Surgical care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage	[56]			[428]		
	Service mix						
	Case mix	[281]			[50, 260]		[224]
	Capacity dimensioning:						
	– operating rooms	[433]			[23]		[282]
	– operating time capacity	[281, 334, 432, 490]			[469]	[334]	[351]
	– presurgical rooms						
	– recovery wards	[300, 301, 302, 432, 433]					[282]
	– ambulatory surgical ward						
	– equipment						
	– staff		[72, 130, 256]		[72, 130]		[282]
	Facility layout	[340]	[387]				[351]
Tactical	Patient routing	[340]	[13]		[13, 398]		[224, 351]
	Capacity allocation	[57, 140, 143, 144, 307, 396, 533]	[31, 32, 33, 462, 501]	[196, 492, 493, 536]	[31, 32, 33, 51, 52, 95, 130, 231, 307, 403, 428, 462, 469, 470, 487, 488, 533]	[536]	[49, 84, 224, 228, 282, 336, 387, 491, 503, 506]
	Temporary capacity change	[140]			[52, 130, 469]		[228, 231, 506]
	Unused capacity (re)allocation	[140, 148]	[148]	[250]			[228]
	Admission control	[57, 138, 291, 478]		[365]	[3, 4]		[49, 224]
	Staff-shift scheduling		[135]		[32, 71, 150]		[231, 404]
Offline operational	Staff-to-shift assignment						
	Surgical case scheduling	[10, 57, 102, 137, 140, 142, 143, 146, 170, 226, 300, 302, 306, 307, 433, 470, 511]	[10, 13, 83, 131, 136, 175, 177, 225, 226, 256, 306, 308, 341, 416, 422, 455, 486]	[196, 228, 365, 381]	[13, 23, 81, 82, 83, 95, 102, 129, 130, 131, 175, 176, 177, 225, 274, 306, 307, 308, 338, 341, 396, 398, 416, 422, 444, 469]	[511]	[49, 84, 229, 336, 351, 387, 451]
Online operational	Emergency case scheduling				[141, 398]		[229]
	Surgical case rescheduling				[3, 338]		[228, 229]
	Staff rescheduling						

Inpatient care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage	[243]			[62, 426]	[47]	
	Service mix						
	Case mix	[221]	[26, 482]				
	Care unit partitioning	[156, 157, 206, 243, 259, 439]	[26, 309, 482]		[376]	[217, 251, 355, 359, 360, 480, 522]	
	Capacity dimensioning:						
	– beds	[7, 19, 104, 108, 112, 156, 157, 206, 221, 238, 242, 243, 244, 259, 292, 295, 347, 375, 371, 372, 373, 415, 439, 478, 497, 517, 519, 523, 524]		[7, 68, 172, 192, 199, 245, 246, 247, 335],	[202, 327, 375, 376]	[11, 29, 66, 104, 105, 123, 152, 188, 208, 217, 240, 251, 258, 292, 327, 415, 480, 484]	
	–equipment						
	–staff	[222, 241, 375]	[309]	[127, 202, 241, 318, 375]			[74, 166, 399, 451]
	Facility layout	[77]	[387]		[77]		
Tactical	Bed reallocation	[242, 284]	[26, 501]		[6]	[284]	
	Temp. bed capacity change	[238, 244, 347, 523]	[26]			[217, 522]	
	Admission control:						
	– static bed reservation	[293, 347, 415, 478]		[46, 167]		[189, 270, 285, 355, 443, 481]	
	– dynamic bed reservation	[293]		[249, 296]	[30]	[30]	
	– overflow rules	[210, 243, 347, 439]		[251]			
	– influence surgical schedule	[3, 156, 170, 210, 238, 291, 293, 469, 502, 524]	[26, 501]	[492, 493]	[3, 4, 469]	[152, 217]	
	Staff-shift scheduling	[222]	[309]		[32, 127, 507, 525]	[463]	[74, 166, 231, 289, 399, 451]
Offline	Admission scheduling				[111]		
	Patient-to-bed assignment		[94, 128]	[94, 128]			
	Discharge planning	[495]				[513]	[441]
	Staff-to-shift assignment		[44, 483]		[44, 273, 423, 483, 525]		[74, 99, 166, 289, 399, 451]

Level	Planning decision	CS	HE	MV	MP	QT	LR
Online operational	Elective adm. rescheduling	[292]	[331]			[292, 442]	
	Acute admission handling	[292]				[292, 442]	
	Staff rescheduling	[222]			[406]		[74, 399, 451]
	Nurse-to-patient assignment	[456]	[367]		[367, 406]		
	Transfer scheduling			[472]			

Home care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Placement policy		[530]	[313]	[96]		[37]
	Regional coverage						[37]
	Service mix						[37]
	Case mix						[37]
	Panel size				[121]		
	Districting		[48]				[37]
	Capacity dimensioning:						
	– staff	[485]		[236]		[75]	[37]
	– equipment	[389]					[37]
	– fleet vehicles						[37]
Tactical	Capacity allocation:						
	– patient group identification					[75]	[37]
	– time subdivision		[53, 305]		[121]		
	Admission control			[313]	[121, 252]	[75]	
	Staff-shift scheduling		[305]				[37]
Offline operational	Assessment and intake		[449, 530]	[313]	[96, 169, 252]		[37]
	Staff-to-shift assignment		[41, 169]		[41, 169]		[37]
	Visit scheduling:						
	– short-term care plan		[28, 168, 169]	[313]	[168, 169, 252]		[37]
	– staff-to-visit assignment		[28, 41, 168, 169]	[313]	[41, 65, 168, 169, 252]		[37]
	– route creation		[28, 41, 63, 64, 168, 169]		[41, 63, 64, 65, 96, 168, 169]		[37]
Online operational	Visit rescheduling		[28, 168, 169, 477]		[28, 168, 169, 477]		

Residential care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Placement policy	[356, 485]	[24, 530]	[103, 173, 174, 191, 236, 268, 343, 344, 345, 346, 356, 357, 358, 359, 360, 364, 394, 438, 440, 464, 465, 466, 467, 505, 528, 529]		[212]	
	Regional coverage				[62, 114, 154]		[399]
	Case mix		[185]				
	Capacity dimensioning:						
	– beds	[132, 161, 294, 356, 392, 485]		[103, 173, 174, 191, 199, 236, 268, 343, 344, 345, 346, 356, 357, 359, 360, 364, 392, 394, 438, 440, 464, 465, 466, 467, 505, 528, 529]		[98, 209, 212, 294, 513]	
	– staff	[132]					
Tactical	Admission control	[392]			[192, 322, 392]		
Offline operational	Treatment scheduling				[430]		

Recall that since the literature on residential care services showed a low variety in addressed planning decisions, we have chosen for residential care services, as opposed to the other care services, to only present planning decisions for which we found references (see Section 3.9).

CHAPTER 4

Tactical planning in care processes with an ILP approach

4.1 Introduction

Tactical planning is a key element of hospital planning and control that concerns the intermediate term allocation of resource capacities and elective patient admission planning. The main objectives are to achieve equitable access and treatment duration for patient groups, to serve the strategically agreed target number of patients (i.e., production targets or quota), to maximize resource utilization and to balance workload.

This research was inspired by many hospitals in the Netherlands. The hospitals we cooperate with have the aim to provide equitable access and treatment duration for patient groups by controlling access times. Access time is the time a patient spends on the waiting list before being served, and controlled access times ensure quality of care for the patient and prevents patients from seeking treatment elsewhere [531]. Access time is incurred at each care stage in a patient's treatment at the hospital, for example before an outpatient clinic visit and before surgery. Also, in some reimbursement systems, hospitals receive payments only after patients have completed their healthcare process. Hence, it can be costly for hospitals when patients have to wait, as resources and materials have already been invested, but revenues are still to come. Furthermore, hospital management may have agreed with insurers or government to serve a target number of patients. Therefore, evaluation and control of the number of patients served helps to ensure that strategic objectives are being reached.

From a clinician's perspective, tactical resource and admission plans break the clinician's time down over separate activities (e.g., consultation time and surgical time) and determine the number of patients to serve from a particular patient group at a particular stage of their care process (e.g, consultation or surgery). We use the term care process in this article to identify a chain of care stages for a patient. These care stages constitute a patient's logistical treatment path. For example, a care process may comprise a visit to an outpatient clinic, a surgery and a revisit to the outpatient clinic. Because care processes connect multiple departments and resources into a network, fluctuations in both patient arrivals (e.g., seasonality) and resource availability (e.g., holidays) result in bullwhip effects in the care chain [427]. From a patient's perspective, this means access times for each separate stage in a care process strongly fluctuate. From a hospital's perspective, this means that resource utilizations and service levels fluctuate. To cope with these fluctuations, intermediate-term re-allocation of hospital resources, taking into account a care chain perspective [80, 231, 401], is required. For example, only optimizing the outpatient clinic capacity may lead to waiting times and congestion downstream at the operating rooms. Likewise, optimizing operating room utilization without considering admission planning in the outpatient clinic may lead to underutilized operating room capacity.

Typically, tactical planning is done for a subset of care processes in a hospital (e.g., one specialty, a subset of specialties), and not for the entire hospital. Tactical planning problems observed at the hospitals we cooperate with typically comprise 6-10 care processes, 4-8 weeks as a planning horizon, and 1-3 resource types. In these hospitals, tactical planning is organized around a biweekly meeting with decision makers involved in developing the tactical plans. These meetings are used to develop and adjust the tactical resource and admission plans for future time periods, based on information subtracted from the hospital information system about waiting lists, resource availability, expected demand and the number of patients served in prior periods. In this way, tactical resource and admission plans are developed in response to anticipated changes in demand or supply on the mid-term, which leads to improved utilization, shorter access times and improved control of the number of patients served. Currently, the decision makers are using spreadsheet solutions to base their tactical resource and admission plans on. Our model provides an optimization step that supports rational decision making in tactical planning. The model can be used to propose a tactical resource and admission plan to the decision makers, and to evaluate particular scenarios with regards to foreseen resource (un)availability, a proposed change in access time targets, expected demand surges, etc.

The available approaches on the development of tactical resource and admission plans in the Operations Research and Management Science literature are myopic, focus on developing long-term cyclical plans, or are not able to provide a solution for real-life sized instances; see Section 4.2 for details. Our aim is to provide a theoretical contribution to the development of tactical resource and admission plans in healthcare. This chapter presents a method to determine intermediate term tactical resource and admission plans to cope with fluctuations in patient arrivals and resource availability. These plans are developed for multiple resources and multiple patient groups with various care processes, thereby integrating decision making for a chain of hospital resources. This chapter is not about developing clinical care pathways, as described in [171], but about methods for the logistical coordination (i.e., allocation and planning) of resource capacities in patient care processes. Clinical care pathways may be used to identify the patient care processes, but the identification of patient care processes is not the main focus of this chapter.

The resource capacity and admission plans are provided for each stage in the care process, this includes for example the outpatient clinic and the operating theater. The method incorporates available knowledge about the state of the waiting lists and the available resource capacities. To test our method, we use it to develop tactical resource and admission plans for generated instances that are inspired by practical problem instances provided by the hospitals we cooperate with. Computational results show that our method can be used to develop tactical resource and admission plans for real-life sized instances and that it improves compliance with strategically set targets for access times, care process duration and the number of patients served. The presented method can also be used to develop tactical plans in other service industries and in manufacturing. However, we restrict the presentation of the model and results in the terms of healthcare.

This chapter is organized as follows. Section 4.2 discusses tactical resource and admission planning in healthcare and industry. Section 4.3 presents our method for tactical resource and admission planning. Section 4.4 discusses our approach to generate instances, based on examples from practice, that are used to run computational experiments. Section 4.5 presents the results of these computational experiments, and Section 4.6 discusses the managerial and practical implications of developing tactical resource and admission plans. Section 4.7 concludes this chapter.

4.2 Background

Due to increasing demand for healthcare and increasing expenditures [385], healthcare organizations are trying to organize processes more efficiently and effectively. Planning and control in healthcare has received an increased amount of attention over the last ten years [61], both in practice and in the literature. Healthcare planning and control can be subdivided in the hierarchical levels of strategic, tactical and operational planning [9, 49, 76, 234]. While strategic planning addresses the dimensioning of resource capacities, tactical planning subdivides the settled resource capacities among patient groups (e.g., identified by specialty) to reach strategically set targets and to facilitate operational planning, and operational planning involves the short-term decision making related to the execution of the healthcare delivery process. In this section, we discuss approaches in the literature for tactical planning in healthcare and in industry.

Tactical resource and admission planning approaches are static or dynamic. Static approaches result in long-term plans that are often cyclical. Dynamic approaches result in intermediate-term plans in response to the variability in demand and supply. These approaches are compared in [498], and their simulation results indicate that the dynamic approach results in lower access times and higher resource utilization.

Tactical resource and admission planning approaches in healthcare are often myopic, which means that they do not consider multiple departments and resources along a care process for patient groups. For example, they focus on the outpatient clinic [89, 162], diagnostic services [219, 498] or operating rooms [4, 31, 93, 130, 488]. Although the benefits of an integrated approach are often recognized [80, 231, 401], relatively few articles integrate decision making for a chain of resources or departments along the patient's care process. To support integrated tactical resource and admission planning, [286] models care processes as Markov chains to derive resource requirements for each stage of a patient's care process. Similar approaches for evaluation of resource requirements are taken in [113, 192, 251, 512]. In order to calculate optimal static, elective patient admission plans for multiple resources and multiple patient groups with various care processes, [374] models the patient process as a Markov Decision Process (MDP). Their experiments show that alternative methods to solve the model should be developed, as the MDP approach is not yet suitable for realistically sized instances.

The process of patients flowing through a network of service units can be compared to a classical job shop in industry [193], which is a network of work stations capable of producing a wide variety of jobs [211]. Hence, methods used in industry for job shop scheduling may be suitable for tactical resource and admission planning in healthcare. Job shop scheduling is applied to surgical care services in [256, 398]. In these articles, mathematical programming is used to allocate surgical resources and to schedule surgical patients. Queueing models can also be used to analyze tactical production plans for a job shop [211, 272]. However, results in queueing theory are often based on steady state assumptions, and therefore, queueing models are not suitable to analyze dynamic plans with a finite planning horizon. Other methods to analyze a network of workstations in industry are in the field of project scheduling. Project scheduling is concerned with small batch production where resources are allocated to production activities over time [69, 70]. Methods to allocate resources to activities in project scheduling are often based on mathematical programming [232, 527, 534] or MDP [43].

Summarizing, existing approaches to tactical resource and admission planning in healthcare are myopic, focus on developing long-term cyclical plans, or do not provide a solution for real-life sized instances. In Section 4.3, we propose a method to develop tactical resource and admission plans on the intermediate term, for multiple resources and multiple care processes.

4.3 Model description

We aim to allocate resource capacities among the various consecutive stages of different care processes. To this end, we propose a Mixed Integer Linear Program (MILP) to compute a patient admission plan for multiple consecutive time periods. Section 4.3.1 provides the constraints to model tactical resource and admission planning. We present our approach to the objective function in Section 4.3.2. In our objective function, a weight reflects the priority to serve patients at a particular stage in a particular care process. The determination of these weights is discussed in Section 4.3.3. Tactical resource and admission planning has multiple objectives. Healthcare organizations may prioritize these objectives differently, resulting in multiple possible objective functions. Hence, we discuss the performance measures that can be calculated within the MILP to form alternative objective functions in Section 4.3.4. In addition, we also discuss other extensions of the model in Section 4.3.4. In the following, we introduce notation and discuss the problem in more detail.

The planning horizon is discretized in consecutive time periods $\mathcal{T} = \{0, 1, 2, \ldots, T\}$. Furthermore, we consider a set of resource types $\mathcal{R} = \{1, 2, \ldots, R\}$ and a set of patient care processes $\mathcal{G} = \{1, 2, \ldots, G\}$. The number of patients that can be served by resource $r \in \mathcal{R}$ is limited by the available resource capacity $\eta_{r,t}$ in time period $t \in \mathcal{T}$. Formally, patients that follow patient care process $g \in \mathcal{G}$ receive care specified by a set of stages $\mathcal{K}_g = \{(g, 1), (g, 2), \ldots, (g, e_g)\}$, where e_g is the number of stages of the care process. Patients following the same care process have the same resource requirements in each stage (e.g., consultation times, surgery duration, number of consultations), and to serve a patient of care process $g \in \mathcal{G}$ in stage $j = (g, a)$ requires $s_{j,r}$ time units of resource $r \in \mathcal{R}$. After service at a certain stage, patients move to the next stage of their care process or leave the system. More precisely, for two stages $i, j \in \mathcal{K}_g$ the value $q_{i,j}$ denotes the fraction of patients that move from stage i to stage j, and the value $1 - \sum_{j \in \mathcal{K}_g} q_{i,j}$ denotes the fraction of patients that leave the system. At each stage, patients may have to queue for service. Hence for each care process, we obtain a set of queues $\mathcal{J}_g$ of cardinality e_g. Although patients of different care processes share resources for service, we model the queues disjointly for different care processes. Consequently, we have a total set of queues $\mathcal{J} = \bigcup\limits_{g \in \mathcal{G}} \mathcal{J}_g$, where $|\mathcal{J}| = \sum_{g \in \mathcal{G}} e_g$.

For each time period $t \in \mathcal{T}$, we determine a patient admission plan, characterized by the decision variable vectors $x_{t,j} = (x_{t,j,0}, x_{t,j,1}, \ldots)$. The decision variable $x_{t,j,u}$ indicates the number of patients to serve in time period $t \in \mathcal{T}$ that have been waiting precisely u time periods at queue $j \in \mathcal{J}$. In order to calculate the decision variables $x_{t,j,u}$, we evaluate for each queue $j \in \mathcal{J}$ the number of patients that are waiting and the time that these patients are waiting. Therefore, we introduce waiting lists $S_{t,j} = (S_{t,j,0}, S_{t,j,1}, \ldots)$, where $S_{t,j,u}$ gives the number of patients that have been waiting precisely u time periods at queue $j \in \mathcal{J}$ at the beginning of time period $t \in \mathcal{T}$. When patients in waiting list entry $S_{t,j,u}$ are not served in time period t, they move to the entry $S_{t+1,j,u+1}$ in period $t + 1$. Figure 4.1 illustrates the dynamics of the waiting list for a single queue.

For ease of notation we summarize the transition rates between the stages/queues in a routing matrix Q of dimension $|\mathcal{J}| \times |\mathcal{J}|$. Furthermore, to be able to take into account a minimum (required) time lag before patients that have been served at one queue, can enter the following queue, we introduce a delay matrix D of dimension $|\mathcal{J}| \times |\mathcal{J}|$, where the entry $d_{i,j}$ denotes the minimum time lag (in time periods) between service from queue i and entrance to queue j ($i, j \in \mathcal{J}$). Such deterministic delay $d_{i,j}$ may for example be specified by

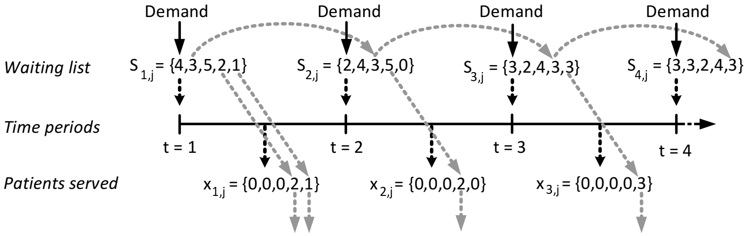


Figure 4.1: The dynamics of the waiting list and patient service for a system with a single queue $j \in \mathcal{J}$.

a doctor, when a given time lag between two stages is medically required in the care process (e.g., such that patients can recover from a procedure). Finally, in addition to demand originating from serving patients from other queues, there is a deterministic demand from outside the system $\lambda_{j,t}$ ($j \in \mathcal{J}, t \in \mathcal{T}$). Together, the number of patients entering queue $j \in \mathcal{J}$ in time period $t \in \mathcal{T}$ is given by:

$$S_{t,j,0} = \lambda_{j,t} + \sum_{i \in \mathcal{J}} \sum_{u=0}^{\infty} q_{i,j} \cdot x_{t-d_{i,j},i,u}. \tag{4.1}$$

Assumptions 4.1 to 4.5 summarize the problem assumptions that underly our modeling approach, which is presented in Section 4.3.1.

Assumption 4.1. *Patient arrivals, delay times, resource requirements and resource capacities are considered to be deterministic and known.*

Assumption 4.2. *All patients arriving to a queue remain in the queue until service completion.*

Assumption 4.3. *All patients in queue j require resources $s_{j,r}$, $r \in \mathcal{R}$.*

Assumption 4.4. *Resource capacity $\eta_{r,t}$ for resource type r and time period t is not transferable from one time period to another time period $s \neq t$; $s, t \in \mathcal{T}$, i.e., when (part of) the resource capacity $\eta_{r,t}$ is unused in time period t, it is 'lost'.*

Assumption 4.5. *Every patient planned according to the decision $x_{t,j}$ will be served in queue j in period t, i.e., there is no deferral to other time periods.*

4.3.1 Constraints to calculate a tactical resource and admission plan

The constraints to model the care processes of patients in the tactical planning problem are given below. Table 4.1 gives the sets, indices, variables and parameters used. Possible extensions are presented in Section 4.3.4.

$$
\begin{aligned}
S_{t,j,0} &= \lambda_{j,t} + \sum_{i \in \mathcal{J}} \sum_{u=0}^{\infty} q_{i,j} \cdot x_{t-d_{i,j},i,u} && \forall j \in \mathcal{J}, t \in \mathcal{T}, && (4.2)\\
S_{t,j,u} &= S_{t-1,j,u-1} - x_{t-1,j,u-1} && \forall j \in \mathcal{J}, t \in \mathcal{T}, u > 0, && (4.3)\\
x_{t,j,u} &\leq S_{t,j,u} && \forall j \in \mathcal{J}, t \in \mathcal{T}, u \geq 0, && (4.4)\\
\sum_{j \in \mathcal{J}^r} s_{j,r} x_{t,j} &\leq \eta_{r,t} && \forall r \in \mathcal{R}, t \in \mathcal{T}, && (4.5)\\
x_{t,j} &= \sum_{u=0}^{\infty} x_{t,j,u} && \forall j \in \mathcal{J}, t \in \mathcal{T}, && (4.6)\\
x_{t,j} &\in \mathbb{N} && \forall j \in \mathcal{J}, t \in \mathcal{T}. && (4.7)
\end{aligned}
$$

Sets		**Indices**	
$\mathcal{J}$	Queues	$i, j \in \mathcal{J}$	Queue
$\mathcal{T}$	Time periods	$t \in \mathcal{T}$	Time period
$\mathcal{R}$	Resource types	$r \in \mathcal{R}$	Resource type
$\mathcal{J}^r$	Queues for resource type r, $\mathcal{J}^r \subseteq \mathcal{J}$	$i, j \in \mathcal{J}^r$	Queue
		u, d	Time periods (to indicate waiting time)
Variables			
Decision variables			
$x_{t,j,u}$	The number of patients served from queue j in time period t, who have been waiting u time periods		
$x_{t,j}$	The total number of patients served from queue j in time period t		
Auxiliary variable			
$S_{t,j,u}$	The number of patients in queue j at the start of time period t, who have been waiting u time periods		
Parameters			
β_j^u	Objective function weight of patients in queue j, who have been waiting u time periods		
$\lambda_{j,t}$	New demand in queue j in time period t		
$\eta_{r,t}$	Capacity of resource type r in time period t in time units		
$q_{i,j}$	Probability that a patient moves from queue i to queue j		
$d_{i,j}$	Number of time periods to move from queue i to queue j		
$s_{j,r}$	Expected capacity requirements from resource type r for a patient in queue j in time units		

Table 4.1: The sets, indices, variables and parameters used.

Constraints (4.2) and (4.3) stipulate that the waiting list variables are consistent. Constraint (4.2) determines the number of patients newly entering a queue. Constraint (4.3) updates the waiting list variables at each time period $t \in \mathcal{T}$. Constraint (4.4) stipulates that not more patients are served than the number of patients on the waiting list. Constraint (4.5) assures that the resource capacity of each resource type $r \in \mathcal{R}$ is sufficient to serve all patients. Constraint (4.6) determines the total number of patients served at a queue in a time period, and Constraint (4.7) is an integrality constraint for the total number of patients served at a queue in a time period.

Remark 4.6. *In Constraint (4.6) the number $x_{t,j}$ of patients that are served from queue j at time period t is calculated. We only require $x_{t,j}$ to be integer and not the entry $x_{t,j,u}$, which expresses the number of patients that are waiting already u time periods. The reason is that the entries $x_{t,j,u}$ are related to $S_{t,j,u}$ by Constraints (4.2) to (4.4), and that the $S_{t,j,u}$ may have noninteger values. Based on the integer constraint on $x_{t,j}$, only 'full' patients are served in the model.*

Remark 4.7. *For numerical purpose, to solve our optimization problem, the number u of time periods that patients are waiting is bounded at some value U. Consequently, Constraints (4.2) to (4.6) require adaptation and a constraint is added to stipulate that the number $S_{t,j,u}$ of patients who are not served in time period $t-1$ and are waiting U time periods, remain on the waiting list in time period t:*

$$S_{t,j,U} = \sum_{m=U-1}^{U} (S_{t-1,j,m} - x_{t-1,j,m}), \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}.$$

4.3.2 Objective function

From our experience with the hospitals we collaborate with, the main objectives of tactical planning are *to achieve equitable access and treatment duration for patient groups* and *to serve the strategically agreed number of patients*. Therefore, we incorporate these two objectives in our objective function (4.8). The other objectives of tactical planning mentioned in Section 4.1; *to maximize resource utilization* and *to balance the workload*, can be captured in alternative objective functions and extensions of the model. We propose starting points for these extensions in Section 4.3.4.

We use the following objective function:

$$\min \sum_{j \in \mathcal{J}} \sum_{u=0}^{\infty} \sum_{t \in \mathcal{T}} \beta_j^u S_{t,j,u}. \tag{4.8}$$

The objective function (4.8) aggregates the weighted number of patients waiting in each queue $j \in \mathcal{J}$ in each time period $t \in \mathcal{T}$. Note that patients appear multiple times in the summation over $S_{t,j,u}$. To illustrate this, consider the following two cases: (1) a served patient may move from $S_{t,j,u}$ to $S_{t+1,i,0}$ (if $q_{ji} > 0$), and (2) a patient that is not served moves from $S_{t,j,u}$ to $S_{t+1,j,u+1}$. If $t, t+1 \in \mathcal{T}$, then the four mentioned waiting lists are in the objective function's summation. Weights β_j^u ($j \in \mathcal{J}$ and $u = 0, 1, 2, \ldots$) are incorporated in the objective function to prioritize the various queues in order to deploy resources where they are most effective. The two objectives, *to achieve equitable access and treatment duration for patient groups* and *to serve the strategically agreed number of patients*, are reflected in these weights. We propose an iterative procedure to determine these weights in Section 4.3.3.

4.3.3 Procedure to determine the weights

The effect of the resource allocation is measured in the MILP's objective. Inspired by the hospitals we collaborate with, we choose to use access time and the number of patients served as performance metrics. The procedure to determine the weights of the objective terms is an iterative one. We initialize the weights, solve the MILP and measure the metrics as we explain below, then update the weights, solve the MILP, etc. In this section we first explain how we measure the performance metrics from the MILP solution, and then explain in detail the iterative procedure of determining the weights.

Access time. As mentioned in Section 4.1, access time is the time a patient spends on the waiting list before being served. The elements $S_{t,j,u}$ in our MILP provide information about the structure of the waiting list for each queue j in each time period t. We get insight into access time by measuring $A^{\alpha}_{t,j}$ from the MILP solution as follows:

$$A^{\alpha}_{t,j} = \min\{u| \sum_{m=0}^{u} S_{t,j,m} > \alpha \sum_{m=0}^{\infty} S_{t,j,m}\}, \qquad j \in \mathcal{J}, t \in \mathcal{T}, \quad (4.9)$$

where α is a given percentile. $A^{\alpha}_{t,j}$ in (4.9) gives the number of periods that the α-th percentile of all patients in a queue j are waiting. In other words, a fraction of $(1-\alpha)$ of all patients in queue j at time period t have been waiting already for at least $A^{\alpha}_{t,j}$ time periods.

Hospital managers aim to control access times by imposing targets $\hat{a}^{\alpha}_{t,j}$ for $A^{\alpha}_{t,j}$. We aim to evaluate the effect of a calculated tactical resource and admission plan on $A^{\alpha}_{t,j}$ in comparison with the target $\hat{a}^{\alpha}_{t,j}$ for each queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$. Hence, we may calculate an access time performance ratio $L^{A}_{\alpha,t,j}$ with:

$$L^{A}_{\alpha,t,j} = \frac{A^{\alpha}_{t,j}}{\hat{a}^{\alpha}_{t,j}}, \qquad j \in \mathcal{J}, t \in \mathcal{T}. \quad (4.10)$$

We use this ratio to evaluate how close to target the performance of the current solution is. For example, if $L^{A}_{\alpha,t,j} > 1$, then $A^{\alpha}_{t,j}$ is above target.

The number of patients served. Healthcare managers aim to control the number $x_{t,j}$ of patients served by imposing a target $\hat{x}_{t,j}$ for the number of patients served. We assume that this target $\hat{x}_{t,j}$ is given for each queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$. In practice, targets may typically be set for care processes, by setting the target for either the first or the last queue in care processes. In our model, we assume that these care process targets can be converted to targets for each stage of a care process.

We aim to evaluate the effect of a calculated tactical resource and admission plan on the number $x_{t,j}$ of patients served in comparison with the target number $\hat{x}_{t,j}$ of patients served for each queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$. Hence, we may calculate a performance ratio $L^C_{t,j}$ for the number of patients served by:

$$L^C_{t,j} = \frac{\hat{x}_{t,j}}{x_{t,j}}, \qquad j \in \mathcal{J}, t \in \mathcal{T}. \tag{4.11}$$

We use the performance ratios (4.10) and (4.11) in the procedure to calculate the weights, which we explain below. The nonnegative weights β^u_j, where $j \in \mathcal{J}$ and $u = 0, 1, 2, \ldots$ indicate the number of time periods waiting, lead to a matrix B. Two assumptions are made regarding the structure of B.

Assumption 4.8. *$\beta^u_j < \beta^{u+1}_j$, for all $j \in \mathcal{J}$ and $u = 0, 1, 2 \ldots$*

Assumption 4.9. *If $q_{i,j} > 0$, then $\max_u \beta^u_i > \min_u \beta^u_j$, for all $i, j \in \mathcal{J}$*

Remark 4.10. *Under Assumption 4.8, $\min_u \beta^u_j = \beta^0_j$, for all $i, j \in \mathcal{J}$*

In the following, we justify these assumptions from a theoretical and practical point of view.

1. When patients are served first-come, first-served (FCFS) at queue $j \in \mathcal{J}$, we want the MILP to have the incentive to first serve the patient who has waited the longest in queue j. This FCFS property leads to monotonically increasing weights β^u_j for each queue $j \in \mathcal{J}$,
2. If a patient moves with positive probability from queue i to queue j ($i, j \in \mathcal{J}$), there is a local incentive to serve the patient at queue i when the maximum weight in row i is larger than the minimum weight in row j. If B is not structured in this way, then even with an infinite resource capacity at queue i, locally there is no incentive to serve a patient at queue i.

We propose the following function to determine B:

$$\beta^u_j = \begin{cases} 0, & \text{if } u = 0, \\ v_j \cdot (m_j)^u, & \text{if } u > 0, \end{cases} \qquad \forall j \in \mathcal{J}. \tag{4.12}$$

This function requires two parameters v_j and m_j per queue $j \in \mathcal{J}$ to determine B. By restricting the parameter m_j to values larger than 1, we satisfy Assumption 4.8. Following Remark 4.10, by setting $\beta^0_j = 0$, we ensure that Assumption 4.9 holds.

Taking into account Assumptions 4.8 and 4.9, the weights in B can be determined with various approaches. For example, one may manually decide on the weights, based on numerous performance measures and perhaps other quantifiable or subjective reasons. These performance measures can be patient oriented, such as access time, medical urgency and pain experience, and organization oriented, such as financial incentives and agreements with insurance companies about the number of patients to serve. In this chapter, we propose to calculate the weights in an iterative manner as follows. First, B is initialized with starting values and the MILP is solved. After that, B is updated based on the solution of the MILP, and the MILP is solved again with the updated B. This iterative way of updating B and solving the MILP is performed until some criterion is met. To design such an iterative procedure, three topics need to be addressed:

1. The initialization of B.

2. The adaptation of B after solving the MILP.

3. The stopping criterion.

The following iterative procedure is used to initialize and update B. In B there are at most $|\mathcal{J}| \times (|\mathcal{U}| + 1)$ elements that require initializing and updating. By using (4.12), we need to adjust at most $2 \times |\mathcal{J}|$ parameters every iteration. The iterative procedure uses the performance ratios $L^A_{\alpha,t,j}$ and $L^C_{t,j}$ to update B by determining new values for v_j and m_j for each queue $j \in \mathcal{J}$. First, the parameters v_j and m_j are initialized by evaluating the performance ratios in previous planning period. Consequently, the performance prior to the planning period influences decision making in the planning period. When no historical data is available, the parameters are assumed to be $v_j = 1$ and $m_j = 1 + \epsilon$, where ϵ is a small number. The weights β^u_j corresponding to the chosen values v_j and m_j are calculated with (4.12) and the MILP is solved. Based on the MILP solution, the parameters v_j and m_j are updated using the performance ratios for this planning period. To avoid strong oscillations of the outcome for the performance ratios over the course of the planning period, we ensure that the number of changes of the parameters gets smaller with increasing number of iterations.

In the following, we formalize the iterative procedure to update B. The iteration number is indicated by n.

Step 1: $n := 1$. Initialize v_j and m_j, for all $j \in \mathcal{J}$, with:

$$v_j(1) = \frac{\hat{x}_{0,j}}{x_{0,j}}, \qquad m_j(1) = 1 + \frac{A^{\alpha}_{1,j}}{\hat{a}^{\alpha}_{1,j}}, \qquad \forall j \in \mathcal{J}, \tag{4.13}$$

where $x_{0,j}$ is the number of patients served from queue $j \in \mathcal{J}$ in the data history, for example the previous planning period. $A^{\alpha}_{1,j}$ gives the evaluation of (4.9) at the start of the planning period. If no history is available, then $v_j(1) = 1$ and $m_j(1) = 1 + \epsilon$, where ϵ is a small number.

Step 2: Determine β^u_j, for all $j \in \mathcal{J}$ and $u = 0, 1, 2, \ldots$, with (4.12). Solve the MILP with the obtained B.

Step 3: $n := n + 1$. Update $v_j(n)$ and $m_j(n)$, for all $j \in \mathcal{J}$, with

$$v_j(n) := \max\{0 + \epsilon, v_j(n-1) + \frac{1}{n}\left(\frac{\sum\limits_{l=0}^{|\mathcal{T}-1} \omega_l \hat{x}_{l,j}}{\sum\limits_{l=0}^{\mathcal{T}-1} \omega_l x_{l,j}} - 1\right)\}, \quad \forall j \in \mathcal{J}, \tag{4.14}$$

$$m_j(n) := \max\{1 + \epsilon, m_j(n-1) + \frac{1}{n}\left(\frac{\sum\limits_{l=1}^{T} \omega_l A^{\alpha}_{l,j}}{\sum\limits_{l=1}^{T} \omega_l \hat{a}^{\alpha}_{l,j}} - 1\right)\}, \quad \forall j \in \mathcal{J}, \tag{4.15}$$

where ω_t are weights for different time periods $t \in \mathcal{T}$.

In (4.14) and (4.15), we subtract 1 from the performance ratio outcome. When the subtraction results in a negative value, queue $j \in \mathcal{J}$ is *overperforming*, i.e., more resource capacities than required are allocated to this queue. This overperformance is mitigated by decreasing the parameters $v_j(n)$ and $m_j(n)$ in (4.14) and (4.15), which causes the weights β^u_j for $u = 0, 1, \ldots$ and queue $j \in \mathcal{J}$ to decrease. This may decrease the allocated resource capacity to this queue, for example when the involved resource capacity can be used to improve performance in other queues. Conversely, when a positive number is the result of subtracting 1 from the performance ratios, queue $j \in \mathcal{J}$ is *underperforming*, and the parameters $v_j(n)$ and $m_j(n)$ are increased. This results in increased weights β^u_j for $u = 0, 1, \ldots$ and queue $j \in \mathcal{J}$, which may increase the resource capacities that are allocated to queue $j \in \mathcal{J}$ to increase performance for queue j. By summing over all time periods in (4.14) and (4.15), we take into account performance over all time periods.

The weights ω_t can be used to emphasize results in particular time periods. For example by letting ω_t increase with t, one emphasizes the results that are obtained later in the planning period. Of course, the objective of these weights ω_t should match the application at hand. For example, a rolling horizon approach may not benefit from an emphasis on later time periods, because those later time periods are not actually implemented.

Step 4: If $\max\{|v_j(n) - v_j(n-1)|, |m_j(n) - m_j(n-1)|\} < \theta$, for all $j \in \mathcal{J}$, where θ is a small number, then stop, else repeat Steps 2-4.

The setup of the above iterative procedure is such that it leads to convergence of the weights in B. This follows from the fact that the terms between brackets in (4.14) and (4.15) are bounded. Changes in both $A^{\alpha}_{t,j}$ and $x_{t,j}$ in (4.14) and (4.15) are bounded by the limited availability of resource capacities. Since these terms are bounded, the changes in parameters ($v_j(n) - v_j(n-1)$ and $m_j(n) - m_j(n-1)$) are converging to 0 as they are multiplied by $\frac{1}{n}$ in (4.14) and (4.15). Therefore, the differences $v_j(n) - v_j(n-1)$ and $m_j(n) - m_j(n-1)$

are also converging to 0 in n. Hence, the stopping criterion is met at some n and therefore, the method converges.

In our approach, the calculation of the weights is separated from the MILP. This separation on the one hand prevents that the objective function of the MILP becomes quadratic. On the other hand, it prevents additional constraints in the MILP with regards to the weights. Another advantage of this separation is the clear distinction between calculating the weights based on explicit performance measures and calculating the patient admission plan with the MILP. This distinction provides the opportunity to determine the weights manually or with the described iterative procedure, which can be easily adapted to incorporate additional requirements.

4.3.4 Alternative performance metrics for tactical resource and admission planning and model extensions

Recall from Section 4.1 that the main objectives of tactical planning are *to achieve equitable access and treatment duration for patient groups, to serve the strategically agreed target number of patients, to maximize resource utilization and to balance workload*. The priority given to different objectives of tactical planning may vary between hospitals and their particular environments. Hence, the model can be adapted and extended in various ways. In this section, we present performance measures that can be used to define alternative objective functions or to initialize and update the weights in the iterative procedure described in Section 4.3.3. We also show how these performance measures can be obtained from the solution of the modeled MILP.

Achieving equitable access and treatment duration for patient groups

- *Number of patients waiting longer than a norm.* The number of patients that wait longer than a certain norm $\hat{a}_{t,j}$ is measured as follows:

$$O_{t,j} = \sum_{u=\hat{a}_{t,j}+1}^{\infty} S_{t,j,u}, \qquad j \in \mathcal{J}, t \in \mathcal{T}.$$

  The number of time periods that patients are waiting longer than the norm $\hat{a}_{t,j}$ may be measured as follows:

$$P_{t,j} = \sum_{u=\hat{a}_{t,j}+1}^{\infty} (u - \hat{a}_{t,j}) S_{t,j,u}, \qquad j \in \mathcal{J}, t \in \mathcal{T}.$$

- *Access time.* With (4.9), the measure $A^{\alpha}_{t,j}$ can be calculated for all α. The average $\bar{A}_{t,j}$ for this measure $A^{\alpha}_{t,j}$ may be calculated by:

$$\bar{A}_{t,j} = \frac{\sum\limits_{u=0}^{\infty} u S_{t,j,u}}{\sum\limits_{u=0}^{\infty} S_{t,j,u}}, \qquad j \in \mathcal{J}, t \in \mathcal{T}.$$

- *Access time performance ratio.* $A^{\alpha}_{t,j}$ may be compared to its target $\hat{a}^{\alpha}_{t,j}$ by calculating the access time performance ratio $L^{A}_{\alpha,t,j}$ with (4.10).

- *Total access time of a complete care process.* We get insight in the total access time of a complete care process (i.e., all queues/stages $\mathcal{J}_g$ in the care process) by summing over $A^{\alpha}_{t,j}$ in each stage as follows:

$$H^{\alpha}_{t,g} = \sum_{j \in \mathcal{J}_g} A^{\alpha}_{t,j}, \qquad g \in \mathcal{G}, t \in \mathcal{T}.$$

The average $\bar{H}_{t,g}$ of this measure may be calculated as follows:

$$\bar{H}_{t,g} = \sum_{j \in \mathcal{J}_g} \bar{A}_{t,j}, \qquad g \in \mathcal{G}, t \in \mathcal{T}.$$

- *Access time performance ratio for a care process.* We may get insight in the access time performance ratio for a care process by aggregating the access time performance ratios in a care process's stages as follows:

$$L^{H}_{\alpha,t,g} = \frac{1}{e_g} \sum_{j \in \mathcal{J}_g} L^{A}_{\alpha,t,j}, \qquad g \in \mathcal{G}, t \in \mathcal{T}. \tag{4.16}$$

Serving the strategically agreed number of patients

- *The number of patients served.* The number $x_{t,j}$ of patients served and a target $\hat{x}_{t,j}$ for the number of patients served are discussed in Section 4.3.3.

- *Performance ratio for the number of patients served.* The number $x_{t,j}$ of patients served in comparison with the target number $\hat{x}_{t,j}$ of patients served may be calculated by performance ratio $L^{x}_{t,j}$ (4.11).

Maximizing resource utilization and balancing workload

- *Fraction of resource capacities that are allocated to care processes.* The fraction $\rho_{r,t}$ of resource capacities that are allocated to care processes may be calculated by:

$$\rho_{r,t} = \frac{\sum_{j \in \mathcal{J}^r} s_{j,r} x_{t,j}}{\eta_{r,t}}, \qquad r \in \mathcal{R}, t \in \mathcal{T}.$$

- *Resource allocation to a set $\mathcal{V}^r \subset \mathcal{J}^r$ of queues.* Hospital management may want to keep resource allocation $\gamma_{\mathcal{V}^r,t}$ to, or the number $\mu_{\mathcal{V}^r,t}$ of patients served in, a subset $\mathcal{V}^r \subset \mathcal{J}^r$ of queues consistent between time periods. These measures may be evaluated by:

$$\gamma_{\mathcal{V}^r,t} = \textstyle\sum_{j \in \mathcal{V}^r} s_{j,r} x_{t,j}, \qquad t \in \mathcal{T},$$

$$\mu_{\mathcal{V}^r,t} = \textstyle\sum_{j \in \mathcal{V}^r} x_{t,j}, \qquad t \in \mathcal{T},$$

where $\mathcal{V}^r \subset \mathcal{J}^r$, for $r \in \mathcal{R}$.

In addition to using alternative metrics in the model, there are also various opportunities for extending the model. Four examples of those opportunities are discussed below.

Constraints to limit variation of patient admissions. Our dynamic approach makes it possible to respond appropriately to expected changes in patient demand or resource availability, but it may also result in varying patient admissions between different time periods. If necessary, this variation may be controlled by introducing additional constraints that limit the variation of the number $x_{t,j}$ of patient admissions between time periods $t \in \mathcal{T}$.

Constraints to limit resource allocation to particular queues. Hospital management may want to bound the amount of resource capacities allocated to particular queues. For example, when doctors serve patients at the outpatient clinic and the operating room, a hospital manager may want to limit the capacity the doctor is allocated to the operating room based on operating room availability. To control or to balance the fraction of resource capacity that is allocated to a queue or a set of queues, constraints can be introduced.

Previously scheduled appointments. Previously scheduled appointments may be included in the MILP. A constraint on the decision variables $x_{t,j}$ can ensure that the number of patients admitted at queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$ is larger or equal to the number of already scheduled appointments. The scheduled patients should also be incorporated in the waiting list to ensure feasibility of the MILP with regards to Constraint (4.4). One can also choose to disregard the already scheduled appointments in the MILP by reducing the resource capacity $\eta_{r,t}$ with the capacity required for the already scheduled appointments. Note that by excluding scheduled patients from the model, they are also omitted from the modeled waiting lists $S_{t,j}$ for $j \in \mathcal{J}$ and $t \in \mathcal{T}$.

Evaluation of given admission plans. Hospital management can evaluate the performance of a given patient admission plan, for example a manual or a cyclical plan, by fixing the decision variables $x_{t,j}$ to the number of planned admissions in the given patient admission plan.

4.4 Test approach

The MILP and iterative method described in Section 4.3 are programmed in AIMMS 3.10, which uses ILOG CPLEX 12.1 to solve the MILP. To test our iterative method, we have implemented an instance generator that allows us to produce test instances with various parameter settings, based on examples from hospitals. Section 4.4.1 discusses the instance generator.

4.4.1 Instance generation

This section describes the instance generation procedure. Various parameter settings can be used to influence the test instances that are generated, in order to align these test instances with the examples from practice. In the hospitals we cooperate with, tactical planning is typically done for a subset of care processes in a hospital (e.g., one specialty, a subset of specialties), and not for the entire hospital. Tactical planning problems at the hospitals we cooperate with typically comprise 6-10 care processes ($|\mathcal{G}|$), 4-8 weeks ($|\mathcal{T}|$) and 1-3 resource types ($|\mathcal{R}|$). For example, the care processes of an orthopedic surgery group may comprise hip surgery, shoulder surgery, knee surgery, etc. Care stages in each care process may for example be described by the initial outpatient clinic visit, preanesthesia visit, surgery, and a follow-up outpatient clinic visit. Typical resources that are involved in each care process are for example a clinician, a nurse, and the allocated operating room time.

Table 4.2 lists the parameters that characterize and influence the complexity of the test instances. Some parameters influence problem size (e.g., the length of the planning horizon, the number of patient groups and the number of resource types), while other parameters influence the solution space (e.g., the initial waiting lists and the resource capacities). In our experiments, we do not take into account the delay matrix D, which has limited influence on the problem size and solution space.

The number $|\mathcal{T}|$ of time periods and the number $|\mathcal{J}|$ of queues principally determine the size of the MILP. The number $|\mathcal{J}|$ of queues is determined by the number of care processes and the number of stages in each care process, as $|\mathcal{J}| = \sum\limits_{g\in\mathcal{G}} e_g$.

For every instance, the values for the parameters in Table 4.2 are uniformly drawn from the possible values given in the third column of Table 4.2. We assume that new demand only arrives to the first queue in care processes. We have three sets of values for the service time $s_{j,r}$, since these vary between different services (e.g., consultations, MRI scans and surgeries). The three sets correspond to a low, medium and high service time respectively.

We first generate $x_{0,j}$, i.e., the number of patients served in queue j in the previous planning period. We start by generating $x_{0,j}$ for the first queue in the care process. For all subsequent queues in the care process, we draw $x_{0,j}$ from $\left[0.75 \sum_{i\in\mathcal{J}} q_{i,j} x_{0,i}, 1.25 \sum_{i\in\mathcal{J}} q_{i,j} x_{0,i}\right]$. A similar approach is applied in generating $\hat{x}_{0,j}$ and $\hat{x}_{t,j}$, for all $t \in \mathcal{T}$. We first generate $\hat{x}_{0,j}$ and $\hat{x}_{t,j}$, for all $t \in \mathcal{T}$, for the first queue in the care process. For all subsequent queues in the care process, we choose $\hat{x}_{0,j} = \sum_{i\in\mathcal{J}} q_{i,j}\hat{x}_{0,i}$ and $\hat{x}_{t,j} = \sum_{i\in\mathcal{J}} q_{i,j}\hat{x}_{t,i}$, for all $t \in \mathcal{T}$.

We then generate the initial waiting list $S_{1,j} = (S_{1,j,0}, S_{1,j,1}, \ldots)$. $S_{1,j}$ represents the waiting list at the start of the planning period, because the waiting

Parameter	*Description*	*Used values for testing*
$\|\mathcal{T}\|$	The number of time periods	$\{8\}$
$\|\mathcal{R}\|$	The number of resource types	$\{2\}$
$\|\mathcal{G}\|$	The number of care processes	$\{6, 8, 10\}$
e_g	The number of stages in care process $g \in \mathcal{G}$	$\{3, 5, 7\}$
$s_{j,r}$	Expected service time from resource type $r \in \mathcal{R}$ for a patient in queue $j \in \mathcal{J}$ in time units (three value sets)	$\{10, 15, 20\}$, $\{100, 120, 140\}$, $\{200, 220, 240\}$
$\lambda_{j,t}$	New demand in queue $j \in \mathcal{J}$ in time period $t \in \mathcal{T}$	$\{2, 6, 10\}$
$q_{i,j}$	The routing probabilities between queue $i, j \in \mathcal{J}$	$\{0, 0.25, 0.5, 0.75, 1\}$
$\hat{a}_{t,j}$	Target for $A^{\alpha}_{t,j}$ for queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$	$\{1, 2, \ldots, 8\}$
$\hat{x}_{t,j}$	Target number of served patients for queue $j \in \mathcal{J}$ and time period $t \in \mathcal{T}$	$\{2, 3, \ldots, 10\}$
$\hat{x}_{0,j}$	Target number of served patients for queue $j \in \mathcal{J}$ in the previous planning period	$\{10, 30, 50\}$
$x_{0,j}$	The number of served patients for queue $j \in \mathcal{J}$ in the previous planning period	$\{10, 30, 50\}$
$\bar{u}_j$	The number of time periods the longest-waiting patients have been waiting on the initial waiting list for queue $j \in \mathcal{J}$	$\{1, 2, \ldots, 16\}$

Table 4.2: The parameters that characterize the test instances.

list $S_{1,j}$ is calculated before patients are served in this time period. First, we draw $\bar{u}_j$, which indicates the number of time periods the longest-waiting patients have been waiting on the initial waiting list of queue $j \in \mathcal{J}$. Then, we determine the number $S_{1,j,u}$ of patients waiting u time periods by:

$$S_{1,j,u} = \frac{b_j}{u}, \qquad j \in \mathcal{J}, 0 < u \leq \bar{u}_j. \tag{4.17}$$

where b_j is calculated as follows. We first generate b_j for the first queue in the care process by:

$$b_j = \frac{\sum_{t \in \mathcal{T}} \lambda_{j,t}}{|\mathcal{T}|}, \qquad j \in \mathcal{J}.$$

For all subsequent queues in the care process, we draw b_j from $\left[0.75 \sum_{i \in \mathcal{J}} q_{i,j} b_i, 1.25 \sum_{i \in \mathcal{J}} q_{i,j} b_i\right]$. By dividing by u in (4.17), the number $S_{1,j,u}$

of patients waiting u time periods decreases as u grows. This structures the initial waiting list $S_{1,j}$ for each $j \in \mathcal{J}$ to resemble waiting lists observed in practice.

To determine the resource capacities $\eta_{r,t}$ for each resource type $r \in \mathcal{R}$ and time period $t \in \mathcal{T}$, we first approximate the amount $\tilde{\eta}_r$ of resources required in the current planning period by summing the amount of resources required by arriving patients $\lambda_{j,t}$, for all $t \in \mathcal{T}$, throughout their care processes. Using $\tilde{\eta}_r$ and a tuning parameter κ_r, we determine $\eta_{r,t}$ by:

$$\eta_{r,t} = \kappa_r \frac{\tilde{\eta}_r}{|\mathcal{T}|}, \qquad r \in \mathcal{R}, t \in \mathcal{T}. \tag{4.18}$$

Unless stated otherwise, we assume $\kappa_r = 1$, for all $r \in \mathcal{R}$. The method's sensitivity to varying capacity dimensions is examined by varying κ_r in the computational experiments.

We bound the computation time for the MILP by 100 seconds. This setting results in an average integrality gap 0.01% for instances with 6 time periods and 50 queues (see Tables 4.3 and 4.4 for more information). For the procedure to determine the weights, we set the following entries $\alpha = 0.9$, $\epsilon = 0.01$, $\theta = 0.01$ and $\omega_t = 1$, for all $t \in \mathcal{T}$. The latter indicates that we give the same weight to each time period.

4.5 Results

We use the performance measures introduced in Section 4.3 to evaluate the proposed method for tactical resource and admission planning. We generate 300 instances following the procedure of Section 4.4. For each queue and time period in the 300 generated instances, we calculate the three performance ratios for access time, the number of patients served and total duration of a care process by (4.10), (4.11) and (4.16) respectively. For each type of performance ratio and each time period, we generate one list of the calculated ratios in all instances. Subsequently, these lists are sorted in ascending order. The sorted lists can be used to evaluate each type of performance ratio at a given percentile for each time period. For example, when there are 3000 ratios on a sorted list, the 300-th entry represents the 10-th percentile. When we curve these percentiles and the curve decreases (increases) for successive time periods, we know that for a given fraction of the queues in all 300 instances, the performance ratio decreases (increases). Below, we present our results for each tactical planning objective.

Achieving equitable access and treatment duration for patient groups
The curves in Figure 4.2 display the percentiles for the access time performance ratios $L^A_{0.9,t,j}$ in all queues in all instances. The curves show that resource capacities are allocated such that the performance ratios $L^A_{0.9,t,j}$ become less variable, as the range between the 20-th and 80-th percentiles decreases and stabilizes over time periods. Hence, we may conclude that resources are more equitably

divided over queues during the planning period, leading to less variation in performance ratios.

The performance ratios tend toward a number above 1, because the total resource capacity $\eta_{r,t}$ per resource $r \in \mathcal{R}$ in time period $t \in \mathcal{T}$ is sufficient to serve new demand, but not the already existing waiting list $S_{0,j}$. When κ_r in (4.18) is increased, more capacity is available to serve new demand and the existing waiting list. As a result, the performance ratios in the graph in Figure 4.3 tend towards a lower number than the performance ratios in the graph in Figure 4.2. In this case, they tend toward 1, which indicates that resource capacities are allocated such that our measures $A^{\alpha}_{t,j}$ for a higher fraction of queues are closer to target. The curves in Figure 4.4 display the percentiles for the access time

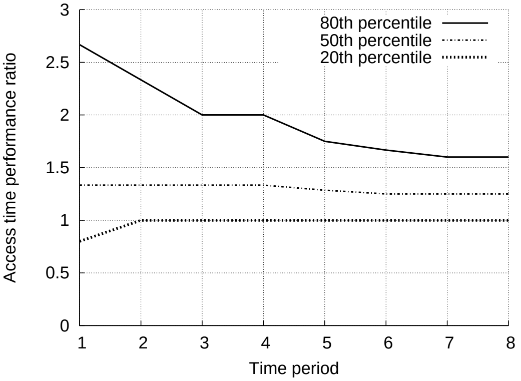


Figure 4.2: The 20-th, 50-th and 80-th percentiles of the access time performance ratios $L^A_{0.9,t,j}$ for all queues in all instances.

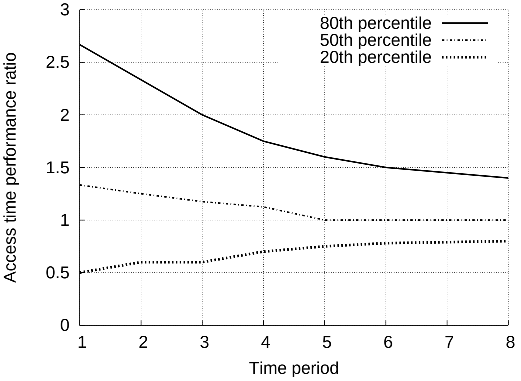


Figure 4.3: The 20-th, 50-th and 80-th percentiles of the access time performance ratios $L^A_{0.9,t,j}$ for all queues in all instances, when $\kappa_r = 1.1$ for all $r \in \mathcal{R}$.

performance ratios $L^{H}_{0.9,t,g}$ for a complete care process, for all care processes in all instances. The method allocates resources such that the performance ratios $L^{H}_{0.9,t,g}$ tend towards 1. We may conclude that the measure $H^{\alpha}_{t,g}$ is closer to target for a larger fraction of care processes.

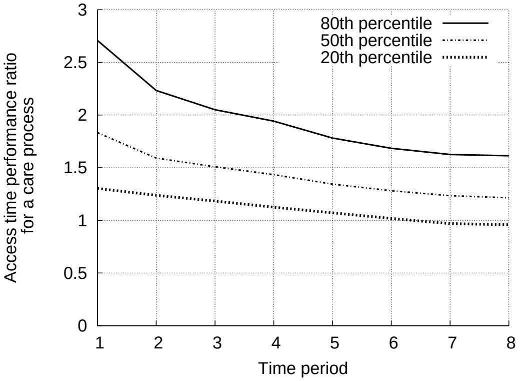


Figure 4.4: The 20-th, 50-th and 80-th percentiles of the access time performance ratios $L^{H}_{0.9,t,g}$ for all care processes in all instances.

Serving the strategically agreed target number of patients
The curves in Figure 4.5 display the percentiles for the performance ratios $L^{C}_{t,j}$ for the number of patients served in all queues in all instances. Resources are allocated such that the performance ratios $L^{C}_{t,j}$ for the number of patients served are less variable and tend toward 1. This indicates that resource capacities are allocated such that the number of patients served for a higher fraction of queues are closer to target.

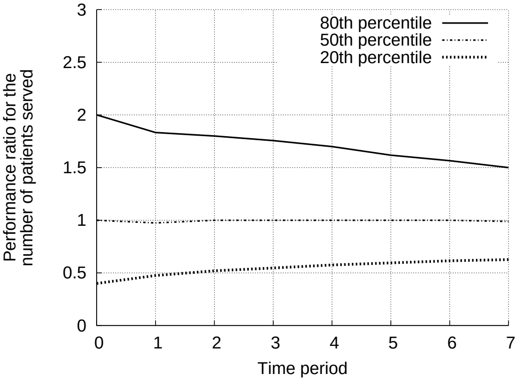


Figure 4.5: The 20-th, 50-th and 80-th percentiles of the performance ratios $L^C_{t,j}$ for the number of patients served for all queues in all instances.

Maximizing resource utilization and balancing workload
The fraction $\rho_{r,t}$ of resource capacities $r \in \mathcal{R}$ that are allocated to care processes in time period $t \in \mathcal{T}$ can be used to identify bottleneck and underutilized resources. Graphing these percentages supports this identification. For example, the histogram in Figure 4.6 shows a decline in the percentage of resource capacity that is allocated to care processes in time periods $t = 3$ and $t = 4$. Hospital management can use these histograms to decide on patient admission policies, or to dimension and allocate resource capacities.

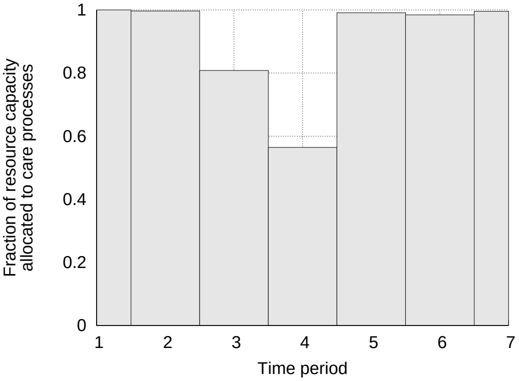


Figure 4.6: Example of the fraction $\rho_{r,t}$ of resource capacities allocated to care processes for a resource type in an instance.

In addition, the fraction $\rho_{r,t}$ of resource capacities that are allocated to care processes can be used to evaluate the workload balance. For example, the workload is significantly lower in time periods $t = 3$ and $t = 4$ for the resource depicted in the graph of Figure 4.6. This can for example be caused by varying demand in different time periods (particular demand patterns), or by allocation decisions for resources in preceding stages and previous time periods. To improve workload balance for specific resources in the planning period, resource allocation constraints may be introduced in the MILP, as discussed in Section 4.3.4.

The average calculation time for relatively large instances ($|\mathcal{G}| = 10$, $e_g = 5, \forall g \in \mathcal{G}$, $|\mathcal{T}| = 6$, $|\mathcal{R}| = 2$) is 4 minutes, which may be assumed to be reasonable for a tactical planning method. Furthermore, the average integrality gap for these instances is 0.01%. The calculation time is principally influenced by the number $|\mathcal{T}|$ of time periods and the number $|\mathcal{J}|$ of queues. Tables 4.3 and 4.4 give more details on average calculation time and average integrality gap for various instances.

Queues	**Time periods**		
	4	6	8
30	43	69	111
50	82	224	1482
70	134	1075	3741

Table 4.3: The average calculation time in seconds for various instances.

Queues	**Time periods**		
	4	6	8
30	0.00%	0.03%	0.03%
50	0.00%	0.01%	0.05%
70	0.01%	0.02%	0.07%

Table 4.4: The average integrality gap for various instances.

4.6 Managerial implications

We collaborate with various hospitals which increasingly implement procedures for tactical resource capacity planning *to achieve equitable access for patient groups, to serve the strategically agreed target number of patients, to maximize resource utilization* and *to balance the workload*. In their tactical planning approaches, some of these hospitals have spreadsheet solutions in place to evaluate for example waiting lists, access time and resource utilization. They use this information for resource allocation decision making, for example to allocate operating time and consultation time. Our method provides an optimization procedure for this step.

The tactical resource and admission plan proposed by our model is implemented using a rolling horizon approach: only near-term decisions are implemented. Every time period, the model is used to evaluate the tactical plan and to set the near-term decisions. The weight determination procedure is performed each time the tactical plan is reevaluated or redeveloped, as the weights are dependent on for example the expected patient arrivals and the selected tactical resource and admission plan. It is out of the scope of this chapter to develop the operational decision rules that address unanticipated events during the execution of a tactical plan, such as a lower or higher demand than forecasted.

Implementation of our method at a particular hospital requires insight in the hospital's performance. Waiting list data, access times, the number of served patients, and expected resource availability should be made available every time period, to be able to propose a tactical plan. Also, the care processes in scope need to be defined (patient groups, the various stages, resource requirements and transition probabilities). The care stages in each care process are defined in close cooperation with medical staff, and by analyzing patient data obtained from the hospital information system. With the information about the care processes and individual patient procedures, methods described in for example [286] can be used to develop the transition probabilities for each stage in the care process. Correct administration of for example patient procedures, the sequence of these procedures, access times, the number of served patients is key in developing the patient care processes and providing the information to develop credible tactical plans.

Introduction of a dynamic tactical planning concept requires flexibility from all involved resources. It requires tactical rules (e.g., how many time periods before implementation is a tactical plan 'cast in stone'?), operational rules (e.g., when are resource capacities reallocated to other care processes?), and organizational changes in the various medical departments to be able to respond to changes in the tactical plan effectively. One particular tactical rule was a prerequisite for participation of the involved medical departments and the successful implementation of dynamic tactical planning in one of the hospitals. The involved decision makers agreed that a decided reduction of allocated resource capacity (in this case operating time) can always be revoked when the resource capacity is required again in the future. Under this agreement, the involved decision makers can be more open for adjustments of the tactical plan, as they are certain that they can always go back to the prior tactical plan. Also, to support the process of tactical planning, agreements are required between the involved decision makers on what should be done (e.g., data analysis, calculating scenarios, discussing proposed plans) and who is involved (e.g., hospital managers, doctors, nurses) in each step of developing a tactical plan.

4.7 Conclusion and discussion

Inspired by multiple hospitals that are investigating the potential use of tactical planning, we have developed an iterative method that can be used dynamically to develop mid-term tactical resource and admission plans for real-life sized instances. These tactical resource and admission plans allocate resource capacity over care processes and determine the number of patients to serve at a particular stage of their care process.

Computational results show that our method improves compliance with access time targets, care process duration and the number of patients served. The method is a tool for hospital management to achieve equitable access and treatment duration for patient groups and to serve the strategically agreed target number of patients. Within this framework, the method can be adapted to maximize resource utilization and/or to balance workload. It may be used to identify bottleneck resources or underutilized resources, and for scenario analysis in anticipation of peaks in patient demand or resource (un)availability. This allows a timely response, such as temporarily increasing or decreasing resource capacities to improve access times and workload balance.

The method integrates decision making for multiple resources, multiple time periods and multiple patient groups with various uncertain care processes. Care processes connect multiple departments and resources into a network and fluctuations in both patient arrivals (e.g., seasonality) and resource availability (e.g., holidays) result in bullwhip effects in the care chain. Therefore, coordinated decision making along a care chain of hospital resources offers improvement potential.

The basic elements of the tactical planning problem in healthcare also occur in other industries. Since our method can be extended and adapted easily, it can be used in other service and manufacturing environments. For example, the model can be useful for tactical planning in a production environment. In such an environment, various products (care processes) are typically produced by multiple resource types. The product goes through different production stages (care stages) and at each stage there is 'work in progress' waiting to be processed (waiting list). The objectives in production may be to use resources effectively, to meet production targets and to have a certain amount of work in progress. These aspects are reflected in the objective function and constraints of our model. Clearly, alternative constraints and objective functions may better fit the objectives of tactical planning of a particular organization. Hence, we have mentioned that various other performance measures can be used to develop alternative objective functions and that various possible extensions of the model may be of interest, including constraints to balance the number of patient admissions and resource capacities allocated to particular care processes over time, and the incorporation of already scheduled patients. These extensions are interesting topics for further research.

CHAPTER 5

Tactical planning in care processes with a stochastic approach

5.1 Introduction

Tactical planning is a key element of hospital planning and control that concerns the intermediate term allocation of resource capacities and elective patient admission planning [234]. Its main objectives are to achieve equitable access and treatment duration for patient groups, to serve the strategically agreed target number of patients (i.e., production targets or quota), to maximize resource utilization and to balance workload [261].

From a clinician's perspective, tactical resource and admission plans break the clinician's time down into separate activities (e.g., consultation time and surgical time) and determine the number of patients to serve from a particular patient group at a particular stage of their care process (e.g, consultation or surgery). We use the term care process to identify a chain of care stages for a patient, for example a visit to an outpatient clinic, a surgery, and a revisit to the outpatient clinic. At each stage in the care process, patients incur access time, which is the time spent on the waiting list before being served. Controlled access times ensure quality of care for the patient and prevent patients from seeking treatment elsewhere [531]. The term care process is not to be confused with "clinical pathway", which is described in [171].

Care processes connect multiple departments and resources together as an integrated network. Fluctuations in both patient arrivals (e.g., seasonality) and resource availability (e.g., holidays) at one department may impact the entire care chain. For patients, this results in varying access times for each separate stage in a care process, and from a hospital's perspective, this results in varying resource utilizations and service levels. To cope with these fluctuations, intermediate-term re-allocation of hospital resources, taking into account a care chain perspective [80, 231, 401], seems necessary.

The tactical planning problem in healthcare is stochastic in nature. Randomness exists in for example the number of (emergency) patient arrivals and the number of patient transitions after being treated at a particular stage of their care process. Several papers have focused on tactical planning problems that span multiple departments and resources in healthcare [192, 286, 374] and other industries [211]. In [261], the literature and various applications is reviews and it is concluded that existing approaches to develop tactical resource and admission plans in the OR/MS literature are myopic, focus on developing long-term cyclical plans, or are not able to provide a solution for real-life instances. The authors develop a deterministic method for tactical planning over multiple departments and resources within a mathematical programming framework.

In this chapter, we develop a stochastic approach for the tactical planning problem in healthcare by modeling it as a Dynamic Programming problem (DP). Due to the properties of the tactical planning problem, with discrete time periods and transitions that depend on the decision being made, DP is a suitable modeling approach. As problem sizes increase, solving a DP is typically intractable due to the 'curse of dimensionality'. To overcome this problem, an alternative solution approach for real-life sized instances of the tactical planning problem is needed. The field of Approximate Dynamic Programming (ADP) provides a suitable framework to develop such an alternative approach, and we use this framework to develop an innovative solution approach. ADP uses approximations, simulations and decompositions to reduce the dimensions of a large problem, thereby significantly reducing the required calculation time. A comprehensive explanation and overview of the various techniques within the ADP framework are given in [402]. The application of ADP is relatively new in healthcare, it has been used in ambulance planning [350, 431] and patient scheduling [391]. Other applications in a wider spectrum of industries include resource capacity planning [163, 436], inventory control [445], and transportation [474].

We aim to contribute to the literature in two ways. First, we provide a theoretical contribution to the development of tactical resource and admission plans in healthcare in the field of Operations Research and Management Science (OR/MS). We develop an approach to develop tactical plans that take randomness in patient arrivals and patient transitions to other stages into account. These plans are developed for multiple resources and multiple patient groups with various care processes, thereby integrating decision making for a chain of hospital resources. The model is designed with a finite horizon, which allows all input to be time dependent. This enables us to incorporate anticipated or forecasted fluctuations between time periods in patient arrivals (e.g., due to seasonality) and resource capacities (e.g., due to vacation or conference visits) in developing the tactical plans. The model can also be used in 'realtime'. If during actual implementation of the tactical plan, deviations from forecasts make reallocation of resource capacity necessary, the developed model can be used to determine an adjusted tactical plan. The model can be extended to include different cost structures, constraints, and additional stochastic elements. Second, the solution approach is innovative as it combines various methods and techniques within the ADP-framework and the field of mathematical programming. Also, the application of ADP is new in tactical resource capacity and patient admission planning, and relatively new in healthcare in general, where it has mainly been applied in ambulance planning [350, 431] and patient scheduling [391].

This chapter is organized as follows. Section 5.2 discusses the mathematical problem formulation, and Section 5.3 describes the exact Dynamic Programming (DP) solution approach for small instances. Section 5.4 introduces the ADP approaches necessary to develop tactical plans for real-life sized instances. Section 5.5 describes how the model can be used to develop or adjust tactical plans in healthcare. Section 5.6 discusses computational results and Section 5.7 concludes this chapter.

5.2 Problem formulation

This section introduces the problem and the patient dynamics in care processes. We provide notation and present the stochastic model formulation of the problem that captures randomness in patient arrivals and patient transitions between queues.

The planning horizon is discretized in consecutive time periods $\mathcal{T} = \{1, 2, \ldots, T\}$. This finite horizon allows all input to the model to be time dependent and enables incorporating anticipated or forecasted fluctuations between time periods in patient arrivals and resource capacities. We consider a set of resource types $\mathcal{R} = \{1, 2, \ldots, R\}$ and a set of patient care processes $\mathcal{G} = \{1, 2, \ldots, G\}$. Each of these care processes consists of a set of stages $\mathcal{K}_g = \{1, ..., e_g\}$, where e_g gives the number of stages in the care process $g \in \mathcal{G}$. To simplify notation, we denote each stage in $\cup_{g\in\mathcal{G}}\mathcal{K}_g$ by a queue j. We introduce the set $\mathcal{J}$ as the set of all queues and $\mathcal{J}^r$ as the set of queues that require capacity of resource $r \in \mathcal{R}$.

Each queue $j \in \mathcal{J}$ requires a given amount of time units from one or more resources, given by $s_{j,r}$, $r \in \mathcal{R}$, and different queues may require the same resource. The number of patients that can be served by resource $r \in \mathcal{R}$ is limited by the available resource capacity $\eta_{r,t}$ in time period $t \in \mathcal{T}$. Because we also allow for queues without resource requirement (dummy queues), $\cup_{r\in\mathcal{R}}\mathcal{J}^r$ is not necessarily equal to $\mathcal{J}$.

After being treated at a queue $j \in \mathcal{J}$, patients either leave the system or join another queue. To model these transitions, we introduce $q_{j,i}$ which denotes the fraction of patients that will join queue $i \in \mathcal{J}$ after being treated in queue $j \in \mathcal{J}$. The value $q_{j,0} = 1 - \sum_{i\in\mathcal{J}} q_{j,i}$ denotes the fraction of patients that leave the system after being treated at queue $j \in \mathcal{J}$. In general, $q_{j,i}$ is positive when $i \in \mathcal{J}$ is immediately succeeding $j \in \mathcal{J}$ in the same care process. However, our modeling framework allows for different types of transitions (for example transitions to any prior or future stage in the same care process, and transitions between queues of different care processes). In addition to demand originating from the treatment of patients at other queues within the system, there is also demand from outside the system. The number of patients arriving from outside the system to queue $j \in \mathcal{J}$ at time $t \in \mathcal{T}$ is given by $\lambda_{j,t}$. As can be observed, our model has a finite horizon.

Within the definition of $q_{j,i}$ lies the major assumption of our model:

Assumption 5.1. *Patients are transferred between the different queues according to transition probabilities $q_{j,i}, \forall j, i \in \mathcal{J}$ independent of their preceding stages, independent of the state of the network and independent of the other patients.*

For practical purposes in which Assumption 5.1 does not hold, we can adjust the various care processes to ensure it does hold. For example, if after some stage within a care process, the remainder of the patient's path depends on the current stage, we create a new care process for the remaining stages and patients flow with a certain probability to the first queue in that new care process.

For the arrival processes, we assume the following.

Assumption 5.2. *Patients arrive at each queue from outside the system according to a Poisson process with rate $\lambda_{j,t}$, $\forall j \in \mathcal{J}, t \in \mathcal{T}$. The external arrival process at each queue $j \in \mathcal{J}$ in time period $t \in \mathcal{T}$ is independent of the external arrival process at other queues and other time periods. Since all arrival processes are independent, we obtain $\lambda_{0,t} = \sum_{j=1}^{|\mathcal{J}|} \lambda_{j,t}$, $\forall t \in \mathcal{T}$.*

We introduce $\mathcal{U} = \{0, 1, 2, ..., U\}$ to represent the set of time periods patients can be waiting. Given Assumption 5.1, patients are characterized by the queue in which they are waiting and the amount of time they have been waiting at this queue. We introduce

$$S_{t,j,u} \quad = \quad \text{Number of patients in queue } j \in \mathcal{J} \text{ at time } t \in \mathcal{T} \text{ with a waiting time of } u \in \mathcal{U}.$$

The state of the system at time period t can be written as S_t, which is a matrix made up of the elements $(S_{t,j,u})$, for all $t \in \mathcal{T}, j \in \mathcal{J}$, and $u \in \mathcal{U}$. We define decisions as actions that can change the state of the system. The decisions are given by

$$x_{t,j,u} \quad = \quad \text{Number of patients to treat in queue } j \in \mathcal{J} \text{ at time } t \in \mathcal{T}\text{, with a waiting time of } u \in \mathcal{U}.$$

The decision at time period t can be written as $x_t = (x_{t,j,u})_{j \in \mathcal{J}, u \in \mathcal{U}}$, in the same way as we write the state description S_t. The cost function $C_t\left(S_t, x_t\right)$ related to our current state S_t and decision x_t can be modeled in various ways. The main objectives of tactical planning are *to achieve equitable access and treatment duration for patient groups* and *to serve the strategically agreed number of patients* [261]. The focus in developing this model is on the patient's waiting time (equitable access and treatment duration), and we assume that the strategically agreed number of patients is set in accordance with patient demand (as the model accepts all patients that arrive). The cost function in our model is set-up to control the waiting time per stage in the care process, so per individual queue ($j \in \mathcal{J}$). It is also possible to adapt the cost function for other tactical planning settings, for example to control the total waiting time per individual care process $g \in \mathcal{G}$ or for all queues that use a particular resource $r \in \mathcal{R}$. We choose the following cost function, which is based on the number of patients for which we decide to wait at least one time unit longer

$$C_t\left(S_t, x_t\right) = \sum_{j \in \mathcal{J}} \sum_{u \in \mathcal{U}} c_{j,u}\left(S_{t,j,u} - x_{t,j,u}\right), \qquad \forall t \in \mathcal{T}. \tag{5.1}$$

The cost component $c_{j,u}$ in (5.1) is set by the hospital to distinguish between queues $j \in J$ and waiting times $u \in \mathcal{U}$. In general, higher $u \in \mathcal{U}$ will have higher costs as it means a patient has a longer total waiting time. This could be modeled in various ways, for example the cost $c_{j,u}$ could be incrementally increasing with $u \in \mathcal{U}$ or the hospital has some target/threshold waiting time after which waiting costs increase significantly.

The Integer Linear Programming (ILP) version of the introduced problem can be written as

$$\min \sum_{t \in \mathcal{T}} C_t\left(S_t, x_t\right) = \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \sum_{u \in \mathcal{U}} c_{j,u}\left(S_{t,j,u} - x_{t,j,u}\right), \tag{5.2}$$

subject to

$$S_{t,j,0} = \lambda_{j,t} + \sum_{i \in \mathcal{J}} \sum_{u \in \mathcal{U}} q_{i,j} x_{t-1,i,u}, \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{5.3}$$

$$S_{t,j,U} = \sum_{u=U-1}^{U} \left(S_{t-1,j,u} - x_{t-1,j,u}\right), \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{5.4}$$

$$S_{t,j,u} = S_{t-1,j,u-1} - x_{t-1,j,u-1}, \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U} \setminus \{0, U\} \tag{5.5}$$

$$x_{t,j,u} \leq S_{t,j,u}, \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U}, \tag{5.6}$$

$$\sum_{j \in \mathcal{J}^r} s_{j,r} \sum_{u \in \mathcal{U}} x_{t,j,u} \leq \eta_{r,t}, \qquad \forall r \in \mathcal{R}, t \in \mathcal{T}, \tag{5.7}$$

$$x_{t,j,u} \in \mathbb{Z}_+, \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U}. \tag{5.8}$$

Constraints (5.3) to (5.5) stipulate that the waiting list variables are consistent. Constraint (5.3) determines the number of patients newly entering a queue. Constraint (5.4) updates the waiting list for the longest waiting patients per queue (which is bounded by U). Constraint (5.5) updates the waiting list variables at each time period for all $u \in \mathcal{U}$ that are not covered by the first two constraints. Constraint (5.6) stipulates that not more patients are served than the number of patients on the waiting list. Constraint (5.7) assures that the resource capacity of each resource type $r \in \mathcal{R}$ is sufficient to serve all patients. Constraint (5.8) is an integrality constraint for the number of patients to serve of each type at each time period.

The above ILP version does not incorporate the different forms of randomness that are apparent in the actual system, such as random patient arrivals and uncertainty in patient transitions to other queues. The ILP uses approximations in the form of the expectation for those processes. More specifically, $\lambda_{i,t}$ and $q_{j,i}$, $i, j \in \mathcal{J}, t \in \mathcal{T}$ in Constraint 5.3 actually are parameters for stochastic processes. To capture all sources of random information, we introduce

$W_t =$ The vector of random variables representing all the new information that becomes available between time $t-1$ and t.

The vector W_t contains all the new information, which consists of new patient arrivals and outcomes for transitions between queues. We distinguish between *exogeneous* and *endogeneous* information in

$$W_t = \left(\widehat{S}_t^e, \widehat{S}_t^o\left(x_{t-1}\right)\right), \qquad \forall t \in \mathcal{T},$$

where the exogeneous $\widehat{S}_t^e = \left(\widehat{S}_{t,j}^e\right)_{\forall j \in \mathcal{J}}$ represents the patient arrivals from outside the system, and the endogeneous $\widehat{S}_t^o\left(x_{t-1}\right) = \left(\widehat{S}_{t,j,i}^o\left(x_{t-1}\right)\right)_{\forall i,j \in \mathcal{J}}$ represents the patient transitions to other queues as a function of the decision vector x_{t-1}. $\widehat{S}_{t,j,i}^o\left(x_{t-1}\right)$ gives the number of patients transferring from queue $j \in \mathcal{J}$ to queue $i \in \mathcal{J}$ at time $t \in \mathcal{T}$, depending on the decision vector x_{t-1}.

Assumptions 5.1 and 5.2 imply that the probability distribution (conditional on the decision) of future states only depends on the current state, and is independent of preceding states in preceding time periods. This means that the described process has the Markov property. We use this property in defining a transition function, S^M, to capture the evolution of the system over time as a result of the decisions and the random information.

$$S_t = S^M\left(S_{t-1}, x_{t-1}, W_t\right), \tag{5.9}$$

where

$$S_{t,j,0} = \widehat{S}_{t,j}^e + \sum_{i \in \mathcal{J}} \widehat{S}_{t,i,j}^o\left(x_{t-1,i}\right), \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{5.10}$$

$$S_{t,j,U} = \sum_{u=U-1}^{U}\left(S_{t-1,j,u} - x_{t-1,j,u}\right), \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{5.11}$$

$$S_{t,j,u} = S_{t-1,j,u-1} - x_{t-1,j,u-1}, \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U} \setminus \{0, U\}, \tag{5.12}$$

are the stochastic counterparts of the first constraints (5.3) to (5.5) in the ILP formulation. The stochastic information is captured in (5.10). All arrivals in time period $t \in \mathcal{T}$ to queue $j \in \mathcal{J}$ from outside the system ($\widehat{S}_{t,j}^e$) and from internal transitions ($\sum_{i \in \mathcal{J}} \widehat{S}_{t,i,j}^o\left(x_{t-1,i}\right)$) are combined in (5.10).

We aim to find a policy (a decision function) to make decisions about the number of patients to serve at each queue. We represent the decision function by

$X_t^\pi(S_t) =$ A function that returns a decision $x_t \in \mathcal{X}_t(S_t)$, under the policy $\pi \in \Pi$.

The set Π refers to the set of potential decision functions or policies. $\mathcal{X}_t$ denotes the set of feasible decisions at time t, which is given by

$$\begin{aligned} \mathcal{X}_t(S_t) = \{ \quad & x_t | & \\ & x_{t,i,u} \leq S_{t,i,u}, & \forall i \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U} \\ & \textstyle\sum_{j \in \mathcal{J}^r} s_{j,r} \sum_{u \in \mathcal{U}} x_{t,j,u} \leq \eta_{r,t}, & \forall r \in \mathcal{R}, t \in \mathcal{T} \\ & x_{t,j,u} \in \mathbb{Z}_+ & \forall i \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U}\}. \end{aligned} \tag{5.13}$$

As given in (5.13), the set of feasible decisions in time period t is constrained by the state space S_t and the available resource capacity $\eta_{r,t}$ for each resource type $r \in \mathcal{R}$. Our goal is to find a policy π, among the set of policies Π, that minimizes the expected costs over all time periods given initial state S_0. This goal is given in

$$\min_{\pi \in \Pi} \mathbb{E} \left\{ \sum_{t \in \mathcal{T}} C_t\left(S_t, X_t^\pi(S_t)\right) | S_0 \right\}, \tag{5.14}$$

where $S_{t+1} = S^M(S_t, x_t, W_{t+1})$. Note that the description in (5.14) is in line with the ILP's objective function in (5.2). The challenge is to find the best policy $X_t^\pi(S_t)$.

Remark 5.3. *Incorporated in the formulation of the model is the assumption that after a treatment decision x_t at the beginning of time t, patients immediately generate waiting costs in the following queue (if they move on to a following care stage, and do not exit the system) after entering that queue in time period $t+1$. In practice, after a treatment, a patient may require to wait a minimum time lag before a follow-up treatment can be initiated. The model can be extended to cover cases with time lags $d_{i,j}$ (time lag in the transition from queue i to queue j) by allowing u to be negative in $S_{t,j,u}$. For example, $S_{t,j,-2}$ then indicates the number of patients that will enter queue j two time periods from now. Incorporating this time lag changes the system dynamics: patients with $u < 0$ cannot be served and we set $C_{t,j,u}(S_{t,j,u}, x_{t,j,u}) = 0$ for $u < 0$, $\forall i \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U}$.*

The various sets, indices, state descriptions and parameters that will be used in following sections are given in Table 5.1

5.3 Dynamic Programming

To solve the problem formulated in Section 5.2, we propose an exact DP approach. The DP approach is suitable due to the finite horizon of the problem, the decision dependent transitions of patients, and the randomness in patient

Sets		**Indices**	
$\mathcal{G}$	Care processes	$g \in \mathcal{G}$	Care process
$\mathcal{J}$	Queues	$i, j \in \mathcal{J}$	Queue
$\mathcal{T}$	Time periods	$t \in \mathcal{T}$	Time period
$\mathcal{R}$	Resource types	$r \in \mathcal{R}$	Resource type
$\mathcal{U}$	Time periods (to indicate waiting time)	$u \in \mathcal{U}$	Waiting time period
$\mathcal{J}^r$	Queues for resource type r	$i, j \in \mathcal{J}^r$	Queue
Decision variables			
$x_{t,j,u}$	Number of patients to treat in queue j in time period t, who have been waiting u time periods		
State description			
S_t	State of the system at time period t		
$S_{t,j,u}$	Number of patients in queue j in time period t with a waiting time of u		
Parameters			
$c_{j,u}$	Costs for queue j and u time periods waiting		
$\lambda_{j,t}$	New demand in queue j in time period t		
$\eta_{r,t}$	Capacity of resource type r in time period t in time units		
$q_{i,j}$	Probability that a patient moves from queue i to queue j		
$s_{j,r}$	Expected capacity requirements from resource type r for a patient in queue j in time units		

Table 5.1: The sets, indices, variables, state description and parameters used.

arrivals and patient transitions. In the following, the DP approach is explained in detail. First, we state the optimality equation, and second, we elaborate the expectation in that optimality equation.

By the principal of optimality [34], we can find the optimal policy by solving

$$V_t\left(S_t\right) = \min_{x_t \in \mathcal{X}_t\left(S_t\right)} \left(C_t\left(S_t, x_t\right) + \mathbb{E}\left\{V_{t+1}\left(S_{t+1}\right) | S_t, x_t, W_{t+1}\right\}\right), \tag{5.15}$$

where $S_{t+1} = S^M\left(S_t, x_t, W_{t+1}\right)$ gives the state S_{t+1} as a function of the current state S_t, the decisions x_t, and the new information W_{t+1}.

The optimal decision minimizes the value that is calculated with the value function $V_t\left(S_t\right)$. In the value function, 'direct' costs are incurred for the decision in the current time period ($C_t\left(S_t, x_t\right)$), and 'future' costs reflect the expected costs in future time periods, *as a result* of the decision in the current time period ($\mathbb{E}\left\{V_{t+1}\left(S_{t+1}\right) | S_t, x_t, W_{t+1}\right\}$). The expectation of the 'future' costs is based on the probability distribution for the arrival of new patients and the transitions of all treated patients in the decision vector x_t of the current time period.

As a next step, we specifiy the expectation in (5.15). We introduce the vector w, consisting of elements w_j representing the number of patients leaving queue j, for all j in $\mathcal{J}$ and arriving from outside the system (w_0). To administer all possible transitions, we introduce the elements, w_{ij} representing a realization for the number of patients that are transferred from queue i to queue j after service at queue i. w_{0j} represents the realization of the number of external arrivals at queue j. w_{j0} represents the number of patients leaving the hospital after treatment at queue j. In addition, we introduce the vector w', which represents a realization of the number of patients arriving at each queue. The element w'_j represents the number of patients arriving at queue j.

Note that under Assumption 5.1, the transition process follows a multinomial distribution with the parameters $q_{j,i}$ with $i, j \in \mathcal{J}$, and $q_{j,0} = 1 - \sum_{i\in\mathcal{J}} q_{j,i}$ for patients leaving the hospital. Enumerating the product of the probability and value associated with all possible outcomes of w', establishes the expectation in (5.15):

$$V_t\left(S_t\right) = \min_{x_t\in\mathcal{X}(S_t)} \left(C_t\left(S_t, x_t\right) + \sum_{w'} P(w'|x_t) V_{t+1}\left(S_{t+1}|S_t, x_t, w'\right) \right),$$

and from [55], we obtain

$$P(w'|x_t) = \sum_{w_0=0}^{\infty} P(w_0) \times$$

$$\sum_{\left\{\begin{array}{c} w_{ij}, i = 0, ..., |\mathcal{J}|, j = 0, ..., |\mathcal{J}| : \\ w_{ij} \geq 0, w_{ij} = 0 \text{ if } p_{i,j} = 0, w_{00} = 0, \\ w_j = \sum_{u\in\mathcal{U}} x_{t,j,u}, j = 1, ..., |\mathcal{J}|, \\ \sum_{j=0}^{|\mathcal{J}|} w_{ij} = w_j, i = 0, ..., |\mathcal{J}|, \\ \sum_{i=0}^{|\mathcal{J}|} w_{ij} = w'_j, j = 0, ..., |\mathcal{J}| \end{array}\right\}} \prod_{i=0}^{|\mathcal{J}|} \binom{w_i}{w_{i0}, ..., w_{i|\mathcal{J}|}} \prod_{j=0}^{|\mathcal{J}|} p_{i,j}^{w_{ij}}.$$

With Assumption 5.2, we obtain

$$P(w_0) = \frac{\lambda_{0,t}^{w_0}}{w_0!} e^{-\lambda_{0,t}}, \quad \text{with } \lambda_{0,t} = \sum_{j=1}^{|\mathcal{J}|} \lambda_{j,t},$$

and

$$p_{i,j} = \begin{cases} q_{i,j}, & \text{when } i = 1, ..., |\mathcal{J}|, j = 0, ..., |\mathcal{J}|, \\ \frac{\lambda_{j,t}}{\lambda_{0,t}}, & \text{when } i = 0, j = 1, ..., |\mathcal{J}|, \\ 0, & \text{when } i = 0, j = 0. \end{cases}$$

Using existing techniques, such as value iteration and backward dynamic programming, (5.15) can be solved to optimality. Proof of the existence of optimal solutions is given in [407].

The exact DP solution method can be used to calculate small instances. These instances particularly do not reflect the complexity and size of a real-life sized instance in a hospital. Computing the exact DP solution is generally difficult and possibly intractable for large problems due to three reasons: (i) The state space $S(t)$ for the problem may be too large to evaluate the value function $V_t\left(S_t\right)$ for all states within reasonable time, (ii) the decision space $\mathcal{X}\left(S_t\right)$ may be too large to find a good decision for all states within reasonable time, and (iii) computing the expectation of 'future' costs may be intractable when the outcome space is large. The outcome space is the set of possible states in time period $t+1$, given the state and decision in time period t. Its size is driven by the random information on the transitions of patients between queues and the external arrivals.

Using dynamic programming to solve real-life size instances of the tactical planning problem seems intractable. To illustrate this, suppose that $\hat{M}$ gives the max number of patients per queue and per number of time periods waiting. The number of states is then given by

$$\hat{M}^{(|\mathcal{J}|\cdot|\mathcal{U}|)}.$$

Consider a system of 8 care processes with an average of 5 stages, resulting in 40 queues, and a maximum number of time periods waiting set to 4. For such a system, the resulting number of states is $\hat{M}^{40\cdot 4} = S_{\max}^{160}$, which is intractable for DP for any $\hat{M} > 1$. Note that in a practical instance at a hospital, $\hat{M}$ may be very large, e.g., $\hat{M} \geq 20$. Additional complexity is added by a large decision space. Assume that in the same system, there is resource capacity available to treat 30 patients in total in one time period. If we assume that they can be treated at 40 queues in the system and that all available resource capacity must be used, this is the same as dividing 30 items over 40 bins. Hence, there are already $\binom{40+32-1}{40-1} = \binom{71}{39} = 1.6 \cdot 10^{20}$ different decisions to evaluate when we are only looking at all decisions that use up maximum resource capacity. In addition, the outcome space may be large, caused by the large number of possible outcomes for the stochastic processes of patient transitions between queues and patient arrivals at each queue.

5.4 Approximate Dynamic Programming

Various alternatives exist to overcome the intractability problems with DP like mentioned in Section 5.3. The problem size can for example be reduced by aggregating information on resource capacities, patients, and/or time periods. We propose an innovative solution approach within the frameworks of ADP and mathematical programming, which can be used to overcome all three mentioned reasons for intractability of DP for large instances. Our solution approach is based on value iteration with an approximation for the value functions. In this section, we explain this approach in more detail.

First, we discuss the use of a 'post-decision' state as a single approximation for the outcome state. Second, we introduce the method to approximate the value of a state and decision, and third, we explain how we use a 'basis functions' approach in the algorithm to approximate that value. This combination uses an approximation for the expectation of the outcome space, thereby reducing complexity significantly. It also enables calculating the value state by state, making the necessity to calculate the entire state space at once, which was the primary reason of intractability of the exact DP approach, obsolete. Fourth, we explain how we overcome the large decision space for large problem instances with an ILP.

5.4.1 Post-decision state

To avoid the problem of a large outcome space and the intractable calculation of the expectation of the 'future' costs, we use the concept of a post-decision state S_t^x [402]. The post-decision state is the state that is reached, directly after a decision has been made, but before any new information W_t has arrived. It is used as a single representation for all the different states the system can be in the following time period, and it is based on the current pre-decision state S_t and the decision x_t. This simplifies the calculation or approximation of the 'future' costs.

The stochastic transitions and external arrivals, captured in W_{t+1}, follow after the post-decision state S_t^x in time period t and before the pre-decision state S_{t+1} of time period $t+1$. The transitions take place as follows. In addition to the transition function (5.9), which gives the transition from pre-decision state S_t to pre-decision state S_{t+1}, we introduce a transition function $S^{M,x}\left(S_t, x_t\right)$, which gives the transition from the pre-decision state S_t to the post-decision state S_t^x. This function is given by:

$$S_t^x = S^{M,x}\left(S_t, x_t\right), \tag{5.16}$$

with

$$S_{t,j,0}^x = \sum_{i\in\mathcal{J}}\sum_{u\in\mathcal{U}} q_{i,j} x_{t,i,u} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{5.17}$$

$$S_{t,j,U}^x = \sum_{u=U-1}^{U} \left(S_{t,j,u} - x_{t,j,u}\right) \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{5.18}$$

$$S_{t,j,u}^x = S_{t,j,u-1} - x_{t,j,u-1} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U}\backslash\left\{0, U\right\}. \tag{5.19}$$

The above constraints are in line with the stochastic transitions between two states in (5.9) to (5.12). The transition in (5.16) to (5.19) is based on the path that patients follow after treatment. There are two differences with the ILP formulation in (5.2) to (5.8). The post-decision state is in the same time-period t as the pre-decision state, and the external arrivals to the system are not included in this formulation, as they are not a result of the decision that is taken. Note that the post-decision state is a direct image of the pre-decision state S_t and the decision x_t.

The actual realizations of new patient arrivals and patient transitions will occur in the transition from the post-decision state in the current time period to the pre-decision state in the next time period. Note that (5.16) can be adapted to include pre-defined priority rules like always treating patients with longest waiting times before selecting others within the same queue. This rule is used in our computational experiments as well. For the remainder of this chapter, whenever we use the word 'state', we are referring to the pre-decision state.

We rewrite the DP formulation in (5.15) as

$$V_t\left(S_t\right) = \min_{x_t\in\mathcal{X}_t\left(S_t\right)} \left(C_t\left(S_t, x_t\right) + V_t^x\left(S_t^x\right)\right),$$

where the value function of the post-decision state is given by

$$V_t^x(S_t^x) = \mathbb{E}\left\{V_{t+1}(S_{t+1})\,|S_t^x\right\}. \tag{5.20}$$

To reduce the outcome space for a particular state and decision, we replace the value function for the 'future costs' of the post-decision state $V_t^x(S_t^x)$ with an approximation based on the post-decision state. We denote this approximation by $\overline{V}_t^n(S_t^x)$, which we are going to learn iteratively, with n being the iteration counter.

We now have to solve

$$\tilde{x}_t^n = \arg\min_{x_t \in \mathcal{X}_t(S_t)} \left(C_t(S_t, x_t) + \overline{V}_t^{n-1}(S_t^x)\right), \tag{5.21}$$

which gives us the decision that minimizes the value $\widehat{v}_t^n$ for state S_t in the n-th iteration. The function $\widehat{v}_t^n$ is given by

$$\widehat{v}_t^n = \min_{x_t \in \mathcal{X}_t(S_t)} \left(C_t(S_t, x_t) + \overline{V}_t^{n-1}(S_t^x)\right). \tag{5.22}$$

Note that $V_t^{n-1}(S_t^x) = 0$ is equivalent to having a standard myopic strategy where the impact of decisions on the future is ignored.

After making the decision $\tilde{x}_t^n$ and finding an approximation for the value in time period t (denoted by $\widehat{v}_t^n$), the value function approximation $\overline{V}_{t-1}^{n-1}(S_{t-1}^x)$ can be updated. We denote this by

$$\overline{V}_{t-1}^n(S_{t-1}^x) \longleftarrow U\left(\overline{V}_{t-1}^{n-1}(S_{t-1}^x), S_{t-1}^x, \widehat{v}_t^n\right). \tag{5.23}$$

In (5.23), we update the value function approximation for time period $t-1$ in the n-th iteration with the 'future' cost approximation for time period $t-1$ in the $n-1$-th iteration, the post-state of time period $t-1$, and the value approximation for time period t. The objective is to minimize the difference between the 'future' cost approximation for time period $t-1$ and the approximation $\widehat{v}_t^n$ for time period t with the updating function, as n increases. This is done by using the algorithm presented in the following section.

5.4.2 The ADP algorithm

We solve (5.21) recursively. Starting with a set of value function approximations and an initial state vector in each iteration, we sequentially solve a subproblem for each $t \in \mathcal{T}$, using sample realizations of W_t, which makes it a Monte Carlo simulation. In each iteration, we update and improve the approximation of 'future' costs with (5.23). Consecutively, the subproblems are solved using the updated value function approximations in the next iteration. This is presented in Algorithm 5.4.

Algorithm 5.4. *The Approximate Dynamic Programming algorithm*

```
Step 0. Initialization

    Step 0a.  Choose an initial approximation $\overline{V}_t^0\left(S_t\right)$ for all $t \in \mathcal{T}$ and $S_t$.
    Step 0b.  Set the iteration counter, $n = 1$, and set the maximum number
              of iterations $N$.
    Step 0c.  Set the initial state to $S_1$.

Step 1. Do for $t = 1, ..., |\mathcal{T}|$:

    Step 1a.  Solve (5.21) to get $\tilde{x}_t$.
    Step 1b.  If $t > 1$, then update the approximation $\overline{V}_{t-1}^n\left(S_{t-1}^x\right)$ for the
              previous post-decision $S_{t-1}^x$ state using

              $\overline{V}_{t-1}^n\left(S_{t-1}^x\right) \longleftarrow U^V\left(\overline{V}_{t-1}^{n-1}\left(S_{t-1}^x\right), S_{t-1}^x, \widehat{v}_t^n\right)$

              where $\widehat{v}_t^n$ is the resulting value of solving (5.22).
    Step 1c.  Find the post-decision state $S_t^x$ with (5.16) to (5.19).
    Step 1d.  Obtain a sample realization $W_{t+1}$ and compute the new pre-
              decision state with (5.9).

Step 2. Increment $n$. If $n \leq N$ go to Step 1.
Step 3. Return $\overline{V}_t^N\left(S_t^x\right)$, $\forall t \in \mathcal{T}$.
```

Using the approximation $\overline{V}_t^N\left(S_t^x\right)$, for all $t \in \mathcal{T}$, we can approximate the value of a post-decision state for each time period. With these approximations, we can find the best decision for each time period and each state, and thus develop a tactical resource capacity and patient admission plan for any given state in any given time period. The difference with the exact DP approach is not only that we now use an value function approximation for the 'future costs', but also that we do not have to calculate the values for the entire state space.

The current set-up of the ADP algorithm is single pass. This means that at each step forward in time in the algorithm, the value function approximations are updated. As the algorithm steps forward in time, it may take many iterations, before the costs incurred in later time periods are correctly transferred to the earlier time periods. To overcome this, the ADP algorithm can also be used with a double pass approach [402], where the algorithm first simulates observations and computes decisions for *all* time periods in one iteration, before updating the value function approximations. This may lead to a faster convergence of the ADP algorithm. We test the use of double pass versus single pass in Section 5.6. More details on double pass can be found in [402].

5.4.3 Basis function approach

The main challenge is to design a proper approximation for the 'future' costs $\overline{V}_t^n\left(S_t^x\right)$ that is computationally tractable and provides a good approximation of the actual value to be able to find a suitable solution for the optimization problem of (5.21). There are various strategies available. A general approximation strategy that works well when the state space and outcome space are large, which generally will be the case in our formulated problem as discussed earlier in this section, is the use of basis functions. We explain the strategy in more detail below.

An underlying assumption in using basis functions is that particular features of a state vector can be identified, that have a significant impact on the value function. Basis functions are then created for each individual feature that reflect the impact of the feature on the value function. For example, we could use the total number of patients waiting in a queue and the waiting time of the longest waiting patient as two features to convert a post-state description to an approximation of the 'future' costs. We introduce

$$\begin{aligned} \mathcal{F} &= \text{set of features,} \\ \phi_f\left(S_t\right) &= \text{basis function for the feature } f \in \mathcal{F} \text{ for the state } S_t. \end{aligned}$$

We now define the value function approximations as

$$\overline{V}_t^n\left(S_t^x\right) = \sum_{f \in \mathcal{F}} \theta_f^n \phi_f\left(S_t^x\right), \qquad \forall t \in \mathcal{T}, \tag{5.24}$$

where θ_f^n is a weight for each feature $f \in \mathcal{F}$, and $\phi_f\left(S_t^x\right)$ is the value of the particular feature $f \in \mathcal{F}$ given the post-decision state S_t^x. The weight θ_f^n is updated recursively and the iteration counter is indicated with n. Note that (5.24) is a linear approximation, as it is linear in its parameters. The basis functions themselves can be nonlinear [402].

Features are chosen that are independently separable. In other words, each basis function is independent of the other basis functions. For our application, we make the assumption that the properties of each queue are independent from the properties of the other queues, so that we can define basis functions for each individual queue that describe important properties of that queue. Example features and basis functions are given in Table 5.3, and we will discuss our selection of basis functions, based on a regression analysis, in Section 5.6.

In each iteration, the value function approximations are updated, as given in (5.23). In the features and basis functions approach, this occurs through the recursive updating of θ_f^n. Several methods are available to update θ_f^n after each iteration. An effective approach is the recursive least squares method, which is a technique to compute the solution to a linear least squares problem [402]. Two types of recursive least squares methods are available. The least squares method for *nonstationary* data provides the opportunity to put increased weight on more recent observations, whereas the least squares method for *stationary* data puts equal weight on each observation.

The method for updating the value function approximations with the recursive least squares method for nonstationary data follows from [402] and is given in Appendix 5.8.2. In this method, the parameter α^n determines the weight on prior observations of the value. Setting α^n equal to 1 for each n would set equal weight on each observation, and implies that the least squares method for stationary data is being used. Setting α^n to values between 0 and 1 decreases the weight on prior observations (lower α^n means lower weight). We define the parameter α^n by

$$\alpha^n = \begin{cases} 1 & \text{, stationary} \\ 1 - \frac{\delta}{n} & \text{, nonstationary} \end{cases} \text{, where } n = 1, 2, ..., N. \tag{5.25}$$

where $1 - \frac{\delta}{n}$ is a function to determine α_n that works well in our experiments. We come back to setting α^n (and δ) in Section 5.6.1.

5.4.4 ILP to find a decision for large instances

In small, toysized problem instances, enumeration of the decision space to find the solution to (5.21) is possible. For real-life sized problem instances, this may become intractable, as explained in Section 5.3. In this case, we require an alternative strategy to enumeration. In case the basis functions are chosen to be linear with regards to the decision being made (or the resulting post-state description), we can apply ILP to solve (5.21). The ILP formulation is given in Appendix 5.8.1, and will be used in Section 5.6.3.

This concludes our theoretical explanation of our solution approach incorporating ADP and ILP. We have formulated an algorithm, an approximation approach involving features to estimate the 'future' costs, a method to update the approximation functions based on new observations, and an ILP formulation to determine the decisions. In the next section, we explain how these methods can be used to develop tactical plans.

5.5 Managerial implications

In the previous section, we developed the ADP algorithm to find the feature weights for the value function approximation. In this section we will explain how this ADP approach can be used to establish the tactical plans.

The ADP approach can be used to establish long-term tactical plans (e.g., three month periods) for real-life instances in two steps. First, N iterations of the ADP algorithm have to be performed to establish the feature weights for each time period $t \in \mathcal{T}$. Implicitly, by determining these feature weights, we obtain and store the value functions as given by (5.22) and (5.24) for each time period. Second, these value functions can be used to determine the tactical planning decision for each state and time period by enumeration of the decision space or the ILP as introduced in Section 5.4.4. In the next paragraph, we explain this in more detail.

For each time period in the time horizon, the value function approximations from the ADP approach are used to establish a tactical planning decision. State transitions are calculated by using the state in the current time period, the decision calculated with the value function approximations, the the expected number of patient arrivals and patient transfers between the queues. Subsequently, the value function approximations are used to determine the tactical planning decision for the new state in the following time period. This is repeated for all successive time periods until the end of the time horizon. The procedure may result in noninteger values for the post-state description, due to the patient transfer probabilities. To implement a tactical planning decision, it requires integer values for the number of patients to be served from each queue. While the ADP-model can contain noninteger values for each entry ($u \in \mathcal{U}$) in the waiting lists and the tactical planning decision, the integer restriction is on the *total* number of patients to be served from each queue (summed over all $u \in \mathcal{U}$ for one queue), like explained in [261]. In case only very few patients are included in the system, causing many queues to have 0 or 1 patient, developing a rounding procedure can be beneficial to ensure integer transfers between queues to obtain integer post-decisions states.

The actual tactical plan is implemented using a rolling horizon approach, in which for example tactical plans are developed for three consecutive months, but only the first month is actually implemented and new tactical plans are developed after this month. The rolling horizon approach is recommended for two reasons. First, the finite horizon approach, apart from the benefits it provides to model time dependent resource capacities and patient demand for example, may cause unwanted and short-term focused behavior in the last time periods. Second, recalculation of tactical plans after several time periods have passed, ensures that the most recent information on actual waiting lists, patient arrivals, and resource capacities is used. As time progresses during the execution of a tactical plan, more information becomes available on the actual realized number of patient arrivals and patient transfers between queues. This information can be used to align the tactical plan with the actual state of the system. The updated tactical planning decision can be calculated with the existing value function approximations and the actual state of the system. If resource requirements, resource capacities, arrival probabilities, or transfer probabilities change, the value functions have to be recalculated using the ADP algorithm.

In the following section, we will determine the features and various other settings for the ADP algorithm, and discuss the algorithm's performance for small and large instances.

5.6 Computational results

In this section, we test the ADP algorithm developed in Section 5.4. One of the methods prescribed by [402] is to compare the values found with the ADP algorithm with the values that result from the exact DP solution for small instances. We will use this method in Sections 5.6.1 and 5.6.2. In Section 5.6.3, we study the performance for large instances, where we compare the ADP algorithm with 'greedy' planning approaches to illustrate its performance. We first discuss setting for the ADP algorithm in Section 5.6.1.

5.6.1 Settings for the ADP algorithm

In this section, we use information from the DP solution to set the basis functions and the parameters for the ADP algorithm. The DP recursions and the ADP algorithm are programmed in Delphi, and for the computational experiments we use a computer with an Intel Core Duo 2.00 GHz processor and 2GB RAM.

First, we explain the parameters used for the problem instance to calculate the exact DP solution. Second, we explain the selected basis functions, and third, we explain general settings for the ADP algorithm.

Parameter settings

Some settings in the ADP algorithm, such as the basis functions and double pass or single pass, can be analyzed by comparing the results from the ADP approach with the results from the exact DP approach. The values of the DP can be calculated for extremely small instances only, due to the high dimensions in states and the expectation of the future value in the tactical planning problem. Only for these small instances, we have the opportunity to compare the ADP approximation with the exact DP values. We do not compare the calculated decision policies from both methods, but compare the obtained values. This comparison provides a clear evaluation of the quality of the approximation in the ADP approach for small instances. Since we use exactly the same ADP algorithm for small and real-life sized large instances, this also provides a strong indication of the quality of the approximation accuracy of the ADP approach for large instances (for which we cannot calculate the exact DP value).

For our experiments with small problems in this section, we use the following instance. The routing probabilities $q_{i,j}$ are: $q_{1,2} = 0.8, q_{2,3} = 0.8, q_{1,1} = q_{2,1} = q_{2,2} = q_{3,1} = q_{3,2} = q_{3,3} = 0$. Hence, a patient that is served at Queues 1 or 2 exits the system with probability 0.2, and a patient that is served at Queue 3 will always exit the system. Since there are 3 queues and there are 2 periods that a patient can wait: 0 and 1 time period, the state description for a time period t becomes:$[S_{t,1,0}, S_{t,1,1}, S_{t,2,0}, S_{t,2,1}, S_{t,3,0}, S_{t,3,1}]$. The exact DP-problem is restricted by limiting the number of patients that can be waiting in each queue to 7. The state holding the most patients is thus $[7, 7, 7, 7, 7, 7]$. If there are transitions or new arrivals that result in a number greater than 7 for a particular queue and waiting time, the number for that particular entry is set to 7. So if, after transitions, we obtain a State $[3, 1, 6, 8, 5, 4]$, this state is truncated to $[3, 1, 6, 7, 5, 4]$. In the same way, the states in the ADP are also restricted for comparison, even though this is not necessary. For large instances, when comparison with an exact DP solution is impossible, this state truncation method is not used. The state truncation may affect the ADP-approximation slightly, as it introduces nonlinearity around the edges of the state space. Using the number of time periods, the truncated state space, the number of queues, and the maximum number of time periods waiting, there are $8 \cdot 8^{(3 \cdot 2)} = 2{,}097{,}152$ entries to be calculated. The weights θ^n in the value function approximations are initialized to $\theta^0 = 1$ for all time periods, and the matrix $B^0 = \epsilon I$ as explained in Section 5.8.2. All other parameters are given in Table 5.2.

Parameter	*Description*	*Used values for testing*
T	The number of time periods	$8, \mathcal{T} = \{1, 2, \ldots, 7, 8\}$
R	The number of resource types	1
G	The number of care processes	1
e_g	The number of stages in each care process	$3, \mathcal{J} = \{1, 2, 3\}$
U	The number of periods waiting	$2, \mathcal{U} = \{0, 1\}$
$s_{j,r}$	Expected service time from resource type $r \in \mathcal{R}$ for a patient in queue $j \in \mathcal{J}$ in time units	1
$\eta r, t$	Resource capacity for resource type $r \in \mathcal{R}$ in time $t \in \mathcal{T}$ in time units	6
$\lambda_{1,t}$	Poisson parameter for new demand in the Queue 1 in time period $t \in \mathcal{T}$	5
$C_{t,j,u}$	Costs per patient waiting in a queue $j \in \mathcal{J}$, for $u \in \mathcal{U}$ time periods, in time period $t \in \mathcal{T}$	$\dfrac{(u+1)}{j}$

Table 5.2: The parameters that characterize the test instance.

Selection of basis functions

In Section 5.4, we introduced basis functions to approximate the future value of a particular decision in a particular state. Basis functions are used because of their relative simplicity. The selection of the features however, requires careful design. The challenge in this careful design is to make sure the choice of basis functions actually contributes to the quality of the solution. The basis functions can be observed as independent variables in the regression literature [402]. Hence, to select a proper set of basis functions that have significant impact on the value function, we use a regression analysis. In the regression analysis, the dependent variables are the computed values in the exact DP approach for the first time period, and the independent variables are the basis functions calculated from the state description.

Table 5.3 shows the regression results on various basis functions. The R^2 depicts the variation in the value that is explained by a regression model that uses the features as mentioned in the table as independent variables. The higher R^2, the better suitable the basis functions are for predicting (and thus approximating) the value. One can observe that the features with high level of detail about the state description score significantly better (are higher in the ordered table). Obviously, in addition to the basis functions in Table 5.3, a significant number of alternatives are available.

Features	**Basis functions**	**# vars**	R^2
The number of patients in queue j that are u periods waiting	$S_{t,j,u}, \forall j \in \mathcal{J}, \forall u \in \mathcal{U}, t = 1$	$\lvert\mathcal{J}\rvert \times \lvert\mathcal{U}\rvert$	0.954
Combination of the total number of patients in queue j and the sum of the number of time periods all patients are waiting in queue j	$\sum_{u=0}^{U} S_{t,j,u}$ and $\sum_{u=0}^{U} u \cdot S_{t,j,u}$, $\forall j \in \mathcal{J}, t = 1$	$2 \times \lvert\mathcal{J}\rvert$	0.954
Combination of the total number of patients in queue j and the longest waiting time currently in queue j	$\sum_{u=0}^{U} S_{t,j,u}$ and $\max_{u \in \mathcal{U}} S_{t,j,u}$, $\forall j \in \mathcal{J}, t = 1$	$2 \times \lvert\mathcal{J}\rvert$	0.954
The total number of patients in queue j	$\sum_{u=0}^{U} S_{t,j,u}$, $\forall j \in \mathcal{J}, t = 1$	$\lvert\mathcal{J}\rvert$	0.950
The sum of the number of time periods all patients are waiting in queue j	$\sum_{u=0}^{U} u \cdot S_{t,j,u}, \forall j \in \mathcal{J}, t = 1$	$\lvert\mathcal{J}\rvert$	0.879
The longest waiting time currently in queue j	$\max_{u \in \mathcal{U}} S_{t,j,u}, \forall j \in \mathcal{J}, t = 1$	$\lvert\mathcal{J}\rvert$	0.199
The average waiting time in queue j	$\frac{\sum_{u=0}^{U} u \cdot S_{t,j,u}}{\sum_{u=0}^{U} S_{t,j,u}}, \forall j \in \mathcal{J}, t = 1$	$\lvert\mathcal{J}\rvert$	0.033

Table 5.3: The basis functions and their R^2 regression on the given value function. In each regression, a constant is added as a variable. All R^2 values are obtained with significance of 0.000, indicating a good fit of the model. The third column '# vars' indicates the number of variables when the particular basis function is used.

For our ADP-model, we choose to use the features 'The number of patients in queue j that are u periods waiting' from the list in Table 5.3. These basis function explain a large part of the variance in the computed values with the exact DP approach ($R^2 = 0.954$), and the basis functions can be straightforwardly obtained from the state or post-state description. We choose these functions as they seize the highest level of detail on the state description, and therefore are likely to provide high quality approximations. In case there is no independent constant in the set of predictors $\mathcal{F}$ in a linear regression model, the model is forced to go through the origin (all dependent and independent variables should be zero at that point). This may cause a bias in the predictors. To prevent this bias, we add a constant term as one of the elements in $\mathcal{F}$. The feature weight θ_f^n may vary, but the feature value $\phi_f\left(S_t^x\right)$ of this constant is always 1, independent of the state S_t^x.

Double pass

In Section 5.4 we introduced the possible use of double pass, where the algorithm first steps through all time periods before updating the value functions. Our experiments confirm that double pass leads to faster convergence of the ADP algorithm than single pass.

To illustrate the effect, we compare the values from the exact DP solution with the found ADP values for 5000 randomly generated states. The ADP algorithm uses the recursive least squares method for nonstationary data, with $\delta = 0.95$ in (5.25). To evaluate the speed of the ADP algorithm, we display the number of iterations required until the algorithm is within 5% of the DP value, so either 95% or 105% of the DP value for a particular state. The average number of iterations before the ADP value is within 5% of the DP value for the 5000 states is 1131.0 when double pass is not used, and 100.3 when double pass is used. Hence, double pass is a significantly faster method to get an accurate approximated value. This effect can also be observed in Figure 5.1 for a single state. Also for other values of δ in (5.25), we find that the use double pass leads to faster convergence to the DP value. For the remainder of our experiments, we use double pass.

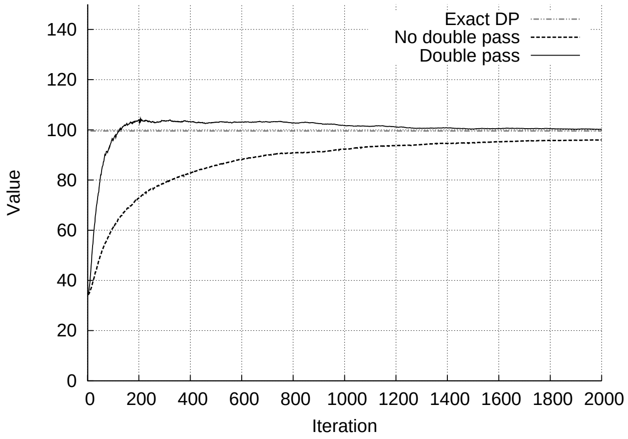


Figure 5.1: The values approximated with the ADP algorithm (Settings: recursive least squares for nonstationary data and $\delta = 0.95$) and calculated with the exact DP approach for initial state [2,7,5,1,7,4]. These graphs illustrate the significantly faster convergence when double pass is used.

Setting α

The parameter α is set in (5.25). When $\alpha = 1$ is chosen, the recursive least square method for stationary data is selected, and equal weight is given to each observation. Because the ADP algorithm is initialized with given arbitrary weights for θ^n and B^n, there is a 'warm-up period' before the weights are properly iterated and getting closer to the actual value. Hence, it seems useful to put less emphasis on the first observations, and more emphasis on later ones. To achieve this the recursive least squares method for nonstationary data is used, as explained in Section 5.4.3.

To find a good value for δ, we compare the values from the exact DP solution with the found ADP values for 5000 randomly generated states. We compare the number of iterations required until the algorithm is within 5% of the DP value, and the average difference between the ADP value and the DP value ((ADP value - DP value)/DP value) for various settings of δ. Figure 5.2 shows the results of these experiments. Note that δ cannot be equal to 1, because this would result in $\alpha^1 = 0$ and a division by 0 in (5.8.2) in the first iteration.

The recursive least squares method for *stationary data* requires 83 runs to reach a value within 5% of the DP value, and has an average difference of 2.2% (standard deviation of 2.7%) after 2500 iterations. The recursive least squares method for *nonstationary data* achieves similar average difference, but in fewer iterations. We explain the results for the recursive least square method for nonstationary data below.

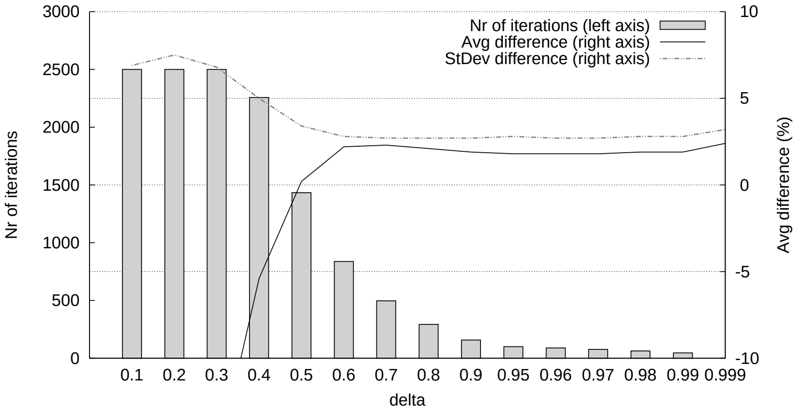


Figure 5.2: The average and standard deviation of the difference between the ADP value and the DP value (ADP value - DP value divided by the DP value) are depicted with lines (on the right axis) and the average number of runs required until the algorithm is within 5% of the DP value is given with block diagrams (on left axis). The number of iterations are bounded to 2500 for this experiment.

The left side of Figure 5.2 shows that when δ is closer to 1 (and α^1 is close to 0), the number of iterations required to reach a value within 5% of the DP value is significantly lower. This is due to the structure of (5.25), where a higher δ causes a lower α, which puts less emphasis on prior observations. In the first iteration, the prior observations are initializations, done by the modeler and independent of the instance or state. Hence, a 'warm-up period' is required to 'forget' the initializations and approximate the actual values. From the experiments it is clear that setting $\delta \geq 0.6$ decreases the warm-up period significantly, and results in stable performance on the average and standard deviation of the difference. Setting $\delta \leq 0.5$ results in unstability in the matrix operations of the nonstationary least squares method, resulting in strongly decreasing average difference (resulting in longer runtimes required to get to proper results). We obtain even more stringent conclusions if we observe the results for average and standard deviation of the difference: 200 iterations after the ADP algorithm finds values within 5% of the DP value. It appears that $\delta \geq 0.8$ gives the best results. Note that with the division of δ by the iteration number n in (5.25), α increases fast. After 10 iterations, $\alpha = 0.901$, with $\delta = 0.99$.

From the above it is clear that setting $\delta = 0.99$ results in stable, relatively good performance. For the remainder of our experiments, we use this setting which approximates the DP value within 5% in an average of 46.1 iterations and accurately with an average difference with the DP value of 1.9% and standard deviation of this difference of 2.8% after 2500 iterations for 5000 states.

5.6.2 Comparison of ADP, DP and greedy approaches for small instances

In this section, the values calculated with the ADP approach are compared with the exact DP solution and two greedy approaches.

Convergence of the ADP algorithm

We have calculated the ADP-algorithm for 5000 random states and found that the values found with the ADP algorithm and the value from the exact DP solution converge. For these 5000 random states, there is an average deviation between the value approximated with the ADP algorithm and the value calculated with the exact DP approach of 2.51%, with standard deviation 2.90%, after 500 iterations. This means the ADP algorithm finds slightly larger values on average than the exact DP approach. This may be caused by the truncated state space, as explained in Section 5.6.1.

For two initial states, Figure 5.3 illustrates that the calculated values with the ADP-algorithm (with $\delta = 0.99$ and double pass) converge to the values calculated with the exact DP approach as the number of iterations grow. In the first iterations, the ADP-values may be relatively volatile, due to the low value for α and thus the high impact of a new observation on the approximation. When the number of iterations increases, the weight on prior observations increases as α increases in (5.25), and the ADP approximations become less volatile.

The calculation time of the ADP algorithm is significantly lower than the calculation of the exact DP solution. Obtaining the DP solution requires over 120 hours. Calculating the ADP solution for a given initial state (with $N = 500$) takes on average 0.439 seconds, which is 0.0001% of the calculation time for the

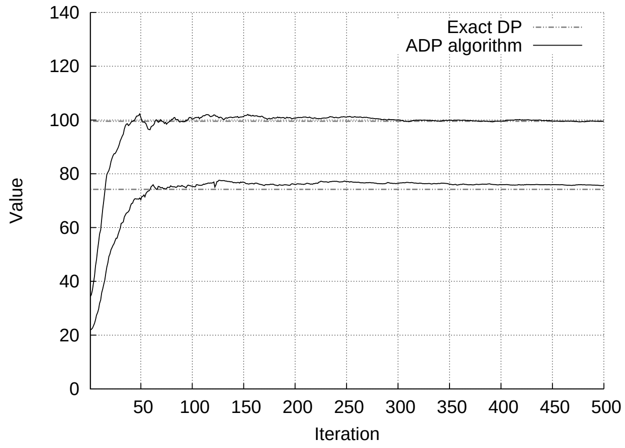


Figure 5.3: Example for two initial states. The values approximated with the ADP algorithm (with $\delta = 0.99$ and double pass) converge to the values from the exact DP approach.

exact DP solution. Obviously the calculation times depend on the used computation power, but these results indicate that solving a toy problem with the exact DP approach is already very time intensive, and solving such a problem with the ADP approximative approach is significantly faster.

Comparing the use of the feature weights, with DP and two greedy approaches

In the sections above, we have evaluated the performance of the ADP algorithm to find the feature weights. After the ADP algorithm has established the feature weights θ^n that accurately approximate the value associated with a state and a decision, these weights for all time periods are fixed and used to calculate planning decisions for each time period, like explained in Section 5.5. In this section, we evaluate the accuracy of the ADP approach by comparing the values obtained with the ADP approach, the DP approach, and two greedy approaches.

The two greedy approaches are rules that can be used to calculate a planning decision for a particular state and time period. We call the two approaches 'HighestNumberOfWaitingPatientsFirst' and 'HighestCostsFirst'. In the greedy approach 'HighestNumberOfWaitingPatientsFirst', the queue with the highest number of waiting patients is served until an other queue has the highest number of waiting patients, or until resource capacity constraints do not allow serving another patient of this queue anymore. After that, the next highest queue is served in the same way, until all queues are served and/or resource capacity constraints do not allow serving another patient anymore. In the greedy approach 'HighestCostsFirst', the queue with the highest costs (calculated with the cost function and the state description) is served until an other queue has the highest costs, or until resource capacity constraints do not allow serving another patient of this queue anymore. After that, the next highest queue is served in the same way, until all queues are served and/or resource capacity constraints do not allow serving another patient anymore. The ADP approach and the two greedy approaches can be used to calculate a tactical plan for a complete time horizon, following the steps explained in Section 5.5.

To compare the value calculated with the four approaches, we calculate a planning decision for each separate time period as follows. As a first step, we generate an initial state for the first time period in time horizon $\mathcal{T}$. We can find the exact DP value associated with this initial state from the already calculated DP solution. To establish the values for the ADP approach and the two greedy approaches, we use simulation as follows. We use the value function approximations from the ADP approach and the described methods from the greedy approaches, to establish a planning decision for the chosen initial state in the first time period. Then simulate the outcomes for patient transfers and patient arrivals. This leads to a particular state in the following time period for which we can establish the planning decision using the ADP approach and greedy approaches. These steps are repeated until the end of the time horizon $\mathcal{T}$. We sum the values associated with each state in each time period in the time horizon $\mathcal{T}$, to obtain the value for the initial state in the first time period. These values are then compared between the different approaches. By following this method, we can properly evaluate and compare the ADP approach in a wide range of possible outcomes for patient transfers and patient arrivals. When one aims to establish a tactical plan for a complete time horizon upfront, the random patient transfers and patient arrivals are replaced by the expectation for these processes, as explained in Section 5.5.

We randomly choose a set of 5000 initial states, that we each simulate with 5000 sample paths for the ADP approach and the two greedy approaches. We calculate the relative difference with the DP value for each of the 5000 initial states. Figure 5.4 displays the average over all initial states. The graph illustrates that the ADP approach provides a relatively accurate approximation for the value of a particular state, and the approximation is significantly better than two greedy approaches. The value resulting from the policy (the value function approximations) obtained with the ADP approach is very close to the values obtained with the optimal policy (found with the exact DP approach). Consequently, the fast and accurate ADP approach is very suitable to determine tactical planning decisions for each time period, and thus to establish a tactical plan for a complete time horizon following the steps explained in Section 5.5.

These results indicate that the ADP algorithm is suitable for the tactical resource capacity and patient admission planning problem.

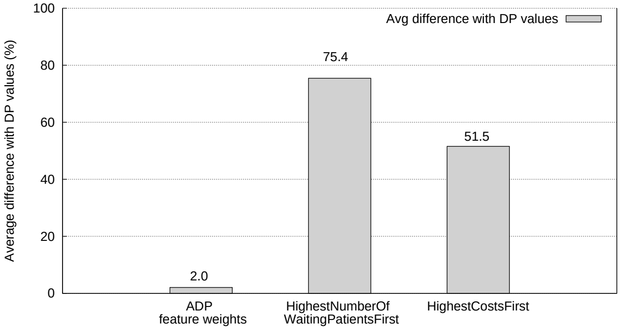


Figure 5.4: The average difference with the DP value when using the feature weights from ADP or the two greedy approaches to develop a tactical plan. The average value calculated with the ADP approach is 92.5.

5.6.3 Performance of the ADP algorithm for large instances

In the previous sections, we analyzed the performance of the ADP algorithm for small, toysized problems to compare the results with the DP approach. In this section, we investigate the performance of the ADP algorithm for large, real-life sized instances. Since for large instances, computation of the exact DP approach is intractable, we evaluate the performance of the ADP algorithm with the two greedy approaches as introduced in Section 5.6.2.

Parameters for large problem instances

As explained in Section 5.3, for large instances, the decision space becomes too large to allow for complete enumeration. Hence, we use an ILP to compute the optimal decision and it is given in Appendix 5.8.1. We use a CPLEX 12.2 callable library for Delphi to solve the ILP, and tolerate solutions with an integrality gap of 0.01%.

The parameters to generate the large instances are given in Table 5.4. When multiple entries are listed, we randomly choose one for each variable. For example, for each initial queue in a care process, we randomly pick the Poisson parameter for new demand from the set 1, 3, or 5 - similar to our instance generator in Chapter 4. The resource capacities $\eta_{r,t}$ for each resource $r \in \mathcal{R}$ and $t \in \mathcal{T}$ are selected from the given set, this means that we can for example have: $\eta_{1,1} = 1200$, $\eta_{1,2} = 0$, $\eta_{1,3} = 1200$, $\eta_{2,1} = 3600$, $\eta_{2,2} = 3600$, and $\eta_{2,3} = 1200$ can be 0. As real-life instances may have changing patient arrivals and changing resource capacities over time, we vary these parameters over the time periods for each queue and each resource respectively. In contrast with the exact DP approach, truncation of the state space is not required for the ADP algorithm, and we will not truncate the state space in the experiments for large instances. We truncate the initial starting state, to ensure that it is in line with the selected resource capacities and resource requirements. To generate the initial states, we

Parameter	*Description*	*Used values for testing*
$\lvert\mathcal{T}\rvert$	The number of time periods	$8, \mathcal{T} = \{1, 2, \ldots, 7, 8\}$
$\lvert\mathcal{R}\rvert$	The number of resource types	4
$\lvert\mathcal{G}\rvert$	The number of care processes	8
e_g	The number of stages in each care process	$\{3, 5, 7\}, \mathcal{J} = \{1, 2, \ldots, 40\}$
$\lvert\mathcal{U}\rvert$	The number of periods waiting	$4, \mathcal{U} = \{0, 1, 2, 3\}$
$s_{j,r}$	Expected service time from resource type $r \in \mathcal{R}$ for a patient in queue $j \in \mathcal{J}$ in time units (four value sets)	$\{10, 15, 20\}, \{30, 45, 60\},$ $\{100, 120, 140\},$ $\{200, 220, 240\}$
$\eta_{r,t}$	Resource capacity for resource type $r \in \mathcal{R}$ in time $t \in \mathcal{T}$ in time units	$\{0, 750, 1000, 1200, 1250,$ $2000, 3600, 5000, 8750,$ $9600, 10000, 17600\}$
$q_{i,j}$	The routing probabilities between queue $i, j \in \mathcal{J}$	$\{0, 0.25, 0.5, 0.75, 1\}$
$\lambda_{1,t}$	Poisson parameter for new demand in the first queue of each care process $g \in \mathcal{G}$ in time period $t \in \mathcal{T}$	$\{1, 3, 5\}$
$C_{t,j,u}$	Costs per patient waiting in a queue $j \in \mathcal{J}$, for $u \in \mathcal{U}$ time periods, in time period $t \in \mathcal{T}$	$\dfrac{(u+1)}{j}$

Table 5.4: The parameters that characterize the large test instances.

randomly pick the number of patients, for each queue and each number of time periods waiting, from the set $[0, 1, \ldots, 4]$. This set is bounded to align the initial state with the generated instance for the available resource capacity with the settings in the Table 5.4.

The weights θ^n in the value functions are initialized to $\theta^0 = 1$ for all time periods, and the matrix $B^0 = \epsilon I$ as explained in Appendix 5.8.2.

Comparison with greedy approaches

After running the ADP algorithm for $N = 100$ iterations, we fix the established feature weights θ^n, and use these to calculate tactical planning decisions in each time period. In this section, we compare the use of the feature weights calculated in the ADP algorithm with the two greedy approaches introduced in Section 5.6.2: 'HighestNumberOfWaitingPatientsFirst' and 'HighestCostsFirst'. For the comparison, we use the same simulation approach as explained in Section 5.6.2, but for larger instances, we simulate 30 initial states over 10 different generated instances. Similar to Section 5.6.2, we perform 5000 simulation runs per initial state. The relative difference between 'HighestNumberOfWaitingPatientsFirst' and the ADP approach is 50.7%, and the relative difference between 'HighestCostsFirst' and the ADP approach is 29.1%. The average value calculated with the ADP approach is 129.0. The lower average value from the ADP approach indicates that the ADP approach develops tactical plans resulting in lower costs than the two greedy approaches for large instances. The lower values indicate that the ADP approach supports and improves tactical planning decision making, and therefore we can conclude that the ADP approach is a suitable method to calculate a tactical plan for real-life sized instances.

Running the ADP algorithm for a given initial state (with $N = 100$) takes approximately 1 hour and 5 minutes for the large instance. This seems reasonable for finding the feature weights that approximate the value functions for 40 queues and 8 time periods. The feature weights that are calculated for the complete time horizon can be used to adjust the tactical plan in later time periods, as time progresses. Hence, the algorithm does not have to be run on a daily or even weekly basis. Since the algorithm converges fast, one may further decrease the number of iterations, resulting in lower runtimes.

Benefit of considering future costs through the ADP approach

Compared to the greedy approach 'HighestCostsFirst', the ADP approach offers an advantage by also considering costs of the future effects of the evaluated decision. The benefits of this advantage seem especially strong, when parameters such as resource capacity and patient arrivals change over time periods. The finite time horizon in the ADP approach allows for setting time dependent parameters for the problem instance, thereby ensuring that changing parameters over time are incorporated in the decision making. To illustrate the additional benefits of considering future costs in instances where parameters change over time, we have conducted the following experiment. We generated instances with Table 5.4, but we limited the number of resource types required for each queue to 1 resource type only. Next, for each resource type separately, we set the resource capacities for each time period to be equal. We obtain the following settings for resource capacities: $\sum_{t=1}^{|\mathcal{T}|} \eta_{1,t} = 6000$, $\sum_{t=1}^{|\mathcal{T}|} \eta_{2,t} = 10000$, $\sum_{t=1}^{|\mathcal{T}|} \eta_{3,t} = 30000$, and $\sum_{t=1}^{|\mathcal{T}|} \eta_{4,t} = 70000$. Leaving these total resource capacities (summed over the full time horizon) for each resource type constant, we set a number of entries for the resource capacities $\eta_{r,t}$ for $r \in \mathcal{R}$ and $t \in \mathcal{T}$ to zero. The total capacity for a resource type is kept constant by increasing the resource capacities for that resource type in time periods where the resource capacities are not set to zero. We compare three scenarios: setting 2, 8 and 14 entries of the total 32 entries for $\eta_{r,t}$ to zero in the complete time horizon $t = 1, \dots, |\mathcal{T}|$. For the comparison of the three scenarios, we use the same simulation approach as explained in Section 5.6.2, but we simulate 8 initial states. We perform 5000 simulation runs per initial state. In each of the three scenarios, we use the same instances and the same 8 initial states for our calculations and simulation. The results in Figure 5.5 illustrate that a higher variation of resource capacities in the time horizon, gives a higher benefit of using the ADP approach compared to the 'HighestCostsFirst' approach. These results indicate that the benefit of considering future costs in making a tactical planning decision increases when volatility in resource capacities increases.

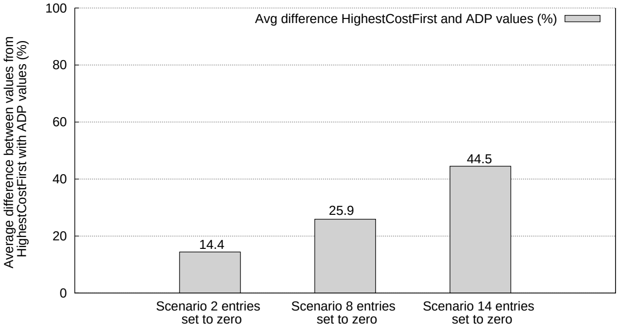


Figure 5.5: The average difference between the value calculated by using the feature weights from ADP and the value calculated by using the greedy approach 'HighestCostsFirst'. The average value calculated with the ADP approach is 165.1.

5.7 Conclusion

We provide a stochastic model for tactical resource capacity and patient admission planning problem in healthcare. Our model incorporates stochasticity in two key processes in the tactical planning problem, namely the arrival of patients and the sequential path of patients after being served. A Dynamic Programming (DP) approach, which can only be used for extremely small instances, is presented to calculate the exact solution for the tactical planning problem. We illustrate that the DP approach is intractable for large, real-life sized problem instances. To solve the tactical planning problem for large, real-life sized instances, we developed an approach within the frameworks of Approximate Dynamic Programming (ADP) and Mathematical Programming.

The ADP approach provides robust results for small, toyproblem instances and large, real-life sized instances. When compared with the exact DP approach on small instances, the ADP algorithm provides accurate approximations and is significantly faster. For large, real-life sized instances, we compare the ADP algorithm with the values obtained with two greedy approaches, as the exact DP approach is intractable for these instances. The results indicate that the ADP algorithm performs better than the two greedy approaches, and does so in reasonable run times.

We conclude that ADP is a suitable technique for developing tactical resource capacity and patient admission plans in healthcare. The developed model incorporates the stochastic processes for (emergency) patient arrivals and patient transitions between queues in developing tactical plans. It allows for time dependent parameters to be set for patient arrivals and resource capacity in order to cope with anticipated fluctuations in demand and resource capacity. The ADP model can also be used as for readjusting existing tactical plans, after more detailed information on patient arrivals and resource capacities are available (for example when the number of patient arrivals were much lower than anticipated in the last week). The developed ADP model is generic, where the objective function can be adapted to include particular targets, such as targets for access times, monthly 'production' or resource utilization. Also, the method can be extended with additional constraints and stochastic elements can be added to suit the hospital situation at hand. It can potentially be used in industries outside healthcare. Future research may involve these extensions, and may also focus on further improving the approximation approach, developing tactical planning methods to adjust a tactical plan when it is being performed, or using the ADP approach for other tactical planning objectives.

5.8 Appendix

5.8.1 The ILP for large instances

$$\min_{x_t \in \mathcal{X}_t(S_t)} \left(C_t\left(S_t, x_t\right) + \sum_{f \in \mathcal{F}} \theta_f \phi_f\left(S_t^x\right) \right),$$

subject to

$$S_{t,j,0}^x = \sum_{i \in \mathcal{J}} \sum_{u \in \mathcal{U}} q_{i,j} x_{t,i,u} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{5.26}$$

$$S_{t,j,U}^x = \sum_{u=U-1}^{U} \left(S_{t,j,u} - x_{t,j,u}\right) \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, \tag{5.27}$$

$$S_{t,j,u}^x = S_{t,j,u-1} - x_{t,j,u-1} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U} \backslash \left\{0, U\right\}, \tag{5.28}$$

$$x_{t,j,u} \leq S_{t,j,u} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U}, \tag{5.29}$$

$$\sum_{j \in \mathcal{J}^r} s_{j,r} \sum_{u \in \mathcal{U}} x_{t,j,u} \leq \eta_{r,t} \qquad \forall r \in \mathcal{R}, t \in \mathcal{T}, \tag{5.30}$$

$$x_{t,j,u} \in \mathbb{Z}_+ \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, u \in \mathcal{U} \tag{5.31}$$

Constraints (5.26) to (5.28) stipulate that the waiting list variables are consistent. Constraint (5.29) stipulates that not more patients are served than the number of patients on the waiting list. Constraint (5.30) assures that the resource capacity of each resource type $r \in \mathcal{R}$ is sufficient to serve all patients, and Constraint (5.31) is an integrality constraint.

5.8.2 Updating method based on regressive least squares for nonstationary data

The method for updating the value function approximations with the recursive least squares method for nonstationary data is explained in detail in [402]. The equations used in our solution approach are given below.

The weights θ_f^n, for all $f \in \mathcal{F}$, are updated each iteration (n is the iteration counter) by

$$\theta_f^n = \theta_f^{n-1} - H_n \phi_f\left(S_t^x\right)\left(\overline{V}_{t-1}^{n-1}\left(S_{t-1}^x\right) - \widehat{v}_t^n\right), \qquad \forall f \in \mathcal{F},$$

where H^n is a matrix computed using

$$H^n = \frac{1}{\gamma^n} B^{n-1}.$$

B^{n-1} is an $|\mathcal{F}|$ by $|\mathcal{F}|$ matrix, which is updated recursively using

$$B^n = \frac{1}{\alpha^n}\left(B^{n-1} - \frac{1}{\gamma^n}\left(B^{n-1}\phi\left(S_t^x\right)\left(\phi\left(S_t^x\right)\right)^T B^{n-1}\right)\right).$$

The expression for γ^n is given by

$$\gamma^n = \alpha^n + \phi\left(S_t^x\right)^T B^{n-1} \phi\left(S_t^x\right).$$

B^n is initialized by using $B^0 = \epsilon I$, where I is the identity matrix and ϵ is a small constant. This initialization especially works well when the number of observations is large [402]. The parameter α^n determines the weight on prior observations of the value, and it is discussed in Sections 5.4 and 5.6.1.

CHAPTER 6

Process redesign and room requirements in outpatient clinics

6.1 Introduction

Demand for outpatient care is growing as a result of increasingly effective ambulatory care treatments and the overall growth of healthcare demand. Hence, managers of outpatient clinics are becoming increasingly aware of the importance of the efficient use of scarce resources, particularly doctor's time and facility space [113]. This results in many hospitals redesigning or rebuilding their outpatient clinics (e.g., the hospitals RIVAS Gorinchem, Reinier de Graaf Gasthuis, Haga Ziekenhuis, and Groene Hart Ziekenhuis).

In many hospitals, outpatient clinics are organized such that doctors remain in one consultation room, while patients visit for individual consultation. In this classic design, each doctor occupies one consultation room, which often doubles as the doctor's office [503]. Patients wait in the waiting room until the doctor is available, and then enter the doctor's office for the consultation. We label this classic design Patient-to-Doctor policy (PtD-policy).

In a different approach, patients prepare themselves in separate, individual consultation rooms. Each patient is then visited by the doctor, who travels from room to room. We label this approach as Doctor-to-Patient policy (DtP-policy). The DtP-policy offers a potential decrease in total service time, given that doctors do not have to be present for patient preparation activities that require a consultation room, but do not require a doctor. We characterize these activities as pre-consultation (e.g., traveling to the room, undressing, blood pressure measures) and post-consultation (e.g., dressing, making appointments, leaving the room, cleaning the room). Nurses or assistants may be involved in these activities. In the DtP-policy, the doctor experiences travel time between each consultation, whilst traveling from room to room. Figure 6.1 illustrates the PtD-policy and the DtP-policy with two rooms.

In search of efficiency improvements in the outpatient clinic, managers are reconsidering the design of the outpatient clinic. Since differences in the outpatient process exist between different (specialties within) outpatient clinics, a policy efficient for one clinic may not be efficient for another. For example, when pre-consultation and/or post-consultation time are non-existent or rela-

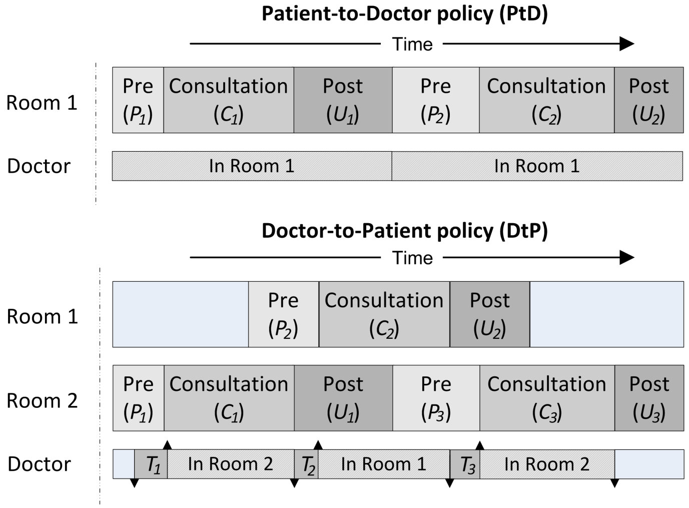


Figure 6.1: An illustration of the PtD-policy and the DtP-policy with two rooms. Pre-consultation, consultation and post-consultation for patient n is indicated by P_n, C_n and U_n respectively. T_n indicates the travel time of the doctor to patient n

tively low in comparison with consultation time in a particular outpatient clinic process (e.g., psychology consultations), the DtP-policy may not result in savings of doctor time. Hence, before deciding to adopt a particular policy, it is important that an outpatient clinic manager understands which policy is most efficient and how many consultation rooms are required for the particular outpatient clinic's parameter settings. To support this decision making, we provide analytical models that can be used to rationally compare the two policies on several performance measures and to determine the required number of consultation rooms in a particular outpatient clinic setting. Our models provide quantitative arguments that facilitate a rational discussion about a proposed decision with stakeholders (e.g., hospital boards, doctors).

In queueing terminology, the PtD-policy resembles a $G/G/1$ queueing model, under the assumption that patients are seen on a first-come, first-served basis (FCFS). The DtP-policy seems to resemble a polling system [324, 461], where the server travels between multiple customer queues. However, as the outpatient clinic has a single queue of patients only, this analogy can not be applied to evaluate the DtP-policy. The queueing model that most closely resembles the DtP-policy is a Production Authorization Card system (PAC-system). In a PAC-system, the number of jobs (patients) at a station (the doctor) is bounded by the number of PACs (rooms). Therefore, the departure of a job (patient exits) initiates demand for new jobs (a patient enters the empty room). The PAC-system, and thus the DtP-policy, is a typical 'pull' system, used in popular management philosophies such as Just-In-Time and Kanban. The PtD-policy is a 'push' system, whereby patients arrive in a buffer (the waiting room) and are pushed through the system. For results in queueing theory on push and pull systems, see [54, 298]. The exact and approximative solution approaches for PAC-systems are based on steady state queueing results [73]. Since appointment schedules have a finite number of customers, and thus do not reach steady state [418, 254, 89], these solution approaches are inappropriate to analyze the DtP-policy and the PtD-policy.

There is a significant body of literature on resource planning in outpatient clinics, particularly related to outpatient scheduling. For a comprehensive review of the literature on outpatient scheduling, see [89]. The design and capacity dimensioning of outpatient clinics has received less attention in the literature. Different process set-ups for an emergency department are compared with a Multi-Class Open Queueing Network (MC-OQN) in [277]. The authors conclude that parallel processing of, for example, treatment and diagnostic tests, rather than serial processing, results in a shorter patient sojourn time under certain conditions. Other examples of successful process redesigns in outpatient clinics are [535, 97]. Simulation is used to find the required number of examination rooms in an outpatient clinic [113], an obstetrics outpatient center [269], a radiology department [279], an emergency department [15, 155] and a family practice [458, 457]. A combination of simulation and function estimation is used to design a transfusion center [122]. All described papers use simulation to find the required number of rooms for a specific setting. In this chapter, we develop analytical models of a generic outpatient clinic to compare the PtD-policy with the DtP-policy, and to determine the required number of rooms in the DtP-policy.

The performance measures we consider are doctor utilization, access time, and patient waiting time. Doctor utilization is the fraction of time the doctor is actually consulting a patient. Access time is the time between the request for an appointment and the realization of the appointment. Patient waiting time is the time between the scheduled starting time of the appointment and the actual starting time of the appointment. Increased doctor utilization leads to decreased access time, but also to increased patient waiting time, given that more patients are scheduled per time unit. Managers of outpatient clinics strive for high doctor utilization and low access times, even at the cost of some patient waiting time [58]. This may be explained by three factors: doctors are considered expensive resources, service level agreements on access times may exist and low access times may attract more patients.

This chapter is organized as follows. Section 6.2 introduces the model and presents expressions for the recursion of the time that the doctor finishes a consultation in both the PtD-policy and the DtP-policy. Section 6.3 compares these recursions analytically, and introduces an expression for the fraction of consultations that are in immediate succession, to calculate the required number of consultation rooms in the DtP-policy. Section 6.4 presents the results for a range of distributions and parameters, and a case study at a medium-size hospital.

Section 6.5 discusses main conclusions.

6.2 Model

In Sections 6.2.1 and 6.2.2, we develop expressions for the time the doctor finishes the consultation of the n-th patient in the PtD-policy (F_n) and the DtP-policy (F'_n). These expressions are used in Section 6.3.1, to compare the PtD-policy and the DtP-policy, and to develop an expression for the fraction of consultations that are in immediate succession to calculate the required number of rooms in the DtP-policy. We first introduce notation and assumptions that apply to both policies.

Assume that at time zero the doctor is free. Patients arrive according to a stochastic process at time points $\{A_n, n = 1, 2, ..., N\}$, thus the first patient arrives at time A_1. The n-th patient leaves the system after finishing pre-consultation (P_n), consultation with the doctor (C_n) and post-consultation (U_n), where P_n, C_n, U_n are random variables with $P_n, C_n, U_n \geq 0$, for $n = 1, 2, ..., N$. The n-th patient leaves at time $D_n = F_n + U_n$ in the PtD-policy, and at time $D'_n = F'_n + U_n$ in the DtP-policy. Let R be the number of rooms and T_n the random variable for the doctor's travel time to the n-th patient. We assume that $T_n, n = 1, 2, ..., N$, is an independent and identically distributed (i.i.d.) sequence of random variables, thus not connected to the sequence with which the doctor visits the rooms, and that the travel time of the doctor (T_n) is not longer than the travel time of the patient (included in P_n). We base the latter assumption on our experience that consultation rooms are located adjacently and the waiting room is at a further distance.

Assumption 6.1. *$T_n \leq P_n$, for $n = 1, 2, ..., N$.*

Throughout this chapter, inequalities in expressions and equations for random variables are with probability one, i.e., $T_n \leq P_n \Leftrightarrow Pr(T_n \leq P_n) = 1$. The following two assumptions imply that patients enter rooms and are consulted by the doctor in the sequence they arrive.

Assumption 6.2. *Patients enter rooms on an FCFS basis. Hence, when a room is empty, the patient who has waited the longest in the queue is admitted.*

Assumption 6.3. *The doctor consults patients on an FCFS basis, thus in the sequence in which the patients enter rooms.*

The following assumption deals with the doctor's travel in the DtP-policy after finishing consultation with a patient.

Assumption 6.4. *When the doctor finishes consultation with the $(n-1)$-th patient, and the n-th patient has not entered a room yet, the doctor travels to an empty room when one becomes available, and waits there for the n-th patient.*

Under Assumption 6.4, the doctor either knows which room to go to after finishing consultation of a patient, or the doctor waits until a patient leaves and a room becomes available.

6.2.1 Recursion of the time the doctor finishes a consultation in the PtD-policy

We obtain the following expression for the recursion of the time that the doctor finishes the consultation of a patient in the PtD-policy.

Lemma 6.5. $F_n = \max\{A_n, F_{n-1} + U_{n-1}\} + P_n + C_n$, *where* $n = 1, 2, ..., N$ *and* $F_0 = 0$.

We prove Lemma 6.5 in Appendix 6.6.1.

6.2.2 Recursion of the time the doctor finishes a consultation in the DtP-policy

Since the processes in the DtP-policy and the PtD-policy are identical when $R = 1$, we focus on $R > 1$ in the DtP-policy. The lemma presented in this section thus holds for any $R > 1$.

The exiting time for patients may not be in the same order as the arrivals, because it is possible for the $(n+1)$-th patient to exit before the (n)-th patient (due to the randomness in U_n). To accommodate this, we define the $s(n)$-th patient as the patient who is succeeded by the n-th patient in a room. Thus when the $s(n)$-th patient exits a room, the n-th patient enters that room. We obtain the following expression for the recursion of the time that the doctor finishes the consultation of a patient in the DtP-policy.

Lemma 6.6. $$F'_n = \begin{cases} \max\{A_n + P_n, F'_{n-1} + T_n\} + C_n & \text{, if } n \le R \\ \max\{\max\{F'_{s(n)} + U_{s(n)}, A_n\} + P_n, & \\ \qquad \max\{F'_{s(n)} + U_{s(n)}, F'_{n-1}\} + T_n\} + C_n & \text{, if } n > R \end{cases},$$
where $n = 1, 2, ..., N$ *and* $F'_0 = 0$.

We prove Lemma 6.6 in Appendix 6.6.2.

6.3 Performance evaluation

We use Lemmas 6.5 and 6.6 obtained in Section 6.2 to compare the DtP-policy with the PtD-policy in Section 6.3.1. In Section 6.3.2 we develop an expression for the fraction of consultations that are in immediate succession to calculate the required number of rooms in the DtP-policy.

6.3.1 Analytical comparison of the recursion of the finishing time for the doctor under both policies

In this section, we show that the time that the doctor finishes the consultation of a patient in the DtP-policy is not later than the time the doctor finishes consultation with that patient in the PtD-policy, under Assumptions 6.1, 6.2, 6.3 and 6.4, i.e.,

Theorem 6.7. *$F'_n \leq F_n$, for $n = 1, 2, ..., N$.*

Since $F'_n \leq F_n$, for $n = 1, 2, ..., N$, this also means $D'_n \leq D_n$, for $n = 1, 2, ..., N$. Therefore, the departure of the n-th patient never occurs later in the DtP-policy than the departure of that same patient in the PtD-policy.

We prove Theorem 6.7 in Appendix 6.6.3.

Remark 6.8. *Under our FCFS assumptions, Assumptions 6.2 and 6.3, the modeled DtP-policy performs worse than a real-life DtP-policy, where the doctor may consult patients according to a dynamic sequence. The FCFS ordering may result in a waste of doctor capacity, since the doctor may be waiting for the n-th patient to finish pre-consultation, while the $(n + 1)$-th patient is already finished with pre-consultation. Additionally, Assumption 6.4 also causes waste of capacity, since the doctor waits until knowing which room to travel to next. This suggests that the ordering of the DtP-policy and the PtD-policy also holds when Assumptions 6.2, 6.3 and 6.4 are relaxed.*

Remark 6.9. *When Assumption 6.1 is replaced by the weaker assumption $Pr(T_n \leq s) \geq Pr(P_n \leq s)$, for $n = 1, 2, \ldots, N$, we can show that $Pr(F'_n \leq t) \geq Pr(F_n \leq t)$, for $n = 1, 2, \ldots, N$, which implies that $\mathbb{E}F' \leq \mathbb{E}F$.*

6.3.2 Analytical expression to calculate the required number of rooms

In a PtD-policy, the required number of rooms per doctor is one. In a DtP-policy, the required number of rooms is more than one. In this section we develop an expression for the fraction of consultations that are in immediate succession to calculate the required number of rooms in the DtP-policy.

To minimize access time of patients, healthcare managers aim to minimize idle time experienced by the doctor. To this end, the doctor's wait for the next available patient should be minimized [239], or in other words, the *fraction of consultations that take place in immediate succession* should be maximized. After leaving a room, the doctor should return to this room after the next patient has finished pre-consultation. During the time that the doctor is away from a specific room ($U_{s(n)} + P_n$), the doctor performs $R - 1$ consultations in the other rooms and R travels (including the travel to the n-th patient). Hence, we obtain the following expression, where the number of rooms (R) is chosen such that the fraction of consultations in immediate succession is larger than α, where $0 \leq \alpha \leq 1$.

$$Pr(\sum_{k=n-R}^{n-1} C_k + \sum_{k=n-R}^{n} T_k \geq U_{s(n)} + P_n) \geq \alpha. \tag{6.1}$$

Examples
We evaluate Equation (6.1) for Gamma and Normal distributed service times. The average duration of a process is given by μ_i and its variance is given by σ_i^2, where $i \in \{P, C, U, T\}$.

The Gamma distribution is a frequently reported distribution for outpatient clinic consultation times [89]. Let the pre-consultation, the post-consultation, and the travel times be deterministic, and the consultation times be i.i.d. Gamma distributed. The convolution of v i.i.d. Gamma distributed variables with parameters (k, θ) is again a Gamma distribution with parameters ($v \cdot k$, θ). Hence, the number of rooms, R, is obtained from

$$\int_{U+P-R\cdot T}^{\infty} x^{(R-1)\cdot(k-1)} \frac{e^{-\frac{x}{\theta}}}{\theta^{(R-1)\cdot k} \cdot \Gamma(R \cdot k)} \, dx \geq \alpha, \tag{6.2}$$

where $\theta = \frac{\sigma_C^2}{\mu_C}$ and $k = \frac{\mu_C}{\theta}$ are parameters of the Gamma distribution and $\Gamma(a)$ is the standard Gamma function with parameter a.

When all service processes are i.i.d. Normal distributed, its convolution results in a Normal distribution with parameters (μ, σ). Hence, the number of rooms, R, is obtained from

$$\int_{0}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2} \, dx \geq \alpha, \tag{6.3}$$

where $\mu = (R-1) \cdot \mu_C + R \cdot \mu_T - \mu_U - \mu_P$ and $\sigma^2 = (R-1) \cdot \sigma_C^2 + R \cdot \sigma_T^2 + \sigma_P^2 + \sigma_U^2$.

6.4 Results

Sections 6.4.1 and 6.4.2 describe the comparison of the two policies and the calculation of the required number of rooms. Section 6.4.3 describes the application of our methods at a pediatric outpatient clinic.

6.4.1 Comparison of the PtD-policy and the DtP-policy

In Theorem 6.7, we showed that the doctor finishes consultation with a patient earlier in the DtP-policy than in the PtD-policy under Assumptions 6.1, 6.2, 6.3

and 6.4. Hence, more patients can be consulted per time unit in the DtP-policy. In Remark 6.8, we indicated that the ordering of the DtP-policy and the PtD-policy may remain the same when Assumptions 6.2, 6.3 and 6.4 are relaxed. Below, we use discrete-event simulation to study the ordering when Assumption 6.1 is relaxed.

The discrete-event simulation is a model of an outpatient clinic, where a consultation session lasts eight hours per day and patients arrive at the time they are scheduled. The Bailey-Welch rule [17] is used for the patient schedule. The rule states that when blocks of the size of the expected consultation time are used to schedule the patients, the last block is deleted and the first block holds two patients. We assume a coefficient of variation ($CV = \frac{\mu}{\sigma}$) of 0.6, which is within the range of 0.35 to 0.85 reported in the literature [89]. The length of each simulation run is one business day. With the replication/deletion approach [319], we find that 1000 replications (days) appear to be sufficient for a confidence level of 99.9% with a relative error of 0.1% with respect to the number of consultations per week.

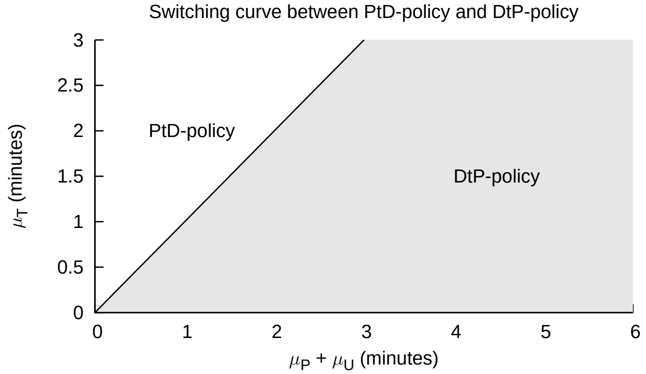


Figure 6.2: The switching curve between the DtP-policy and the PtD-policy, where all processes are Gamma distributed with $CV = 0.6$. A policy is superior to the other policy, when average doctor utilization is higher. The number of rooms is chosen with Equation (6.1), with $\alpha = 0.90$

Figure 6.2 shows the switching curve when all processes are Gamma distributed. The switching curve from the PtD-policy to the DtP-policy depends on the ratio of doctor travel time to pre-consultation time and post-consultation time, and is insensitive to changes in the average consultation time and the CV. Also, the ratio pre-consultation to post-consultation has only negligible impact on the choice for a policy; it is their sum that influences the superiority of a policy.

When ρ is varied ($\rho = \lambda E[C]$, where λ is the number of patients scheduled per time unit, and $E[C]$ is the expected consultation time), the switching curve for the DtP-policy is identical to the curve in Figure 6.2 for $\rho \geq 0.7$. For

$\rho < 0.7$, the DtP-policy performs better at even higher average travel times, but the potential benefit of the DtP-policy is relatively low, as can be seen in Figure 6.3. Also, Figure 6.3 illustrates that the potential benefit of the DtP-policy decreases as the ratio of consultation time versus pre-consultation time and post-consultation time decreases. This is caused by the fact that decreasing pre-consultation and post-consultation time per patient while keeping consultation time constant, leads to lower potential savings of doctor time.

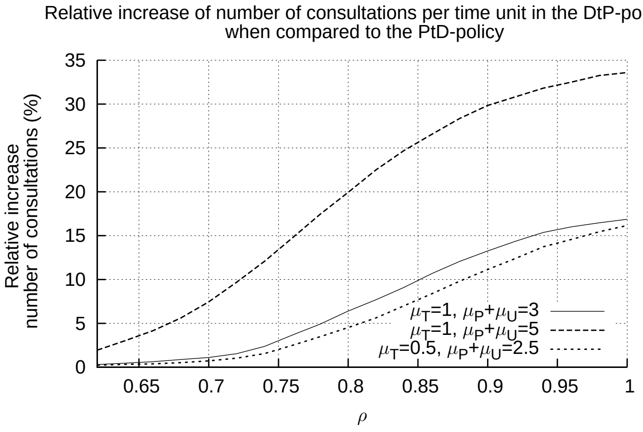


Figure 6.3: The effect of varying ρ on the relative increase of the number of consultations per time unit in the DtP-policy, when compared to the PtD-policy. All processes are Gamma distributed with $CV = 0.6$, $\mu_C = 10$, and $R = 2$

6.4.2 Evaluation of the required number of rooms

The fraction (P_{succ} in Table 6.1) of consultations that are in immediate succession, left-hand side in Equation (6.1), is evaluated numerically with Monte Carlo simulation for the Gamma, Lognormal and Exponential distribution. For the Normal distribution, we use Equation 6.3. To compare the fraction with a performance measure, such as doctor utilization (*Util.* in Table 6.1), we use the discrete-event simulation introduced in Section 6.4.1. Table 6.1 presents both the fraction results and the doctor utilization for a given number of rooms, and it shows the effect of choosing a certain α. For example, when $\alpha = 0.90$, four rooms are required when $\mu_P = \mu_U = 9$ and all processes Lognormal distributed. In that case, the doctor utilization found with the simulation is 90.4%. The results in Table 6.1 show that doctor utilization increases with the fraction of consultations that are in immediate succession.

The stochastic nature of the consultation process should be considered when the required number of rooms is determined. When all processes are considered to be deterministic, three rooms are required in the example of Figure 6.4. The

	$(\mu_P, \mu_C,$	Gamma		Lognormal		Normal		Exponential	
R	$\mu_U)$	P_{succ}	*Util.*	P_{succ}	*Util.*	P_{succ}	*Util.*	P_{succ}	*Util.*
2	(3,15,3)	0.920	91.2%	0.946	91.4%	0.879	91.1%	0.781	86.6%
3	(3,15,3)	0.998	91.6%	0.996	91.6%	0.981	91.8%	0.960	87.6%
4	(3,15,3)	1.000	91.6%	0.997	91.6%	0.997	91.8%	0.993	87.7%
2	(3,15,6)	0.800	89.8%	0.823	90.3%	0.791	89.6%	0.674	84.6%
3	(3,15,6)	0.985	91.5%	0.979	91.6%	0.963	91.7%	0.909	87.4%
4	(3,15,6)	0.999	91.6%	0.988	91.6%	0.993	91.8%	0.976	87.7%
2	(3,15,9)	0.675	87.0%	0.687	87.6%	0.680	87.0%	0.591	81.9%
3	(3,15,9)	0.953	91.4%	0.945	91.5%	0.933	91.5%	0.852	87.0%
4	(3,15,9)	0.995	91.6%	0.973	91.6%	0.987	91.8%	0.949	87.6%
2	(6,15,3)	0.800	89.4%	0.823	89.7%	0.791	89.1%	0.674	84.1%
3	(6,15,3)	0.985	91.0%	0.979	91.0%	0.963	91.2%	0.909	86.8%
4	(6,15,3)	0.999	91.0%	0.988	91.0%	0.993	91.3%	0.976	87.0%
2	(6,15,6)	0.671	86.8%	0.684	87.3%	0.685	86.6%	0.580	81.5%
3	(6,15,6)	0.958	90.8%	0.952	90.9%	0.937	91.0%	0.853	86.4%
4	(6,15,6)	0.997	91.0%	0.978	91.0%	0.988	91.2%	0.953	87.0%
2	(6,15,9)	0.551	83.0%	0.550	83.5%	0.571	83.0%	0.509	78.2%
3	(6,15,9)	0.911	90.5%	0.903	90.6%	0.896	90.6%	0.794	85.6%
4	(6,15,9)	0.989	90.9%	0.961	91.0%	0.978	91.2%	0.921	86.9%
2	(9,15,3)	0.675	86.1%	0.687	86.6%	0.680	85.9%	0.591	81.0%
3	(9,15,3)	0.953	90.2%	0.945	90.3%	0.933	90.4%	0.852	85.8%
4	(9,15,3)	0.995	90.3%	0.973	90.4%	0.987	90.7%	0.949	86.4%
2	(9,15,6)	0.551	82.6%	0.550	83.1%	0.571	82.5%	0.509	77.8%
3	(9,15,6)	0.911	89.9%	0.903	90.0%	0.896	89.9%	0.794	85.0%
4	(9,15,6)	0.989	90.3%	0.961	90.4%	0.978	90.6%	0.921	86.3%
2	(9,15,9)	0.449	78.3%	0.434	78.7%	0.466	78.4%	0.446	74.3%
3	(9,15,9)	0.853	89.2%	0.844	89.4%	0.843	89.2%	0.739	84.0%
4	(9,15,9)	0.975	90.3%	0.944	90.4%	0.963	90.6%	0.887	86.1%
5	(9,15,9)	0.996	90.4%	0.960	90.4%	0.992	90.7%	0.953	86.4%

Table 6.1: The results for the fraction of consultations that are in immediate succession, where $\mu_T = 1$ and $CV = 0.6$ for the Gamma, Lognormal and Normal distributions, and $CV = 1$ for the Exponential distribution. The half-length of the 99.9% confidence interval for the doctor utilization is between 0.011% and 0.096%

graph shows that more rooms are required when CV increases.

6.4.3 Case study at a medium-sized hospital

We apply our methods at the pediatric outpatient clinic of the 'Groene Hart Ziekenhuis' hospital (GHZ) in Gouda, the Netherlands. GHZ has 450 beds and over 2000 employees [198], and the seven doctors at the pediatric outpatient clinic consult 12000 patients per year. We focus on a single doctor, who consults patients for nine hours per week. Patients are planned in time slots of 15 min-

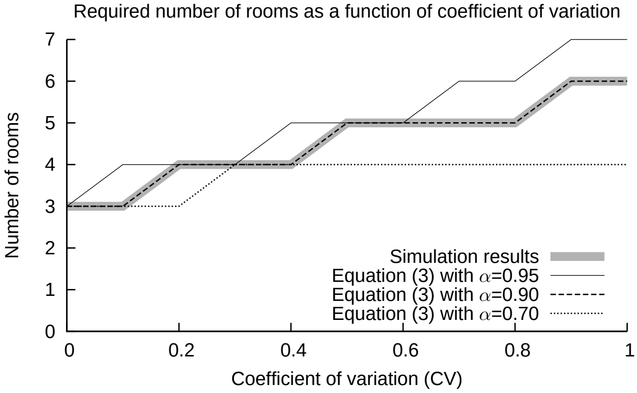


Figure 6.4: The required number of rooms when CV increases. All processes are Gamma distributed, with $\mu_P = 10$, $\mu_C = 10$, $\mu_U = 10$, $\mu_T = 1$. The number of rooms in the simulation is chosen such that the doctor utilization can not increase more than 0.5% with an additional room. The simulation results coincide with Equation (6.1), where $\alpha = 0.90$

utes. The parameters in Table 6.2 are the result of extensive data gathering. We know that the DtP-policy outperforms the PtD-policy if we assume that the doctor's travel time is always lower than the patient's travel time. The simulation results indicate that the DtP-policy outperforms the PtD-policy, when the average travel time does not exceed 6 minutes. In estimating the number of rooms, we assume that travel time is 0.5 minute on average, with $CV = 0.6$. Table 6.3 shows that three rooms are required, if $\alpha = 0.90$. The fraction of consultations that are in immediate succession (P_{succ} in Table 6.3) is evaluated numerically with Monte Carlo simulation, and the doctor utilization (*Utilization* in Table 6.3) is found with our discrete-event simulation.

Process	*Distribution*	*Average*	*Std. deviation*
Pre-consultation	Gamma	5.90	6.06
Consultation	Gamma	15.57	8.12
Post-consultation	-	-	-

Table 6.2: Duration parameters (minutes), retrieved from data for 1875 patients of the pediatric outpatient clinic in 2009

6.5 Conclusion

Inspired by the hospitals ‘RIVAS Gorinchem’, ’Reinier de Graaf Gasthuis’ and ‘Groene Hart Ziekenhuis’, which were in the process of redesigning their out-

Number of rooms	P_{succ}	*Utilization*
2	0.883	92.5%
3	0.984	93.3%
4	0.998	93.3%

Table 6.3: Results to determine the required number of rooms in the case study. The half-length of the 99.9% confidence interval for the doctor utilization is between 0.053% and 0.057%

patient clinic, this chapter has developed analytical and simulation models to compare different parameter settings in two policies for the organization of outpatient clinics. In the first policy, doctors remain in one consultation room, while patients visit for consultation. We call this the Patient-to-Doctor policy (PtD-policy), and in this policy, the doctor attends the complete patient process: pre-consultation, consultation and post-consultation. In the second policy, patients prepare themselves in individual consultation rooms, with or without the aid of a nurse, while the doctor travels from room to room. We call this the Doctor-to-Patient policy (DtP-policy), and in this policy, the doctor only attends the consultation, and experiences travel time to go from room to room.

We use the models to evaluate the two policies on doctor utilization, patient access time and patient waiting time. The models provide insight in the ordering of the PtD-policy and the DtP-policy in different parameter settings for different outpatient clinics. As a result, we show that an outpatient clinic should choose the DtP-policy, when for each patient the doctor's travel time is lower than the patient's pre-consultation time. We extend this result with a discrete-event simulation, which indicates that a DtP-policy should be chosen when the average doctor travel time is lower than the sum of the average pre-consultation time and the average post-consultation time.

We developed an expression for the fraction of consultations that are in immediate succession to calculate the required number of rooms in the DtP-policy. Using the developed expression as described in this chapter results in choosing the required number of rooms such that the fraction of consultations in immediate succession is maximized and the idle time of the doctor is minimized.

To support decision making in outpatient clinics, we provide analytical models that can be used to compare the two policies on several performance measures and to determine the required number of consultation rooms in a particular outpatient clinic setting. Our experience in applying this research showed that our models are valuable for providing quantitative arguments to support the discussion of a proposed decision with stakeholders (e.g., hospital boards, doctors).

For the aforementioned hospitals we have successfully applied the insights obtained with our methods in the redesign of their outpatient clinics, based on data from their outpatient clinics. For the hospital managers, our results provided quantitative measures and formal proof to support their decision to redesign the outpatient clinic from a PtD-policy to a DtP-policy. With our models and the data, we also helped the hospitals to determine the required number of consultation rooms for each doctor in the DtP-policy.

6.6 Appendix

6.6.1 Proof of Lemma 6.5

Consider the recursion of the departure process. We distinguish two cases:

i. When $A_n \geq D_{n-1}$, the n-th patient comes in after the $(n-1)$-th patient has left, thus the doctor is available immediately upon arrival of the n-th patient at A_n. Hence, $D_n = A_n + P_n + C_n + U_n$.

ii. When $A_n < D_{n-1}$, the n-th patient comes in while the doctor is occupied. The n-th patient can start pre-consultation upon departure of the $(n-1)$-th patient. Hence, $D_n = D_{n-1} + P_n + C_n + U_n$.

Combining (*i*) and (*ii*) obtains

$$D_n = \max\{A_n, D_{n-1}\} + P_n + C_n + U_n. \tag{6.4}$$

Since $D_n = F_n + U_n$, we have proven Lemma 6.5. □

6.6.2 Proof of Lemma 6.6

The recursion of the finishing time for the doctor is explained by examining the time both the patient and the doctor are ready for consultation. The n-th patient is available for consultation after finishing pre-consultation. The doctor is available for the n-th patient, after the consultation of the $(n-1)$-th patient plus the travel to the n-th patient. We distinguish two cases:

i. When $n \leq R$, the number of customers in the system is smaller than the number of rooms. Hence, the n-th patient enters a room immediately upon arrival and is ready for consultation after pre-consultation $(A_n + P_n)$. The doctor consults the patient after finishing consultation of the $(n-1)$-th patient and the travel time $(F'_{n-1} + T_n)$. The moment consultation can start if $n \leq R$ is thus: $\max\{A_n + P_n, F'_{n-1} + T_n\}$.

ii. When $n > R$, the n-th patient may have to wait for the exit of the $s(n)$-th patient($F'_{s(n)} + U_{s(n)}$) before entering a room, or the patient can enter a room immediately upon arrival (A_n), if a room is available. After entering a room, pre-consultation has to be finished before consultation can start. Hence, the patient is ready for consultation at $\max\{F'_{s(n)} + U_{s(n)}, A_n\} + P_n$. The doctor is ready for consultation after traveling to the room (T_n). The doctor can start traveling after the consultation of the $(n-1)$-th patient (F'_{n-1}), and, due to Assumption 6.4, the $s(n)$-th patient must have exited $(F'_{s(n)} + U_{s(n)})$. Therefore, the doctor is available for the consultation of the n-th patient at $\max\{F'_{s(n)} + U_{s(n)}, F'_{n-1}\} + T_n$. The moment consultation can start if $n > R$ is thus $\max\{\max\{F'_{s(n)} + U_{s(n)}, A_n\} + P_n, \max\{F'_{s(n)} + U_{s(n)}, F'_{n-1}\} + T_n\}$.

We combine (*i*) and (*ii*) to obtain Lemma 6.6. □

6.6.3 Proof of Theorem 6.7

We prove Theorem 6.7 by induction. Clearly $F'_1 \leq F_1$, since in an initial (empty) system, the process is identical, because we have Assumption 6.1. The induction hypothesis is $F'_j \leq F_j$, for $j = 1, 2, ..., n-1$. It remains to prove that $F'_n \leq F_n$.

Observe from Assumptions 6.2 and 6.3 that $F_{n-1} \leq F_n$ and $F'_{n-1} \leq F'_n$. Additionally, the $s(n)$-th patient is the patient that leaves a room before the n-th patient can enter that room. Therefore, it is certain that the $s(n)$-th patient has entered a room before the n-th patient, so that

$$F'_{s(n)} + U_{s(n)} \leq F'_{n-1} + U_{n-1}, \text{ for } n = 1, 2, ..., N. \tag{6.5}$$

It is sufficient to consider the case $n > R$, since for $n \leq R$ by definition we have $F'_{s(n)} + U_{s(n)} = 0$.

For the case $A_n \leq F'_{s(n)} + U_{s(n)}$, we obtain:

$$\begin{aligned}
F'_n &= \max\{F'_{s(n)} + U_{s(n)} + P_n, \\
&\qquad \max\{F'_{s(n)} + U_{s(n)}, F'_{n-1}\} + T_n\} + C_n && \text{(Lemma 6.6, } n > R\text{)} \\
&\leq \max\{F'_{n-1} + U_{n-1} + P_n, \\
&\qquad \max\{F'_{n-1} + U_{n-1}, F'_{n-1}\} + T_n\} + C_n && \text{(Equation (6.5))} \\
&= F'_{n-1} + U_{n-1} + \max\{P_n, T_n\} + C_n \\
&\leq F_{n-1} + U_{n-1} + \max\{P_n, T_n\} + C_n && \text{(Induction hypothesis)} \\
&\leq F_{n-1} + U_{n-1} + P_n + C_n && \text{(Assumption 6.1)} \\
&\leq \max\{A_n, F_{n-1} + U_{n-1}\} + P_n + C_n = F_n && \text{(Lemma 6.5)}
\end{aligned}$$

For the case $A_n \geq F'_{s(n)} + U_{s(n)}$, we obtain:

$$\begin{aligned}
F'_n &= \max\{A_n + P_n, \max\{F'_{s(n)} + U_{s(n)}, F'_{n-1}\} + T_n\} + C_n && \text{(Lemma 6.6, } n > R\text{)} \\
&\leq \max\{A_n + P_n, \max\{A_n, F'_{n-1}\} + T_n\} + C_n \\
&= \max\{A_n + \max\{P_n, T_n\}, F'_{n-1} + T_n\} + C_n \\
&\leq \max\{A_n + \max\{P_n, T_n\}, F_{n-1} + T_n\} + C_n && \text{(Induction hypothesis)} \\
&\leq \max\{A_n + P_n, F_{n-1} + P_n\} + C_n && \text{(Assumption 6.1)} \\
&\leq \max\{A_n, F_{n-1} + U_{n-1}\} + P_n + C_n = F_n && \text{(Lemma 6.5)}
\end{aligned}$$

From the above, it follows that if $F'_j \leq F_j$, for $j = 1, 2, ..., n-1$, then $F'_n \leq F_n$. This proves Theorem 6.7. □

Epilogue

As discussed in Chapter 1, one of the main causes why healthcare planning and control lags behind manufacturing planning and control is the fragmented nature of healthcare planning and control. Healthcare organizations are typically organized as a cluster of autonomous departments, where planning and control is also often functionally dispersed. In this fragmented reality of healthcare planning and control, healthcare professionals face the challenging task to design and organize the healthcare delivery process more effectively and efficiently. We argue that an integrated approach to healthcare planning and control is likely to bring improvements, as the clinical course of patients traverses multiple departments and thus requires to coordinate decision making across multiple departments and resources.

This thesis aims to contribute to integrated decision making in healthcare in two ways. First, we develop a framework (Chapter 2), taxonomy, and extensive literature review (Chapter 3). Second, we propose planning approaches to develop tactical resources allocation and patient admission plans (Chapters 4 and 5), and provide models for a tactical planning problem in an outpatient setting (Chapter 6).

This Epilogue is organized as follows. We will highlight the two contributions and discuss their implications for practice. We conclude by providing our views on directions for future research.

The first contribution of our thesis concerns a conceptual framework, taxonomy, and extensive literature review within the field of resource capacity planning and control. Our framework consists of four hierarchical, or temporal, levels and four managerial areas to classify planning decisions in healthcare. Our taxonomy is a further specification of this framework in the managerial area of resource capacity planning for six identified care services. This taxonomy is filled through our literature review, which gives a comprehensive overview of the typical decisions to be made in resource capacity planning and control in healthcare and a structured review of relevant articles within the field of Operations Research and Management Science (OR/MS) for each planning decision.

The first contribution of our thesis can be subdivided in three potential uses in practice: (1) a framework and taxonomy to position planning decisions and map their interrelations, (2) a comprehensive overview of the planning decisions on all hierarchical levels for six care services, (3) a common language for managers, medical staff and planning experts. We discuss these in more detail below.

The conceptual framework and taxonomy support healthcare professionals in hierarchically structuring planning and control functions in multiple managerial areas and care services. This facilitates a structural breakdown and analysis of planning and control functions and their interaction. Moreover, it helps to identify managerial problems regarding, for example, planning functions that are deficient or inappropriate, that lack coherence, or have conflicting objectives. Overlooked or poorly addressed managerial functions are mostly found on the tactical level of control. Tactical planning is less tangible than operational planning and strategic planning. Inundated with operational problems, managers are inclined to solve problems at hand. Strategic measures such as increasing capacity are sometimes claimed as a solution for these operational problems. In such cases, it is often overlooked that instead of such a drastic strategic measure, tactically allocating and organizing the available resources may be more effective and cheaper.

Our literature review identifies relationships and dependencies in concrete planning decisions both between care services and between different hierarchical levels. Healthcare professionals can use our categorization of planning decisions to identify interactions between different care services, to detect conflicting objectives, and to discover opportunities for coordinated decision making. Also, the hierarchical relationships between planning decisions can be used to understand where increasing flexibility in planning may provide additional benefits. Increasing flexibility enables specifying and adjusting planning decisions closer to the time of actual healthcare delivery such that more detailed and accurate information can be incorporated. This provides healthcare professionals with opportunities to make planning and control decisions better match with supply and demand.

The framework and taxonomy offer a common language for involved decision makers in healthcare planning and control: clinical staff, managers, and experts on planning and control. This allows coherent formulation and realization of objectives on all levels and in all managerial areas [126]. Such common language may also contribute to interconnecting healthcare professionals and OR/MS researchers so that the potential of OR/MS can be discovered and exploited in improving healthcare delivery performance.

The second contribution of this thesis concerns planning decisions on the *tactical level* of planning and control. We develop planning approaches to create tactical resource allocation and patient admission plans for multiple departments, multiple resources and multiple patient types, thereby integrating decision making for a chain of healthcare resources. In addition, we provide analytical models to evaluate the performance of two different policies for patient routing in an outpatient clinic.

In their tactical planning approaches, some of the hospitals we cooperated with have spreadsheet solutions in place to evaluate for example waiting lists, access time and resource utilization. They use this information for resource allocation decision making, for example to allocate operating time and consultation time. Our method provides an optimization procedure for this step.

Hospital managers can use our methods to create tactical plans that allocate resource capacities to patient processes and activities, and determine the number of patients of a particular patient group to admit. Some hospitals have implemented or are implementing tactical planning in their hospitals. Implementing tactical planning can take relatively long (6-18 months), as it requires a different way of planning resource capacities and tracking systems for key performance indicators for tactical planning. The new planning methodology means that doctors and departments have to cooperate intensively and thus give up part of their planning autonomy. Trust from the involved stakeholders, doctors, healthcare managers and the hospital board, is therefore a key aspect in the implementation and the use of tactical planning methodologies. Implementation of tactical planning comprises several steps that we explain in detail below.

The first step is an initial assessment of the potential benefits of integrated and coordinated tactical planning in a hospital. This initial assessment ensures that the involved doctors, healthcare managers and hospital board members become familiar with the current performance of their department or specialty. In this review, they will get a better understanding of the benefits and disadvantages of current planning procedures and the restrictions that reduce planning flexibility. Also, this step is used to estimate the potential benefits of integrated and coordinated planning, such as cost savings, improved access times and/or improved resource utilization. By completing this step, the involved stakeholders gain more insight in current performance, which often is a strong motivation and *case for change*, and become aware of the potential benefit of improved coordination and integrated tactical planning.

After the first step is successfully completed, and the involved doctors and healthcare managers are convinced that tactical planning may be beneficial for their hospital, it is crucial to have the support of a sponsor in the hospital's board in a second step. As the full implementation of tactical planning requires investments, changes to information systems, cooperation of multiple departments, doctors and healthcare managers, it is important to have someone who can lead and support the implementation with the authority to invest and settle discussions between the various involved stakeholders. In addition to a sponsor, in implementations at some hospitals, it was key to set-up a fulltime project team dedicated to tactical planning. Implementation of tactical planning requires a change of the planning methodology and for example obtaining and analyzing data to understand the key performance indicators. Hence, without a fulltime project team, the chances of success appear much smaller. It is important that the tactical planning team has analytical skills, as the translation of performance indicators to information that can be used to develop a tactical plan often requires analysis, calculations and the presentation of data to doctors and healthcare managers.

The third step concerns data validation, setting up a system to track key performance indicators on a regular basis, and setting up the forecasted stages in each care process. Implementation of tactical planning methods requires insight in the hospital's performance. The key performance indicators that are required to develop the tactical plan should be available and correctly tracked. Typical key performance indicators to track are access times, capacity utilization, a forecast of demand and available resource capacity. Also, this step concerns 'cleaning up' the existing database and getting the available data up to the quality that it can be used for decision making. Data validation and a review of the tracked key performance indicators are important to gain and maintain the trust of involved stakeholders. If mistakes are found or data is regarded incorrect, the trust of doctors and healthcare managers in the tactical planning process may decrease. This third step in the implementation of tactical planning is a broad and complicated engagement, and may take several months, depending on the current state of information systems, databases and performance indicator tracking at the hospital.

The fourth step concerns setting up the tactical planning process and organization. To support the process of tactical planning, agreements are required between the involved decision makers on what should be done (e.g., data analysis, calculating scenarios, discussing proposed plans) and who is involved (e.g., hospital managers, doctors, nurses) in each step of developing a tactical plan. In some hospitals tactical planning is for instance organized around a two-week tactical planning meeting with the involved stakeholders, which is prepared by the tactical project team. In the tactical planning meeting, the key performance indicators are reviewed and the tactical planning project team advises on a tactical plan for the next few weeks. The decision making software to support tactical planning is chosen and connected to the information systems holding the key performance indicators. In addition, the involved stakeholders decide on the agreements under which the tactical planning process is implemented. One particular tactical agreement was a prerequisite for participation of the involved medical departments and the successful implementation of dynamic tactical planning in one of the hospitals. The involved decision makers agreed that a decided reduction of allocated resource capacity (in this case operating time) can always be revoked when the resource capacity is required again in the future. These types of tactical planning agreements ensure that doctors and healthcare managers are open to engage in an initial start-up period of tactical planning.

The fifth step concerns a start-up period for actual implementation of tactical planning. In these first months, the tactical planning meetings are used to 'train' the doctors and healthcare managers to understand the implications of past decisions on the performance indicators. They get more insight in what drives volatility in resource utilization and access times, and learn what the key trade-offs are in tactical planning decisions. The meetings are only used to track the performance indicators and discuss the effects of the tactical plan that is in place, no changes to the existing tactical plan are made in this start-up period. These first months are important to gain the confidence and trust of the involved doctors and healthcare managers, and to gain an understanding of what potential effects a change in the tactical plan may have. In all their analyses and tactical planning preparations, the tactical planning project team analyses the data objectively and should not appear biased or subjective. The project team is the centralized authority that provides the analyses and supports decision making in the tactical planning meeting. Hence, the tactical planning team should stay independent in order to not loose trust from the involved healthcare managers and doctors.

After this start-up period, the discussions in the tactical planning meetings about the key performance indicators are used to set the tactical plan for the next time periods. The project team remains involved in the analyses and preparing the tactical planning meetings. Key in implementation and use of tactical planning methodologies is the confidence and trust of doctors and healthcare managers in the tactical planning method, and the conviction that integrated and coordinated planning supports the involved stakeholders in providing healthcare to patients in an improved, more effective and efficient way.

Our literature review reconfirms the conclusions drawn in prior articles [282, 491] that there is a lack of contributions in the literature focusing on integrated models for complete healthcare processes. At various points in our overview, we have observed that taking an integrated approach in decision making is beneficial. We are convinced that developing such integrated models for healthcare planning and control remains an important direction for future research.

For our integrated planning models, numerous directions for future research can be identified. A first direction is practical implementation of the developed theoretical models. A first step involves testing the developed methods in an actual hospital setting, by running a shadow planning. This requires adapting the various parameters and objective functions in the developed methods to the application at hand. In a second step, these methods could be programmed into commercial decision making software to support the development of tactical plans in hospital settings. As argued in Chapter 4 implementation of such software does require proper data gathering, a transformation in the tactical planning process, cooperation between all involved parties, etc.

A second direction for future research is in the theoretical development of the tactical planning models. The size of the tactical planning benefit could be researched by comparing our tactical finite time horizon models with tactical planning models that incorporate infinite time horizons and thus provide fixed tactical plans for each time period. Also, the developed models can be further refined by researching different alternatives for the objective functions, constraints, and incorporated stochastic elements for example. The development of different solution methods to further advance speed and accuracy of the proposed methods is also an area that may lead to further benefits.

As we argue in this thesis, the OR/MS community has shown a rapid increase of attention to healthcare operations management in recent years. We believe that this field will continue to flourish in the coming decades, just like manufacturing operations management has in the last decades. This thesis illustrates that healthcare operations management contributes to more rational, integrated decision making in healthcare and that it can provide strong benefits in reorganizing healthcare more effectively and efficiently. We are convinced that healthcare managers and doctors are increasingly aware of these benefits of healthcare operations management. Their belief in the practical need for integrated solutions from healthcare operations management is a key driver to turn theory into practice and to make the contributions of the OR/MS community achieve real societal value.

Bibliography

[1] E.H.L. Aarts and J.K. Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 2003.

[2] W.J. Abernathy and J.C. Hershey. A spatial-allocation model for regional health-services planning. *Operations Research*, 20(3):629–642, 1972.

[3] I. Adan, J. Bekkers, N. Dellaert, J. Jeunet, and J. Vissers. Improving operational effectiveness of tactical master plans for emergency and elective patients under stochastic demand and capacitated resources. *European Journal of Operational Research*, 213(1):290–308, 2011.

[4] I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009.

[5] L.H. Aiken, S.P. Clarke, D.M. Sloane, J. Sochalski, and J.H. Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association*, 288(16):1987–1993, 2002.

[6] E. Akcali, M.J. Coˆté, and C. Lin. A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science*, 9(4):391–404, 2006.

[7] R. Akkerman and M. Knip. Reallocation of beds to reduce waiting time for cardiac surgery. *Health Care Management Science*, 7(2):119–126, 2004.

[8] T. Andersson and P. Värbrand. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2006.

[9] R.N. Anthony. *Planning and control systems: a framework for analysis*. Division of Research, Graduate School of Business Administration, Harvard University, 1965.

[10] J.P. Arnaout. Heuristics for the maximization of operating rooms utilization using simulation. *Simulation*, 86(8-9):573–583, 2010.

[11] M. Asaduzzaman, T.J. Chaussalet, and N.J. Robertson. A loss network model with overflow for capacity planning of a neonatal unit. *Annals of Operations Research*, 178(1):67–76, 2010.

[12] R. Ashton, L. Hague, M. Brandreth, D. Worthington, and S. Cropper. A simulation-based study of a NHS walk-in centre. *Journal of the Operational Research Society*, 56(2):153–161, 2005.

[13] V. Augusto, X. Xie, and V. Perdomo. Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. *Computers & Industrial Engineering*, 58(2):231–238, 2010.

[14] M. Babes and G.V. Sarma. Out-patient queues at the Ibn-Rochd health centre. *Journal of the Operational Research Society*, 42(10):845–855, 1991.

[15] F.F. Baesler, H.E. Jahnsen, and M. DaCosta. Emergency departments I: the use of simulation and design of experiments for estimating maximum capacity in an emergency room. In S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morrice, editors, *Proceedings of the 35th Conference Winter Simulation*, pages 1903–1906, New York, NY, USA, 2003. ACM.

[16] A. Bagust, M. Place, and J.W. Posnett. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *British Medical Journal*, 319(7203):155–158, 1999.

[17] N.T.J. Bailey. 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 (Methodological)*, 14(2):185–199, 1952.

[18] M.O. Ball and F.L. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1):18–36, 1993.

[19] J. Barado, J.M. Guergué, L. Esparza, C. Azcárate, F. Mallor, and S. Ochoa. A mathematical model for simulating daily bed occupancy in an intensive care unit. *Critical Care Medicine*, 40(4):1098–1104, 2011.

[20] J.F. Bard and H.W. Purnomo. Hospital-wide reactive scheduling of nurses with preference considerations. *IIE Transactions*, 37(7):589–608, 2005.

[21] E.R. Barthel, J.R. Pierce, C.J. Goodhue, H.R. Ford, T.C. Grikscheit, and J.S. Upperman. Availability of a pediatric trauma center in a disaster surge decreases triage time of the pediatric surge population: a population kinetics model. *Theoretical Biology and Medical Modelling*, 8(1):38, 2011.

[22] A. Başar, B. Çatay, and T. Ünlüyurt. A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. *Journal of the Operational Research Society*, 62(4):627–637, 2010.

[23] S. Batun, B.T. Denton, T.R. Huschka, and A.J. Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing*, 23(2):220–237, 2011.

[24] K.S. Bay, P. Leatt, and S.M. Stinson. A patient-classification system for long-term care. *Medical Care*, 20(5):468–488, 1982.

[25] H. Beaulieu, J.A. Ferland, B. Gendron, and P. Michelon. A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Management Science*, 3(3):193–200, 2000.

[26] R. Beech, R.L. Brough, and B.A. Fitzsimons. The development of a decision-support system for planning services within hospitals. *Journal of the Operational Research Society*, 41(11):995–1006, 1990.

[27] M.A. Begen and M. Queyranne. Appointment scheduling with discrete random durations. *Mathematics of Operations Research*, 36(2):240–257, 2011.

[28] S.V. Begur, D.M. Miller, and J.R. Weaver. An integrated spatial DSS for scheduling and routing home-health-care nurses. *Interfaces*, 27(4):35–48, 1997.

[29] R. Bekker and A.M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, 2010.

[30] R. Bekker and P.M. Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):1–13, 2011.

[31] J. Beliën and E. Demeulemeester. Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185 – 1204, 2007.

[32] J. Beliën and E. Demeulemeester. A branch-and-price approach for integrating nurse and surgery scheduling. *European Journal of Operational Research*, 189(3):652–668, 2008.

[33] J. Beliën, E. Demeulemeester, and B. Cardoen. A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling*, 12(2):147–161, 2009.

[34] R.E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton NJ, USA, 1957.

[35] J.C. Bennett and D.J. Worthington. An example of a good but partially successful OR engagement: Improving outpatient clinic operations. *Interfaces*, 28(5):56–69, 1998.

[36] R. Benveniste. Solving the combined zoning and location problem for several emergency units. *Journal of the Operational Research Society*, 36(5):433–450, 1985.

[37] E. Benzarti, E. Sahin, and Y. Dallery. A literature review on operations management based models developed for home health care services. Technical report, Cahier d'Études et de Recherche, Ecole Centrale Paris, 2010.

[38] P. Beraldi and M.E. Bruni. A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research*, 196(1):323–331, 2009.

[39] P. Beraldi, M.E. Bruni, and D. Conforti. Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research*, 158(1):183–193, 2004.

[40] G.N. Berlin, C. ReVelle, and D.J. Elzinga. Determining ambulance-hospital locations for on-scene and hospital services. *Environment and Planning A*, 8(5):553–561, 1976.

[41] S. Bertels and T. Fahle. A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem. *Computers & Operations Research*, 33(10):2866–2890, 2006.

[42] J.W.M. Bertrand, J.C. Wortmann, and J. Wijngaard. *Production control: a structural and design oriented approach*. Elsevier Science Inc. New York, NY, USA, 1990.

[43] D. Bertsimas and J. Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1):80–90, 2000.

[44] M.J. Bester, I. Nieuwoudt, and J.H. Van Vuuren. Finding good nurse duty schedules: a case study. *Journal of Scheduling*, 10(6):387–405, 2007.

[45] G. Bianchi and R.L. Church. A hybrid fleet model for emergency medical service system design. *Social Science & Medicine*, 26(1):163–171, 1988.

[46] J.F. Bithell. A class of discrete-time models for the study of hospital admission systems. *Operations Research*, 17(1):48–69, 1969.

[47] E.L. Blair and C.E. Eric. 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, 1981.

[48] M. Blais, S.D. Lapierre, and G. Laporte. Solving a home-care districting problem in an urban setting. *Journal of the Operational Research Society*, 54(11):1141–1147, 2003.

[49] J.T. Blake and M.W. Carter. Surgical process scheduling: a structured review. *Journal of the Society for Health Systems*, 5(3):17–30, 1997.

[50] J.T. Blake and M.W. Carter. A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140(3):541–561, 2002.

[51] J.T. Blake, F. Dexter, and J. Donald. Operating room managers' use of integer programming for assigning block time to surgical groups: A case study. *Anesthesia & Analgesia*, 94(1):143–148, 2002.

[52] J.T. Blake and J. Donald. Mount Sinai hospital uses integer programming to allocate operating room time. *Interfaces*, 32(2):63–73, 2002.

[53] D. Boldy and N. Howell. The geographical allocation of community care resources – a case study. *Journal of the Operational Research Society*, 31(2):123–129, 1980.

[54] R. J. Boucherie, X. Chao, and M. Miyazawa. Arrival first queueing networks with applications in kanban production systems. *Performance Evaluation*, 51(2-4):83–102, 2003.

[55] R.J. Boucherie and N.M. Van Dijk. Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability*, 23(1):152 – 187, 1991.

[56] J. Bowers and G. Mould. Concentration and the variability of orthopaedic demand. *Journal of the Operational Research Society*, 53(2):203–210, 2002.

[57] J. Bowers and G. Mould. Managing uncertainty in orthopaedic trauma theatres. *European Journal of Operational Research*, 154(3):599–608, 2004.

[58] M. Brahimi and D.J. Worthington. Queueing models for out-patient appointment systems – a case study. *Journal of the Operational Research Society*, 42(9):733–746, 1991.

[59] S.C. Brailsford, P.R. Harper, B. Patel, and M. Pitt. An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3(3):130–140, 2009.

[60] S.C. Brailsford, V.A. Lattimer, P. Tarnaras, and J.C. Turnbull. Emergency and on-demand health care: modelling a large complex system. *Journal of the Operational Research Society*, 55(1):34–42, 2004.

[61] S.C. Brailsford and J.M.H. Vissers. OR in healthcare: A European perspective. *European Journal of Operational Research*, 212(2):223 – 234, 2011.

[62] M.L. Brandeau, F. Sainfort, and W.P. Pierskalla, editors. *Operations Research and Health Care: a Handbook of Methods and Applications*. International Series in Operations Research & Management Science, Vol. 70. Kluwer Academic Publishers, 2004.

[63] O. Bräysy, W. Dullaert, and P. Nakari. The potential of optimization in communal routing problems: case studies from Finland. *Journal of Transport Geography*, 17(6):484–490, 2009.

[64] O. Bräysy, P. Nakari, W. Dullaert, and P. Neittaanmäki. An optimization approach for communal home meal delivery service: a case study. *Journal of Computational and Applied Mathematics*, 232(1):46–53, 2009.

[65] D. Bredström and M. Rönnqvist. Combined vehicle routing and scheduling with temporal precedence and synchronization constraints. *European Journal of Operational Research*, 191(1):19–31, 2008.

[66] K.M. Bretthauer, H.S. Heese, H. Pun, and E. Coe. Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management*, 20(3):375–391, 2011.

[67] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[68] J.R. Broyles, J.K. Cochran, and D.C. Montgomery. A statistical Markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657, 2010.

[69] P. Brucker, A. Drexl, R. Möhring, K. Neumann, and E. Pesch. Resource-constrained project scheduling: Notation, classification, models, and methods. *European Journal of Operational Research*, 112(1):3–41, 1999.

[70] P. Brucker and S. Knust. Resource-constrained project scheduling. *In: Complex Scheduling. GOR-Publications Springer Germany (P. Brucker and S. Knust (authors))*, 2nd edition:117–238, 2012.

[71] J.O. Brunner, J.F. Bard, and R. Kolisch. Midterm scheduling of physicians with flexible shifts using branch and price. *IIE Transactions*, 43(2):84–109, 2011.

[72] J.O. Brunner and G. Edenharter. Long term staff scheduling of physicians with different experience levels in hospitals using column generation. *Health Care Management Science*, 14(2):189–202, 2011.

[73] R. Buitenhek. *Performance evaluation of dual resource manufacturing systems*. PhD thesis, University of Twente, The Netherlands, 1998.

[74] E.K. Burke, P. De Causmaecker, G.V. Berghe, and H. Van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7(6):441–499, 2004.

[75] C.R. Busby and M.W. Carter. A decision tool for negotiating home care funding levels in Ontario. *Home Health Care Services Quarterly*, 25(3-4):91, 2006.

[76] T.W. Butler, K.R. Karwan, and J.R. Sweigart. Multi-level strategic evaluation of hospital plans and decisions. *Journal of the Operational Research Society*, 43(7):665–675, 1992.

[77] T.W. Butler, K.R. Karwan, J.R. Sweigart, and G.R. Reeves. An integrative model-based approach to hospital layout. *IIE transactions*, 24(2):144–152, 1992.

[78] T.W. Butler, G.K. Leong, and L.N. Everett. The operations management role in hospital strategic planning. *Journal of Operations Management*, 14(2):137–156, 1996.

[79] A.B. Calvo and D.H. Marks. Location of health care facilities: an analytical approach. *Socio-Economic Planning Sciences*, 7(5):407–422, 1973.

[80] B. Cardoen and E. Demeulemeester. Capacity of clinical pathways: A strategic multi-level evaluation tool. *Journal of Medical Systems*, 32(6):443–452, 2008.

[81] B. Cardoen and E. Demeulemeester. A decision support system for surgery sequencing at UZ Leuven's day-care department. *International Journal of Information Technology and Decision Making*, 10(3):435, 2011.

[82] B. Cardoen, E. Demeulemeester, and J. Beliën. Optimizing a multiple objective surgical case sequencing problem. *International Journal of Production Economics*, 119(2):354–366, 2009.

[83] B. Cardoen, E. Demeulemeester, and J. Beliën. Sequencing surgical cases in a day-care environment: An exact branch-and-price approach. *Computers and Operations Research*, 36(9):2660–2669, 2009.

[84] B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201(3):921 – 932, 2010.

[85] A.P. Carpenter, L.M. Leemis, A.S. Papir, D.J. Phillips, and G.S. Phillips. Managing magnetic resonance imaging machines: support tools for scheduling and planning. *Health Care Management Science*, 14(2):158–173, 2011.

[86] G.M. Carter, J.M. Chaiken, and E. Ignall. Response areas for two emergency units. *Operations Research*, 20(3):571–594, 1972.

[87] M. Carter. Diagnosis: mismanagement of resources. *OR/MS Today*, 29(2):26–33, 2002.

[88] M.W. Carter and S.D. Lapierre. Scheduling emergency room physicians. *Health Care Management Science*, 4(4):347–360, 2001.

[89] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[90] T. Cayirli, E. Veral, and Rosen H. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1):47–58, 2006.

[91] R. Ceglowski, L. Churilov, and J. Wasserthiel. Combining data mining and discrete

event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society*, 58(2):246–254, 2006.

[92] Centraal Plan Bureau (CPB Netherlands Bureau for Economic Policy Analysis). Toekomst voor de zorg [in Dutch]. *Retrieved July 26, 2013, from: http://www.cpb.nl/en/publication/toekomst-voor-de-zorg*, 2013.

[93] E. Cerdá, L. Pablos, and M. Rodriguez. Waiting lists for surgery. *In: Patient flow: reducing delay in healthcare delivery (Hall, R.W. (editor))*, International Series in Operations Research & Management Science, Vol. 91:151–187, 2006.

[94] S. Ceschia and A. Schaerf. Local search and lower bounds for the patient admission scheduling problem. *Computers and Operations Research*, 38(10):1452–1463, 2011.

[95] S. Chaabane, N. Meskens, A. Guinet, and M. Laurent. Comparison of two methods of operating theatre planning: Application in Belgian hospital. *Journal of Systems Science and Systems Engineering*, 17(2):171–186, 2008.

[96] S. Chahed, E. Marcon, E. Sahin, D. Feillet, and Y. Dallery. Exploring new operational research opportunities within the home care context: the chemotherapy at home. *Health Care Management Science*, 12(2):179–191, 2009.

[97] S. Chand, H. Moskowitz, J.B. Norris, S. Shade, and D.R. Willis. Improving patient flow at an outpatient clinic: study of sources of variability and improvement factors. *Health Care Management Science*, 12(3):325–340, 2009.

[98] T.J. Chaussalet, H. Xie, and P. Millard. A closed queueing network approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*, 45(5):492–497, 2006.

[99] B. Cheang, H. Li, A. Lim, and B. Rodrigues. Nurse rostering problems–a bibliographic survey. *European Journal of Operational Research*, 151(3):447–460, 2003.

[100] C.F. Chien, F.P. Tseng, and C.H. Chen. An evolutionary approach to rehabilitation patient scheduling: A case study. *European Journal of Operational Research*, 189(3):1234–1253, 2008.

[101] CHOIR. Center for Healthcare Operations Improvement and Research (CHOIR) at the University of Twente. *Retrieved June 22, 2013, from: http://www.utwente.nl/choir/en/*, 2013.

[102] V.S. Chow, M.L. Puterman, N. Salehirad, W. Huang, and D. Atkins. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management*, 20(3):418–430, 2011.

[103] G. Christodoulou and G.J. Taylor. Using a continuous time hidden Markov process, with covariates, to model bed occupancy of people aged over 65 years. *Health Care Management Science*, 4(1):21–24, 2001.

[104] J.K. Cochran and A. Bharti. Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*, 9(1):31–45, 2006.

[105] J.K. Cochran and K. Roche. A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society*, 59(11):1471–1482, 2007.

[106] J.K. Cochran and K.T. Roche. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, 36(5):1497–1512, 2009.

[107] T. Coelli, D.S. Prasada Rao, and G.E. Battese. *An introduction to efficiency and productivity analysis*. Kluwer Academic Publishers, 2005.

[108] M.A. Cohen, J.C. Hershey, and E.N. Weiss. Analysis of capacity decisions for progressive patient care hospital facilities. *Health Services Research*, 15(2):145, 1980.

[109] D. Conforti, F. Guerriero, and R. Guido. Optimization models for radiotherapy patient scheduling. *4OR: A Quarterly Journal of Operations Research*, 6(3):263–278, 2008.

[110] D. Conforti, F. Guerriero, and R. Guido. Non-block scheduling with priority for radiotherapy treatments. *European Journal of Operational Research*, 201(1):289–296, 2010.

[111] D. Conforti, F. Guerriero, R. Guido, M.M. Cerinic, and M.L. Conforti. An optimal decision making model for supporting week hospital management. *Health Care Management Science*, 14(1):74–88, 2011.

[112] A.X. Costa, S.A. Ridley, A.K. Shahani, P.R. Harper, V. De Senna, and M.S. Nielsen. Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia*, 58(4):320–327, 2003.

[113] M.J. Côté. Patient flow and resource utilization in an outpatient clinic. *Socio-Economic Planning Sciences*, 33(3):231–245, 1999.

[114] M.J. Côté, S.S. Syam, W.B. Vogel, and D.C. Cowper. A mixed integer programming model to locate traumatic brain injury treatment units in the department of veterans affairs: a case study. *Health Care Management Science*, 10(3):253–267, 2007.

[115] S. Creemers and M. Lambrecht. An advanced queueing model to analyze appointment-driven service systems. *Computers & Operations Research*, 36(10):2773–2785, 2009.

[116] J. Dale, H. Lang, J.A. Roberts, J. Green, and E. Glucksman. Cost effectiveness of treating primary care patients in accident and emergency: a comparison between general practitioners, senior house officers, and registrars. *British Medical Journal*, 312(7042):1340–1344, 1996.

[117] A Dash. Lost + found: making the right choice in equipment location systems. *Health Facilities Management*, 22(11):19, 2009.

[118] M.S. Daskin and E.H. Stern. A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137, 1981.

[119] Data from 2011 from the website of Organisation of Economic Co-operation and Development (OECD). *Retrieved July 12, 2011, from: http://www.oecd.org/health*, 2011.

[120] R.W. Day, M.D. Dean, R. Garfinkel, and S. Thompson. Improving patient flow in a hospital through dynamic allocation of cardiac diagnostic testing time slots. *Decision Support Systems*, 49(4):463–473, 2010.

[121] V. De Angelis. Planning home assistance for AIDS patients in the city of Rome, Italy. *Interfaces*, 48(3):75–83, 1998.

[122] V. De Angelis, G. Felici, and P. Impelluso. Integrating simulation and optimisation in health care centre management. *European Journal of Operational Research*, 150(1):101–114, 2003.

[123] A.M. de Bruin, A.C. van Rossum, M.C. Visser, and G.M. Koole. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137, 2007.

[124] M.L. De Grano, D.J. Medeiros, and D. Eitel. Accommodating individual preferences in nurse scheduling via auctions and optimization. *Health Care Management Science*, 12(3):228–242, 2009.

[125] G. De Vries, J.W.M. Bertrand, and J.M.H. Vissers. Design requirements for health care production control systems. *Production Planning & Control*, 10(6):559–569, 1999.

[126] L. Delesie. Bridging the gap between clinicians and health managers. *European*

Journal of Operational Research, 105(2):248–256, 1998.

[127] N. Dellaert, J. Jeunet, and G. Mincsovics. Budget allocation for permanent and contingent capacity under stochastic demand. *International Journal of Production Economics*, 131(1):128–138, 2011.

[128] P. Demeester, W. Souffriau, P. De Causmaecker, and G. Vanden Berghe. A hybrid tabu search algorithm for automatically assigning patients to beds. *Artificial Intelligence in Medicine*, 48(1):61–70, 2010.

[129] B.T. Denton and D. Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016, 2003.

[130] B.T. Denton, A.J. Miller, H.J. Balasubramanian, and T.R. Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations research*, 58(4-Part-1):802–816, 2010.

[131] B.T. Denton, J. Viapiano, and A. Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1):13–24, 2007.

[132] M.S. Desai, M.L. Penn, S. Brailsford, and M. Chipulu. Modelling of Hampshire adult services – gearing up for future demands. *Health Care Management Science*, 11(2):167–176, 2008.

[133] F. Dexter. 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, 1999.

[134] F. Dexter. Bibliography of operating room management articles. *Retrieved May 10, 2012, from: http://www.franklindexter.com/*, 2012.

[135] F. Dexter, R.H. Epstein, and H.M. Marsh. A statistical analysis of weekday operating room anesthesia group staffing costs at nine independently managed surgical suites. *Anesthesia & Analgesia*, 92(6):1493–1498, 2001.

[136] F. Dexter and J. Ledolter. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. *Anesthesiology*, 103(6):1259–1267, 2005.

[137] F. Dexter and A. Macario. Decrease in case duration required to complete an additional case during regularly scheduled hours in an operating room suite: a computer simulation study. *Anesthesia & Analgesia*, 88(1):72–76, 1999.

[138] F. Dexter, A. Macario, and D.A. Lubarsky. The impact on revenue of increasing patient volume at surgical suites with relatively high operating room utilization. *Anesthesia & Analgesia*, 92(5):1215–1221, 2001.

[139] F. Dexter, A. Macario, and L. O'Neill. A strategy for deciding operating room assignments for second-shift anesthetists. *Anesthesia & Analgesia*, 89(4):920–924, 1999.

[140] F. Dexter, A. Macario, and L. O'Neill. 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, 2000.

[141] F. Dexter, A. Macario, and R.D. Traub. Optimal sequencing of urgent surgical cases. *Journal of Clinical Monitoring and Computing*, 15(3):153–162, 1999.

[142] F. Dexter, A. Macario, and R.D. Traub. Which algorithm for scheduling add-on elective cases maximizes operating room utilization?: Use of bin packing algorithms and fuzzy constraints in operating room management. *Anesthesiology*, 91(5):1491–1500, 1999.

[143] F. Dexter, A. Macario, R.D. Traub, M. Hopwood, and D.A. Lubarsky. An operat-

ing room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesthesia & Analgesia*, 89(1):7–20, 1999.

[144] F. Dexter, A. Macario, R.D. Traub, and D.A. Lubarsky. Operating room utilization alone is not an accurate metric for the allocation of operating room block time to individual surgeons with low caseloads. *Anesthesiology*, 98(5):1243–1249, 2003.

[145] F. Dexter and R.D. Traub. Statistical method for predicting when patients should be ready on the day of surgery. *Anesthesiology*, 93(4):1107, 2000.

[146] F. Dexter and R.D. Traub. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia & Analgesia*, 94(4):933–942, 2002.

[147] F. Dexter, R.D. Traub, and P. Lebowitz. Scheduling a delay between different surgeons' cases in the same operating room on the same day using upper prediction bounds for case durations. *Anesthesia & Analgesia*, 92(4):943–946, 2001.

[148] F. Dexter, R.D. Traub, and A. Macario. How to release allocated operating room time to increase efficiency: predicting which surgical service will have the most underutilized operating room time. *Anesthesia & Analgesia*, 96(2):507–512, 2003.

[149] F. Dexter, R. E. Wachtel, M.W. Sohn, J. Ledolter, E. U. Dexter, and A. Macario. Quantifying effect of a hospital's caseload for a surgical specialty on that of another hospital using multi-attribute market segments. *Health Care Management Science*, 8(2):121–131, 2005.

[150] F. Dexter, R.E. Wachtel, R.H. Epstein, J. Ledolter, and M.M. Todd. Analysis of operating room allocations to optimize scheduling of specialty rotations for anesthesia trainees. *Anesthesia & Analgesia*, 111(2):520–524, 2010.

[151] G. Dobson, S. Hasija, and E.J. Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011.

[152] G. Dobson, H.H. Lee, and E. Pinker. A model of ICU bumping. *Operations Research*, 58(6):1564–1576, 2010.

[153] V. F. Dokmeci. Planning ambulatory health care delivery systems. *Omega*, 4(5):617–622, 1976.

[154] K. Dong-Guen and K. Yeong-Dae. A branch and bound algorithm for determining locations of long-term care facilities. *European Journal of Operational Research*, 206(1):168–177, 2010.

[155] C. Duguay and F. Chetouane. Modeling and improving emergency department systems using discrete event simulation. *Simulation*, 83(4):311–320, 2007.

[156] M.B. Dumas. Simulation modeling for hospital bed planning. *Simulation*, 43(2):69, 1984.

[157] M.B. Dumas. Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels. *Health Services Research*, 20(1):43, 1985.

[158] D.J. Eaton. Determining ambulance deployment in Santo Domingo, Dominican Republic. *Journal of the Operational Research Society*, 37(2):113–126, 1986.

[159] EBSCOhost. *Retrieved May 10, 2012, from: http://www.ebscohost.com/*, 2012.

[160] G.M. Edward, S.F. Das, S.G. Elkhuizen, P.J.M. Bakker, J.A.M. Hontelez, M.W. Hollmann, B. Preckel, and L.C. Lemaire. Simulation to analyse planning difficulties at the preoperative assessment clinic. *British Journal of Anaesthesia*, 100(2):195–202, 2008.

[161] E. El-Darzi, C. Vasilakis, T. Chaussalet, and PH Millard. A simulation modelling

approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149, 1998.

[162] S.G. Elkhuizen, S.F. Das, P.J.M. Bakker, and J.A.M. Hontelez. Using computer simulation to reduce access time for outpatient departments. *British Medical Journal*, 16(5):382–386, 2007.

[163] A. Erdelyi and H. Topaloglu. Approximate dynamic programming for dynamic capacity allocation with multiple priority levels. *IIE Transactions*, 43(2):129–142, 2010.

[164] G. Erdogan, E. Erkut, A. Ingolfsson, and G. Laporte. Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society*, 61(4):543–550, 2010.

[165] E. Erkut, A. Ingolfsson, and G. Erdoğan. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–58, 2008.

[166] A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: a review of applications, methods and models. *European Journal of Operational Research*, 153(1):3–27, 2004.

[167] A.O. Esogbue and A.J. Singh. A stochastic model for an optimal priority bed distribution problem in a hospital ward. *Operations Research*, 24(5):884–898, 1976.

[168] P. Eveborn, P. Flisberg, and M. Ronnqvist. Laps care–an operational system for staff planning of home care. *European Journal of Operational Research*, 171(3):962–976, 2006.

[169] P. Eveborn, M. Rönnqvist, H. Einarsdóttir, M. Eklund, K. Lidén, and M. Almroth. Operations research improves quality and efficiency in home care. *Interfaces*, 39(1):18–34, 2009.

[170] J.E. Everett. A decision support simulation model for the management of an elective surgery waiting system. *Health Care Management Science*, 5(2):89–95, 2002.

[171] N.R. Every, J. Hochman, R. Becker, S. Kopecky, and C.P. Cannon. Critical pathways: a review. *Circulation*, 101(4):461–465, 2000.

[172] M.J. Faddy, N. Graves, and A. Pettitt. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314, 2009.

[173] M.J. Faddy and S.I. McClean. Markov chain modelling for geriatric patient care. *Methods of Information in Medicine-Methodik der Information in der Medizin*, 44(3):369–373, 2005.

[174] M.J. Faddy and S.I. McClean. Using a multi-state model to enhance understanding of geriatric patient care. *Australian Health Review*, 31(1):91–97, 2007.

[175] H. Fei, C. Chu, and N. Meskens. Solving a tactical operating room planning problem by a column-generation-based heuristic procedure with four criteria. *Annals of Operations Research*, 166(1):91–108, 2009.

[176] H. Fei, C. Chu, N. Meskens, and A. Artiba. Solving surgical cases assignment problem by a branch-and-price approach. *International Journal of Production Economics*, 112(1):96–108, 2008.

[177] H. Fei, N. Meskens, and C. Chu. A planning and scheduling problem for an operating theatre using an open scheduling strategy. *Computers & Industrial Engineering*, 58(2):221–230, 2010.

[178] R.B. Fetter and J.D. Thompson. The simulation of hospital systems. *Operations Research*, 13(5):689–711, 1965.

[179] R.B. Fetter and J.D. Thompson. Patients' waiting time and doctors' idle time in the

outpatient setting. *Health Services Research*, 1(1):66–90, 1966.

[180] J.A. Fitzsimmons. A methodology for emergency ambulance deployment. *Management Science*, 19(6):627–636, 1973.

[181] A. Fletcher, D. Halsall, S. Huxham, and D. Worthington. The DH accident and emergency department model: a national generic model used locally. *Journal of the Operational Research Society*, 58(12):1554–1562, 2006.

[182] D. Fone, S. Hollinghurst, M. Temple, A. Round, N. Lester, A. Weightman, K. Roberts, E. Coyle, G. Bevan, and S. Palmer. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health*, 25(4):325, 2003.

[183] B.E. Fries. Bibliography of operations research in health-care systems. *Operations Research*, pages 801–814, 1976.

[184] B.E. Fries and V.P. Marathe. Determination of optimal variable-sized multiple-block appointment systems. *Operations Research*, 29(2):324–345, 1981.

[185] B.E. Fries, D.P. Schneider, W.J. Foley, M. Gavazzi, R. Burke, and E. Cornelius. Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). *Medical Care*, 32(7):668, 1994.

[186] O. Fujiwara, T. Makjamroen, and K. K. Gupta. Ambulance deployment analysis: a case study of Bangkok. *European Journal of Operational Research*, 31(1):9–18, 1987.

[187] P.H.P. Fung Kon Jin, M.G.W. Dijkgraaf, C.L. Alons, C. van Kuijk, L.F.M. Beenen, G.M. Koole, and J.C. Goslings. Improving CT scan capabilities with a new trauma workflow concept: simulation of hospital logistics using different CT scanner scenarios. *European Journal of Radiology*, 80(2):504–509, 2011.

[188] S. Gallivan and M. Utley. A technical note concerning emergency bed demand. *Health Care Management Science*, 14(3):1–3, 2011.

[189] S. Gallivan, M. Utley, T. Treasure, and O. Valencia. Booked inpatient admissions and hospital capacity: mathematical modelling study. *British Medical Journal*, 324(7332):280, 2002.

[190] S. Ganguli, J.C. Tham, and B.M.J. d'Othee. Establishing an outpatient clinic for minimally invasive vein care. *American Journal of Roentgenology*, 188(6):1506–1511, 2007.

[191] JA Garcia-Navarro and WA Thompson. Analysis of bed usage and occupancy following the introduction of geriatric rehabilitative care in a hospital in huesca, spain. *Health Care Management Science*, 4(1):63–66, 2001.

[192] L. Garg, S. McClean, B. Meenan, and P. Millard. A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. *Health Care Management Science*, 13(2):155–169, 2010.

[193] P. Gemmel and R. Van Dierdonck. Admission scheduling in acute care hospitals: does the practice fit with the theory? *International Journal of Operations and Production Management*, 19(9):863–878, 1999.

[194] M. Gendreau, G. Laporte, and F. Semet. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, 57(1):22–28, 2005.

[195] N. Geng, X. Xie, V. Augusto, and Z. Jiang. A Monte Carlo optimization and dynamic programming approach for managing MRI examinations of stroke patients. *IEEE Transactions on Automatic Control*, 56(11):2515–2529, 2011.

[196] Y. Gerchak, D. Gupta, and M. Henig. Reservation planning for elective surgery

under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334, 1996.

[197] N. Geroliminis, K. Kepaptsoglou, and M.G. Karlaftis. A hybrid hypercube-genetic algorithm approach for deploying many emergency response mobile units in an urban network. *European Journal of Operational Research*, 210(2):287–300, 2011.

[198] GHZ Website. Website of Groene Hart Ziekenhuis (In Dutch). *Retrieved April 16, 2011, from: http://www.ghz.nl*, 2012.

[199] J. Gillespie, S. McClean, B. Scotney, L. Garg, M. Barton, and K. Fullerton. Costing hospital resources for stroke patients using phase-type models. *Health Care Management Science*, 14(13):1–13, 2011.

[200] S. Glouberman and H. Mintzberg. Managing the care of health and the cure of disease-part i: Differentiation. *Health care management review*, 26(1):56–69, 2001.

[201] S. Glouberman and H. Mintzberg. Managing the care of health and the cure of disease-part ii: Integration. *Health care management review*, 26(1):70–84, 2001.

[202] A. Gnanlet and W.G. Gilland. Sequential and simultaneous decision making for optimizing health care resource flexibilities. *Decision Sciences*, 40(2):295–326, 2009.

[203] Y. Gocgun, B.W. Bresnahan, A. Ghate, and M.L. Gunn. A Markov decision process approach to multi-category patient scheduling in a diagnostic facility. *Artificial Intelligence in Medicine*, 53(2):73–81, 2011.

[204] J. Goldberg, R. Dietrich, J.M. Chen, M. Mitwasi, T. Valenzuela, and E. Criss. A simulation model for evaluating a set of emergency vehicle base locations: development, validation, and usage. *Socio-Economic Planning Sciences*, 24(2):125–141, 1990.

[205] J. Goldberg, R. Dietrich, J.M. Chen, M.G. Mitwasi, T. Valenzuela, and E. Criss. Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ. *European Journal of Operational Research*, 49(3):308–324, 1990.

[206] J. Goldman, H.A. Knappenberger, and J.C. Eller. Evaluating bed allocation policy with computer simulation. *Health Services Research*, 3(2):119–129, 1968.

[207] N. Görmez, M. Köksalan, and F.S. Salman. Locating disaster response facilities in Istanbul. *Journal of the Operational Research Society*, 62(7):1239–1252, 2010.

[208] F. Gorunescu, S.I. McClean, and P.H. Millard. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24, 2002.

[209] F. Gorunescu, S.I. McClean, and P.H. Millard. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4):307–312, 2002.

[210] D. Gove, D. Hewett, and A. Shahani. Towards a model for hospital case-load decision support. *Mathematical Medicine and Biology*, 12(3–4):329–338, 1995.

[211] S.C. Graves. A tactical planning model for a job shop. *Operations Research*, 34(4):522–533, 1986.

[212] S.C. Graves, H.S. Leff, J. Natkins, and M. Senger. A simple stochastic model for facility planning in a mental health care system. *Interfaces*, 13(5):101–110, 1983.

[213] L.V. Green. Queueing analysis in healthcare. *In: Patient flow: reducing delay in healthcare delivery (Hall, R.W. (editor))*, International Series in Operations Research & Management Science, Vol. 91:281–307, 2006.

[214] L.V. Green and P.J. Kolesar. Improving emergency responsiveness with management science. *Management Science*, 50(8):1001–1014, 2004.

[215] L.V. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing

service systems that have cyclic demands. *Operations Research*, 49(4):549–564, 2001.

[216] L.V. Green, P.J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

[217] L.V. Green and V. Nguyen. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research*, 36(2):421–442, 2001.

[218] L.V. Green and S. Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[219] L.V. Green, S. Savin, and B. Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 54(1):11–25, 2006.

[220] L.V. Green, J. Soares, J.F. Giglio, and R.A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[221] J. Griffin, S. Xia, S. Peng, and P. Keskinocak. Improving patient flow in an obstetric unit. *Health Care Management Science*, 15(1):1–14, 2012.

[222] J.D. Griffiths, N. Price-Lloyd, M. Smithies, and J.E. Williams. Modelling the requirement for supplementary nurses in an intensive care unit. *Journal of the Operational Research Society*, 56(2):126–133, 2005.

[223] R. Grol, P. Giesen, and C. Van Uden. After-hours care in the United Kingdom, Denmark, and the Netherlands: new models. *Health Affairs*, 25(6):1733–1737, 2006.

[224] F. Guerriero and R. Guido. Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, 14(1):89–114, 2011.

[225] A. Guinet and S. Chaabane. Operating theatre planning. *International Journal of Production Economics*, 85(1):69–81, 2003.

[226] S. Gul, B.T. Denton, J.W. Fowler, and T. Huschka. Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management*, 20(3):406–417, 2011.

[227] E.D. Güneş. Modeling time allocation for prevention in primary care. *Central European Journal of Operations Research*, 17(3):359–380, 2009.

[228] D. Gupta. Surgical suites' operations management. *Production and Operations Management*, 16(6):689–700, 2007.

[229] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

[230] D. Gupta and L. Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research-Baltimore*, 56(3):576–592, 2008.

[231] R.W. Hall, editor. *Patient Flow: Reducing Delay in Healthcare Delivery*. International Series in Operations Research & Management Science, Vol. 91. Springer, 2006.

[232] E. W. Hans. *Resource loading by branch-and-price techniques*. PhD thesis, University of Twente, the Netherlands, 2001.

[233] E.W. Hans, W.S. Herroelen, R. Leus, and G. Wullink. A hierarchical approach to multi-project planning under uncertainty. *Omega*, 35(5):563–577, 2007.

[234] E.W. Hans, M. van Houdenhoven, and P.J.H. Hulshof. A framework for healthcare planning and control. *In: Handbook of Healthcare System Scheduling (Hall, R.W. (editor))*, International Series in Operations Research & Management Science, Vol. 168:303–320, 2012.

[235] E.W. Hans, G. Wullink, M. van Houdenhoven, and G. Kazemier. Robust surgery loading. *European Journal of Operational Research*, 185(3):1038–1050, 2008.

[236] W.L. Hare, A. Alimadad, H. Dodd, R. Ferguson, and A. Rutherford. A determin-

istic model of home and community care client counts in British Columbia. *Health Care Management Science*, 12(1):80–98, 2009.

[237] S.I. Harewood. Emergency ambulance deployment in Barbados: a multi-objective approach. *Journal of the Operational Research Society*, 53(2):185–192, 2002.

[238] P.R. Harper. A framework for operational modelling of hospital resources. *Health Care Management Science*, 5(3):165–173, 2002.

[239] P.R. Harper and H.M. Gamlin. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, 25(2):207–222, 2003.

[240] P.R. Harper, V.A. Knight, and A.H. Marshall. Discrete conditional phase-type models utilising classification trees: Application to modelling health service capacities. *European Journal of Operational Research*, 219(3):522–530, 2011.

[241] P.R. Harper, N.H. Powell, and J.E. Williams. Modelling the size and skill-mix of hospital nursing teams. *Journal of the Operational Research Society*, 61(5):768–779, 2009.

[242] P.R. Harper and A.K. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, 53(1):11–18, 2002.

[243] P.R. Harper, A.K. Shahani, J.E. Gallagher, and C. Bowie. Planning health services with explicit geographical considerations: a stochastic location-allocation approach. *Omega*, 33(2):141–152, 2005.

[244] R.A. Harris. Hospital bed requirements planning. *European Journal of Operational Research*, 25(1):121–126, 1986.

[245] G.W. Harrison and G.J. Escobar. Length of stay and imminent discharge probability distributions from multistage models: variation by diagnosis, severity of illness, and hospital. *Health Care Management Science*, 13(3):268–279, 2010.

[246] G.W. Harrison and P.H. Millard. Balancing acute and long-term care: the mathematics of throughput in departments of geriatric medicine. *Methods of Information in Medicine*, 30(3):221, 1991.

[247] G.W. Harrison, A. Shafer, and M. Mackay. Modelling variability in hospital bed occupancy. *Health Care Management Science*, 8(4):325–334, 2005.

[248] A.C. Hax and H.C. Meal. Hierarchical integration of production planning and scheduling. *In: TIMS Studies in the management sciences: logistics (Geisler, M. (editor))*, North Holland-American Elsevier, Amsterdam:53–69, 1975.

[249] J.E. Helm, S. AhmadBeygi, and M.P. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374, 2011.

[250] W.L. Herring and J.W. Herrmann. The single-day surgery scheduling problem: sequential decision-making and threshold-based heuristics. *OR Spectrum*, 34(2):429–459, 2012.

[251] J.C. Hershey, E.N. Weiss, and M.A. Cohen. A stochastic service network model with application to hospital facilities. *Operations Research*, 29(1):1–22, 1981.

[252] A. Hertz and N. Lahrichi. A patient assignment algorithm for home care services. *Journal of the Operational Research Society*, 60(4):481–495, 2009.

[253] F. Hillier. *Introduction to Operations Research*. McGraw-Hill; 9th edition, 2009.

[254] C.J. Ho and H.S. Lau. Minimizing total cost in scheduling outpatient appointments. *Management Science*, 38(12):1750–1764, 1992.

[255] C.J. Ho and H.S. Lau. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of*

Operational Research, 112(3):542–553, 1999.

[256] V.N. Hsu, R. de Matta, and C.Y. Lee. Scheduling patients in an ambulatory surgical center. *Naval Research Logistics*, 50(3):218–238, 2003.

[257] R. Huang, S. Kim, and M.B.C. Menezes. Facility location for large-scale emergencies. *Annals of Operations Research*, 181(1):271–286, 2010.

[258] X.M. Huang. A planning model for requirement of emergency beds. *Mathematical Medicine and Biology*, 12(3-4):345, 1995.

[259] X.M. Huang. Decision making support in reshaping hospital medical services. *Health Care Management Science*, 1(2):165–173, 1998.

[260] W.L. Hughes and S.Y. Soliman. Short-term case mix management with linear programming. *Hospital & Health Services Administration*, 30(1):52–60, 1985.

[261] P. J.H. Hulshof, R.J. Boucherie, E.W. Hans, and J.L. Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science*, 16(2):152–166, 2013.

[262] P.J.H. Hulshof, R.J. Boucherie, J.T. van Essen, E.W. Hans, J.L. Hurink, N. Kortbeek, N. Litvak, P.T. Vanberkel, E. Van der Veen, B. Veltman, I.M.H. Vliegen, and M.E. Zonderland. ORchestra: an online reference database of OR/MS literature in health care. *Health Care Management Science*, 14(4):383–384, 2011.

[263] P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, and P.J.M. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012.

[264] P.J.H. Hulshof, P.T. Vanberkel, R.J. Boucherie, E.W. Hans, M. Van Houdenhoven, and J.C.W. van Ommeren. Analytical models to determine room requirements in outpatient clinics. *OR Spectrum*, 34(2, SI):391–405, 2012.

[265] INFORMS Website. *Retrieved May 10, 2012, from: http://www.informs.org/*, 2012.

[266] A. Ingolfsson, S. Budge, and E. Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11(3):262–274, 2008.

[267] A. Ingolfsson, E. Erkut, and S. Budge. Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society*, 54(7):736–746, 2003.

[268] V. Irvine, S. McClean, and P. Millard. Stochastic models for geriatric in-patient behaviour. *Mathematical Medicine and Biology*, 11(3):207, 1994.

[269] M. W. Isken, T. J. Ward, and T. C. McKee. Simulating outpatient obstetrical clinics. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, editors, *Proceedings of the 31st Conference Winter Simulation*, pages 1557–1563, ACM, New York, NY, USA, 1999.

[270] M.W. Isken, T.J. Ward, and S.J. Littig. An open source software project for obstetrical procedure scheduling and occupancy analysis. *Health Care Management Science*, 14(1):56–73, 2011.

[271] E.P. Jack and T.L. Powers. A review and synthesis of demand management, capacity management and performance in health-care services. *International Journal of Management Reviews*, 11(2):149–174, 2009.

[272] J.R. Jackson. Jobshop-like queueing systems. *Management Science*, 50(12):1796–1802, 2004.

[273] B. Jaumard, F. Semet, and T. Vovor. A generalized linear programming model for nurse scheduling. *European Journal of Operational Research*, 107(1):1–18, 1998.

[274] A. Jebali, H. Alouane, B. Atidel, and P. Ladet. Operating rooms scheduling. *International Journal of Production Economics*, 99(1-2):52–62, 2006.

[275] P.A. Jensen and J.F. Bard. *Operations Research models and methods*. Wiley, 2003.

[276] H. Jia, F. Ordóñez, and M. Dessouky. A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions*, 39(1):41–55, 2007.

[277] L. Jiang and R. E. Giachetti. A queueing network model to analyze the impact of parallelization of care on patient cycle time. *Health Care Management Science*, 11(3):248–261, 2008.

[278] G. Johnson, K. Scholes, and R. Whittington. *Exploring corporate strategy*. Prentice Hall, New Jersey, 8th edition, 2008.

[279] M. J. Johnston, P. Samaranayake, A. Dadich, and J. A. Fitzgerald. Modelling radiology department operation using discrete event simulation. In *The 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*, 2009. Working paper.

[280] P.E. Joustra, J. de Wit, V.M.D. Struben, B.J.H. Overbeek, P. Fockens, and S.G. Elkhuizen. Reducing access times for an endoscopy department by an iterative combination of computer simulation and Linear Programming. *Health Care Management Science*, 13(1):17–26, 2010.

[281] P.E. Joustra, J. de Wit, N. Van Dijk, and P.J.M. Bakker. How to juggle priorities? an interactive tool to provide quantitative support for strategic patient-mix decisions: an ophthalmology case. *Health Care Management Science*, 14(4):348–360, 2011.

[282] J.B. Jun, S.H. Jacobson, and J.R. Swisher. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, 50(2):109–123, 1999.

[283] G.C. Kaandorp and G. Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3):217–229, 2007.

[284] E.P.C. Kao and G.G. Tung. Bed allocation in a public health care delivery system. *Management Science*, 27(5):507–520, 1981.

[285] A.S. Kapadia and Y.K.C.M. Fasihullah. Finite capacity priority queues with potential health applications. *Computers & Operations Research*, 12(4):411–420, 1985.

[286] A.S. Kapadia, S.E. Vineberg, and C.D. Rossi. Predicting course of treatment in a rehabilitation hospital: a Markovian model. *Computers & Operations Research*, 12(5):459–469, 1985.

[287] K. Katsaliaki and N. Mustafee. Applications of simulation within the healthcare context. *Journal of the Operational Research Society*, 62(8):1431–1451, 2010.

[288] J.H. Katz. Simulation of outpatient appointment systems. *Communications of the ACM*, 12(4):215–222, 1969.

[289] D.L. Kellogg and S. Walczak. Nurse scheduling: from academia to implementation or not? *Interfaces*, 37(4):355, 2007.

[290] K. Khoumbati, M. Themistocleous, and Z. Irani. Evaluating the adoption of enterprise application integration in health-care organizations. *Journal of Management Information Systems*, 22(4):69–108, 2006.

[291] S.C. Kim and I. Horowitz. Scheduling hospital services: the efficacy of elective-surgery quotas. *Omega*, 30(5):335–346, 2002.

[292] S.C. Kim, I. Horowitz, K.K. Young, and T.A. Buckley. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research*, 115(1):36–46, 1999.

[293] S.C. Kim, I. Horowitz, K.K. Young, and T.A. Buckley. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, 18(4):427–443, 2000.

[294] N. Koizumi, E. Kuno, and T.E. Smith. Modeling patient flows using a queuing

network with blocking. *Health Care Management Science*, 8(1):49–60, 2005.

[295] A. Kokangul. A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. *Computer Methods and Programs in Biomedicine*, 90(1):56–65, 2008.

[296] P. Kolesar. A Markovian model for hospital admission scheduling. *Management Science*, 16(6):384–396, 1970.

[297] R. Kolisch and S. Sickinger. Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum*, 30(2):375–395, 2008.

[298] A. Kopzon, Y. Nazarathy, and G. Weiss. A push–pull network with infinite supply of work. *Queueing Systems*, 62(1):75–111, 2009.

[299] J. Kros, S. Dellana, and D. West. Overbooking increases patient access at East Carolina University's student health services clinic. *Interfaces*, 39(3):271–287, 2009.

[300] P.J. Kuzdrall, N.K. Kwak, and H.H. Schmitz. Simulating space requirements and scheduling policies in a hospital surgical suite. *Simulation*, 36(5):163–172, 1981.

[301] N.K. Kwak, P.J. Kuzdrall, and H.H. Schmitz. Simulating the use of space in a hospital surgical suite. *Simulation*, 25(5):147–151, 1975.

[302] N.K. Kwak, P.J. Kuzdrall, and H.H. Schmitz. The GPSS simulation of scheduling policies for surgical patients. *Management Science*, 22(9):982–989, 1976.

[303] L.R. LaGanga and S.R. Lawrence. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276, 2007.

[304] M. Lagergren. What is the role and contribution of models to management and research in the health services? A view from Europe. *European Journal of Operational Research*, 105(2):257–266, 1998.

[305] N. Lahrichi, SD Lapierre, A. Hertz, A. Talib, and L. Bouvier. Analysis of a territorial approach to the delivery of nursing home care services based on historical data. *Journal of Medical Systems*, 30(4):283–291, 2006.

[306] M. Lamiri, F. Grimaud, and X. Xie. Optimization methods for a stochastic surgery planning problem. *International Journal of Production Economics*, 120(2):400–410, 2009.

[307] M. Lamiri, X. Xie, A. Dolgui, and F. Grimaud. A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research*, 185(3):1026–1037, 2008.

[308] M. Lamiri, X. Xie, and S. Zhang. Column generation approach to operating theater planning with elective and emergency patients. *IIE Transactions*, 40(9):838–852, 2008.

[309] T.P. Landau, T.R. Thiagarajan, and R.S. Ledley. Cost containment in the concentrated care center: a study of nursing, bed and patient assignment policies. *Computers in Biology and Medicine*, 13(3):205–238, 1983.

[310] D.C. Lane and E. Husemann. System dynamics mapping of acute patient flows. *Journal of the Operational Research Society*, 59(2):213–224, 2007.

[311] D.C. Lane, C. Monefeldt, and JV Rosenhead. Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society*, 51(5):518–531, 2000.

[312] J.R. Langabeer. *Health Care Operations Management: A Quantitative Approach to Business and Logistics*. Jones & Bartlett Publishers, Sudbury, MA, 2007.

[313] E. Lanzarone, A. Matta, and G. Scaccabarozzi. A patient stochastic model to support human resource planning in home care. *Production Planning and Control*, 21(1):3–25, 2010.

[314] R.C. Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23(5):845–868, 1975.

[315] V. Lattimer, S. Brailsford, J. Turnbull, P. Tarnaras, H. Smith, S. George, K. Gerard, and S. Maslin-Prothero. Reviewing emergency care systems I: insights from system dynamics modelling. *Emergency Medicine Journal*, 21(6):685–691, 2004.

[316] V. Lattimer, J. Turnbull, A. Burgess, H. Surridge, K. Gerard, J. Lathlean, H. Smith, and S. George. Effect of introduction of integrated out of hours care in England: observational study. *British Medical Journal*, 331(7508):81–84, 2005.

[317] K.C. Laudon and J.P. Laudon. *Management information systems*. Prentice Hall, New Jersey, 11th edition, 2010.

[318] M.S. Lavieri and M.L. Puterman. Optimizing nursing human resource planning in British Columbia. *Health Care Management Science*, 12(2):119–128, 2009.

[319] A. M. Law. *Simulation modeling and analysis*. McGraw-Hill, 4th edition, 2009.

[320] D.K.K. Lee and S.A. Zenios. Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations Research*, 57(4):852–865, 2009.

[321] S. Lee. The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, 62(10):1888–1897, 2010.

[322] H.S. Leff, M. Dada, and S.C. Graves. An LP planning model for a mental health community support system. *Management Science*, 32(2):139–155, 1986.

[323] B. Lehaney, S.A. Clarke, and R.J. Paul. A case of an intervention in an outpatients department. *Journal of the Operational Research Society*, 50(9):877–891, 1999.

[324] H. Levy and M. Sidi. Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communcations*, 38(10):1750–1760, 1990.

[325] L.X. Li and W.C. Benton. Performance measurement criteria in health care organizations: Review and future research directions. *European Journal of Operational Research*, 93(3):449–468, 1996.

[326] L.X. Li, W.C. Benton, and G.K. Leong. The impact of strategic operations management decisions on community hospital performance. *Journal of Operations Management*, 20(4):389–408, 2002.

[327] X. Li, P. Beullens, D. Jones, and M. Tamiz. An integrated queuing and multi-objective bed allocation model with application to a hospital in China. *Journal of the Operational Research Society*, 60(3):330–338, 2009.

[328] X. Li, Z. Zhao, X. Zhu, and T. Wyatt. Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, 74(3):1–30, 2011.

[329] C.J. Liao, C.D. Pegden, and M. Rosenshine. Planning timely arrivals to a stochastic production or service system. *IIE Transactions*, 25(5):63–73, 1993.

[330] C.S. Lim, R. Mamat, and T. Bräunl. Impact of ambulance dispatch policies on performance of emergency medical services. *Intelligent Transportation Systems, IEEE Transactions on*, 12(2):624–632, 2011.

[331] S.J. Littig and M.W. Isken. Short term hospital occupancy prediction. *Health Care Management Science*, 10(1):47–66, 2007.

[332] L. Liu and X. Liu. Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, 49(12):1254–1259, 1998.

[333] N. Liu, S. Ziya, and V.G. Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2):347–364, 2010.

[334] W.S. Lovejoy and Y. Li. Hospital operating room capacity expansion. *Management*

Science, 48(11):1369–1387, 2002.

[335] M. Mackay. Practical experience with bed occupancy management and planning systems: an Australian view. *Health Care Management Science*, 4(1):47–56, 2001.

[336] J.M. Magerlein and J.B. Martin. Surgical demand scheduling: a review. *Health Services Research*, 13(4):418–433, 1978.

[337] E. Marcon and F. Dexter. Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Science*, 9(1):87–98, 2006.

[338] E. Marcon, S. Kharraja, and G. Simonnet. The operating theatre planning by the follow-up of the risk of no realization. *International Journal of Production Economics*, 85(1):83–90, 2003.

[339] V. Marianov and C. ReVelle. The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research*, 93(1):110–120, 1996.

[340] R.A. Marjamaa, P.M. Torkki, E.J. Hirvensalo, and O.A. Kirvelä. What is the best workflow for an operating room? A simulation study of five scenarios. *Health Care Management Science*, 12(2):142–146, 2009.

[341] I. Marques, M.E. Captivo, and M. Vaz Pato. An integer programming approach to elective surgery scheduling. *OR Spectrum*, 34(2):407–427, 2011.

[342] H.B. Marri, A. Gunasekaran, and R.J. Grieve. Computer-aided process planning: a state of art. *The International Journal of Advanced Manufacturing Technology*, 14(4):261–268, 1998.

[343] A.H. Marshall and S.I. McClean. Using Coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Science*, 7(4):285–289, 2004.

[344] A.H. Marshall, S.I. McClean, and P.H. Millard. Addressing bed costs for the elderly: a new methodology for modelling patient outcomes and length of stay. *Health Care Management Science*, 7(1):27–33, 2004.

[345] A.H. Marshall, S.I. McClean, C.M. Shapcott, and P.H. Millard. Modelling patient duration of stay to facilitate resource management of geriatric hospitals. *Health Care Management Science*, 5(4):313–319, 2002.

[346] A.H. Marshall, B. Shaw, and S.I. McClean. Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution. *Statistics in Medicine*, 26(13):2716–2729, 2007.

[347] B.J. Masterson, T.G. Mihara, G. Miller, S.C. Randolph, M.E. Forkner, and A.L. Crouter. Using models and data to support optimization of the military health system: A case study in an intensive care unit. *Health Care Management Science*, 7(3):217–224, 2004.

[348] M.E. Matta and S.S. Patterson. Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. *Health Care Management Science*, 10(2):173–194, 2007.

[349] R.S. Maull, PA Smart, A. Harris, and A.A.F. Karasneh. An evaluation of 'fast track' in A&E: a discrete event simulation approach. *The Service Industries Journal*, 29(7):923–941, 2009.

[350] M.S. Maxwell, M. Restrepo, S.G. Henderson, and H. Topaloglu. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22(2):266–281, 2010.

[351] J.H. May, W.E. Spangler, D.P. Strum, and L.G. Vargas. The surgical scheduling problem: Current research and future opportunities. *Production and Operations*

Management, 20(3):392–405, 2011.

[352] L. Mayhew and D. Smith. 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, 2008.

[353] A. Maynard. Developing the health care market. *The Economic Journal*, 101(408):1277–1286, 1991.

[354] A. Maynard. Can competition enhance efficiency in health care? lessons from the reform of the uk national health service. *Social science & medicine*, 39(10):1433–1445, 1994.

[355] J.O. McClain. A model for regional obstetric bed planning. *Health Services Research*, 13(4):378–394, 1978.

[356] S. McClean, M. Barton, L. Garg, and K. Fullerton. A modeling framework that combines markov models and discrete-event simulation for stroke patient care. *ACM Transactions on Modeling and Computer Simulation*, 21(4):25, 2011.

[357] S. McClean and P. Millard. Patterns of length of stay after admission in geriatric medicine: an event history approach. *The Statistician*, 42(3):263–274, 1993.

[358] S. McClean and P. Millard. Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care? *Journal of the Operational Research Society*, 58(2):255–261, 2007.

[359] S.I. McClean, B. McAlea, and P.H. Millard. Using a Markov reward model to estimate spend-down costs for a geriatric department. *Journal of the Operational Research Society*, 49(10):1021–1025, 1998.

[360] S.I. McClean and P.H. Millard. A three compartment model of the patient flows in a geriatric department: a decision support approach. *Health Care Management Science*, 1(2):159–163, 1998.

[361] D.B. McLaughlin and J.M. Hays. *Healthcare operations management*. Health Administration Press; AUPHA Press, 2008.

[362] Medical Subject Headings (MeSH). *Retrieved May 10, 2012, from: http://www.nlm.nih.gov/mesh/*, 2012.

[363] B. Mielczarek and J. Uziałko-Mydlikowska. Application of computer simulation modeling in the health care sector: a survey. *Simulation*, 88(2):197–216, 2012.

[364] P.H. Millard, G. Christodoulou, C. Jagger, G.W. Harrison, and S.I. McClean. Modelling hospital and social care bed occupancy and use by elderly people in an english health district. *Health Care Management Science*, 4(1):57–62, 2001.

[365] D. Min and Y. Yih. An elective surgery scheduling problem considering patient priority. *Computers & Operations Research*, 37(6):1091–1099, 2010.

[366] D. Min and Y. Yih. Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, 206(3):642–652, 2010.

[367] C. Mullinax and M. Lawley. Assigning patients to nurses in neonatal intensive care. *Journal of the Operational Research Society*, 53(1):25–35, 2002.

[368] N. Mustafee, K. Katsaliaki, and S.J.E. Taylor. Profiling literature in healthcare simulation. *Simulation*, 86(8-9):543, 2010.

[369] K. Muthuraman and M. Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9):820–837, 2008.

[370] NBII. Website of the National Biological Information Infrastructure (NBII). *Retrieved May 10, 2012, from: http://www.nbii.gov*, 2012.

[371] J.M. Nguyen, P. Six, D. Antonioli, P. Glemain, G. Potel, P. Lombrail, and P. Le Beux. A simple method to optimize hospital beds capacity. *International Journal of Medical*

Informatics, 74(1):39–49, 2005.

[372] J.M. Nguyen, P. Six, T. Chaussalet, D. Antonioli, P. Lombrail, and P. Le Beux. An objective method for bed capacity planning in a hospital department – a comparison with target ratio methods. *Methods of Information in Medicine*, 46(4):399–405, 2007.

[373] J.M. Nguyen, P. Six, R. Parisot, D. Antonioli, F. Nicolas, and P. Lombrail. A universal method for determining intensive care unit bed requirements. *Intensive Care Medicine*, 29(5):849–852, 2003.

[374] L.G.N. Nunes, S.V. de Carvalho, and R.C.M. Rodrigues. Markov decision process applied to the control of hospital elective admissions. *Artificial Intelligence in Medicine*, 47(2):159–171, 2009.

[375] J.P. Oddoye, D.F. Jones, M. Tamiz, and P. Schmidt. Combining simulation and goal programming for healthcare planning in a medical assessment unit. *European Journal of Operational Research*, 193(1):250–261, 2009.

[376] J.P. Oddoye, M.A. Yaghoobi, M. Tamiz, D.F. Jones, and P. Schmidt. A multi-objective model to determine efficient resource levels in a medical assessment unit. *Journal of the Operational Research Society*, 58(12):1563–1573, 2006.

[377] OECD. *Health at a glance 2011: OECD indicators*. OECD Publishing, 2011.

[378] OECD. *OECD Health Data 2013*. OECD, 2013.

[379] OECD Glossary. *Retrieved June 11, 2012, from: http://http://stats.oecd.org/glossary/*, 2012.

[380] H.C. Oh and W.L. Chow. Scientific evaluation of polyclinic operating strategies with discrete-event simulation. *International Journal of Simulation Modelling*, 10(4):165–176, 2011.

[381] M. Olivares, C. Terwiesch, and L. Cassorla. Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science*, 54(1):41–55, 2008.

[382] ORAHS - Operational Research Applied to Health Services. *Retrieved July 12, 2011, from: http://orahs.di.unito.it/*, 2011.

[383] ORchestra. Developed by Center for Healthcare Operations and Improvement Research (CHOIR) at the University of Twente. *Retrieved May 10, 2012, from: http://www.utwente.nl/choir/en/orchestra/*, 2012.

[384] Organisation of Economic Co-operation and Development (OECD). Data for 2011 from the website of oecd. *Retrieved May 10, 2011, from: http://www.nlm.nih.gov/mesh/*, 2011.

[385] Organisation of Economic Co-operation and Development (OECD). *Data retrieved May 10, 2012, from: http://www.oecd.org/health*, 2012.

[386] J. Orlicky. *Material requirements planning*. McGraw-Hill, London, 1975.

[387] Y.A. Ozcan. *Quantitative methods in health care management: techniques and applications*. Jossey Bass/Wiley, 2nd edition, 2009.

[388] C.H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Dover Publications, 1998.

[389] F. Pasin, M.H. Jobin, and J.F. Cordeau. An application of simulation to analyse resource sharing among health-care organisations. *International Journal of Operations and Production Management*, 22(4):381–393, 2002.

[390] J. Patrick. A Markov decision model for determining optimal outpatient scheduling. *Health Care Management Science*, 15(2):91–102, 2012.

[391] J. Patrick, M.L. Puterman, and M. Queyranne. Dynamic multi-priority patient

scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525, 2008.

[392] Jonathan Patrick. Access to long-term care: The true cause of hospital congestion? *Production and Operations Management*, 20(3):347–358, 2011.

[393] S.A. Paul, M.C. Reddy, and C.J. DeFlitch. A systematic review of simulation studies investigating emergency department overcrowding. *Simulation*, 86(8-9):559–571, 2010.

[394] C. Pelletier, T.J. Chaussalet, and H. Xie. A framework for predicting gross institutional long-term care cost arising from known commitments at local authority level. *Journal of the Operational Research Society*, 56(2):144–152, 2004.

[395] E. Pérez, L. Ntaimo, C. Bailey, and P. McCormack. Modeling and simulation of nuclear medicine patient service management in DEVS. *Simulation*, 86(8-9):481–501, 2010.

[396] M.J. Persson and J.A. Persson. Analysing management policies for operating room planning using simulation. *Health Care Management Science*, 13(2):182–191, 2010.

[397] D. Petrovic, M. Morshed, and S. Petrovic. Multi-objective genetic algorithms for scheduling of radiotherapy treatments for categorised cancer patients. *Expert Systems with Applications*, 38(6):6994–7002, 2011.

[398] D.N. Pham and A. Klinkert. Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, 185(3):1011–1025, 2008.

[399] W.P. Pierskalla and D.J. Brailer. Applications of operations research in health care delivery. *In: Handbooks in OR & MS (S.M Pollock, M.H. Rotkopf, A. Barnett (eds))*, 6:469–505, 1994.

[400] M.E. Porter. *Competitive advantage*, volume 15. Free Press New York, 1985.

[401] M.E. Porter and E.O. Teisberg. How physicians can change the future of health care. *Journal of the American Medical Association*, 297(10):1103–1111, 2007.

[402] W.B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 2011.

[403] C. Price, B. Golden, M. Harrington, R. Konewko, E. Wasil, and W. Herring. Reducing boarding in a post-anesthesia care unit. *Production and Operations Management*, 20(3):431–441, 2011.

[404] Z.H. Przasnyski. Operating room scheduling. a literature review. *Association of PeriOperative Registered Nurses Journal*, 44(1):67–82, 1986.

[405] PubMed. *Retrieved May 10, 2012, from: http://www.pubmed.gov/*, 2012.

[406] P. Punnakitikashem, J.M. Rosenberger, and D. Buckley Behan. Stochastic programming for nurse assignment. *Computational Optimization and Applications*, 40(3):321–349, 2008.

[407] M.L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. Wiley, NY, USA, 1994.

[408] X. Qu, R.L. Rardin, J.A.S. Williams, and D.R. Willis. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2):812–826, 2007.

[409] X. Qu and J. Shi. Modeling the effect of patient choice on the performance of open access scheduling. *International Journal of Production Economics*, 129(2):314–327, 2011.

[410] A. Rais and A. Viana. Operations research in healthcare: a survey. *International Transactions in Operational Research*, 18(1):1–31, 2011.

[411] T.A. Reilly, V.P. Marathe, and B.E. Fries. A delay-scheduling model for patients using a walk-in clinic. *Journal of Medical Systems*, 2(4):303–313, 1978.

[412] J. F. Repede and J. J. Bernardo. Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research*, 75(3):567–581, 1994.

[413] C. ReVelle. Review, extension and prediction in emergency service siting models. *European Journal of Operational Research*, 40(1):58–69, 1989.

[414] D.M. Rhyne and D. Jupp. Health care requirements planning: A conceptual framework. *Health Care Management Review*, 13(1):17–27, 1988.

[415] J.C. Ridge, S.K. Jones, M.S. Nielsen, and A.K. Shahani. Capacity planning for intensive care units. *European Journal of Operational Research*, 105(2):346–355, 1998.

[416] A. Riise and E.K. Burke. Local search for the surgery admission planning problem. *Journal of Heuristics*, 17(4):389–414, 2011.

[417] E.J. Rising, R. Baron, and B. Averill. A systems analysis of a university-health-service outpatient clinic. *Operations Research*, 21(5):1030–1047, 1973.

[418] L.W. Robinson and R.R. Chen. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, 35(3):295–307, 2003.

[419] L.W. Robinson and R.R. Chen. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Operations Management*, 12(2):330–346, 2010.

[420] T.R. Rohleder, D.P. Bischak, and L.B. Baskin. Modeling patient service centers with simulation and system dynamics. *Health Care Management Science*, 10(1):1–12, 2007.

[421] T.R. Rohleder, P. Lewkonia, D.P. Bischak, P. Duffy, and R. Hendijani. Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. *Health Care Management Science*, 14(2):135–145, 2011.

[422] B. Roland, C. Di Martinelly, F. Riane, and Y. Pochet. Scheduling an operating theatre under human resource constraints. *Computers & Industrial Engineering*, 58(2):212–220, 2010.

[423] E. Rönnberg and T. Larsson. Automating the self-scheduling process of nurses in swedish healthcare: a pilot study. *Health Care Management Science*, 13(1):35–53, 2010.

[424] S. M. Ross. *Introduction to probability models*. Academic Press, 9 edition, 2007.

[425] A.V. Roth and R.V. Dierdonck. Hospital resource planning: Concepts, feasibility, and framework. *Production and Operations Management*, 4(1):2–29, 1995.

[426] R.J. Ruth. A mixed integer programming model for regional planning of a hospital inpatient service. *Management Science*, 7(5):521–533, 1981.

[427] C. Samuel, K. Gonapa, P.K. Chaudhary, and A. Mishra. Supply chain dynamics in healthcare services. *International Journal of Health Care Quality Assurance*, 23(7):631–642, 2010.

[428] P. Santibáñez, M. Begen, and D. Atkins. Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority. *Health Care Management Science*, 10(3):269–282, 2007.

[429] E. S. Savas. Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science*, 15(12):608–627, 1969.

[430] K. Schimmelpfeng, S. Helber, and S. Kasper. Decision support for rehabilitation hospital scheduling. *OR Spectrum*, 32(2):1–29, 2010.

[431] V. Schmid. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*,

219(3):611–621, 2012.

[432] H.H. Schmitz and N.K. Kwak. Monte Carlo simulation of operating-room and recovery-room usage. *Operations Research*, 20(6):1171–1180, 1972.

[433] H.H. Schmitz, N.K. Kwak, and P.J. Kuzdrall. Determination of surgical suite capacity and an evaluation of patient scheduling policies. *RAIRO - Operations Research*, 12(1):3–14, 1978.

[434] A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Springer Verlag, 2003.

[435] E. Schut, S. Sorbe, and J. Hoj. Health care reform and long-term care in the netherlands. *Organisation for Economic Co-operation and Development (OECD) - Economics Department Working Paper*, 1010:1–36, 2013.

[436] H.J. Schütz and R. Kolisch. Approximate dynamic programming for capacity allocation in the service industry. *European Journal of Operational Research*, 218(1):239 – 250, 2012.

[437] Scopus. *Retrieved May 10, 2012, from: http://www.scopus.com/*, 2012.

[438] D.G. Seymour. Health care modelling and clinical practice. theoretical exercise or practical tool? *Health Care Management Science*, 4(1):7–12, 2001.

[439] A.K. Shahani, S.A. Ridley, and M.S. Nielsen. Modelling patient flows as an aid to decision making for critical care capacities and organisation. *Anaesthesia*, 63(10):1074–1080, 2008.

[440] B. Shaw and A.H. Marshall. Modeling the health care costs of geriatric inpatients. *Information Technology in Biomedicine, IEEE Transactions on*, 10(3):526–532, 2006.

[441] S. Shepperd, J. McClaran, C.O. Phillips, N.A. Lannin, L.M. Clemson, A. McCluskey, I.D. Cameron, and S.L. Barras. Discharge planning from hospital to home. *Cochrane Database of Systematic Reviews*, 1:CD000313, 2010.

[442] A. Shmueli, C.L. Sprung, and E.H. Kaplan. Optimizing admissions to an intensive care unit. *Health Care Management Science*, 6(3):131–136, 2003.

[443] W. Shonick and J.R. Jackson. An improved stochastic model for occupancy-related random variables in general-acute hospitals. *Operations Research*, 21(4):952–965, 1973.

[444] D. Sier, P. Tobin, and C. McGurk. Scheduling surgical procedures. *Journal of the Operational Research Society*, 48(9):884–891, 1997.

[445] H. Simao and W.B. Powell. Approximate dynamic programming for management of high-value spare parts. *Journal of Manufacturing Technology Management*, 20(2):147–160, 2009.

[446] M. Singer and P. Donoso. Assessing an ambulance service with queuing theory. *Computers & Operations Research*, 35(8):2549–2560, 2008.

[447] D. Sinreich and O. Jabali. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science*, 10(3):293–308, 2007.

[448] D. Sinreich, O. Jabali, and N.P. Dellaert. Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *IIE Transactions*, 44(3):163–180, 2012.

[449] C.E. Smith, S.V.M. Kleinbeck, K. Fernengel, and L.S. Mayer. Efficiency of families managing home health care. *Annals of Operations Research*, 73(0):157–175, 1997.

[450] K.R. Smith, A.M. Over Jr, M.F. Hansen, F.L. Golladay, and E.J. Davenport. Analytic framework and measurement strategy for investigating optimal staffing in medical practice. *Operations Research*, 24(5):815–841, 1976.

[451] V.L. Smith-Daniels, S.B. Schweikhart, and D.E. Smith-Daniels. Capacity management in health care services: review and future research directions. *Decision Sciences*, 19(4):889–919, 1988.

[452] B.G. Sobolev, V. Sanchez, and C. Vasilakis. Systematic review of the use of computer simulation modeling of patient flow in surgical care. *Journal of medical systems*, 35(1):1–16, 2011.

[453] A. Sonnenberg. How to overbook procedures in the endoscopy unit. *Gastrointestinal endoscopy*, 69(3-Part-2):710–715, 2009.

[454] E.F. Stafford Jr and S.C. Aggarwal. Managerial analysis and decision-making in outpatient health clinics. *Journal of the Operational Research Society*, 30(10):905–915, 1979.

[455] M.C. Su, S.C. Lai, P.C. Wang, Y.Z. Hsieh, and S.C. Lin. A SOMO-based approach to the operating room scheduling problem. *Expert Systems with Applications*, 38(12):15447–15454, 2011.

[456] D. Sundaramoorthi, V.C.P. Chen, J.M. Rosenberger, S.B. Kim, and D.F. Buckley-Behan. A data-integrated simulation model to evaluate nurse–patient assignments. *Health Care Management Science*, 12(3):252–268, 2009.

[457] J. R. Swisher and S. H. Jacobson. Evaluating the design of a family practice healthcare clinic using discrete-event simulation. *Health Care Management Science*, 5(2):75–88, 2002.

[458] J.R. Swisher, S.H. Jacobson, J.B. Jun, and O. Balci. Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers and Operations Research*, 28(2):105–125, 2001.

[459] C. Swoveland, D. Uyeno, I. Vertinsky, and R. Vickson. Ambulance location: a probabilistic enumeration approach. *Management Science*, 20(4):686–698, 1973.

[460] H.A. Taha. *Operations research: an introduction*. Prentice Hall; 9th edition, 2010.

[461] H. Takagi. Queueing analysis of polling models: progress in 1990-1994. In *Frontiers in queueing: models and applications in science and engineering*, pages 119–146. CRC Press, Inc., Boca Raton, FL, USA, 1998.

[462] E. Tànfani and A. Testi. A pre-assignment heuristic algorithm for the Master Surgical Schedule Problem (MSSP). *Annals of Operations Research*, 178(1):105–119, 2010.

[463] H. Tarakci, Z. Ozdemir, and M. Sharafali. On the staffing policy and technology investment in a specialty hospital offering telemedicine. *Decision Support Systems*, 46(2):468–480, 2009.

[464] G. Taylor, S. McClean, and P. Millard. Geriatric-patient flow-rate modelling. *Mathematical Medicine and Biology*, 13(4):297, 1996.

[465] G.J. Taylor, S.I. McClean, and P.H. Millard. Continuous-time Markov models for geriatric patient behaviour. *Applied Stochastic Models and Data Analysis*, 13(3-4):315–323, 1997.

[466] G.J. Taylor, S.I. McClean, and P.H. Millard. Using a continuous-time Markov model with Poisson arrivals to describe the movements of geriatric patients. *Applied Stochastic Models and Data Analysis*, 14(2):165–174, 1998.

[467] G.J. Taylor, S.I. McClean, and P.H. Millard. Stochastic models of geriatric patient bed occupancy behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):39–48, 2000.

[468] I.D.S. Taylor and J.G.C. Templeton. Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service. *Operations Research*, 28(5):1168–1188, 1980.

[469] A. Testi and E. Tànfani. Tactical and operational decisions for operating room planning: Efficiency and welfare implications. *Health Care Management Science*, 12(4):363–373, 2009.

[470] A. Testi, E. Tànfani, and G. Torre. A three-phase approach for operating theatre schedules. *Health Care Management Science*, 10(2):163–172, 2007.

[471] S.J. Thomas. Capacity and demand models for radiotherapy treatment machines. *Clinical Oncology*, 15(6):353 – 358, 2003.

[472] S. Thompson, M. Nunez, R. Garfinkel, and M.D. Dean. Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations research*, 57(2):261–273, 2009.

[473] H.C. Tijms. *A first course in stochastic models*. John Wiley & Sons Inc, 2003.

[474] H. Topaloglu and W.B. Powell. Dynamic-programming approximations for stochastic time-staged integer multicommodity-flow problems. *INFORMS Journal on Computing*, 18(1):31–42, 2006.

[475] C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

[476] P. Toth and D. Vigo. *The Vehicle Routing Problem*. SIAM, Philadelphia, 2001.

[477] A. Trautsamwieser, M. Gronalt, and P. Hirsch. Securing home health care in times of natural disasters. *OR Spectrum*, 33(3):1–27, 2011.

[478] P.M. Troy and L. Rosenberg. Using simulation to determine the need for ICU beds for surgery patients. *Surgery*, 146(4):608–620, 2009.

[479] W.J.C. Tunnicliffe. A review of operational problems tackled by computer simulation in health care facilities. *Health and Social Service Journal*, 90(4702):73–80, 1980.

[480] M. Utley, S. Gallivan, K. Davis, P. Daniel, P. Reeves, and J. Worrall. Estimating bed requirements for an intermediate care facility. *European Journal of Operational Research*, 150(1):92–100, 2003.

[481] M. Utley, S. Gallivan, T. Treasure, and O. Valencia. Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. *Health Care Management Science*, 6(2):97–104, 2003.

[482] M. Utley, M. Jit, and S. Gallivan. Restructuring routine elective services to reduce overall capacity requirements within a local health economy. *Health Care Management Science*, 11(3):240–247, 2008.

[483] C. Valouxis and E. Housos. Hybrid optimization techniques for the workshift and rest assignment of nursing personnel. *Artificial Intelligence in Medicine*, 20(2):155–175, 2000.

[484] N.M. Van Dijk and N. Kortbeek. Erlang loss bounds for OT–ICU systems. *Queueing Systems*, 63(1):253–280, 2009.

[485] E. Van Gameren and I. Woittiez. Transitions between care provisions demanded by Dutch elderly. *Health Care Management Science*, 8(4):299–313, 2005.

[486] M. Van Houdenhoven, J.M. van Oostrum, E.W. Hans, G. Wullink, and G. Kazemier. Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. *Anesthesia & Analgesia*, 105(3):707–714, 2007.

[487] M. Van Houdenhoven, J.M. van Oostrum, G. Wullink, E. Hans, J.L. Hurink, J. Bakker, and G. Kazemier. Fewer intensive care unit refusals and a higher capacity utilization by using a cyclic surgical case schedule. *Journal of Critical Care*, 23(2):222–226, 2008.

[488] J.M. van Oostrum, M. Van Houdenhoven, J.L. Hurink, E.W. Hans, G. Wullink, and G. Kazemier. A master surgical scheduling approach for cyclic scheduling in oper-

ating room departments. *OR Spectrum*, 30(2):355–374, 2008.

[489] C.J.T. van Uden, A.J.H.A. Ament, G.B.W.E. Voss, G. Wesseling, R.A.G. Winkens, O.C.P. van Schayck, and H.F.J.M. Crebolder. Out-of-hours primary care. implications of organisation on costs. *BMC Family Practice*, 7(1):29, 2006.

[490] P.T. Vanberkel and J.T. Blake. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health Care Management Science*, 10(4):373–385, 2007.

[491] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, and N. Litvak. A survey of health care models that encompass multiple departments. *International Journal of Health Management and Information*, 1(1):37–69, 2010.

[492] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, W.A.M. van Lent, and W.H. van Harten. Accounting for inpatient wards when developing master surgical schedules. *Anesthesia & Analgesia*, 112(6):1472–1479, 2011.

[493] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, W.A.M. van Lent, and W.H. van Harten. An exact approach for relating recovering surgical patient workload to the master surgical schedule. *Journal of the Operational Research Society*, 62(10):1851–1860, 2011.

[494] S. Vanderby and M.W. Carter. An evaluation of the applicability of system dynamics to patient flow modelling. *Journal of the Operational Research Society*, 61(11):1572–1581, 2009.

[495] C. Vasilakis and E. El-Darzi. A simulation study of the winter bed crisis. *Health Care Management Science*, 4(1):31–36, 2001.

[496] C. Vasilakis, B.G. Sobolev, L. Kuramoto, and A.R. Levy. A simulation study of scheduling clinic appointments in surgical care: individual surgeon versus pooled lists. *Journal of the Operational Research Society*, 58(2):202–211, 2007.

[497] G. Vassilacopoulos. A simulation model for bed allocation to hospital inpatient departments. *Simulation*, 45(5):233–241, 1985.

[498] I.B. Vermeulen, S.M. Bohte, S.G. Elkhuizen, H. Lameris, P.J.M. Bakker, and H.L. Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine*, 46(1):67–80, 2009.

[499] S. Villa, M. Barbieri, and F. Lega. Restructuring patient flow logistics around patient care needs: implications and practicalities from three critical cases. *Health Care Management Science*, 12(2):155–165, 2009.

[500] J. M. H. Vissers, J. W. M. Bertrand, and G. de Vries. A framework for production control in health care organizations. *Production Planning & Control*, 12(6):591–604, 2001.

[501] J.M.H. Vissers. Patient flow-based allocation of inpatient resources: a case study. *European Journal of Operational Research*, 105(2):356–370, 1998.

[502] J.M.H. Vissers, I.J.B.F. Adan, and N.P. Dellaert. Developing a platform for comparison of hospital admission systems: An illustration. *European Journal of Operational Research*, 180(3):1290–1301, 2007.

[503] J.M.H. Vissers and R. Beech. *Health operations management: patient flow logistics in health care*. Routledge Health Management. Routledge, London, 2005.

[504] J.M.H. Vissers and J. Wijngaard. The outpatient appointment system: Design of a simulation study. *European Journal of Operational Research*, 3(6):459–463, 1979.

[505] M. Von Korff. A statistical model of the duration of mental hospitalization: The mixed exponential distribution. *Journal of Mathematical Sociology*, 6(2):169–175, 1979.

[506] R.E. Wachtel and F. Dexter. Tactical increases in operating room block time for capacity planning should not be based on utilization. *Anesthesia & Analgesia*, 106(1):215, 2008.

[507] L.M. Walts and A.S. Kapadia. Patient classification system: an optimization approach. *Health Care Management Review*, 21(4):75, 1996.

[508] J. Wang, S. Quan, J. Li, and A. Hollis. Modeling and analysis of work flow and staffing level in a computed tomography division of University of Wisconsin Medical Foundation. *Health Care Management Science*, 15(2):108–120, 2012.

[509] W.Y. Wang and D. Gupta. Adaptive appointment systems with patient preferences. *Manufacturing and Service Operations Management*, 13(3):373–389, 2011.

[510] Web of Science (WoS). *Retrieved May 10, 2012, from: http://www.isiknowledge.com/*, 2012.

[511] E.N. Weiss. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions*, 22(2):143–150, 1990.

[512] E.N. Weiss, M.A. Cohen, and J.C. Hershey. An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. *Operations Research*, 30(6):1082–1104, 1982.

[513] E.N. Weiss and J.O. McClain. Administrative days in acute care facilities: A queueing-analytic approach. *Operations Research*, 35(1):35–44, 1987.

[514] J.D. Welch and N.T.J. Bailey. Appointment systems in hospital outpatient departments. *The Lancet*, 259(6718):1105–1108, 1952.

[515] G. Werker, A. Sauré, J. French, and S. Shechter. The use of discrete-event simulation modelling to improve radiation therapy planning processes. *Radiotherapy and Oncology*, 92(1):76–82, 2009.

[516] G.P. Westert, J.S. Burgers, and H. Verkleij. The Netherlands: regulated competition behind the dykes? *British Medical Journal*, 339:839–842, 2009.

[517] F. Wharton. On the risk of premature transfer from coronary care units. *Omega*, 24(4):413–423, 1996.

[518] M.J.B. White and M.C. Pike. Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care*, 2(3):133–145, 1964.

[519] S.V. Williams. How many intensive care beds are enough? *Critical Care Medicine*, 11(6):412, 1983.

[520] W.L. Winston. *Operations research: applications and algorithms*. Duxbury Press, 2003.

[521] R.W. Wolff. *Stochastic modeling and the theory of queues*. Prentice Hall, 1989.

[522] D.J. Worthington. Queueing models for hospital waiting lists. *Journal of the Operational Research Society*, 38(5):413–422, 1987.

[523] D.J. Worthington. Hospital waiting list management models. *Journal of the Operational Research Society*, 42(10):833–843, 1991.

[524] M.B. Wright. The application of a surgical bed simulation model. *European journal of operational research*, 32(1):26–32, 1987.

[525] P.D. Wright, K.M. Bretthauer, and M.J. Côté. Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages. *Decision Sciences*, 37(1):39–70, 2006.

[526] C.H. Wu and K.P. Hwang. Using a discrete-event simulation to balance ambulance availability and demand in static deployment systems. *Academic Emergency Medicine*, 16(12):1359–1366, 2009.

[527] G. Wullink. *Resource loading under uncertainty*. PhD thesis, University of Twente, the Netherlands, 2005.

[528] H. Xie, T.J. Chaussalet, and P.H. Millard. A continuous time Markov model for the

length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):51–61, 2005.

[529] H. Xie, T.J. Chaussalet, and P.H. Millard. A model-based approach to the analysis of patterns of length of stay in institutional long-term care. *Information Technology in Biomedicine, IEEE Transactions on*, 10(3):512–518, 2006.

[530] H. Xie, T.J. Chaussalet, W.A. Thompson, and P.H. Millard. A simple graphical decision aid for the placement of elderly people in long-term care. *Journal of the Operational Research Society*, 58(4):446–453, 2006.

[531] R.Y.T. Yeung, G.M. Leung, S.M. McGhee, and J.M. Johnston. Waiting time and doctor shopping in a mixed medical economy. *Health Economics*, 13(11):1137–1144, 2004.

[532] S. Zeltyn, Y.N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, et al. Simulation-based models of emergency departments: operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation*, 21(4):24, 2011.

[533] B. Zhang, P. Murali, M.M. Dessouky, and D. Belson. A mixed integer programming approach for allocating operating room capacity. *Journal of the Operational Research Society*, 60(5):663–673, 2009.

[534] W.H.M. Zijm. Towards intelligent manufacturing planning and control systems. *OR Spectrum*, 22(3):313–345, 2000.

[535] M. E. Zonderland, F. Boer, R. J. Boucherie, A. de Roode, and J. W. van Kleef. Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. *Anesthesia & Analgesia*, 109(5):1612–1614, 2009.

[536] M. E. Zonderland, R. J. Boucherie, N. Litvak, and C. L. A. M. Vleggert-Lankamp. Planning and scheduling of semi-urgent surgeries. *Health Care Management Science*, 13(3):256–267, 2010.

Acronyms

Acronym	Description
ADP	Approximate Dynamic Programming
CHOIR	Center for Healthcare Operations Improvement and Research
CS	Computer Simulation
CT	Computerized Tomography
CV	Coefficient of Variation
DEA	Data Envelopment Analysis
DP	Dynamic Programming
DRG	Diagnosis-Related Group
DtP	Doctor to Patient
ED	Emergency Department
FCFS	First Come, First Served
GDP	Gross Domestic Product
GHZ	Groene Hart Ziekenhuis
GP	General Practitioner
HCMS	Health Care Management Science
HE	Heuristics
ICT	Information Communications Technology
ICU	Intensive Care Unit
ILP	Integer Linear Program
INFORMS	The Institute for Operations Research and the Management Sciences
LPTF	Longest Processing Time First
LR	Literature Review

Acronym	Description
MC-OQN	Multi-Class Open Queueing Network
MCU	Medium Care Unit
MDP	Markov Decision Problem
MeSH	Medical Subject Headings
MILP	Mixed Integer Linear Program
MP	Mathematical Programming
MPC	Manufacturing Planning and Control
MRI	Magnetic Resonance Imaging
MRP	Materials Requirements Planning
MRP-II	Manufacturing Resource Planning
MSS	Master Surgical Schedule
MV	Markov Processes
OR/MS	Operations Research and Management Science
ORAHS	Operational Research Applied to Health Services
PAC	Production Authorization Card
PACU	Post Anesthesia Care Unit
PCC	Primary Care Cooperative
PtD	Patient to Doctor
QT	Queueing Theory
RUG	Resource Utilization Group
SFA	Stochastic Frontier Analysis
WoS	Web of Science

Summary

Healthcare professionals face the challenging task to design and organize their processes more effectively and efficiently. Designing and organizing processes is known as planning and control. Healthcare planning and control lags behind manufacturing and control for various reasons. One of the main causes is the fragmented nature of healthcare planning and control. Healthcare organizations such as hospitals are typically formed as a cluster of autonomous departments, where planning and control is also often functionally dispersed. A more integrated approach to healthcare planning and control is likely to bring improvements, as healthcare planning and control in one department is frequently dependent on decision making in other departments in the patient's care chain.

Driven by the lack of frameworks for healthcare planning and control, an integreated framework for healthcare planning and control is discussed in this thesis. The developed framework integrates all managerial areas involved in healthcare delivery operations and all hierarchical levels of control, to ensure completeness and coherence of responsibilities for every managerial area. The framework is built upon the "classical" hierarchical decomposition often used in manufacturing planning and control, which discerns *strategic*, *tactical*, and *operational* levels of control. This decomposition is extended by discerning between *offline* and *online* on the operational level. This distinction reflects the difference between "in advance" decision making and "reactive" decision making.

The integrated framework for planning and control in healthcare can be used to structure the various planning and control functions, and their interaction. It is applicable broadly, to an individual department, an entire healthcare organization, and to a complete supply chain of cure and care providers. The framework can be used to identify and position various types of managerial problems, to demarcate the scope of organization interventions, and to facilitate a dialogue between clinical staff and managers. (**Chapter 2**)

To position the research in this thesis and to investigate integrated planning and control in the literature, Chapter 3 provides a comprehensive overview of the typical decisions to be made in resource capacity planning and control in healthcare, and a structured review of relevant articles within the field of Operations Research and Management Science (OR/MS) for each

planning decision.

First, to position the planning decisions, a taxonomy is presented. This taxonomy provides healthcare managers and OR/MS researchers with a method to identify, break down and classify planning and control decisions. It is based on the integrated planning and control framework for healthcare, presented in Chapter 2. Second, following the taxonomy, for six health-care services, an exhaustive specification of planning and control decisions in resource capacity planning and control is provided. For each planning and control decision, the key OR/MS articles and the OR/MS methods and techniques that are applied in the literature to support decision making are reviewed and discussed.

Prior literature reviews conclude that there is a lack of models for complete healthcare processes. Although a body of literature focusing on two-departmental interactions was identified, very few contributions were found on complete hospital interactions, let alone on complete healthcare system interactions. The literature review in this thesis reconfirms these observations. (**Chapter 3**)

The second part of this thesis describes methods and models for integrated planning and control in healthcare, developed with techniques from OR/MS. The integrated models are developed for the tactical level of planning and control. Tactical planning of resources in hospitals concerns elective patient admission planning and the intermediate term allocation of resource capacities. Its main objectives are to achieve equitable access for patients, to meet production targets/to serve the strategically agreed number of patients, and to use resources efficiently.

A method is proposed to develop a tactical resource allocation and elective patient admission plan. These tactical plans allocate available resources to various care processes and determine the selection of patients to be served that are at a particular stage of their care process. The method is developed in a Mixed Integer Linear Programming (MILP) framework and copes with multiple resources, multiple time periods and multiple patient groups with various uncertain treatment paths through the hospital, thereby integrating decision making for a chain of hospital resources.

Computational results indicate that the developed method leads to a more equitable distribution of resources and provides control of patient access times, the number of patients served and the fraction of allocated resource capacity. The developed approach is generic, as the base MILP and the solution approach allow for including various extensions to both the objective criteria and the constraints. Consequently, the proposed method is applicable in various settings of tactical hospital management. (**Chapter 4**)

In Chapter 5 of this thesis, a method is proposed to develop tactical resource allocation and elective patient admission plans taking stochastic elements into consideration, thereby potentially providing more robust tactical plans. The stochastic formulation of the tactical planning problem is stated, and an exact Dynamic Programming (DP) solution approach is provided. As the exact DP approach is only tractable for extremely small instances, and it does not allow solving real-life sized instances of the tactical planning problem, a solution approach using an alternative technique within the framework of Approximate Dynamic Programming (ADP) is developed.

The developed ADP approach copes with multiple resources, multiple time periods and multiple patient groups with various uncertain treatment paths through the hospital. It incorporates the stochastic processes for (emergency) patient arrivals and patient transitions between queues in developing tactical plans. Moreover, it integrates decision making for a chain of hospital resources while taking stochastic elements into consideration.

Computational results show that the developed ADP approach is suitable for the tactical planning problem in healthcare and that it provides accurate results (close the exact results obtained with DP approach). The ADP approach performs significantly better than two alternative greedy planning approaches for large reallife-sized instances. (**Chapter 5**)

Chapter 6 discusses a tactical planning problem in the outpatient clinic. Outpatient clinics traditionally organize processes such that the doctor remains in a consultation room while patients visit for consultation. This is called the Patient-to-Doctor policy in this thesis. An alternative approach is the Doctor-to-Patient policy, whereby the doctor travels between multiple consultation rooms, in which patients prepare for their consultation. In the latter approach, the doctor saves time by consulting fully prepared patients.

Using a queueing theoretic and a discrete-event simulation approach, generic models are developed that enable performance evaluations of the two policies for different parameter settings. These models can be used by managers of outpatient clinics to compare the two policies and choose a particular policy when redesigning the patient process. In addition, methods are developed to calculate the required number of consultation rooms in the Doctor-to-Patient policy. In the discussed computational experiments, the developed methods are applied to a range of distributions and parameters, and to a case study in one of the general hospitals that inspired this research. (**Chapter 6**)

Samenvatting

Zorgmanagers staan voor de uitdagende taak om zorgprocessen efficiënter en effectiever in te richten en te organiseren. Het ontwerpen en organiseren van processen wordt in de literatuur planning en besturing genoemd. De ontwikkelingen in planning en besturing van zorgprocessen lopen achter op planning en besturing van productieprocessen door verschillende oorzaken. Een belangrijke oorzaak is de gefragmenteerde inrichting van zorgprocessen. Zorgaanbieders, zoals ziekenhuizen, zijn typisch georganiseerd als een netwerk van verschillende, autonome zorgafdelingen, en ook de functies voor planning en besturing zijn vaak autonoom georganiseerd. Doordat de patiënt vaak meerdere afdelingen aandoet in het behandelpad, hebben beslissingen in planning en besturing van zorgprocessen in een bepaalde afdeling sterke invloed op de zorgprocessen in een andere afdeling. Door deze onderlinge afhankelijkheden biedt een geïntegreerde planning en besturing van zorgprocessen potentieel verbetering.

In dit proefschrift wordt een geïntegreerd raamwerk voor planning en besturing van zorgprocessen gepresenteerd, omdat een dergelijk geïntegreerd raamwerk nog ontbreekt in de literatuur. Dit raamwerk integreert alle managementgebieden die bij een zorgproces betrokken (kunnen) zijn met alle hiërarchische niveaus van planning en besturing. Door deze integratie is het mogelijk om de coherentie tussen verschillende plannings- en besturingsfuncties te analyseren en ontbrekende functies te identificeren. Het raamwerk is gestoeld op de klassieke hiërarchische indeling gebruikt in planning en besturing van productieprocessen. Deze hiërarchische indeling onderscheidt functies op *"strategisch"*, *"tactisch"* en *" operationeel"* niveau. Deze indeling wordt uitgebreid door binnen het operationele niveau onderscheid te maken tussen *"offline"* en *"online"*. Dit onderscheid reflecteert het verschil tussen "vooruit plannen" en "reactief" plannen op het operationele niveau.

Het geïntegreerde raamwerk voor planning en besturing van zorgprocessen kan worden gebruikt door zorgmanagers om de verschillende planningsfuncties en hun onderlinge afhankelijkheden in kaart te brengen en te structureren. Het raamwerk kan worden toegepast op een afdeling, een complete zorgorganisatie, en een keten van zorgaanbieders. Het raamwerk kan worden ingezet om managementproblemen te identificeren en te positioneren, om interventies in de planning en besturing van zorgprocessen in te kaderen, en om een dialoog te faciliteren tussen medische staf en managers in de zorg. (**Hoofdstuk 2**)

In Hoofdstuk 3 van dit proefschrift wordt een uitgebreid literatuuronderzoek beschreven om de bestaande literatuur in geïntegreerde planning en besturing van zorgprocessen in kaart te brengen en dit proefschrift in deze literatuur te positioneren. In het literatuuronderzoek behandelt dit proefschrift vraagstukken omtrent capaciteitsplanning in de zorg.

In een eerste stap wordt een taxonomie gepresenteerd om de vraagstukken betreffende capaciteitsplanning in de zorg te positioneren, te identificeren en te classificeren. Deze taxonomie is gebaseerd op het raamwerk ontwikkeld in Hoofdstuk 2. In een tweede stap wordt deze taxonomie gevuld met alle vraagstukken in capaciteitsplanning voor zes verschillende zorgclusters (o.a. polikliniek, OK, thuiszorg). Voor elk vraagstuk worden de belangrijkste artikelen en technieken uit de literatuur voor mathematische besliskunde (vanaf nu aangeduid met de Engelse term: Operations Research and Management Science (OR/MS)) en de gebruikte wiskundige technieken in die artikelen behandeld. Deze artikelen zijn gevonden via een gestructureerde zoekmethode, die wordt besproken in het hoofdstuk.

Eerdere literatuuronderzoeken concluderen dat er een gebrek is aan OR/MS artikelen die een compleet zorgproces modelleren. Ondanks dat er artikelen zijn gevonden die interacties tussen twee afdelingen modelleerden, zijn er nauwelijks artikelen gevonden die complete zorgprocessen van begin tot eind in een ziekenhuis in kaart brengen, laat staan een compleet zorgproces over verschillende zorgaanbieders. Het literatuuronderzoek in Hoofdstuk 3 bevestigt deze conclusies. (**Hoofdstuk 3**)

Het tweede deel van dit proefschrift beschrijft modellen en methoden voor geïntegreerde planning en besturing van zorgprocessen. Deze modellen en methoden zijn ontwikkeld met behulp van wiskundige technieken uit de OR/MS, en zijn ontwikkeld voor het eerder besproken *tactische niveau* van planning en besturing. Een tactisch capaciteitsplan in ziekenhuizen bepaalt welke patiëntgroepen wanneer worden behandeld, en wijst zorgcapaciteit (tijd van artsen, verpleegkundigen, OK, etc.) toe aan patiëntgroepen en zorgactiviteiten voor de middellange termijn. De belangrijkste doelen van tactisch plannen zijn een evenwichtige toegang tot zorg voor alle patiëntgroepen, het behalen van strategisch vastgestelde productiedoelen, en een effectieve en efficiënte inzet van beschikbare zorgcapaciteit.

Hoofdstuk 4 presenteert een methode om een tactisch patiënt- en capaciteitsplan te ontwikkelen. Dit tactische plan wijst de beschikbare zorgcapaciteit toe aan meerdere zorgprocessen (zorgketens) en berekent het aantal patiënten om te behandelen in elke stap van elke zorgproces. De methode is ontwikkeld binnen het raamwerk van “Mixed Integer Linear Programming” (MILP) en kan meerdere capaciteitstypes, meerdere tijdsperioden, en meerdere patiëntengroepen met onzekere paden door het zorgproces aan. Hierdoor integreert de ontwikkelde methode capaciteitsplanning voor een complete keten van ziekenhuisafdelingen en zorgcapaciteiten.

Resultaten van experimenten laten zien dat de ontwikkelde methode leidt tot een meer evenwichtige toegang tot de zorg voor alle patiëntgroepen en de mogelijkheid biedt om toegangstijden, productieaantallen en bezettingsgraden te beheersen. De methode is generiek, omdat de onderliggende MILP en algoritmiek de mogelijkheid biedt om verschillende wijzigingen en uitbreidingen door te voeren. Dit heeft tot resultaat dat de methode in meerdere praktische toepassingen van tactisch plannen in de zorg te gebruiken is. (**Hoofdstuk 4**)

Hoofdstuk 5 van dit proefschrift ontwikkelt een methode voor tactisch plannen in de zorg die rekening houdt met stochastisch elementen. In het hoofdstuk wordt de stochastische uitdrukking voor het maken van een tactisch plan geformuleerd en een exacte uitwerking met behulp van de wiskundige techniek "Dynamic Programming" (DP) gegeven. De exacte oplossing met DP werkt alleen voor zeer kleine instanties van het tactisch plannen probleem, veel kleiner dan instanties die voorkomen in de praktijk. Om die reden wordt in het hoofdstuk een alternatieve oplossingsmethode ontwikkeld met behulp van technieken uit het wiskundige veld voor "Approximate Dynamic Programming" (ADP).

De ontwikkelde ADP methode kan meerdere capaciteitstypes, meerdere tijdsperioden, en meerdere patiëntengroepen met onzekere paden door het zorgproces aan. De methode houdt rekening met onzekerheid over bepaalde elementen in het model, zoals de onzekere patiëntvraag in elke tijdsperiode en de onzekerheid over het vervolgpad van iedere patiënt nadat deze een onderdeel van de zorgketen heeft doorlopen. Hierdoor integreert deze methode de capaciteitsplanning voor een complete zorgketen van ziekenhuisafdelingen en zorgcapaciteiten en het houdt tegelijkertijd rekening met stochastische elementen.

Rekenresultaten uit de experimenten laten zien dat de ontwikkelde ADP-methode geschikt is voor tactische planning in de zorg, dat deze methode accuraat is (zeer dicht bij de resultaten van de exacte DP-methode), en dat de ADP-methode significant beter presteert dan twee alternatieve methoden voor tactisch plannen. (**Hoofdstuk 5**)

Hoofdstuk 6 behandelt een vraagstuk op het *tactische niveau* van planning en besturing in poliklinische zorg. Traditioneel zijn processen in een polikliniek zo ingericht dat de arts in een spreekkamer verblijft en de patiënten deze spreekkamer betreden wanneer hun consult begint. Deze traditionele inrichting wordt in dit proefschrift het Patient-to-Doctor systeem genoemd. Een alternatieve inrichting is het Doctor-to-Patient systeem, waarbij de arts patiënten consulteert in meerdere spreekkamers en van spreekkamer wisselt tussen de verschillende afspraken. Patiënten kunnen al voor hun consult begint in een lege spreekkamer om zich voor te bereiden op het consult (bloeddruk opnemen, temperatuur opnemen, gaan zitten, etc.). Artsen bezoeken de patiënt pas als deze volledig is voorbereid op het consult van de arts (evt. door een assistent(e)). Hierdoor bespaart de arts potentieel kostbare tijd.

Het hoofdstuk ontwikkelt generieke modellen met behulp van technieken uit de wachtrijtheorie en simulatie om beide systemen in verschillende settings te kunnen evalueren op verschillende prestatie-indicatoren. Deze modellen kunnen worden gebruikt door managers van poliklinieken en artsen om de twee systemen te vergelijken en met die vergelijking een keuze te maken voor één van de twee systemen bij de inrichting van het proces in de polikliniek. Naast deze modellen biedt Hoofdstuk 6 ook een generieke methode om het benodigd aantal spreekkamers in het Doctor-to-Patient systeem te bepalen. De modellen en methoden worden getest op een brede set aan verdelingen en parameterinstellingen. Ter illustratie worden de modellen toegepast op een praktische case uit één van de ziekenhuizen waarmee is samengewerkt tijdens dit onderzoek. (**Hoofdstuk 6**)

Acknowledgements

During my PhD, I was very fortunate to be supervised by two great mathematicians and professors. Also I've had the privilege to work with superb researchers and passionate colleagues. I thank all the people that helped or supported me during my PhD. I thank the people in my writing below in particular.

I thank Erwin Hans for his relentless energy and enthusiasm, mathematical and practical insight yielding many interesting discussions and fruitful ideas, his amazing writing skills, and his commitment. It was my privilege to work with you.

I thank Richard Boucherie for his commitment, and his frank analysis of my research, striving for the best, and always professional and supporting. My two supervisors provided a combination of personality traits that were important in, among others, publishing the articles from this thesis in renowned journals.

I thank the thesis committee for their participation and careful reading of my thesis: prof. dr. K.I. van Oudenhoven - van der Zee, prof. dr. P.R. Harper, prof. dr. G. Kazemier, dr. ir. M.R.K. Mes, prof. dr. M. Uetz, prof. dr. B. Werners, and prof. dr. W.H.M. Zijm.

In addition to my supervisors and graduation committee, I wish to thank five co-authors specifically. Martijn Mes, Nikky Kortbeek, Johann Hurink, Jan-Kees van Ommeren, and Peter Vanberkel, it was great to work with you.

I thank the Reinier de Graaf hospital for their contribution to my research. In particular I thank Jaap van den Heuvel, John Verver, David Lange, Paul Duyndam, Cees van der Zande, Kees Broekman, Stephan Vehmeijer, Kim van Leeuwen, and Hilde van der Vegte. I thank Technology Foundation STW and University of Twente for financing and supporting our research.

My former colleagues from CHOIR, IEBIS, SOR and DMMP, thank you. I thank the following people for their contribution to this thesis: Mark van Houdenhoven, Piet Bakker, Warren Powell, Maartje Zonderland, Theresia van Essen, Egbert van der Veen, Aleida Braaksma, Nelly Litvak, Bart Veltman, Ingrid Vliegen, Marco Schutten, Matthieu van der Heijden, Leo van der Wegen, Thyra Kamphuis, Elke van der Veen, Paula van Brakel, Pieter Wolbers, Leon Nijenhuis, Marcel van Eijken, Richelle Rijntjes, Jasper Quik, Gerwen Apenhorst, and all LogiDOC group members.

My family and friends provided indispensable joy and support. I thank them a lot for that. Kirsten, thank you for always embarking, enjoying and completing each of our journeys together with me.

Peter Hulshof

Amsterdam, November 2013

About the author

Peter Hulshof was born in Groenlo, the Netherlands, on June 19th, 1983. In 2001, he obtained his VWO degree at the Staring College in Lochem, with a strong interest in the combination of Mathematics, Physics and Economics. As a result of these combined interests, Peter started his studies Industrial Engineering and Management at the University of Twente in the same year.

During his studies, Peter focused on operations management and logistics, with an emphasis on healthcare, and he participated in three internships, in New York, Paris and in the Netherlands. He obtained his Master of Science degree with a thesis research on vehicle routing and order planning in the oil and gas sector at the decision software company ORTEC. In this research period, Peter became enthousiastic about creating insight by research, and he applied for a PhD-position at the Center for Healthcare Operations and Improvement Research at the University of Twente. He started his PhD with a combined position at the University of Twente and the Reinier de Graaf hospital in 2008, under the supervision of prof. dr. ir. Erwin W. Hans and prof. dr. Richard J. Boucherie.

Currently, Peter is a consultant at The Boston Consulting Group (BCG).

List of publications

P.J.H. Hulshof, M.R.K. Mes, R.J. Boucherie, and E.W. Hans. Tactical planning in healthcare using Approximate Dynamic Programming. *Memorandum 2014*, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2013.
(Basis for Chapter 5).

P.J.H. Hulshof, R.J. Boucherie, E.W. Hans, and J.L. Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science* 16(2):152-166, 2013.
(Basis for Chapter 4).

P.J.H. Hulshof, N. Kortbeek, R.J. Boucherie, E.W. Hans, and P.J.M. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems* 1(2):129-175, 2012.
(Basis for Chapters 1 and 3, and the Epilogue).

P.J.H. Hulshof, P.T. Vanberkel, R.J. Boucherie, E.W. Hans, M. van Houdenhoven, and J.C.W. van Ommeren. Analytical models to determine room requirements in outpatient clinics. *OR Spectrum* 34(2):391-405, 2012.
(Basis for Chapter 6).

E.W. Hans, M. van Houdenhoven, P.J.H. Hulshof. A framework for healthcare planning and control. *In: Handbook of Healthcare System Scheduling*, Randolph Hall (editor), *International Series in Operations Research & Management Science* 168:303-320, 2012.
(Basis for Chapters 1 and 2, and the Epilogue).

P.J.H. Hulshof, R.J. Boucherie, J.T. van Essen, E.W. Hans, J.L. Hurink, N. Kortbeek, N. Litvak, P.T. Vanberkel, E. van der Veen, B. Veltman, I.M. Vliegen, and M.E. Zonderland. ORchestra: an online reference database of OR/MS literature in health care . *Health Care Management Science* 14(4):383-384, 2011.

P.J.H. Hulshof. Minimizing costs of oil and gas distribution. Master's thesis at the University of Twente, 2008. *Published as a book by LAP LAMBERT Academic Publishing in 2012, ISBN: 978-38-443-0690-3.*

The pressure on healthcare systems rises as both demand for healthcare and expenditures are increasing steadily. As a result, healthcare professionals face the challenging task to design and organize the healthcare delivery process more effectively and efficiently. Healthcare planning and control lags behind manufacturing and control for various reasons. One of the main causes is the fragmented nature of healthcare. Healthcare organizations such as hospitals are typically organized as a cluster of autonomous departments, where planning and control is also often functionally dispersed. As the clinical course of patients traverses multiple, thus interdependent, departments, an integrated approach to healthcare planning and control is likely to bring improvements.

This thesis aims to contribute to integrated decision making in healthcare in two ways. First, it develops a framework, taxonomy, and extensive literature review to support healthcare professionals in structuring and positioning planning and control decisions in healthcare. Second, this thesis proposes planning approaches to develop resource allocation and patient admission plans for multiple resources and patient types, thereby integrating decision making for a chain of healthcare departments. These planning approaches are developed with techniques from Operations Research and Management Science.

UNIVERSITY OF TWENTE.

Beta
Research School for Operations
Management and Logistics

School of Management and Governance
Department of Industrial Engineering and Business Information Systems
Center for Healthcare Operations Improvement and Research