INTEGRAL CAPACITY MANAGEMENT & PLANNING IN HOSPITALS

Thomas Schneider

INTEGRAL CAPACITY MANAGEMENT & PLANNING IN HOSPITALS

Anton Johan (Thomas) Schneider

Dissertation committee

Chairman & secretary:	Prof. dr. J.N. Kok *University of Twente, Enschede, the Netherlands*
Promotors:	Prof. dr. R.J. Boucherie *University of Twente, Enschede, the Netherlands* Prof. dr. ir. E.W. Hans *University of Twente, Enschede, the Netherlands*
Co-promotor:	Dr. ir. M.E. Zonderland *University of Twente, Enschede, the Netherlands*
Members:	Prof. dr. N.M. van Dijk *University of Twente, Enschede, the Netherlands* Prof. dr. J.L. Hurink *University of Twente, Enschede, the Netherlands* Prof. dr. M.J. Schalij *Leiden University Medical Center, Leiden, the Netherlands* Dr. J. Tamsma *University of Twente, Enschede, the Netherlands* *Leiden University Medical Center, Leiden, the Netherlands* Prof. dr. J.T. van der Vaart *University of Groningen, Groningen, the Netherlands*

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Technical Medical Centre
Center for Healthcare Operations Improvement & Research

This research was in part conducted at and financially supported by the Leiden University Medical Center.

The distribution of this thesis is financially supported by Center for Healthcare Operations Improvement & Research and the Leiden University Medical Center.

Typeset in LaTeX. Printed by Ridderprint, Alblasserdam, the Netherlands.
Cover design: Lydia van der Spek.

ISBN 978-90-365-5034-5
DOI 10.3990/1.9789036550345

INTEGRAL CAPACITY MANAGEMENT & PLANNING IN HOSPITALS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
Prof. dr. T.T.M. Palstra,
volgens het besluit van het College voor Promoties,
in het openbaar te verdedigen
op vrijdag 18 september 2020 om 14:45 uur

door

Anton Johan Schneider

geboren op 17 oktober 1984
Ouderkerk aan de Amstel, Nederland

Dit proefschrift is goedgekeurd door:
Prof. dr. R.J. Boucherie (promotor)
Prof. dr. ir. E.W. Hans (promotor)
Dr. ir. M.E. Zonderland (co-promotor)

Mam, weet je nog

Voorwoord

Weet je nog, Maartje? Toen je mij bijna tien geleden de vraag stelde of promoveren iets voor mij zou zijn? Het antwoord was destijds resoluut *'Nee'*. Ik dacht dat ik mijzelf, tijdens het onderzoek voor mijn Master thesis, voldoende had uitgedaagd op het gebied van onderzoek. Wetende dat ik hierna zou gaan werken in een academische omgeving, namenlijk het Leids Universitair Medisch Centrum. Na het afronden van mijn Master Technische Bedrijfskunde op de Universiteit Twente, had ik dan ook niet verwacht ooit nog een carrièrestap te maken in het onderzoek. Tijdens mijn eerste jaren als adviseur zorglogistiek binnen Divisie 2 in het LUMC, bleef ik literatuur lezen over de onderwerpen die voor mijn werk actueel waren. Zodoende bleef ik op de hoogte van recente ontwikkelingen binnen het onderzoeksgebied van, wat tegenwoordig beter bekend is als, integraal capaciteitsmanagement en capaciteitsplanning.

Terugkijkend ben ik trots op dit proefschrift en wat ik heb bereikt. Zoals je in een ziekenhuis nooit iets in isolement doet, geldt dit ook voor een proefschrift. Zonder de ondersteuning en inzet van velen had ik dit nooit kunnen bereiken. Ik ben dankbaar voor de kansen die mij geboden zijn de afgelopen 5 jaar, de ervaringen die ik op heb kunnen doen en de samenwerkingen die heb ik kunnen opzetten. Zonder tekort te doen aan mijn dankbaarheid voor anderen, wil ik een aantal mensen in het bijzonder bedanken.

Allereerst wil ik mijn copromotor, Maartje, bedanken. Toen ik aan de vooravond van mijn promotietraject nog twijfels had of dit wel iets voor mij was, heb je mij eerlijk verteld wat jouw eigen promotie voor jou heeft betekend en wat het van een promovendus vraagt. Jouw oprechte verhaal heeft bij mij uiteindelijk de doorslag gegeven. Je hebt mij echt op weg geholpen en was altijd bereikbaar. Jouw nuchtere en rustige persoonlijkheid, ervaring en kennis, en het oneindige vertrouwen in mij, hebben mij hier nu gebracht. Ook tijdens 'de pauze' van mijn onderzoek vanwege de vroeggeboorte van mijn tweeling, bleef je altijd contact houden. Waar de rode draad in mijn onderzoek op sommige momenten lastig te vinden was, is deze bij onze samenwerking duidelijk terug te zien. Je hebt mij begeleid tijdens mijn Master thesis en nu ook tijdens mijn promotieonderzoek. Ik waardeer onze samenwerking en vriendschap zeer. Ik weet dat we elkaar zullen blijven opzoeken.

Richard, mijn eerste ervaring met jou was tijdens het vak Stochastische Modellen in Operations Management. Als TBK-er waren jouw colleges een nieuwe ervaring. Onze eerste kennismaking ter voorbereiding op het promotietraject heeft wederom een nieuwe ervaring opgeleverd. Aan het begin van het traject moesten wij beiden

wennen aan de parttime opzet. Gelukkig is er altijd vertrouwen geweest in een goede afloop. Ik heb mij gewaardeerd gevoeld binnen de CHOIR groep en daar heb jij een grote rol in gespeeld. Je directe communicatie past mij erg goed. Je kritische blik en scherpe, maar duidelijke feedback vind ik uitermate prettig. Je hebt mij uitgedaagd en richting gegeven om zodoende steeds een stapje verder in de juiste richting te zetten. Ook je humor waardeer ik, ik heb vaak met je lachen!

Erwin, jouw colleges bevestigden mijn keuze om logistieke kennis in de zorg toe te gaan passen. De ORAHS congressen die we samen hebben bezocht gaven ruimte om gezamenlijk onze visie op ICM te vormen en jou en je gezin beter te leren kennen. Onze samenwerking was aan het eind zeer intensief en je hebt mij toen ook meermaals gevraagd of ik het nog wel zag zitten? En dat we toch echt stappen aan het maken waren! Ik heb toen heel koeltjes aangeven dat ik wel voor grotere uitdagingen heb gestaan. Je feedback op en gesprekken over ICM op onmogelijke tijdstippen van de dag en dagen van de week hebben mij gepusht dit uitdagende onderwerp verder te brengen. Dat je zelfs je vishengel weer opborg om met mij te discussieren over ICM spreekt voor jou. Ik ben trots op onze samenwerking en de passie die we delen voor ICM.

Ik wil hierbij ook mijn dank overbrengen naar mijn commissieleden, Nico van Dijk, Johann Hurink, Martin Schalij, Jouke Tamsma en Taco van der Vaart, voor de tijd die jullie genomen hebben voor de waardevolle feedback op mijn dissertatie en de verdediging.

De samenwerkingen die hebben geleid tot de hoofdstukken van dit proefschrift, ben ik ook dank verschuldigd. Ik zie samenwerkingen als versterking tussen twee partijen en daar zijn de hoofdstukken van dit proefschrift een voorbeeld van.

Voor hoofdstuk 1 bedank ik Maartje, Erwin en Richard. Het was wat persen en malen en schuiven met secties, maar het staat.

Hoofdstuk 2 is een mooie symbiose tussen ons geweest, Erwin. In het huidige tijdperk van werken op de zolderkamers, door middel van korte intensieve sprints en veel mooie discussies over uiteenlopende onderwerpen is dit hoofdstuk tot stand gekomen. Uiteindelijk ligt er nu een goede basis voor verder onderzoek naar ICM. Je bent nog lang niet van mij af.

MaartjeV, je had bij de start van ons eerste gezamenlijke boekhoofdstuk waarschijnlijk een andere werkverdeling in gedachte. Ik prijs je discipline en kennis. Ons eerste boekhoofdstuk en het hoofdstuk voor het CHOIR boek, heeft gezamenlijk geleid tot hoofdstuk 3 van mijn proefschrift. Onze samenwerking vond ik prettig en inspirerend. Dat we ook nog korte tijd collega's zijn geweest, heeft onze samenwerking en vriendschap nog verder versterkt. Je mag trots zijn op de dappere strijd die je nu aangaat.

Hoofdstuk 4 is mijn eerste publicatie. Hiervoor bedank ik Luuk, Paul, Jaap, Ton, Job en Wilbert. Jullie input heeft een vliegende start gegeven aan mijn onderzoek.

Mijke, dankjewel voor de succesvolle samenwerking bij hoofdstuk 5. Jouw Master thesis heeft geleid tot dit prachtige onderzoek. Ik ben ook trots dat je je kennis en ervaring nog steeds inzet binnen de zorg. Rhythm was er dan ook als de kippen bij om jou te werven. Theresia, ik wil jou uiteraard ook bedanken voor dit hoofdstuk. Te

midden van mijn prille vaderschap hebben we dit toch mooi samen gedaan. Verder wil ik je ook bedanken voor de prettige samenwerking aan de verschillende opdrachten voor studenten.

Zou je een keer met mij mee willen denken? Je haalde laatst nog deze vraag van mij aan. Niet wetende dat dit het startschot zou zijn voor onze langdurige samenwerking. Maarten, wat een levenswerk is dit geworden en wat hebben we gehuild van het lachen (of andersom?). Ik wil je bedanken voor je intellect, inzet, doorzettingsvermogen, gezucht en de ellenlange discussies. Prachtig vond ik het, als je een uurtje met rust gelaten moest worden, zodat je even rustig kon nadenken. Uiteindelijk zijn we gekomen tot een fraaie oplossing van het THOMAS-probleem in hoodstuk 6. Richard, ook in dit hoofdstuk heb jij een aanzienlijke rol gespeeld. Dank voor de scherpe theoretische discussies en je pragmatische sturing. Martin, je hebt mij het vertrouwen gegeven dit prachtige probleem vanuit de praktijk te analyseren en ook echt voor jouw polikliniek op te lossen. Dit is dan ook de inspiratie geweest voor dit hoofdstuk, dank daarvoor.

Ook wil ik mijn collega's binnen het LUMC bedanken. Ondanks dat ik elke dag weer nieuwe collega's leer kennen, en ken ik er inmiddels een heleboel, zijn er een aantal die betrokken zijn geweest bij dit onderzoek. Allereerst het (ex-) bestuur van divisie 2; Ton, Paul, Koos en Wouter. Dank voor jullie vertrouwen en ruimte om werk, onderzoek en privé te kunnen combineren. Hierbij wil ik ook mijn waardering kenbaar maken voor jullie steun rondom de geboorte van Vik en Len. Paul en Ton, ook jullie bijdrage aan hoofdstuk 4 heeft gezorgd voor de vliegende start van mijn onderzoek.

Job, dank dat je mijn onderzoek mede gefaciliteerd hebt. Ik waardeer de discussies over de verschillen en overeenkomsten tussen dit onderzoek en de medische praktijk. Samen met Wilbert, is jullie input bij hoodstuk 3 en 4 zeer waardevol gebleken. Fred, jij hebt het stokje van Job overgenomen. Ook jouw input en de ervaring van het eerdere traject met Maartje was waardevol.

Guillaine, Martin en Wouter, om samen met jullie tijdens het laatste deel van mijn promotie nu echt werk te maken van ICM in het LUMC, is tot nu toe het mooiste hoofdstuk uit mijn carrière. De vorming van het LUMC Capaciteitscentrum is hiervan een prachtig resultaat waar wij trots op mogen zijn. Dit heeft ook geleid tot waardevolle input voor hoofdstuk 2. En nu weer gewoon aan het werk, er is nog genoeg te doen.

Mijn team van het LUMC Capaciteitscentrum, Els, Fieke, Ilse, Iwona, Mirkan en Viktor wil ik ook bedanken. Jullie enorme inzet tijdens de COVID-19 crisis bewonder ik. Dit gaf mij de energie voor de laatste loodjes. We zijn met iets unieks bezig!

Daarnaast wil ik mijn CHOIR collega's en kamergenoten op de UT bedanken. Het was mooi om onderdeel te zijn geweest van zo'n grote groep onderzoekers binnen dit onderzoeksgebied. Dit heeft elkaars onderzoek versterkt en kenmerkt CHOIR misschien wel het meest. Ook van de uitjes, barbecues, congressen, lunchwandelingen en koffiemomentjes heb ik erg genoten. Verder wil ik ook de andere collega's op de UT van MOR bedanken. Thyra, we hebben elkaar sporadisch gezien en gesproken. Je betrokkenheid rondom de geboorte van de jongens, het eerste welkomstbloemetje en andere ondersteuning heb ik altijd gewaardeerd. Nico, wat een mooie discussies heb-

ben wij gehad en wat liepen die altijd uit. Ook de samenwerking tijdens de begeleiding van een aantal studenten heb ik als prettig ervaren. Dank ook voor het meedenken bij ons MDP probleem. Joost, Maarten, Jasper, Eline en Robin, heel veel succes met de afronding van jullie promotie. Ook wil ik alle studenten die ik begeleid heb bedanken. Chantal, Mijke, Laurien, Jitske, Ivan, Bjarty en Guusje, het was fantatisch om te zien welke ontwikkeling we samen doormaakten en ik hoop dat jullie er met net zoveel plezier op terugkijken als ik.

Tot slot wil ik mijn gezin, familie en vrienden bedanken. Jullie steun, ondersteuning en afleiding was essentieel. Martijn, voor jou zijn de laatste twee typeringen in de vorige zin van toepassing. Het is uniek hoe sterk onze band is, hoeveel overeenkomsten er zich tussen ons leven voltrekken. We wonen nu weer vlak bij elkaar, dus dit houdt nog wel even aan. Zonder jouw steun op alle vlakken, zouden wij niet gezamenlijk mijn proefschrift verdedigen.

Mam, ik zal nooit vergeten toen je vroeg hoe lang ik nog moest promoveren en ik je zag rekenen. Wat is het snel gegaan het laatste jaar. Het geeft rust, dat je nu zo op je plek bent. Ook al weet je niet meer wie ik ben, je bent nog steeds bij mij. Je hebt mij gevormd tot wie ik nu ben en ik herken veel terug in de manier hoe ik nu zelf de opvoeding van mijn jongens invul. Ik weet dat je trots op mij bent. En ik ook op jou. Hoe jij Willemien en mij hebt opgevoed toen papa ziek was geeft mij het doorzettingsvermogen als ik het nu zwaar heb. Pap, dank voor de ritjes van en naar het station. Dank ook voor de tips vanuit je eigen promotieonderzoek en hoe je werk en promoveren kan combineren. Je hebt mij laten inzien dat je alles kan bereiken als je iets echt wilt en doorzet. Wil, dank dat je nooit geklaagd hebt wanneer ik weereens druk was met de jongens of mijn onderzoek. Hoe we samen de zorg voor mama hebben opgepakt en ingevuld, zegt veel over ons. Ik kijk uit naar jullie kleine 'Trix', Kasper en Wil. Peter en Nancy, Hotel Amstelstraat geef ik vijf sterren. Dank voor alle steun en de rustige nachten. Nancy, ook bedankt voor je onvoorwaardelijke steun bij de opvoeding van Vik en Len. Dit gaf mij net het beetje extra ruimte om dit proefschrift af te ronden. Luuk en Nienke, zo ver weg en toch zo dichtbij. Het fijnst is toch wel om jullie weer in ons midden te hebben.

Lieve Anne, jij verdient misschien nog wel meer lof voor dit proefschrift dan ik. Die dagen, nachten, weekenden en weken dat ik niet thuis was of zat te werken aan dit proefschrift of Leiden. Zonder jouw onvoorwaardelijke steun en begrip lag dit proefschrift er nu niet. Hoe we samen, maar ook in dit geval vooral jij, rondom de geboorte van onze tweeling alles hebben opgepakt, kunnen alleen de sterksten. Het heeft ons geleerd het leven te nemen zoals het komt en altijd positief te blijven. Naast je eigen fulltime baan, was je er altijd voor de jongens, vrienden en familie. Én was er ook nog tijd voor een strikt sportschema. Ik snap niet waar jij deze energie vandaan haalt.

Lieve Vik, lieve Len, wat ben ik trots op jullie. De start was pittig en het duurde dan ook even voordat we samen alles op de rit hadden. Nu word ik uitgedaagd en behendig uitgespeeld door twee doodnormale peuters. Als ik kijk naar wat jullie allemaal in de eerste maanden van jullie leven te verduren hebben gekregen, mag ik nooit meer klagen.

Thomas Ouderkerk aan de Amstel, 2020

Contents

CHAPTER 1

Introduction

1.1 Motivation for this research

Good health contributes to quality of life, and therefore societies are willing to invest an increasing amount of their gross domestic product in healthcare [182]. An improved health status prolongs life expectancy [142]. Since healthcare costs are strongly age dependent [9], with improved life expectancy comes a greater number of years during which people are in need of care, leading to ever-increasing healthcare costs. In addition, medical and technological innovations drive healthcare costs further as they increase the number of treatment options.

If the current trend of growing demand for healthcare continues, by the end of this decade 25% of the available workforce in the Netherlands should be working in healthcare [235], putting society under pressure. While demand and thus costs are increasing, Western countries are confronted with a shrinking workforce as a result of an ageing society [44]. This unbalances healthcare demand and supply even further.

For half a century, the problem of an imbalance between healthcare demand and supply was solved, to a large extent, by increasing capacity [180]. Innovations that might have led to lower hospital demand were not scalable and therefore had marginal impact [129]. Although a vast amount of research has generated multiple options for changing course, there has been no "game changer" to solve the healthcare productivity challenge.

To bridge this productivity gap, six sources of waste should be eliminated [28]: 1) failures of care delivery, 2) failures of care coordination, 3) overtreatment or low-value care, 4) administrative complexity, 5) pricing failures and 6) fraud and abuse. Four key steps can be identified to help eliminate waste sources 1 to 4 [35]: prevention, early diagnostics, active treatment and care coordination. Currently health care costs can be reduced 30% by optimizing capacity planning and management [74]. Therefore, care coordination offers substantial potential for increasing productivity. This thesis focuses on care coordination through optimizing hospital capacity planning and management.

1.2 Capacity Planning and Management in Hospitals

Hospital are characterized by many dependencies between their departments and other healthcare organizations as a result of the flows of patients and healthcare professionals. This complicates the organization of hospital processes, as the effects of pulling one string may resonate in many different places inside and outside the organization. Thus, optimization of a single resource is myopic by definition. Whereas new medical treatments are strictly regulated and initially tested on patients in randomized controlled trials (RCTs), the implementation of new organizational designs is not regulated and effects are rarely analyzed [150]. Using RCTs for testing new policies would be unethical as (adverse) effects are difficult to predict and therefore it would be difficult not to expose patients to negative and adverse effects during tests of organizational experiments. However, there are many other methodologies for analyzing new organizational designs and policies.

Operations management is the field of expertise that studies the process of making decisions about resources [215]. A design approach is one options for organizing decision-making processes. Operations research (OR) supports decision-making about new organizational designs [253] and encompasses many different methodologies, such as queuing theory and discrete event simulation. With OR, the effects of interventions on trade-offs between key performance indicators can be analyzed and optimized in a safe environment, helping minimize counter productive interventions in healthcare delivery processes. For over 50 years, OR has been applied to healthcare related problems [16], and during that time a vast amount of research has been published this topic. However, the actual implementation in practice of solutions following from these modelling efforts is rarely described in the literature [43]. This is striking, as implementation is the final step in improvement and would be a valuable topic to study. One explanation for this lack is that actual implementation requires a different set of skills and expertise.

Working from both operations management and OR perspectives, this thesis focuses on improving decision-making processes related to hospital capacity.

1.3 Capacity Planning and Management in the Leiden University Medical Center

The research presented in this thesis is inspired by practices in the Leiden University Medical Center (LUMC). Founded in 1636, the LUMC was the first Dutch academic hospital and was part of the first Dutch university. At that time, the LUMC consisted of an anatomical theater, a botanical garden and several beds in the Caecilia Gasthuis. Currently, the main focus of the LUMC is top clinical care and highly specialized care in oncology, regenerative medicine and cardiology, and population health. As an academic center it fulfils three social responsibilities: patient care, training and education for healthcare professionals and medical students, and research.

The LUMC is the smallest academic center in The Netherlands and is situated close

(< 50km from) two of the largest academic centers: Amsterdam University Medical Center and Erasmus Medical Center. As quality standards in terms of minimum volumes for clinical procedures increase, it becomes difficult for smaller hospitals to meet these standards. The LUMC has therefore started to attain a clear strategic patient care portfolio. To align strategy and operations and to improve both efficiency and quality in healthcare delivery, the LUMC developed a program to design and implement integral capacity management (ICM). ICM is now anchored within the chain of command of the hospital and is widely adopted and accepted within the organization. It is also being implemented in the hospital's capacity center.

Since 2007 the LUMC has cooperated with the Center for Healthcare Operations Improvement and Research of the University of Twente to improve capacity planning and management both in practice and in research. In numerous projects, theory and practice have been connected to improve capacity in various areas (e.g. outpatient clinics, inpatient wards, operating rooms, emergency department,and more) evolving from local to hospital-wide improvements and organizational designs. This research is currently embedded in the LUMC Capacity Center and ICM program.

1.4 Recent Developments for Hospital Capacity Management and Planning

The recent SARS-CoV-2 (i.e. COVID-19) pandemic has disrupted healthcare like no past outbreak [13]. Hospital capacity was completely given to patients recovering from this virus. As a result, there was little capacity available for other patients. This scarcity required hospital-wide (i.e. integral) decision-making for capacity planning, as personal preferences and myopic decision-making were, to a large extent, no longer valid. On the positive side, a vast majority of healthcare professionals did experience the added value of alignment as convenient and are committed to integrally organizing capacity management when this pandemic is over. On the negative side, apparently a crisis of this size is necessary to enable integral capacity decision-making while steps into this directions could be taken much earlier and without a crisis. Furthermore, as most hospitals had to build up other treatments and diagnostics from scratch as almost no patients other than COVID-19 patients were hospitalized and gave rise to opportunities for redesign breaking stuck routines. This pandemic can therefore been seen as "game changer" for ICM implementation. Furthermore, this pandemic has also resulted in compatibility improvements in information systems for data exchange to analyze complete care pathways within the healthcare network.

Another development emerging from this pandemic, is the remote monitoring of patients, which means patients are invited to the hospital only when necessary. Remote monitoring, may raise interesting questions to analyze using OR methodologies as the patient arrival rate will be more predictable. These new innovations will also reduce waste from overtreatment and low-value care.

1.5 Thesis outline

This thesis aims to connect theory and practice on integral capacity management and planning by presenting several case studies for the presented theoretical results. In fact, several of the theoretical results have actually been implemented in practice. We elaborate on the challenges experienced when implementing research results, and provide several factors for successful implementation. The thesis is organized in two parts, briefly introduced below.

Part I concerns the integral management of hospital capacity. As mentioned in Section 1.2, capacity and process improvements in practice require two elements: (1) the (near) optimal decision and (2) the organization of decision-making processes. Part I focuses on the second element, and we refer this to as capacity management. In *Chapter 2* we conceptualize ICM to initiate a research agenda on this topic. ICM aims to satisfy hospital stakeholders requirements by integrally managing patient care pathways. This means that for patients access and flow are maximized and for employees workload variability is minimized. We present three organizational integration dimensions along which hospitals can align capacities: (1) hierarchical, (2) patient-centered care and (3) managerial domains. This research should be seen as a first step towards theoretical understanding. We present directions for further research in the discussion. *Chapter 3* starts with an overview of performance measures (Section 3.2) and OR methodologies applied to hospital ward occupancy modelling. Next, literature on hospital ward occupancy is reviewed (Section 3.4). Based on logistical characteristics and patient flows, we distinguish the following ward types: intensive care, acute medical units, obstetric wards, weekday wards, and general wards. We then derive typical trade-offs between performance measures for each ward type and elaborate on managing ward-related capacities: beds and workforce. We also discuss what kinds of models can be used to analyze trade-offs in these decision-making processes (see Section 3.5). Finally, we present three case studies that use OR to analyze practical decisions and discuss the implementations in Section 3.7.

Part II presents three chapters that focus on integral capacity planning considering multiple patient flows and multiple resources. *Chapter 4* analyzes the flow of emergency admissions. Patient flow involves beds in three departments: the emergency department, the acute medical unit(AMU) and general wards. To improve the emergency admission flow, some hospitals introduce AMUs. Without integral capacity coordination, AMUs do not solve flow problems comparable to emergency department overcrowding. We develop a discrete event simulation model to analyze different capacity allocations related to the number beds in each ward type. We use two heuristics to derive feasible solutions for the distribution of beds among each ward type. This simulation model has been used repeatedly to support tactical capacity decisions within the LUMC.

In *Chapter 5*, we analyze the surgical patient flow for operating rooms, intensive care units and general wards. We combine multiple data analytics: we first use clustering techniques to generate surgery groups consisting of comparable surgical procedures, and then optimally schedule surgery groups within a master surgery schedule

using a mixed integer linear programming model. We demonstrate multiple variants of our model with minor modifications for managerial insights.

The final chapter of this part, *Chapter 6*, analyzes the online multi-appointment scheduling problem. When patients have multiple appointments during the same day, appointment schedules become increasingly vulnerable to delays and are therefore more fragile. We present a decomposition approach to deal with fragility and optimize both patient waiting time and resource utilization. First, we analyze this problem using a Markov decision process model to derive optimal policies for accepting or rejecting new arrivals. Next, we develop an integer linear programming model to schedule patients. Finally we compare performances of our approach and an heuristic. The results show the great potential of online multi-appointment scheduling optimization.

Part I

Integral Capacity Management in Hospitals

CHAPTER 2

Integral Capacity Management in Hospitals[1]

In this chapter we introduce a systematic approach to integrally organize all capacities involved in healthcare delivery at hospitals. Thus, this chapter focuses on one the main research topics of this thesis: managing hospital capacity.

2.1 Introduction

Hospitals are continuously challenged to improve their healthcare delivery on both outcomes and output. When demand for care increases, healthcare delivery workforce becomes increasingly scarce, and therefore the gap between demand and supply grows rapidly. Both trends are spurred by an ageing society, and by the increasing capabilities and diversification of the healthcare system that result from innovations. In this risk-averse sector, such challenges have long been addressed by increasing capacity and expenditures, but this is not sustainable and is arguably largely ineffective. The previously mentioned productivity challenge can be overcome by clinical innovation (treating patients more efficiently, without affecting quality) or by organizing more efficient capacity use. Our focus is on the latter.

Many hospitals organize capacity management (CM) as silos [148], or even as single cost centers, with their own operations management systems. The operating theater (OT) department is often considered to be the "most important" [100], as there, the greatest costs are accrued, the most income is earned, and clinically, the most interventions take place. However, from an operations point of view, and also patient's point of view, it is merely one step in the care pathway. Nevertheless, making the OT department the leader (i.e. by making its utilization the foremost performance indicator), and other departments followers causes bullwhip effects in the care chains. This common practice is not patient centered, and offers to, in our opinion, the greatest potential for productivity improvement. We aim to realize this potential by breaking through the siloed system and optimizing flow, rather than myopically optimizing

[1]This chapter is based on A.J. Schneider and E.W.Hans. Integral Capacity Management in Hospitals. *Working paper*.

utilization by aligning capacity in care pathways. To this end, in this chapter we propose 'Integral Capacity Management' (ICM) as the successor to CM. ICM strives to optimize integral care pathways for all stakeholders. ICM aims to improve equitable access and flow, in terms of speed and variability, in care pathways by making capacity agile. Implementation of ICM in hospitals is a comprehensive organizational change. This requires a systems design approach, starting from strategy development. Systems design is the process of defining elements of a system and their interfaces to satisfy the specific needs and requirements of a business or organization [251].

Productivity can be further improved by optimizing operational processes using an operational excellence approach (e.g. Lean, Six Sigma, Theory of Constraints). Although there is much evidence of successful implementations of such programs in hospitals, they rarely lead to comprehensive changes in organizational structures, and instead focus on operational processes. By contrast, ICM focuses on flow at all levels of control. An operational excellence program does not lead to systems redesign (such as ICM), and it is difficult to get staff support for systems redesign when operational processes are ailing. Therefore, ICM and operational excellence approaches reinforce each other [115].

ICM is gaining increased attention in hospitals, despite the lack of literature about what ICM is and how it works. Although CM has been used in hospitals for two decades [216] and ICM for over 10 years, universal definition and theoretical understanding are lacking for both and hospitals having difficulties during implementation. The contribution of this chapter is two fold: (1) we give a theoretical introduction of ICM and challenge researchers to further conceptualize the concept and (2) to guide hospitals how ICM may be approached for implementation.

This chapter is structured as follows. Section 2.2 discusses the problems of current CM practices in hospitals. In Section 2.3, we present ICM as a systems design approach for optimizing patient flows through integration along three decision-making dimensions for capacity. Finally, in Section 2.4 we discuss future options for research and implementation based on our approach.

2.2 Capacity Management in Hospitals

CM encompasses decision-making related to the acquisition, use and allocation of three types of renewable resources: workforce, equipment, and facilities. Its purpose is to satisfy stakeholders' (e.g. customer and staff) requirements [94, 106, 216]. Nonrenewable resources (e.g. materials) are relatively more flexible, whereas renewable resources often require longer commitments and are therefore more difficult to manage. In literature, capacity planning and capacity management are used interchangeably. However, they are not the same. Capacity planning concerns all planning activities, while capacity management focuses on organizing capacity. CM requires two elements: (1) infrastructure consisting of non-renewable resources, facilities and layout and capabilities, and (2) a management system to ensure an efficient care delivery process design, equitable access, and financial stability.

Hospitals use top-down decision processes [208] over multiple hierarchical levels: strategic, tactical and operational. Furthermore, staff (e.g. clinicians and nurses) are

highly educated and trained and therefore have high degrees of autonomy to design processes and their own schedule [122, 156]. From a CM view, this highly educated staff is a result of labor division, specialization and standardization to improve productivity [73, 183]. Therefore, many hospitals have their functional departments manage their own budgets and planning. This top-down, decentralized decision process (e.g. siloed management structure) often results in myopic optimization of resource utilization, poor alignment of interdependent resources that are adjacent in care pathways and large fluctuations in upstream and downstream departments [122]. In our opinion, this is one of the main problems of CM one we elaborate on in Section 2.3.2.

Dealing with CM problems necessitates a management structure that aligns capacity decision-making along multiple dimensions. We therefore coin the term *Integral Capacity Management (ICM)* for hospitals, which we explain further in the following section.

2.3 Integral Capacity Management in Hospitals

Since the 1960's, well known concepts for CM in the field of manufacturing and services are extensively covered. For example, the author of [11] decomposes CM decisions through their hierarchical nature: strategic, tactical and operational. The author of [263] combines the hierarchical decomposition with nonrenewable and renewable resource planning and technological planning. And the authors of [247] translate the hierarchical decomposition approach of [11] into CM decisions in healthcare organizations and [106] combines the hierarchical decomposition with four managerial areas of healthcare organizations: clinical, nonrenewable resources, renewable resources and financial. Thus, as mentioned in Section 2.2, CM literature focuses mainly on the hierarchical dimension without integrating other dimensions and in hospitals it is mainly executed top-down [208].

Since the nineties, it has been stated that the future challenges for CM will be to integrally manage and plan capacity (i.e. the next step of CM is ICM) [46, 216] to optimize flow, thereby ensuring equitable access and optimal waiting times for all patients [122]. Organizational integration is defined as the extent to which distinct and interdependent units, departments and management levels, including business processes, people, and technology involved, share a unified purpose [18]. More specific for ICM, integration is seen as the coordinated management of information, operations, and logistics through a common set of principles, strategies, policies, and performance metrics [55]. This is not limited by organizational boundaries and can also be formed across organizations. As healthcare is organized within networks, integration should also be sought between healthcare organizations. There is a considerable amount of research available that analyzes the effects of integrally managing capacity to manage process flows [124]. However, little research is available on the effects of integrally managing capacity in hospitals [124].

ICM aims to operationalize patient-centered care by incorporating patient flow optimization in capacity decisions. Therefore, ICM integrates existing dimensions for CM decisions of manufacturing literature and translates these dimensions to a hospital setting. ICM integrates the following three dimensions from existing manufacturing

literature: (1) hierarchical alignment of strategy and operations, (2) patient-centered care that considers care pathways and patient flows and (3) alignment of managerial domains. To design and implement an ICM system is extremely difficult as it involves many features of the organization and production. Many products and services in hospitals are 'engineered-to-order'. This complicates the design and implementation of ICM, as it makes processes difficult to predict. Furthermore, implementing ICM is also caused by unfulfilled preconditions needed for ICM, among others: missing management information, ambiguous decision-making, and inaccurate forecasts. Therefore, ICM must consider the specific identity of a hospital in terms of value proposition, available infrastructure, professional autonomy and its environment. No "one size fits all" implementation approach exists.

We will further present each dimension in the following sections. We start by the hierarchical alignment of strategy and operations in Section 2.3.1). We then present the patient-centered care dimension in Section 2.3.2 and discuss in Section 2.3.3 the alignment of managerial domains.

2.3.1 Hierarchical Integration

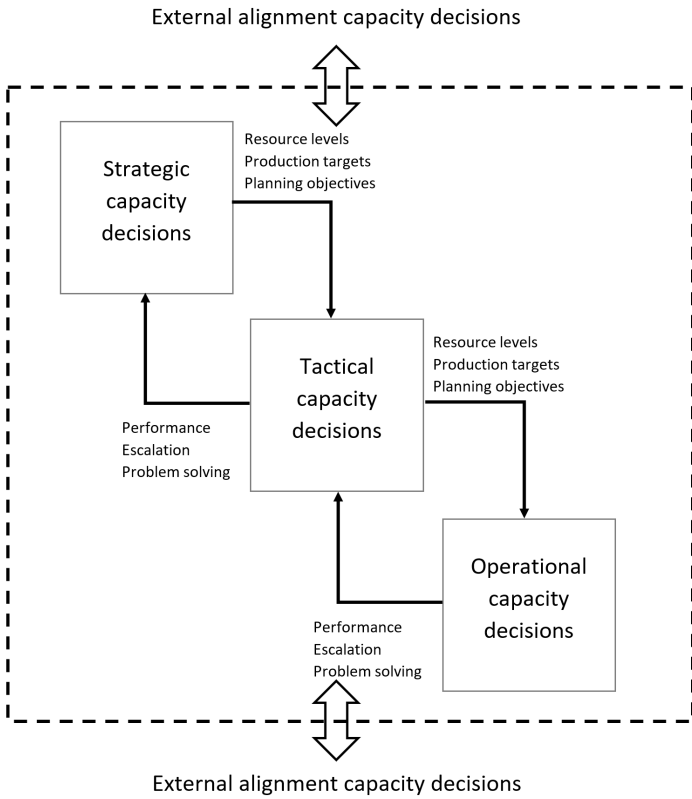
Integration in this dimension concerns the alignment of the hierarchical of nature of decisions in manufacturing [11] and hospitals [106, 247] (see Figure 2.1. From an ICM perspective, hospitals are characterized by multiple services, a wide variety of equipment and operations in several departments and therefore capacity decisions affect multiple resources and processes. To understand these decisions from the hierarchical dimension, we use the decomposition approach based on early literature in management control systems [11]: strategic decisions, tactical decisions and operational decisions. The rationale behind this decomposition is that higher levels set boundaries, targets and planning objectives (i.e. increasingly disaggregated information) for lower levels. As higher hierarchical levels involve decision-making with larger horizons and therefore larger portions of patient pathways, decision-making on higher levels spans more capacities. Furthermore, care pathways are not bounded by an organization and therefore ICM must encompass capacity decisions across multiple organizations.

Strategic capacity decisions. ICM addresses every step of the healthcare delivery process and starts with translating strategy to strategic ICM (e.g. operations strategy). When an ICM strategy is incorporated in strategy development, it can ultimately be translated into operational capacity management by providing clear goals for planning at all levels. ICM therefore supports strategy execution and will lead to improved performance [156].

Strategic capacity decisions concern the structural design, dimensioning, and development of the healthcare delivery process. Typical decisions on this level include decisions about the service or case-mix (e.g. patient types and volumes) translated into the required level infrastructure to realize goals and objectives. This can results in the acquisition of new infrastructures. Information used for these type of decisions is highly aggregated, drawn from many external sources and both demand (e.g. patient case-mix) and supply are characterized by a high degree of uncertainty.

When mission statements and strategy are not translated to operational goals and objectives and therefore are disconnected, staff often do not realize that they are bound

Figure 2.1: Hierarchical Integration Dimension for Integral Capacity Management in Hospitals



and committed to corporate strategy. ICM facilitates translation of mission statements into measurable and achievable goals [181], as it makes them normative and, to a large extent, manageable. Such goals can then be used for strategic capacity planning (e.g. case-mix planning and forecasting). Without key performance indicators (KPIs) (e.g. measurable goals and objectives), goals and objectives can become ambiguous, and therefore ICM can not be accomplished. In other words, ICM demands a mission with a clear value proposition (e.g. desired patient case-mix and service levels). Furthermore, these measurable goals can be operationalized and thus can ensure process quality (ultimately supporting quality of care), process safety and equitable customer access. In addition ensuring quality for customers, ICM also can ensure quality of labor [207]. In short, ICM can deliver productivity and efficiency improvements for all stakeholders [123]. We therefore challenge hospitals to think about making their mission and strategy normative, measurable and manageable.

Tactical capacity decisions. These type of decisions concern the organization of healthcare delivery at the highest level [5, 106]. Typical decisions on this level focus on periodical capacity dimensioning through *allocation* of resources (e.g. blueprint/master scheduling) and workforce. Information used is moderately aggregated, coming from both external and internal sources and while most of resource levels are determined at this level, demand is characterized by a moderate level of uncertainty as a result of emerging demand and/or increasing urgency driven by disease progression in known patients. Other decisions typical for this level concern the expansion or reduction of overtime and temporary capacity. As the authors of [106, 200] already state, we observe that these types of decisions are still not systematically managed leading hospitals to jump from strategic decision-making to operational fire fighting almost without considering tactical decision-making. This is increasingly time consuming as direct alignment is difficult, for example, strategic horizons of at least one year have to be directly translated to operational horizons of at most a couple of months).

Operational capacity decisions. This level encompasses day-to-day capacity *matching* decisions. While for tactical decisions, capacity levels can be temporarily adjusted (e.g. extra shifts, overtime, or capacity reallocation), for operational decisions capacity levels are given. Inherently, capacity decisions on this level concern a short term horizon and where both demand and capacity are known (e.g. low uncertainty) leaving little flexibility. This level can be further disaggregated into offline (e.g. in advance) and online (e.g. instant) decisions [106]. Typical decisions here are patient-to-nurse scheduling at the beginning of a shift and rostering adjustments as a result of sickness.

Summarizing, these levels can be uniquely defined by their extent of capacity flexibility - high (strategic), moderate (tactical), or low (operational) - and further explained by the following aspects:

- Length of planning horizon.
- Detail level of information and type of sources (e.g. external and/or internal sources of information).
- Authority and responsibility of involved management.
- Level of uncertainty in demand and/or supply.

Top-down and bottom-up

Top-down integration is important to translate strategy into operations as each level sets resource levels, production targets and planning objectives for underlying levels. Bottom-up integration is important to provide feedback to improve higher-level aggregated information, and the quality of higher-level decision-making based on performance monitoring, timely escalation, and structural problem solving. As mentioned in the introduction, this dimension is already in place in most hospitals. However, most CM decision-making processes in hospitals are focused on top-down deployment, while bottom-up feedback loops are often lacking. This may be explained by the strongly decentralized autonomy of departments and healthcare professionals [122]. As a result, problems are often not escalated to higher hierarchical levels and therefore are not solved structurally. This leads to myopic optimization. A dysfunctional hierarchical integration leads to mismatches between demand and supply, resulting in stop-and-go operations, increased waiting times and inefficient resource utilization. We often observe such dysfunctional hierarchical integration in hospitals where tactical decisions are rarely taken (e.g. institutions with infinitely repeating cyclical blueprints or master schedules for capacity allocations). Moreover, inundated in operational problem fighting, management is in a real-time problem engagement (e.g. managing solely depends on the availability of real-time information), while problems can structurally be solved on higher levels through adjusting master schedules. However, when problem are escalated, which may occur periodically, they quickly plead for "more capacity", which requires a long-term (e.g. strategic) decision. Thus, hierarchical integration is an iterative process, in which a hybrid adoption (e.g. top-down and bottom-up) forms ICM on this dimension [133] as bottom-up integration feeds new strategy development and top-down integration facilitates strategy execution. Furthermore, once bottom-up capacity decision-making is in place, it may reduce healthcare professionals' resistance to increased coordination [208].

Performance management is a crucial part of the hierarchical dimension. Through performance management, targets and objectives can be cascaded to other hierarchical levels and monitored as they are realized. Unfortunately, in both literature and practice universal definitions and standardized sets of performance indicators for ICM are missing [228]. The lack of ICM performance management facilitates the aforementioned myopic optimization, as departments are not accountable for such indicators. One cause for the lack of such indicators could be that most hospitals only recently emerged from the digitization era, and are now discovering the value of process information in improving ICM. Incompatibility of information systems hinders

this progress.

2.3.2 Patient-Centered Care Integration

Patient-centered care integration is *the coordination and alignment of capacity across departments and organizations to optimize care pathways.*. This integration dimension considers the perspective of the patient. It is also referred to as horizontal integration. The latter term is more ambiguous, as it is defined in the literature as "concerted practices between companies operating at the same level(s) in the market" [63].

A consequence of increasing clinical specialization is that patients will face an increasing number of healthcare professionals involved in their care pathways. This also means that patients may have to visit an increasing number of departments and even organizations. From a clinical point of view and that of the patient, these are all necessary steps in the care pathway. From an ICM point of view, this creates increasing interdependencies between departments and therefore requires coordination and integral capacity decisions and planning (e.g. multi-appointment scheduling).

From a hierarchical dimension, patient-centered care integration should be realized on every level. On a strategic level, long-term collaborations are formed between healthcare organizations that are adjacent in care pathways, to optimize transfers and minimize blocking (i.e. creating flow). An example is the collaboration between hospitals and nursing homes, or between hospitals and rehabilitation centers. This requires strategic investments in information architectures to ensure optimal information sharing, preventing work duplication or even loss of information.

Patient flows are managed on a tactical level. As in industry supply chains, demand propagates through the hospital from outpatient clinic to diagnostics, to preoperative screening, to operating room, to inpatient ward. It is well known from supply chain management theory and from queuing theory, that myopic optimization of capacity utilization leads to bullwhip effects [248]. This means that without coordination, downstream departments observe fluctuations in demand. Trying to deal with these fluctuations, these downstream departments try to increase capacity availability of their downstream departments and so on with increasing variation of capacity downstream. Although, covered extensively in the literature [49], We observe this is still the case in hospitals, where operating room scheduling "leads"‘ (i.e. where utilization needs to be maximized), and downstream departments, for example the ICU and inpatient wards, observe an increase in the number of admissions and therefore they try to increase their bed capacity. The resulting bullwhip effects causes patients to wait and staff to experience stop-and-go operations. ICM encompasses alignment to optimize flow over all capacities involved (i.e. complete care pathways). This implies that fluctuating demand is propagated along all capacities, by adjusting capacity levels (e.g. flexible capacity sharing) so that capacity matches demand as closely as possible. This results in minimal waiting for patients (i.e. optimal flow), stable workload and considerable capacity utilization throughout the system. Stabilization has a limited and to further balance demand and supply capacity may be flexibilized. Tactical planning should improve downstream forecasts based on upstream information, which allows for timely capacity adjustments. The difficulty of tactical decisions lies in determining the aggregation level of information that is required to make such decisions. For

instance, which level of detail on care pathways should be considered? Management should balance the trade-off between the information loss through aggregation and the complexity of the decision. Carefully making such capacity decisions could improve efficiency, as it will decrease costly operational fire-fighting [106].

On an operational level, inter-unit coordination of capacity should focus on both online and offline multi-appointment scheduling, promoting one-stop-shopping to reduce the number of hospital visits. Patients can be offered a choice of preference to further enhance patient-centeredness. News of a downstream blockage (i.e. one of the patient's appointments being delayed or cancelled) should be swiftly communicated between involved capacities to prevent waiting and enable flexible scheduling adjustments and capacity reallocations.

2.3.3 Domain Integration

Hospitals also have a fragmented management structure in which management decisions are limited based on information from other managerial domains (e.g. financial decisions are functionally dispersed from capacity decisions) and are often functionally dispersed [106]. This may lead to unbalanced capacity allocations, resulting in unbalanced workloads, with an overloaded workforce in some departments, and an idle workforce in others.

As explained in Section 2.2, there are three managerial domains related to capacity decisions: clinical, nonrenewable resources and financial. Therefore, the last dimension of ICM is managerial domain integration. This integration aims to align decision-making processes between domains such that the impact on capacity is integrally analyzed. For this, we build upon the framework presented in [106].

Clinical Domain

In hospitals, the role of technological planning in capacity planning is performed by clinicians [106, 263] and is referred to as clinical planning. This role encompasses the design of production processes (e.g. clinicians design treatment plans). ICM serves clinical decision-making and therefore capacity should be aligned to clinical planning. For instance, on a strategic level, designs of new treatment plans impact capacity requirements. On a tactical level, the clinical decision of selecting treatment plans impacts required capacity. And on an offline operational level, the selected anesthesia and surgical protocols affect surgery durations. Lastly, an online operational decision example is the triage of patients at an emergency department influence required capacity. However, also capacity decisions have influence on clinical decision making. When demand exceeds available capacity, treatment plans may be adjusted such that demand and supply are balanced.

In hospital governance, clinical leadership plays a crucial role in hospital decision-making on all hierarchical levels [26]. Clinical leaders have to embrace and design ICM in co-creation for successful implementation. This has been recognized for decades: "Practitioners have to develop greater appreciation of the managerial processes, and managers as well as community representatives have to reflect a deeper understanding of the clinical operations" [90]. However, we observe that capacity and clinical

decisions are still, to a large extent, functionally dispersed. This is striking, as, in the end, capacity decisions are clinical decisions (e.g. matching capacity enables clinical practice and clinical prioritization determines capacity usage). ICM serves clinical decision making, as it can assign clear boundaries to what can be done within a desired time interval (i.e. available capacity) and therefore creates realizable treatment plans and planning objectives or what level of additional capacity is required to meet treatment plans. This dispersed decision-making may be explained by the blind spots that both clinicians and administrators have for the other's practice. We are therefore convinced that clinical and administrative leadership at all levels should embrace and have knowledge of each other's motives so they can understand each others behavior in decision-making processes [34]. Ultimately, hospital leadership should facilitate training of both disciplines to successfully implement ICM [40].

As both formal and informal clinical leaders are involved in the execution of ICM, they should be aware how their clinical decisions impact capacity. The challenge for clinical and nurse leadership dealing with ICM is that they constantly balance clinical or nursing objectives (e.g. to negotiate for and represent the interests of clinical or nurse staff) against organizational objectives to ensure both the quality and efficiency of care [26, 205]. Therefore, ICM may give insights into KPI trade-offs for decision-making. With these insights, clinical and nurse leadership can explain, from their perspective, counter-intuitive decisions to their staff and thereby create support for the realization of plans.

Financial Domain

Financial planning should align with clinical planning and as hospital expenditures are, to a large extent, capacity related (e.g. workforce and facilities) it should also be aligned with capacity decisions. For instance, a hospital's desired patient case-mix portfolio decisions should be aligned with financial and capacity decisions, such that portfolio decisions can be translated to their financial and capacity impact (e.g. negotiations with healthcare insurers, which involves financial decision-making, are translated into case-mix portfolio planning and capacity levels). Furthermore, information required for capacity decisions (e.g. durations of procedures) is also valuable when making financial decisions, as costs can then be allocated to procedures (e.g. workforce or material costs) and insights into operational spending become available.

Non-renewable Resources Domain

There are considerable dependencies between nonrenewable and renewable resources for healthcare delivery. Without nonrenewable resources, many processes (e.g. surgical sutures or diagnostic isotopes) cannot be executed. For efficient supply chain management of equipment, it is essential to align nonrenewable resource planning to capacity planning (e.g. patient scheduling and workforce rostering) such that required equipment is available at the right time and in the right place. Therefore, renewable and nonrenewable resource management should be aligned.

2.4 Discussion

This research aims to theoretically introduce ICM and should therefore been seen as a first step for theoretical conceptualization of the subject. Based on the literature and our own observations, we have developed an approach to design ICM in hospitals. The approach consists of decision integration in three dimensions: hierarchical, patient-centered care and managerial domain. Integration is not unequivocally a good thing as it complicates decision-making by increasing alignments and necessitates adjusted autonomies (beyond clinical specialties and departments). Increasing alignment and coordination comes with characteristic problems, such as sharp dilemmas related to process, commitment, empowerment, disclosure, and escalation. Therefore, integration should only be realized when there are strong interdependencies among the aforementioned decision-making dimensions for capacity. Due to the multi-factorial dimensions of our approach and the many involved features of a hospital organization, it will be challenging to implement ICM.

In regulated healthcare markets, institutions and companies try to cap healthcare expenditure by taking rigorous measurements [205], resulting in volatile reimbursements levels and consequently less predictable income for hospitals. Hospitals have great difficulties incorporating external dynamics into operational adjustments, as capacities are fragmentarily managed and hierarchical alignment is lacking [208]. ICM creates agility in capacity decisions and thus allocations, such that changing environments can quickly be responded to on all levels. Systems design starts with strategy development. Therefore it is important to incorporate ICM in strategy development. Unfortunately, we often observe discrepancies between mission statement and capacity related strategic decisions. As a result, hospitals must exert great effort to fulfill strategic goals and to deal with external dynamics that affect capacity. The mission statements and strategies of hospitals show us a broad scope of value propositions, where everything should be achieved for nothing less than 100%. Hospitals often fail to achieve the broad set of goals within their mission statement, as resources are scarce and therefore the options for fulfilling mission statements are limited. This indicates that hospitals do not consider ICM while developing new strategies and mission statements. However, the development of a hospital strategy often results in a new course with new products and/or services or the even disposal of existing ones. This means that clinician practices may be impacted immediately (with new treatment options or less treatment options) and could result in resistance against these new strategic directions,making operationalization difficult. Hospital organizations are characterized by a decentralized governance structure and therefore coordination of strategy development and strategy execution is challenging [89].

Healthcare is a labor- and knowledge-intensive service where available resources are, to a large extent, defined by the level of available workforce. Workforce management, therefore, has great impact on the strategic course and is an important condition for ICM. Increased integration requires more centrally organized control and alignment and often means a decrease in decentralized autonomy[89, 122, 156]. This decreasing autonomy may raise resistance as the span of control of clinicians and nurses will be reduced [122, 156]. For ICM to be successful in hospitals, one should be aware of this potential resistance and should challenge clinical leadership to embrace ICM as this

may minimize resistance and strengthen adoption of ICM [105].

Healthcare delivery is also characterized by a large degree of process uncertainty, as treatment plans can differ by patient (i.e. engineer-to-order). With the implementation of electronic medical record systems, data becomes available to evaluate the performance of realized plans from both a system and a local perspective. Data availability and awareness are therefore important enablers of ICM. Observe that logistical performance is a proxy for quality performance. Performance management therefore facilitates bottom-up decision-making as it gives clear targets and indicators for monitoring performance. Analytics can give insight into implementations and can be used to improve planning and forecasting (e.g. expected durations of procedures). Furthermore, EMR data can be used to digitally monitor performance, allowing timely detection of deviations, such that fire-fighting adjustments will decrease. Performance management creates feedback loops to evaluate and continuously improve ICM and capacity planning. Combining analytics and EMRs may ultimately facilitate automated clinical planning using patient characteristics and historical [88]. In the future, this automated treatment planning may be used as input for automated capacity planning.

ICM implementation requires a starting point. The selection of this starting point will depend on many contextual and situational factors. ICM is not a one-size-fits-all approach and there are many ways to implement it (e.g. from strategic systems design to operational excellence). Many hospitals approach ICM as tooling, as an increasing number of software vendors claim that tooling will result in improvements. As mentioned in Section 2.3, ICM is a management approach for which performance management is an important condition and data availability and awareness (and thus tooling) are enablers to implement performance management. A potential starting point for implementation could be to follow the organizational sense of urgency that focuses on a particular part of ICM. Another starting point could be to integrate the annual budget planning with production and capacity planning. Once such a process is in place, it may enable implementation of ICM at other hierarchical levels.

A relatively new organizational format for embedding ICM in hospitals, is the establishment of a command center [126]. Command centers originate from military, transport, and aerospace sectors and centralize previously local administrative processes and performance initiatives. This systems view is essential to prioritize projects, share best practices, and standardize work across the hospital [126]. The center described in [126] focuses on operational capacity decisions. This may be extended by centralizing capacity decisions made by higher hierarchical levels such that deployments (e.g. hierarchical integration) and alignments (e.g. patient centered care integration) are easily formed. Furthermore, centralizing capacity management-related activities and decisions creates corporate visibility of ICM, adds to ICM knowledge. and prevents local initiatives that result in myopic optimizations.

Value-based healthcare (VBHC) operationalizes patient-centered care through integrated practice units (IPUs), where all healthcare providers involved are jointly delivering care. The IPU concept goes beyond the fact that shared capacity will always be necessary and therefore it cannot be dedicated to specific IPUs. This makes implementation of IPUs difficult. ICM may therefore enable successful implementation of VBHC, as it integrates the patient-centered care dimensions into other dimensions [136].

Further research is needed to gather empirically evidence for the presented ICM dimensions. To conduct such research, reliable and valid methods of assessing ICM must to be developed. This can be achieved in at least two ways: (1) measures could be obtained from observations of hospitals that are known for their management system integration and/or ICM maturity or (2) ICM could be assessed by determining the hospital's ability to effectively respond to various external incentives.

CHAPTER 3

Ward Capacity Planning & Management[1]

3.1 Introduction

During hospitalization, patients spend most of their time in wards. These wards are also referred to as inpatient care facilities, and they provide care to hospitalized patients by offering a room, a bed and board [118]. Wards are strongly interrelated with upstream hospital services such as the OT and the ED. Given this interrelation, it is essential that hospital wards be readily available in order to achieve efficient patient flows. Of course, hospital management desires that resources be used efficiently and therefore they try to optimize the trade-off between availability (e.g. measured in access time and refusal rate) and occupancy (e.g. measured in utilization). Hospital ward management often aims for the *golden standard* of 85% occupancy rates to maximize the number of admissions. This simplified objective is often not optimal for the availability-occupancy trade-off, and achieving it also depends on the definitions of these measurements. An optimal bed occupancy rate depends on several factors such as: inflow, number of beds available and length of stay. For wards with a small number of beds (e.g. up to 12 beds), it will be difficult to attain an occupancy rate of at least 85% given the fluctuations in the number of arrivals each day, and therefore patients have to be refused, deferred or rescheduled. Targeting occupancy rates of 85% as a golden standard for all ward may thus be counterproductive.

This chapter aims to give an overview for both researchers and hospital management of available literature on ward related OR models by discussing the KPIs for wards, the type of ward and the models used to analyze these wards, the type of decision and the models used to analyze these decisions, and how these models can have impact in practice. We open this chapter by discussing various concepts of logistical KPIs for wards, how they are related, and how they, together, can give ward management realizable targets (Section 3.2). In Section 3.3, we introduce different types

[1]This chapter is based on N.M. van de Vrugt, A.J. Schneider, M.E. Zonderland, D.A. Stanford and R.J. Boucherie. Operations Research for Occupancy Modeling at Hospital Wards and Its Integration into Practice In C. Kahraman, Y. Topcu, editors, *Operations Research Applications in Health Care Management*, chapter 5, Springer International Publishing, Cham, United States, 2018. 101-137
And A.J. Schneider and N.M. van de Vrugt. Applications of Hospital Bed Optimization. *Working paper*.

of wards based on these logistical KPIs. OR can provide managerial insights about trade-offs between these KPIs and therefore in Section 3.4 we present OR models for the types of wards we have defined. Next, in Section 3.5 we take a broader scope and discuss ward-related capacity management decisions and how these decisions are related to each other from a hospital perspective. We then, in Section 3.6 show how OR models could support ward capacity management decisions making. In Section 3.7, we discuss illustrative cases in which OR models have had a practical impact on ward capacity decisions. Finally, in Section 3.8, we discuss factors critical to having an impact in practice.

3.2 Key Performance Indicators for Wards

The logistical performance of wards is generally assessed by three KPIs: throughput, blocking probability and occupancy. Optimizing only one of these KPIs, will reduce performance on the others. Therefore, each of these KPIs should be balanced with the others. Based on the type of ward, these KPIs can be targeted differently. In general, occupancy is the most important KPI for ward management. Therefore, we begin this section with definitions of these performance indicators. Next, we define various ward types from a logistical perspective and show how OR models are used to analyze these types of wards. We also illustrate some OR models used for various ward types.

3.2.1 Terminology

Although the concept of occupancy may seem simple initially, researchers and healthcare practitioners use different definitions of it. This may result in false comparisons, if the definitions used are not clearly stated. Therefore, we provide the frequently used definitions in the following paragraphs. We first define different concepts of capacity (based on [246]), then define throughput and blocking probability, and conclude this section with the various definitions of occupancy.

Each ward has a certain capacity, which is expressed in terms of the number of patients and the care intensity that the ward can accommodate. The capacity of a ward is measured by the number of beds and nurses, and there are different types of capacity. The *physical capacity* is the number of beds in the ward. Each nurse can take care of a certain number of patients, determined by the nurse-to-patient ratio, which determines the *structural available capacity*. Additionally, temporary capacity changes can occur: for example, bed closures in holiday periods or beds that are used during shortages but are not officially staffed. The structural capacity and temporary changes together determine the *realized available capacity*.

Example:
Suppose, a hospital ward has 15 beds. There are always three nurses scheduled to work on the ward, and each nurse can take care of at most four patients at the same time. Each summer and during Christmas holidays, the ward experiences decreased in numbers of patient, and decides to schedule only two nurses. The holiday periods together last 8 weeks. Then, for this ward the physical capacity

is 15 beds, and the structural capacity is 3 (nurses) $\times$4 (patients per nurse) = 12 beds. Due to the holidays, each year has 8 weeks during which only eight beds are open, so the average realized capacity is:

$$\frac{8 \text{ (weeks)} \times 8 \text{ (beds)} + (52 - 8) \text{ (weeks)} \times 12 \text{ (beds)}}{52 \text{ (weeks)}} \approx 11.4 \text{ beds}.$$

As mentioned in the introduction of this section, the logistical performance of a ward is assessed by three performance indicators: *blocking probability* (i.e. the opposite of availability), *occupancy*, and *throughput* (which is a result of the first two KPIs). The *throughput* of a ward can be measured as the number of admissions or discharges per time unit. However, this KPI is subject to the variance in the LoS. When the LoS is highly variable, the throughput of a ward is also. The *blocking probability* of a ward is the percentage of patients who request a bed in the ward at a moment when there are no available beds:

$$P_b = \frac{\text{Number of patients not accommodated on ward}}{\text{Total number of patients requesting a bed on ward}} \times 100\%. \tag{3.1}$$

Blocked patients are either accommodated in a different ward, deferred to another hospital or delayed if possible. Furthermore, this KPI can also be used as to estimate the time a ward is fully occupied.

In contrast to throughput and blocking probability, *bed occupancy* can be quantified by three definitions: based a on bed census at given time, based on real length of stay (LoS) or based on the number of hospitalization days. Here we aim to give an overview of the most commonly used definitions.

One of the definitions of bed occupancy includes the bed census measured once a day at a specified point in time: for example, every morning. Then, dividing the average of these measurements by the structural available capacity, the occupancy is:

$$O_{bc}(t) = \frac{\text{average bed census at time } t}{\text{structural available capacity}} \times 100\%. \tag{3.2}$$

Note that for the occupancy it also matters how the capacity of a ward is calculated; in most hospitals the structural available capacity is used. A slightly different occupancy measure is obtained by taking the average of multiple bed census measurements throughout each day: for example, each hour. We denote this measure by $\bar{O}_{bc}$. The advantage of taking more measurements is that the average better reflects actual bed usage.

Hospitals may also define the occupancy of a ward as the ratio of the sum of all patient LoSs combined to the total time available:

$$O_{LoS}(T) = \frac{\text{sum of all LoSs for all admissions in time period } T}{\text{structural available capacity} \times \text{ length } T} \times 100\%. \tag{3.3}$$

This measure is calculated using admission and discharge time stamps for a certain period, or by multiplying the average LoS by the number of patients accommodated on the ward. By taking the sum of all LoSs of admissions during time period T, a fraction of these LoSs is not realized within this time period (e.g. for admissions at

the end of T, the LoS exceeds T). The opposite is in place at the start the period, where a fraction of admissions is not considered as these admissions started before the period. One remark, is that this occupancy measure reflects the actual time the beds are used, but does not incorporate unavailability due to cleaning of beds.

Until recently, it was common in Dutch hospitals to determine the bed occupancy using the hospitalization days declared to insurance companies:

$$O_{hd}(T) = \frac{\text{sum of hospitalization days for all admissions in } T}{\text{structural available capacity} \times \text{ length } T} \times 100\%. \tag{3.4}$$

Again, by taking the sum of all hospitalization days of admissions in time period T, a fraction of these hospitalization days is not realized within this time period (e.g. for admissions at the end of T, hospitalizations days exceed T). Therefore this approach gives some over estimation of the actual occupancy. Financial hospitalization days are counted in integers, and can be invoiced if the patient is in a bed before 8:00 pm and discharged after 7:00 am the next day. This implies that the occupancy can be over 100% as beds can be reused if patients are discharged early in the day and new patients are admitted in the afternoon. A drawback of this measure is that it cannot be used as a targeted occupancy for all ward types. For example, it cannot be used in wards in which patients generally stay for only a part of a day so that multiple patients can be served by the same bed on the same day (e.g. gynecology). In this system, these wards can achieve occupancy targets over 100%, while wards in which patients have much longer stays (e.g. geriatrics) suffer severe bed shortages if the occupancy is over 90%.

Example:
Here we will illustrate the different definitions of occupancy. Consider a ward with three beds that is empty at the start of our observation period. We choose to observe the ward from 8:00 am on day 1, until 5:00 pm on day 4. In this period the following patients arrive:

Figure 3.1: Occupancy Chart for Example

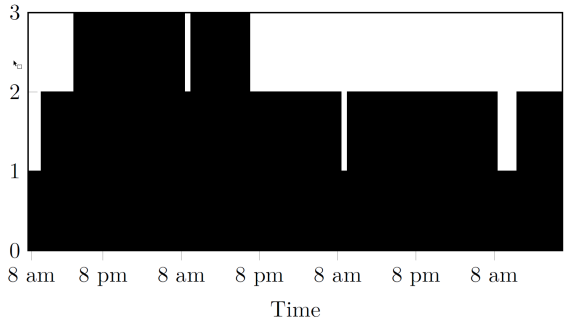


Table 3.1: Occupancy Data for Example

	Arrival		Discharge			
	Day	Time	Day	Time	LoS	Hosp. days
Patient 1	1	8:00 am	2	6:00 pm	1.42	2
Patient 2	1	10:00 am	4	8:00 am	2.92	4
Patient 3	1	3:00 pm	2	8:00 am	0.71	2
Patient 4	2	3:00 am	Patient is blocked		-	-
Patient 5	2	9:00 am	3	8:00 am	0.96	2
Patient 6	3	9:00 am	After day 4		1.33	2
Patient 7	4	10:00 am	After day 4		0.29	1

In this example, patient 4 is blocked as patients 1, 2 and 3 fill all available beds and the first patient discharged (patient 3) is not discharged before patient 4 arrives. Note that the LoS for patients 6 and 7 in the table is not their exact LoS but only the part until the end of the observation period. The bed census for this ward is depicted in Figure 3.1. The blocking probability for this time period equals $1/7 \approx 15\%$. The different occupancy measures are calculated as follows.

The bed census at 10:00 am for days 1 to 4 is 2, 3, 2, and 1, respectively, so the average equals 2. Therefore $O_{bc}(10\text{ am}) = 2/3 \approx 66.7\%$. The average hourly bed census is 2.2, so $\bar{O}_{bc} = 2.2/3 \approx 74.8\%$.

The sum of the LoS for all patients on this ward in this observation period, T, equals 7.63 days. The length of the observation period is 3.38 days. Therefore, $O_{LoS}(T) = 7.63/(3 \times 3.38) \approx 75.3\%$.

The sum of the hospitalization days declared for these patients is 13, and the total number of days in this observation period is 4. Therefore, $O_{hd}(T) = 13/(3 \times 4) \approx 108.3\%$.

Hospital management determines which of the aforementioned occupancy measures is used and sets a seperate target throughput level for each ward. A high occupancy usually results in a high blocking probability [16]. Therefore it is important for management to balance these three performance indicators. Adequate targets for the performance indicators depend on many factors: for example, the capacity of a ward, the fraction of admissions that is acute, the possibility of deferring admissions, the cost per bed, and the ward layout. Large wards have economies of scale, so a higher bed occupancy can be achieved with a lower blocking probability. If a ward has mostly acute admissions, occupancy targets need to be set lower; elective admissions can be rescheduled in the case of a bed shortage, while acute admissions cannot be. If the deferral of an arriving patient could give rise to a life-threatening situation (e.g. in the case of an intensive care unit), a ward has to lower the target occupancy to produce a lower blocking probability. However, such wards usually have high costs per staffed bed, driving the occupancy targets upwards. Finally, if a ward has many rooms with multiple beds, the bed assignment is less flexible than in wards with many single-bed rooms; for example, a patient with an infectious disease cannot share a room with others. In conclusion, it can be said that determining adequate occupancy, blocking probability and throughput targets is a challenging task.

3.3 Ward Taxonomy

We distinguish different ward types based on logistical characteristics: the type of inflow and outflow, typical LoS, resources, equipment, and planning problems the wards face. Based on the literature cited in this chapter and our own experience, we distinguish the following types of wards:

- Intensive care unit (ICU)
- Acute medical unit (AMU)
- Obstetrics ward (OBS)
- Weekday ward (WDW)
- General ward (GEN)

We describe each type of ward in terms its logistical characteristics in Table 3.2, in which "0" denotes average occurrence, occupancy or costs, and "+" or "-" denotes increase or decrease, respectively,compared to average.

Table 3.2: Summary of Characteristics per Ward Type

	ICU	AMU	WDW	OBS	GEN
Long LoS	+	−	−	0	+
Short LoS	+	+	+	+	0
Acute admissions	+	+	−	+	+
Elective admissions	+	−	+	+	+
Bed occupancy	+	−	+	−	0
Staff / bed ratio	+	+	−	0	0
Equipment	+	0	−	0	0

Intensive care unit For this category in our taxonomy we group several ward types with similar logistical characteristics: traditional ICUs, specialized ICUs, and critical, high- or medium-care units. Specialized ICUs included, for example, stroke units, cardiac care units and neonatal ICUs. High care and medium care units are sometimes combined and are often referred to as step-down units between ICUs and general wards. In the United Kingdom and the United States these wards are also referred to as "Critical Care Units". The difference between high and medium care is generally the need for breathing support. The ICU of a hospital accommodates the most severely ill patients who require constant close monitoring and support from advanced medical equipment and staff (nurses mostly on a 1:1 basis, and intensivists, who are readily available) [159]. In the remainder of this chapter we refer to the ward types discussed in this section as ICUs.

Due to the used equipment and the available staff, the ICU has the highest costs per bed of all hospital wards. An ICU preferably does not defer patients, as this would imply serious mortality risks. However, the costs per bed do not allow for a large buffer in the number of available beds. Therefore, ICUs tend to be fully occupied, and they discharge the least-ill patient, although this patient may not be medically indicated for transfer, when a bed needs to be freed for a newly arriving patient, or they cancel an elective procedure in the OT which that would require ICU capacity afterwards.

Patients typically have either a short LoS or a very long LoS, and they arrive from the OT, ED, wards or surrounding hospitals.

Acute medical unit There is no universally accepted definition. We believe the following definition best covers AMUs: "an AMU is a designated hospital ward specifically staffed and equipped to receive medical inpatients presenting with acute medical illness from EDs and outpatient clinics for expedited multidisciplinary and medical specialist assessment, care and treatment for up to a designated period (typically between 24 and 72 hours) prior to discharge or transfer to medical wards" [206]. Often, AMUs serve as a buffer for both the ED and inpatient wards. Since an AMU treats only urgent patients and should alleviate ED congestion, management is more focused on throughput and LoS, and the target utilization of the AMU beds is typically lower than for general wards. AMUs are also known as "emergency observation and assessment ward", "acute assessment unit" and "acute medical assessment units". The systematic reviews [60, 206] provide a comprehensive overview of definitions and concepts for AMUs. The inflow mainly consists of acute patients from the ED, outpatient clinics, surrounding hospitals, or general practitioners.

Weekday ward Weekday wards (WDWs) are wards admitting patients with an expected LoS between 2 and 5 days, and they usually open on weekdays [58]. WDW-type hospitals are also sometimes referred to as "Monday to Friday clinic" or "Week Hospital". Most patients at WDWs are elective, and can be transferred to regular wards without any health risks. Only patients with a highly predictable LoS may be admitted, which is why WDWs mostly treat patients for whom strict treatment protocols apply. Scheduling patients at a WDW is complicated by each patient's different LoS and urgency level, which implies a deadline by which the patient should be treated. The requirement that the ward should be closed during weekends also complicates patient scheduling. Most admissions arrive directly from home.

Obstetrics ward Obstetrics and gynecology wards (OBS) provide care for women during pregnancy and during and after labor, and they also take care of their newborns [57]. Additionally, gynecology wards accommodate women who have problems with their reproductive organs. These patients often require brief surgical intervention, and typically a short hospitalization. Some hospitals group these types of wards under names such as "birthing center", "maternity clinic", or "women's and child's Center". Most patients arrive from home, outpatient clinics or other hospitals.

General wards General wards in hospitals are often dedicated to a single medical specialty such as neurology, geriatrics, or hematology. As these wards are generally equipped with similar resources and accommodate both acute and elective admissions which differ in LoS, we aggregate these ward types. General wards can be either surgical or medical and certain wards, such as psychiatric or geriatric wards, are closed, implying that patients cannot leave without approval. Other wards are equipped with a specific type of resource, such as dialysis machines and heart monitors. The nurse-to-patient ratio is often 1:5 to 1:6. Patients with a particular medical specialty are typically not all accommodated in the same ward, but may also be admitted at, for

example, a WDW or an ICU. Patient inflow is mainly from referrals of outpatient clinics, ICUs, general practitioners or other hospitals.

3.4 Ward-related OR Models

In the previous section we distinguished different ward types and their logistical similarities and differences. In this section, we review the OR literature for each ward type, emphasizing the main questions or problems the literature tries to solve, the context of the problems (e.g. ward type), and the type of models invoked for each paper. Table 3.3 shows the number of papers found for each ward type along with the OR model/method used. If a paper invoked multiple OR models, we categorized the paper in all applicable categories. We review the literature related to the specific ward types in Sections Sections 3.4.1 to 3.4.5.

Table 3.3: Literature Categorized by Applied Models and Ward Type.

OR model or method	ICU	AMU	OBS	General	WDW	**Total**
Algorithms	1	0	0	3	0	**4**
Dynamic Programming	1	0	0	0	1	**2**
Markov processes	4	0	2	11	0	**17**
Mathematical programming	3	2	3	5	1	**15**
Queueing theory	13	2	3	15	0	**34**
Regression	1	0	0	1	0	**2**
Simulation	21	1	2	21	0	**45**
Time series	1	0	0	2	0	**3**
Total	**46**	**5**	**10**	**56**	**2**	**119**

3.4.1 Intensive Care Unit

The ICU has both elective and emergency patient arrivals. Emergency patients mostly come from the ED or surrounding hospitals and elective patients mainly arrive after surgery. Since significant costs are involved, management tends to maximize utilization. This has resulted in increasing numbers of refusals or severely ill patients being transferred from the ICU to high, medium or regular care wards. This could lead to situations in which quality of care is at stake and possibly to disruptions in the OT schedule. These are also the main problems the literature of this section focuses on: admission and discharge control.

In [33], a queueing model ($M/G/s/s$ queue, see Appendix A for explanation) was used to analyze the minimum number of ICU beds required for burn care for the state of New York. The authors start by determining the number of beds at an aggregate level given a maximum blocking probability of 5%, and then they apply a heuristic allocating these ICU beds among several regional units, while trying to maintain the same blocking probability. This model is extended to analyze an overflow model in [154]. Here, each ICU reserves bed capacity for regional emergency patients, which may be used as overflow beds in a given region. To approximate the blocking probability

Table 3.4: Literature on ICUs Categorized by Applied Models.

OR model/method	References
Algorithms	[33]
Dynamic programming	[54]
Markov processes	[33, 45, 72, 85]
Mathematical programming	[134, 159, 160]
Queueing theory	[33, 69, 98, 131, 154, 160, 171, 210, 229, 252, 255, 264]
Regression	[159]
Simulation	[37, 38, 61, 66, 131, 132, 134, 135, 154, 159, 160, 164, 165, 167, 176, 177, 198, 209, 214, 221, 255]
Time series	[85]

of this overflow model, the equivalent random method is used, whereas a simulation model is used to validate the results of this queueing model with historical data. Another modified $M/M/s/s$ model is used to analyze different admission policies and their relation to survival gains [210]. The following policies are analyzed: (1) the standard first come first served (FCFS) discipline; (2) arrivals are served if and only if a bed is available and the survival gain is greater than an arbitrary threshold value; and (3) arrivals are served if and only if a bed is available and the survival gain threshold value is met, with this policy the threshold value dependent on the number of beds available. If fewer beds are available, the threshold value for survival gain is increased. The results show significant increases for survival gain in both the second and third policies compared to the first policy. The third policy demonstrates only marginal survival gain compared to the second policy, while it significantly increases the number of rejected patients. Another application of the $M/M/s/s$ queue is used to analyze an ICU [171]. This model is validated with observed data and it is proved that the calculated blocking probabilities from the queueing model were accurate.

In addition to queueing models, discrete-time Markov chains are also applied to ICUs. The authors in [72] develop a Markov chain to analyze so called bumping (patient transfers from the ICU to free capacity for new arrivals who are more severely ill). Another application of Markov chains by [54], is used to evaluate the effect of ICU discharge strategies and bed census on patient mortality and total readmission load.

Simulation is also often applied to analyze the required number of ICU beds. Several scenarios are analyzed in which ICU beds are reserved for emergency arrivals [134, 165, 198]. The authors of [131] simulate several ICU arrival processes and compares these results with theoretical results using an $M/M/s$ queue. Based on the simulation model, the authors also determine the blocking probability for the current capacity. Another study [132] analyzes several scenarios to minimize the number of elective surgery patients refused at the ICU. The efficient frontier method is used to plot the trade-off between the number of canceled surgeries and the average waiting time per scenario.

In [159] multiple OR techniques are used to analyze the ICU: first, a regression model is proposed for modeling the ICU LoS; second, a comprehensive simulation model is developed for analyzing system behavior and blocking probabilities; and last,

mathematical programming is used to model the triage problem (which current and arriving patients are most in need of ICU capacity?) for early or delayed discharges from the ICU that are based on high or low utilization of ICU capacity.

When patient logistics at the ICU are analyzed, there is a clear distinction between the type of models used and the type of problems solved. Because a significant fraction of arrivals at the ICU are unscheduled, queueing theory gives accurate and representative results. In analyzing ICU dynamics, this technique is typically used to achieve general insights about blocking probability, occupancy, ICU capacity, and the trade-offs among these. Markov chains are used to analyze bed census probabilities and the probability of bumping. Simulation is generally used to analyze multiple scenarios where particular details are involved or case-specific dynamics need to be studied.

3.4.2 Acute Medical Unit

The systematic reviews on AMUs mentioned in Section 3.3 conclude that AMUs may have many advantages but that evidence of economic effectiveness is not clear. One review finds AMU "performance is dependent on good management and availability of diagnostic services", and asserts that there is no proof of cost-effectiveness of AMUs [60]. An extensive list of success factors for AMUs is also available [206]. From an OR perspective, if a hospital does not add beds or staff to its current capacity for opening an AMU, the improved performance reported in the reviews is disputable. The beds assigned to the AMU are taken from other wards, decreasing the benefits of economies of scale and affecting other patients at those wards. Additionally, patients who require inpatient care after their stay at the AMU encounter more process steps than if they had been admitted directly. Therefore, the effects of opening an AMU cannot be predicted beforehand without the use of appropriate mathematical models. Perhaps partly because AMUs are a relatively new concept, the OR literature applied to AMU is somewhat sparse. In this section, we review this available literature.

Depending on the performance measures of interest and the research goals, several models can be applied to AMUs. We describe a goal programming approach used to minimize the delay from moving a patient from the ED to AMU, and two different queueing networks to evaluate blocking probability and bed censuses.

A goal programming approach is used to determine the required additional resources (beds, doctors, and nurses) for each hour of the day to minimize the delay patients experience on an AMU staffed with eight beds, two nurses and three doctors [178]. Goal programming is an extension of mathematical programming in which for each objective, a target (or goal) is set, and deviations from these targets are minimized. In the model, each patient requires a bed, and a specific treatment by a nurse, a doctor, or both. A patient is delayed if there are no beds available upon arrival or if the doctors and nurses are seeing other patients at the moment the patient requires care. For the case studied, the average LoS was 5 hours, and the run time of the model equaled a day and a half. The conclusion is that only two doctors are required, and a third nurse should be on standby in the afternoon and at midnight to cope with peak demand.

In a follow-up study for a larger AMU (58 beds), a simulation is used to analyze

14 scenarios with different numbers of beds [179]. Here, each resource type (beds, nurses, and doctors) has its own queue, and patients wait in these queues until the resource they require is available. Initial targets for each queue length are fed into a goal programming model, together with targets for total LoS and the number of beds. The authors minimize weighted positive deviations from these targets. The model output comprises the resource levels that minimize patient' delays at the AMU, and a trade-off between higher utilization of resources and patient- and staff-related objectives is provided.

Another study analyzes a network with one AMU and an aggregated regular ward, in which patients are transferred between the wards if their care requirements change [223]. The authors use an infinite server queueing network to determine the probability that the bed occupancy on either ward exceeds a certain number of beds. Based on this probability, they determine the optimal assignment of the available beds to either the AMU or the regular ward. For the case in which total mean bed occupancy is 85%, and 91% of the patients require acute care, they conclude that 60% to 65% of the available beds should be designated for acute care.

For a network comprising an ED, two aggregated wards, and an AMU, the study reported in [265] determines the blocking probability by invoking a network of Erlang loss queues in which the AMU both has direct patient arrivals and serves as an overflow ward. The authors consider both urgent patients (arriving from the ED) and elective patients. The hospital is only allowed to reallocate existing beds from the wards to the AMU. The equivalent random method is used to analyze the network with overflows, since overflow traffic does not follow a Poisson distribution. This method approximates the original network by truncating an infinite server network. The authors conclude that opening an AMU is beneficial for accommodating urgent patients, but the blocking probability for elective patients increases significantly.

The advantage of a simulation or goal programming approach over queueing networks is that time-dependent arrivals can be incorporated relatively easily. However, the size of the state space in a goal programming model increases with the time horizon considered and will explode when several departments of realistic sizes are considered. The drawback of simulation models is that they are not easily applied to other hospitals. The advantage of considering infinite server queues is that straightforward formulas for the analysis exist in the literature.

3.4.3 Obstetrics Ward

The literature includes several OR models applied to OBS wards and maternity clinics. We describe different queueing theory approaches, a simulation model, a discrete-time conditional phase type model, and a discrete time Markov model.

In research conducted almost 40 years ago, the bed occupancy at an OBS ward using an infinite server queue is modeled [170]. The ward may also admit gynecology patients to achieve higher occupancy rates, but those patients are transferred to other wards if an OBS patient has no available bed upon arrival. The gynecology patients may only be admitted to the OBS ward when the bed census is lower than a certain threshold. The authors use an infinite server queue to represent the situation in which patients are placed in unstaffed beds as a temporary measure when no official beds are

available upon arrival. The results are compared for multiple hospitals and incorporate the national guidelines on the admittance of gynecology patients to OBS wards. The authors state which thresholds are best for given ward sizes.

Another study calculates the probability of delay for example, the probability that there is no bed available upon patient arrival, using an $M/M/s$ queue [96]. Key to this model is that arriving patients who find all beds occupied wait at the clinic until a bed becomes available. During their waiting time, patients are not treated, as their "service" commences as soon they are placed in a bed. Inputs are the average LoS found in hospital data and different arrival rates. The authors compare the probability of delay for different occupancy targets and different arrival rates.

For a maternity clinic consisting of different wards, including a neonatal ward and ICU, the Queueing Network Analyzer is used (cf. [264]) to model the bed occupancy [57]. The authors evaluate all possible bed arrangements among the wards for the peak arrival rate of the clinic. The best arrangements are then evaluated in a system with an inhomogeneous arrival rate in a discrete event simulation. The authors report that the hospital has implemented some of their recommendations, but instead of reassigning beds the hospital chose to add 15 beds to the ward with the greatest bed shortage as identified by the simulation and the queueing model.

To model different types of wards in a network of multiple maternity clinics independent $M/M/s/s$ queues are also used [186]. The general Erlang loss formulas for the blocking probability are then fed into a mixed integer linear program (MILP) to determine strategic bed-assignment policies. Each year the clinic may reassign, open and close beds at the wards and clinics, and each decision entails certain costs. The authors incorporate long-term planning, since it is undesirable to have a ward close beds and fire nursing staff one year, only to reopen these beds and recall staff the following year. The objective of the optimization program is to minimize the costs over the decision horizon. One of the authors' conclusions is that efficiency could be improved if resources were transferred among units that experience different demographic changes (such as an increase or decrease in the number of women giving birth).

In an attempt to improve the occupancy rate of an obstetric clinic, this next study investigates different scenarios by means of discrete event simulation [97]. Inflow and LoS of the model are based on hospital data; patients in the model follow one of the predefined care pathways through the clinic. When they compare the results of both approaches to hospital data, the authors conclude that the care pathway-based approach reflects reality better than a transition probability-based approach. One of the investigated scenarios includes "swing rooms", which are rooms that can be used by multiple wards of the clinic, but not at the same time. The clinic subsequently implemented the swing rooms, which proved useful for balancing utilization throughout the clinic during bed census peaks.

A discrete-time Markov model is developed to mimic a maternity clinic consisting of four wards in [121]. Patients can flow among units, with the routes patients take depent on their type. The authors define 11 patient types and six arrival streams (e.g. natural birth or cesarean), and the LoS has an empirical discrete distribution. All input is derived from hospital data. Since the model assumes infinite capacity, the authors derive the mean and variance of the bed occupancy at the units in case no

patients would be deferred to other clinics. These can be used to approximate the bed census by fitting a normal distribution with the same mean and variance. The normal approximation is included in an integer linear programming (ILP) optimization model to optimize the scheduled arrivals at the clinic. Several of the assumptions are validated by means of a discrete event simulation. One of the conclusions is that scheduling some patients on Saturdays smooths the bed census significantly. The authors report that their model has supported multiple clinics in the United States.

The next study focuses on predicting the LoS of women arriving at a maternity clinic [109]. The authors define a phase-type distributed LoS for two labor types: spontaneous and scheduled. For both types a decision tree based on patient characteristics, such as age and weight, further specifies the LoS parameters. The prediction of the LoS is then included in a simple continuous time Markov model to calculate bed occupancy for the labor ward of the clinic, using a homogeneous arrival rate. The model uses the LoS distribution and transition probabilities that women experience in each phase of labor. The steady state of the model reflects the bed census at different phases, which require the inclusion of different wards at the clinic.

In the literature on OBS wards we found two attempts at increasing bed occupancy, either by admitting non-OBS patients or by using swing rooms. Interestingly,the authors of [109] conclude that the hospital data they obtained does not show a specific time-dependent arrival distribution, while other researchers [57, 97, 121] do model time-dependent arrival rates. Arguably, scheduled arrivals (scheduled cesarean births) likely occur only during office hours, which implies a time dependent arrival rate. Queueing models are more difficult to use in a time dependent system, since the simple formulas for waiting and blocking probability do not hold in a time dependent system. The drawback of using simulation models is that most models are case-specific, applicable only to the clinic they were designed for. However, the advantage of a simulation model's graphical visualization is that practitioners can easily see the implications of different interventions, which often means that results of the research are more easily implemented into practice. An advantage of the discrete time Markov models is that these models are able to mimic reality better than queueing models can, and are still more general than simulation models. However, a potential drawback could be a rapidly increasing state space for average sized clinics consisting of multiple wards. Others propose an approximation of the bed census by a Normal distribution, and based on their simulation results this seems a reasonable proposition [121].

3.4.4 General Ward

In most of the literature included in this section, general concepts are analyzed that are applicable to many types of wards, or the studies take multiple departments into account. Given this breadth, most literature discussed here focuses on strategic or tactical planning by evaluating capacity dimensioning decisions or predicting demand.

The models for analyzing general bed census concepts cover a wide range of OR techniques and are applied on different levels. The techniques reviewed in this section are listed in Table 3.5. We highlight these models and their conclusions below by discussing a selection of the studies in Table3.5.

A queueing model is used to determine the bed demand at the community level,

Table 3.5: Literature on General Wards Categorized by Applied Models.

OR model/method	References
Algorithms	[29, 113, 230]
Markov processes	[8, 93, 128, 141, 197, 211, 217, 219, 224, 225, 244]
Mathematical programming	[7, 21, 29, 152, 230]
Queueing theory	[20, 21, 29, 69, 82, 86, 91–93, 96, 99, 111, 152, 243, 264]
Regression	[139]
Simulation	[8, 15, 75, 77, 78, 93, 102, 110, 111, 113, 128, 135, 139, 140, 143, 145, 236, 243, 244, 261, 262]
Time series	[145, 158]

focusing on high occupancy rates, while keeping refusal rates for emergency patients low and waiting lists short [211]. The bed census is modeled using an infinite server queue incorporating two classes (elective and emergency) of arrival streams. This model elaborates on earlier research applying an infinite server queue by adding a threshold parameter (B). To balance the elective and emergency arrival streams, the model blocks elective admissions if the occupancy rate is higher than or equal to B. This model provides policy makers useful insights into the relationships between bed census, length of the waiting list and emergency refusals. Another queueing model incorporates predictable fluctuations in the average number of arrivals [20]. This time-dependent queue, an $M(t)/H/s/s$ model (in which $M(t)$ indicates a time-dependent Poisson process, see Appendix A for information on the notation), is evaluated by using approximations based on the infinite server queue. It is shown that daily fluctuations have limited impact on the bed census, whereas weekly patterns do have a significant impact on both the bed census and the number of refused admissions. Finally, the authors present a method to determine the required number of beds across the week. In[91], an $M/PH/s$ queue is used to determine the optimal bed census for a hospital, in which the LoS is phase-type distributed (which is denoted by the abbreviation PH). The authors in [69], employ the Erlang loss model ($M/G/s/s$ queue) to relate the blocking probability to the occupancy. Additionally, a broad introduction of various applications of queueing networks in healthcare is also available [264].

Several studies use a discrete Markovian approach to predict the short-term bed census. These predictions are mainly based on the current bed census at day t, the expected elective and emergency admissions, and the expected discharges at day $t+j$. In these models the LoS is often empirically distributed. The census distribution is approximated from their Markov model by a normal distribution [224], showing that this relatively easy approximation performs satisfactory when applied to hospital wards. Markov models are also applied to obtain the distribution of the number of patients in each phase of a care pathway, for geriatric [93, 219] or stroke patients [244], to determine the required resources in each phase of the pathway.

Simulation is used by [75] to analyze bed allocation and usage policies for all beds in a hospital based on hospitalizations days (e.g. 24-hour bed occupancy) per specialty, average daily bed census at a certain time, bed occupancy over a time period, patient misplacements and annual misplaced patient-days. Another simulation analyzes the so called “winter bed crisis”, a yearly bed shortage during mid winter [243]. The results

show that discharge delays during mid winter were the main reason for high bed census counts. The following study analyzes waiting times for surgical procedures by means of simulation [236]. Another simulation study was performed to balance emergency and elective admissions for the available bed capacity [143]. The last simulation study focuses on the overflows between wards (in which patients are transferred to another ward because the designated ward is fully occupied), and find that the occupancy of wards is a good predictor for the frequency of overflows [128].

Time series models are also used to predict bed census demand. An hourly bed census prediction is modeled with a time series model [145]. The results are used to reallocate beds between ward types such as medical, surgical or obstetrics. A different but related approach involves the use of mixed exponential equations to obtain the probability distribution of patients in different phases of their care pathways. In [157, 244] the model is applied to mimic the bed census, allocating emergency admissions on both a regional and hospital level. Results show that this type of model mimics the bed occupancy accurately. The first study analyzes the accuracy of these mixed exponential equations based on a case study and compares equations by evaluating the effect of adding more parameters [158]. The latter study relates the blueprint schedule of the OT, in which each subspecialty gets a fraction of the available OT time, to the hourly bed census distribution in the postoperative wards [138].

A nonlinear mixed integer mathematical programming model is used to allocate the number of available beds among hospital services over a finite planning horizon [7]. The decisions are based on patients' waiting time before admission and budget limits. A similar technique is employed in [230], where integer programming assists in clustering the clinical departments and assigning these clustered departments to available wards. These assignments are such that capacity is sufficient to guarantee a maximum blocking probability.

In conclusion, the choice of a modeling technique depends on the desired output. Queueing theory is suitable for determining the capacity or census distribution of a single ward, preferably with mostly unscheduled patient admissions, when a maximum blocking probability or target occupancy must be achieved. Markov models and time series models are accurate for determining the census distribution or certain percentiles, but might be tedious to analyze as the state space may become large. Simulation models can be developed as detailed or macro-level as desired but are generally suitable for obtaining average performance measures. Mathematical programming can be used to optimize the reallocation of beds to wards.

3.4.5 Weekday Ward

Although most Dutch hospitals have a WDW, and the optimization potential is significant, we were able to find only two references that study WDWs. This may be explained by the lack of capacity issues in this ward types. Since all patients are elective, they can be scheduled at a time when beds are available, and patients who cannot be admitted can be accommodated on the general ward. Still, we are confident that WDWs have a substantial logistical potential; large efficiency gains can be achieved if the number of beds is adequate and patient scheduling is optimized.

Due to the lack of modeling work on WDWs and the sparsity of scheduling work for this type of ward, we describe below two models for optimizing the patient scheduling that are relevant to the present discussion.

For a Monday-to-Friday rheumatology clinic, admissions from a waiting list are optimized in this model [58]. An introductory meeting determines a patient's medical priority, resource requirements and LoS. LoS is maximally 5 days. The authors develop an ILP model, in which they decide on a schedule time slot, if any, for each resource the patient requires (e.g. beds or diagnostic tests). Patients are assigned a weight according to their medical priority and time spent on the waiting list, and the objective is to maximize the weighted number of admissions. The authors conclude that the number of available beds is the bottleneck, and their optimized schedule can accommodate twice as many patients as manually composed schedule.

The last study we found on WDWs considers an online appointment scheduling version of the WDW patient scheduling problem: a patient's request arrives and should be assigned to a date and time immediately, without knowledge of future patient arrivals [41]. The authors develop an approximate dynamic programming model to obtain the optimal scheduling policy. This technique is often invoked when dynamic programming models suffer from the curse of dimensionality, and encompasses aggregating the state space and approximating the value function.

3.5 Ward Capacity Management

Thus far, we have elaborated on specific ward types and models. From this section on, the focus is on hospital-wide ward capacity decisions. The term capacity management is most often used for decisions on the acquisition and usage of renewable resources to efficiently satisfy customer demands [94]. Efficient realization of organizational goals (e.g., satisfied and healthy patients and employees) requires hospital-wide coordination of capacity and flows, by continuously balancing demand and supply. OR can give managerial insights for capacity management to improve the efficiency of capacity and flows.

To demarcate the scope of capacity management decisions and optimization analyses in wards, we use the four-by-four framework of [106]. The framework hierarchically decomposes managerial levels on one axis: strategic, tactical and operational (offline and online). It covers different managerial areas on the other axis: medical planning, renewable resource planning, nonrenewable resource planning and financial planning. An important step in planning and control is to set the length of the scheduling horizon for the different hierarchical levels. Strategic planning encompasses the longest horizon, while operational planning has the shortest horizon. On the other axis, the framework integrates the managerial planning areas in healthcare: medical, resource capacity, materials and financial planning. In this chapter we focus on resource capacity planning for both planning and control and OR models for wards. Following this framework, we use a top-down approach explaining all planning and control decisions at the different hierarchical levels, as higher levels create capacity boundaries for lower levels based on forecasts. Nevertheless, in practice bottom-up feedback from actual planning realizations should also occur, so detected deviations and problems can be

lifted one hierarchical level higher for problem solving.

3.5.1 Strategic Ward Capacity Management

At the strategic level, the hospital administration decides on the hospital's long-term mission and strategy - the areas on which the hospital aims to focus and excel. Important strategic decisions included: the desired case-mix, hospital layout, performance targets, bed capacity and workforce plan.

The desired case-mix of the hospital

Based on the hospital's mission and strategy, the case-mix is formed: a case-mix is the collection of patient groups a hospital treats. Some hospitals choose a very specific patient group to treat - for example a breast cancer clinic - while general hospitals have a more diverse case-mix. The case-mix of a hospital determines to a large extent the required capacity.

The case-mix of a hospital can be adjusted by attracting certain patient groups and defferring others. As healthcare organizations and their professionals have a duty to care for their patients, a hospital is not allowed to defer a patient group until other regional or national providers agree to treat this patient group. Furthermore, the patient case-mix can also change when a new doctor with a different specialization is hired. The same can happen when specialized doctors leave the hospital.

Adjusting a hospital's case-mix is a complex process as many factors should be taken into account. For example, case-mix decisions can affect not only the capacity needed for care delivery, but also the education and research possibilities. Another example is that patients are not often treated only by a single specialty. Thus, the decision to stop treatment of a specific patient group by a medical specialty could affect the case-mix of many other medical specialties.

Hospital layout planning

Based on the mission and strategy, a hospital board decides at the strategic level which type of wards and rooms will be available in the hospital. One factor they consider is the mix between single- and multi-person rooms. Single-person rooms ensure privacy for patients and their families, but require more space and increase walking distance for the staff, which may complicate the monitoring of patients. On the other hand, multi-person rooms are inefficient when patients have infectious diseases and when all rooms are same-sex.

Another aspect of strategic planning for wards is the decision to establish wards for special types of care; AMUs, ICUs, CCUs, and surgical admission lounges are examples of such dedicated wards. These wards serve specific patient groups based on severity, urgency, treatment or flow (e.g. elective versus acute admissions) and are mainly introduced to improve the quality of care, efficiency, or both. As such decisions have high economic impact and a long time frame, they will affect the hospital logistics for multiple years.

After the global lay-out of the hospital is determined, all patient groups in the case-mix must be assigned to wards. An important decision at this level is how many

staffed beds to assign to individual wards. From a logistical viewpoint, larger wards will result in economies of scale, leading to a higher occupancy with an acceptable blocking probability (see [226]). For this decision, the trade-off between medical and logistical perspectives should be considered. Purely from a logistical perspective, if all patient groups could be treated at any bed in the hospital, the bed census in this single ward system could be optimally balanced. A result of this would be that nurses in this hospital would have to be multi-skilled (which is impossible for highly complex care), and doctors would spend more time visiting their patients at different areas of the hospital. From a medical perspective, however, a more differentiated distribution of patient groups over wards would be optimal, with patient types clustered according to the skills required for their treatment. A balance between these two perspectives should be found.

Setting performance targets

On a strategic level, the hospital board may set performance targets for the hospital. In hospital wards, the logistical performance indicators used are often the bed census or occupancy, with occupancy measurable in many different ways [226]. Setting high occupancy targets for wards will, certainly for smaller wards, result in deferring more patients to other wards or hospitals. Different performance targets would be remaining below an upper-bound on the percentage of deferred patients, and achieving the desired case-mix.

The number of beds

Once a hospital's desired case-mix is determined, this information is used for strategic planning to forecast the demand for care for the hospital. This forecast is based on aggregated data, trends, and forecasts of the patient population. Using the forecasted demand for care and the set performance targets, a hospital can determine the required capacity to treat these patients. The required capacity is determined on an aggregated scale, such as the total number of required OT hours, outpatient clinic hours and ward hours for the upcoming years. Based on the aggregated data, the required ward capacity is re-evaluated yearly.

Typically, the number of physical beds in a ward is higher than the average number of used beds. Each ward should have buffer capacity, to accommodate unexpected peaks in patient arrivals. Often, not all physical beds at a ward are staffed: there is no nurse available to treat a patient in an unstaffed bed. These extra beds ensure, for example, that patients with infection risks can be treated in isolation when the ward has multi-person rooms. Moreover, these beds form a buffer of clean beds if the time between one patient's discharge and another patient's admission is short.

A ward may also have over-capacity in the number of staffed beds, when the nurse-to-patient ratios do not perfectly match the expected bed census. The ratio depends on the average workload each patient represents, and denotes the number of patients one nurse can take care of. For example, consider a ward with an expected bed census of 17 beds and ratios for each shift that are as follows: day $1:3$, afternoon $1:5$ and night $1:8$. As a result, the shifts requires at least six, four and three nurses, respectively. On a strategic level this slack should be taken into account when the

expected bed census is translated into the required number of full-time-equivalent contracted nurses.

Workforce planning

Often, the number of nurses rather then the physical bed capacity determines the available beds at wards. Once the demand for beds has been determined by the case-mix plan, the workforce should be aligned to it. An important capacity management decision that determines how many nurses are required, is the nurse-to-bed ratio. For example, an ICU patient has a relatively high workload and the nurse-to-patient ratio is 1:1, while for a general ward during a night shift the ratio may be 1:16. Workload lacks a universally accepted definition but is generally considered to be the relation between the demand (of patients) and the capacity available to fulfill this demand. Workload can be divided into objective (e.g. patient acuity metrics) and subjective (e.g. nurse workload perception) factors ([213, 234]). Patient acuity metrics are generally activities of daily living, cognitive support, communication support, emotional support, safety management, patient assessment, injury or wound management, observational needs and medication preparation. Perceived workload is also not universally accepted as a measurement for workload, but is mainly related to staff characteristics such as age, experience and educational level. The total workload in a ward is based on the patients' acuity, the shift (day, afternoon or night), and the bed census.

Nurses are mostly assigned to a single ward, or to a few wards that accommodate patients with the same care requirements. At a strategic level, a hospital can decide to flexibly allocate a part of its capacity. For wards, this implies that not all beds are assigned to a certain medical specialty or specific patient group (e.g. organ transplantation patients), but part of the bed capacity will be assigned based on the actual demand for care for each patient group. Flexibility may also imply that a hospital creates a flex-pool of nurses; these nurses are often multiskilled and are allocated on a short-term (for example each morning) to the busiest ward. The advantage of flexible capacity is that a hospital can better adapt to stochastic patient demand. The determination of the level of flexible capacity, and the selection of which KPIs the allocation is based on, is a strategic capacity management decision.

A hospital's desired case-mix also determines the quantity and capabilities of the required staff, to a large extent. However, the translation from case-mix to the number of staff members is not the same for each hospital. Continuous technological and medical innovations require greater specialization from practitioners. In general, more specialization increases the number of specialists involved during diagnosis and treatment, as all specialists are focused on a small part of a human body or specific diseases. Additionally, a hospital's policy on education affect how the case-mix is translated to number of required staff-members; junior residents usually decrease the capacity by increasing supervision duties of the staff, while residents ending their training often work with minimal supervision. Moreover, staff involved involved in research projects, typically decreases the available capacity for treating patients. Additionally, each hospital has differences in its workforce, which implies that strategic workforce planning should incorporate these factors: influx (training, education and immigrants) and outflux (retirement or retention) of staff, level of task differentiation

(e.g. highly trained versus basic trained staff) and the in- or outsourcing of training and education programs.

Strategic capacity management decisions are always long-term and often require major monetary adjustments to accomplish. For economic value and employee satisfaction, yearly changes to the ward layout of the hospital are not desirable. However, small changes in the case-mix may make ward sizes inadequate over time. Here, both under- and overcapacity are problems (see section 3.7.1). Strategic decisions define the directions for tactical and operational planning, and are thus, to a large extent, accountable for the performance of the hospital.

3.5.2 Tactical Ward Capacity Management

At the tactical level, capacity management decisions focus on organizing the desired case-mix, controlling patient access times, and efficiently using capacity via generating master schedules, allocation of flexible capacity and scheduling bounds, which we explain in this section. As mentioned earlier, tactical capacity management decisions concern the organization of operations and processes over the medium term.

Master Schedules

A major topic in tactical management is the division of the capacity among stakeholders and over the weeks of the year, resulting in a master schedule. This is called rough-cut capacity planning. Often a master schedule is set for an entire year, but when the scheduling horizon is shorter, hospitals can gain flexibility to adjust capacity to patient demand. For wards a master schedule may drive a weekly schedule in which the capacity of each ward changes over time, and beds are divided among different patient groups (e.g., elective and emergency admissions). Typically, a master schedule handles each season differently. Temporary leaves of staff may also require alterations to the master schedule. Holidays, training, education and internships are examples of reasons for temporary leaves and should be planned on the tactical level.

The master schedule of a ward should be aligned with the master schedules of other capacities, such as OTs and outpatient clinics, to create a stable flow of patients. Especially for wards that accommodate surgical patients, aligning the master schedules of the operating rooms and wards results in increased efficiency. This alignment is twofold: aligning holiday weeks and balancing bed censuses. Typically, a hospital has several holiday weeks per year in which elective patients and staff are not available and capacity is reduced for elective (i.e. scheduled) care. To prevent a shortage or surplus of beds, the holiday weeks of the OT and wards should be aligned. Additionally, by optimizing the operating room master schedule, the post-operative bed census can be improved [81]. Balancing the bed census implies that a ward requires less buffer capacity, and thus it increases efficiency in a ward and reduces the risk that a surgery will be canceled due to a lack of postoperative beds.

Flexible Allocation of Capacity

Hospital demand usually fluctuates, thus it may be beneficial to adjust part of the capacity throughout the year. To create flow (e.g. minimize variation) in the hospital,

admitting a similar number and type of patients each week is optimal. This is a difficult to achieve, however, given staff holidays and the stochasticity in patient arrivals and durations (i.e. LoSs). It is beneficial to periodically evaluate the available capacity, for each patient group (or aggregated for each medical specialty) and make minor adjustments where necessary and possible. For wards, this could mean asking nurses to work on a different ward for several weeks, or asking doctors to, for example, help on the ward instead of working in the outpatient clinic.

Hospital management can decide on the strategic level to reserve a portion of capacity and allocate this on a regular basis - for example, monthly. At the tactical level, this capacity may be allocated several weeks in advance. When applying flexible capacity allocation it is important to have consensus among all stakeholders about the parameters and performance indicators upon which the allocation will be based and on the scheduling horizon for allocating the flexible capacity. For wards, flexible capacity could imply that nurses from the flex-pool are assigned to a specific ward. Moreover, downstream resources should be considered when allocating staff from a flex-pool to align patient loads. For example, allocating flexible operating room time affects the bed census in postoperative wards, thus these decisions may require additional nurses in those wards.

In most hospitals, staff rosters are generated several months in advance, and thus staff planning is performed on tactical level too. These rosters indicate only which shifts and days an employee will work and do not specify the department, bed, or patients. The scheduling of these details, is performed on the operational level. This provides the departments additional flexibility in staff allocation.

Regulating the demand for care

Tactical capacity management decisions are crucial to efficiently organizing patient care and flows, especially at the interface between different types of resources in a hospital. From our own experience, this level is still underdeveloped in many hospitals. To balance patient flow and optimize efficiency, at the tactical level hospital management can decide to formulate rules for patient scheduling. Such rules may for example state the minimum and maximum numbers of elective patients that may be admitted to a ward per day within a week. Another example is a rule imposing a maximum on the number of ICU bed-requiring surgeries that may be scheduled on the same day. These rules can be ward-specific or medical specialty-specific, or they may hold for the entire hospital.

3.5.3 Operational Ward Capacity Management

The operational level is, by definition, divided into offline (service at a later point in time) and online (instant service) capacity management decisions [106]. Compared to the strategic and tactical level, the operational level has very limited possibilities for adjusting capacity based on patient demand. The online level comprises the actual patient-(room and bed)-to -staff scheduling and ad hoc decisions, such as replacing ill staff members and responding to admissions of emergency patients.

Patient scheduling

Patient scheduling on the offline operational level consists of deciding each elective patient's admission date and ward. This scheduling should take into account that each admission and discharge implies a workload peak for the nurses. Additionally, it is important to consider the expected urgent patient admissions, as scheduling too many elective patients results in deferral of emergency patients. Moreover, a patient schedule should minimize the number of in-hospital patient transfers, as each transfer can be a risk for the quality of care. Therefore, a patient schedule should, for example, account that for wards that close beds on the weekends and thus do not accept admissions expected to stay beyond Friday. This means accurate LoS predictions are crucial. Some hospitals and wards adjust the patient schedule one week in advance, based on the actual bed census and LoS predictions for the currently admitted patients.

On the online level, a ward manager may decide to transfer a patient in relatively good health to a ward with a lower care-level or to another hospital, to reserve capacity for high-care patients. In practice, patients are typically admitted to their medically preferred ward when there is available capacity, and ward managers start to transfer patients to 'second-best' alternative (also called 'overflow') wards when capacity runs out. As a consequence, patients may wait for a long time in wards that are not medically preferred before a bed is available, as another patient needs to be transferred first. To decrease the time until a patient is assigned to a bed, ward managers may transfer patients even when there is still available capacity, to reserve enough capacity for new patients. Focusing on the expected discharge date at the moment of arrival decreases the LoS (this is also called discharge management). Transfer decision are often difficult, but optimizing this decision-making process may improve patient waiting time, quality of care, and even hospital revenues significantly. Hospitals may also apply admission control to ensure enough capacity is available for those patients that need it the most or benefit from it the most, especially when there are other hospitals nearby.

Staff scheduling

On the operational offline level, staff is assigned to a specific ward several weeks in advance. When the hospital has a flex-pool of nurses, these nurses may be allocated to specific wards at the operational level. Some hospitals allocate these nurses several weeks in advance, based on long-term illness of staff, short-term staff leaves or forecasted patient demand. Hospitals may also decide to assign nurses from the flex-pool in an online manner, with each a nurse assigned to a ward at the beginning of the week or even of each shift.

On the online level, nurses are assigned to patients. This scheduling task is performed before each shift starts. The number and type of patients assigned to each nurse is optimized. As patients acuities and staff characteristics vary over time, nurse-patient assignments should be optimized by distributing the workload among available nurses on the operational level.

Although there is little room to adjust capacity to actual demand on the operational level, we have shown many capacity management decisions on this level that can further optimize patient care delivery. As improvement on this level require relatively

small adjustments in terms of work routine or investments, change is relatively easily to implement. Therefore, both management and staff can fulfill the potential of these improvements at any time.

3.5.4 Feedback between the hierarchical levels

As is common in the literature, we have used a top-down approach for discussing all hierarchical control levels in wards. As mentioned earlier, healthcare processes and planning deal with stochasticity, and therefore unforeseen situations often occur. Monitoring systems should be in place to detect deviations from scheduled care processes. Using data from electronic health records, software can easily detect, present and even predict these deviations. It is important to note that some data has to be entered manually (e.g. the expected discharge date) to make accurately detect deviations. When a deviation is detected or predicted, planners and ward management can pro-actively make adjustments in capacity, demand, or both.

When detected deviations cannot be resolved within the managerial boundaries of the level at which the unforeseen situation occurred, the deviations should be escalated. Bottom-up feedback loops provide escalation channels to shift problem solving to higher hierarchical levels. For each level, it must be clear when detected deviations have to be escalated. An example of a situation that may require escalation would be regularly occurring peaks in postoperative elective patient arrivals; the master schedule of the OT should then be revisited to balance the postoperative arrivals in wards. In general, recurring problems may require structural redesign of processes and may thus require decision-making on a higher hierarchical level. As such, escalation channels are an important component of the planning and control cycle for resource capacity planning.

3.6 Ward Capacity Planning

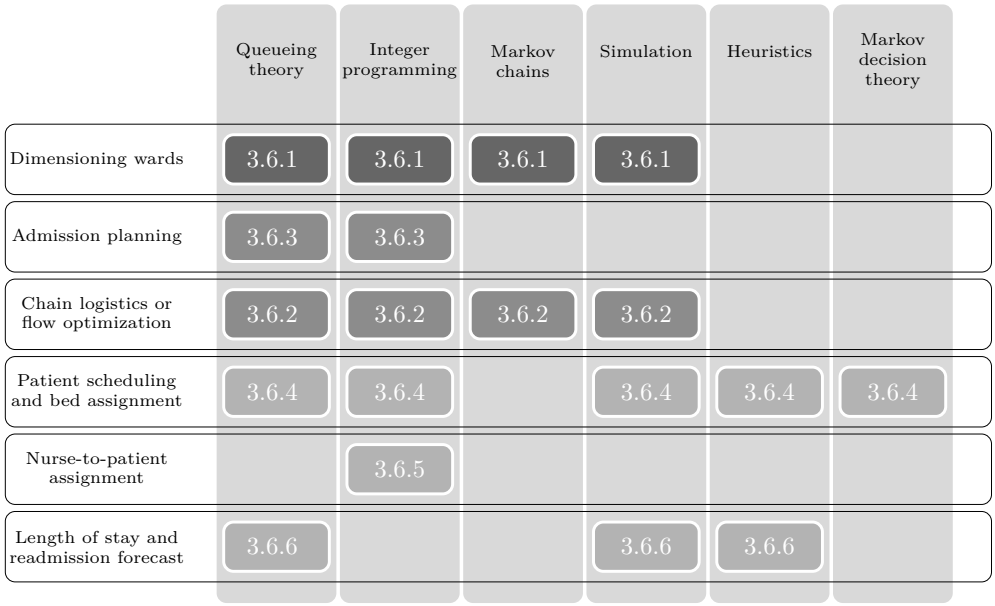
In this section we reflect on OR models that can be used to analyze ward capacity management decisions. We follow the same hierarchical approach as the previous section and in Figure 3.2 we show the capacity management decisions covered in this chapter and used OR techniques from literature to analyze these types of decisions.

One important optimization problem in wards is nurse staffing. Although the physical capacity of a ward is determined by the number of beds that present, in most hospitals the number of nurses present in the ward determines, to a large extent, the number of patients that can be accommodated. Many departments schedule the same number of nurses each shift or marginally adapt the nurse schedule to the bed demand. In this chapter we focus on bed dimensioning and patient scheduling and do not cover workforce scheduling.

3.6.1 Dimensioning wards

Finding the optimal capacity of a ward by allocating patient groups or types to wards is a typical strategic decision. In the literature, dimensioning decisions are based on queueing models, Markov chains, simulations, goal programming, or mixed integer

Figure 3.2: Overview of Section 3.6.



programming (MIP) models. Below we evaluate these approaches and provide some examples from the literature.

Queueing theory

The Erlang loss and infinite server queuing models are by far the most-used models for determining the best dimensioning of hospital wards. With easy-to-use tools available, such as the Queuing Network Analyzer (see [264]), hospital practitioners are able to analyze decisions with queueing models. The examples provided by [226] and in the case study presented later in this chapter in Section 3.7.2 demonstrate the value of these basic models for dimensioning hospital wards. An other advantage of the Erlang loss queue and infinite server queue is that these models are insensitive to the distribution of the LoS; obtaining an average LoS from hospital data is enough for the analysis. Sophisticated data analysis to generate input data is therefore not required for these models. The basic queueing models do not include all hospital ward dynamics. For example, they do not encompass non-homogeneous arrival and discharge rates, although in reality scheduled patients arrive and are discharged only during the day. Another example of misrepresentation is that in practice patients are often not 'blocked and lost' if all beds in their medically preferred ward are occupied upon their arrival, which implies that queueing models underestimate the bed occupancy. [82] demonstrate that for an infinite server queue with piecewise-stationary Poisson arrivals, the resulting model is easy to analyze. However, most queueing models become intractable with time-varying arrival or service rates. Additionally, feedback and overflow are typically difficult to analyze, as shown by, for example, [229] for a

small network of an OT and an ICU. To increase the predictive value of the model, the authors of [252] consider an Erlang loss queue in which the arrival rate depends on the number of occupied beds, to reflect that fewer patients are admitted to the ward when it is almost full. In [20], the authors analyze an infinite server queue with time-dependent arrival rate, and use the square-root staffing rule to dimension an ICU.

Integer Programming

Queueing models require a trial-and-error approach to find optimal capacity. To overcome this problem, queueing models can be incorporated into a MIP approach (e.g. a MILP). The authors of [230] analyze three approaches assigning patient group clusters to wards. The exact approach uses the Erlang loss model to determine bed capacity given a blocking probability and an ILP model is used to determine which patient groups should be clustered and assigned to a ward. The second approach uses an approximation of the Erlang loss model by using a linear function for the required number of beds followed by an ILP for the clustering process. This last approach uses the exact formulation of the Erlang loss model for the number of required beds and a local search heuristic to form the clusters. Another example of combining queueing models with optimization models is given in [186], whose authors use approaches similar to those use in [230] to determine the bed capacity for a network of maternity clinics. The authors in [186] also linearize the blocking probability and bed census of the Erlang loss model, and they analyze interactions between clinics with a mathematical model. The queueing model formulas can also be incorporated in a goal programming approach, which is the method used in [152], to allocate a number of beds in each ward and to ultimately optimize multiple objectives set by the hospital management. Finally, the authors of [179] use simulation to relate the capacity (beds, nurses and doctors) of a medical assessment unit to queue lengths for patients, and they incorporate this into a goal programming model.

Markov chains

Predicting a bed census using Markov chains may result in higher accuracy than a queueing approach, as time-varying arrival and discharge rates may be incorporated in such models. The authors of [137] invoke a Markov chain to predict the hourly bed census, which includes post-operative surgical patients, emergency admissions and overflow patients to and from other wards. Using the steady-state distribution, the authors obtain an expression for the 95th percentile of the bed census.

Markov chain models are also applicable in transient analyses; for example, the authors in [45] predict the ICU bed census by invoking a transient Markov chain analysis with maximum likelihood regression.

With Markov chains almost every desired detail can be modeled. However, adding more details to the model can quickly make a Markov chain intractable.

Simulation

With simulations, all features of hospital wards imaginable can be incorporated, which makes this type of modeling sometimes the best or only option for modeling a ward. In [113], a simulation model is used to relate the ward capacity to the bed census for several wards for all possible numbers of beds and to heuristically assign beds to the wards using these relationships. The model is evaluated using data from a university medical center. The authors of [236] analyze multiple scenarios to solve waiting list issues, one of which is re-distributing beds among wards.

Transient analyses are also possible while using simulation models: the authors of [259] present a simulation model they use to obtain short-term predictions based on the specific characteristics of the current patient population present in a ward.

Simulation models are labor-intensive, they require data analyses to generate input parameters, developing time, often require complete enumeration and output analyses. Furthermore, strategic analyses do not always require all details and therefore simulation models are not an obvious first choice to analyzed optimal ward capacity decisions. When using simulation techniques to analyze capacity management decisions, both academics and professionals should keep in mind the trade-off between required level of detail, the time needed to build and run a model, and the value of the outcomes.

3.6.2 Chain logistics or flow optimization

Thus far we have highlighted research that mainly focuses on a particular step in the patient care pathway: clinical treatment in inpatient wards. However, the inflow of inpatients is often determined by other hospital departments. Especially for wards that accommodate many surgical patients, the OT schedule determines, to a great extent, the ward's bed census. For wards that accommodate many urgent patients, the ED and acute admission unit, influence the bed census. The authors of [237] survey health-care models that encompass multiple departments. Interestingly, achieving optimal logistical flow through a hospital may result in suboptimal use of the capacity of individual resources. In this section we highlight literature on queueing, simulation, MIP, and Markov chain models that mimic the interaction between multiple departments.

Queueing theory

Queueing networks are useful in relating capacity levels to certain performance measures such as blocking probability. For example, the authors of [265] analyze multiple scenarios for a network with an ED, AMU and two wards, in which the acute admission unit may function as an overflow for the other three departments. They observe that with the setting they use, the arrivals of urgent patients can be increased at the cost of decreasing elective arrivals (the increase in emergency patients is greater than the decrease in elective arrivals).

A problem involving the deferring of intensive care patients because of capacity problems is also analyzed via queueing models [154]. They show that regional cooperation between ICUs results in higher acceptance rates for these patients. The authors approximate the blocking probabilities in an overflow network using the equivalent random method and the Erlang loss queue. Setting a threshold for this blocking probability they determine how many beds each ICU in a given region should reserve (for regional patients) so that all acute intensive care patients in the region can be accommodated promptly.

Simulation

Simulation models are also used to analyze patient flows through multiple hospital departments, and to determine the effect of changes in, for example, the capacity. Optimizing patient flows of individual hospital departments may lead to disturbed patient flows in other departments, as the bottleneck in the patient flow may shift to another department [135].

The authors of [204] use simulation to analyze the flow of emergency patients among three departments: the ED, the AMU and inpatient wards. The hospital in their case study has great difficulty accommodating all emergency admissions. Using heuristics they optimize the number of allocated beds per inpatient ward for emergency patients who need to stay longer than was intended in the AMU.

Another example of a simulation study in this area is [176], which is an investigation of strategies to keep patients from occupying high-care beds longer than necessary because lower care beds are unavialable. The authors in [68] investigate the effects of adding capacity and changing the discharge policy on the patient flow at a pediatric surgical center.

Mixed Integer programming

MIP models are often developed to optimize the OT's master surgery schedule (MSS). The authors of [81] optimize the MSS while minimizing the probability that overcapacity will be necessary to accommodate all patients in the postoperative wards. They analytically express the bed census distribution function for each ward based on which surgical specialty is assigned to which time slot in the MSS.

An operational offline approach is taken in [87], which invokes an MIP model to determine admission dates for patients who require care in multiple departments.

Markov chains

A Markov chain approach is invoked by [121] to model multiple patient pathways in an obstetrics department with multiple wards. The expressions obtained are incorporated into a MIP model to optimize the schedule of elective patients.

3.6.3 Admission planning

After the capacity dimensions are set for wards, demand and capacity can be optimized on a medium-term horizon through admission planning and nurse rostering, respectively. As mentioned in the introduction, admission planning generates a blueprint schedule that schedules the different patient groups and not individual patients or treatments. For this type of optimization, MIP models are preferred in the literature.

Mixed Integer programming

The authors in [226] use a mixed integer programming approach to develop a tactical schedule for a weekday ward. Weekday wards admit elective patients only on weekdays and close on the weekends. All patient care is delivered according to strict protocols, that result in highly accurate treatment times and LoS predictions. Typically, treatments with a longer LoS are scheduled at the beginning of a week and shorter treatments later during the week, to ensure that all patients are discharged before the weekend.

The authors of [112] develop two infinite-server queueing models (one for emergency arrivals and one for elective arrivals) to determine the bed census that results from any admission plan for regular wards. Based on these bed censuses, a MIP model minimizes the blocking probability of emergency arrivals, the cancellation probability of elective arrivals and the average number of boarders (patients who have to wait for their preferred ward) in the tactical admission plan.

The authors of [21] use a quadratic program to obtain a daily quota for the number of admissions to a ward to minimize the variability in the bed census (i.e. quota planning). In the quadratic program, the bed census is modeled using a $GI/G/\infty$ queue with a heavy traffic approximation, and the authors present an approximation for the bed census of a ward that experiences a non-Poisson arrival process. The aim is to generate rules of thumb for management and planners based on the model results. They conclude that quota planning has the greatest impact on bed census variability (e.g. smooth bed census during the planning horizon). Using quota planning makes the admissions arrival process at wards more stable. A stable arrival process results in a more stable bed census compared to highly variable arrival rates. An additional rule of thumb is to schedule arrivals given the number of available (or closed beds) during the planning horizon. Based on the absence of admissions during weekends, beds can be closed or patients with a longer LoS can be scheduled on Friday to improve the bed census on weekends.

Typically, during their hospitalization, patients require other types of resources, such as the OT or diagnostic facilities. Thus, the patient schedules for these resources affect each other. Incorporating many different capacity types and patients following uncertain treatment paths, the authors of [119] invoke an MIP approach to optimize the number of admitted patients per time period. The tractability of the MIP approach appears insufficient for optimizing realistic scenarios and therefore the authors turn to approximate dynamic programming [120].

Queueing theory

Queueing approaches may be used to determine the number of beds that should be reserved for a given patient type. For example, the authors of [154] use the equivalent random method to investigate a network of ICUs that all reserve some capacity to admit emergency patients in the region. As ICU capacity is scarce and costly, it is typically utilized maximally, which results in blocked emergency patients and cancellations of scheduled patients. The analysis shows that when multiple regional ICUs cooperate as a network, they can increase the acceptance level of emergency patients with a smaller total number of beds compared to the setting of reserve capacity by

individual ICUs. The authors of [163] present another application of queueing models to balance the bed censuses of wards with a similar level of care by considering algorithms for routing patients from the ED to wards.

3.6.4 Patient scheduling and bed assignment

Tactical admission planning results mainly in a blueprint schedule for patient admissions at wards. In the next planning phase (i.e. operational planning), actual patients are scheduled and assigned to available beds. The tactical blueprint serves as a guideline for scheduling patients. In some circumstances (e.g. based on patients who have been scheduled or on availability of staff), management and planners can deviate from this blueprint. This is not ideal as downstream resources must also accommodate this deviation. Using optimization, patient admission dates and bed assignments can be chosen such that the number of beds required to treat all patients is minimized or the variation in bed usage is minimized. Additionally or alternatively, the number of patients who receive treatment within their preferred access time window can be maximized.

Optimizing bed assignments has the greatest impact when the medically preferred ward has multi-person rooms or when there are several wards offering an adequate level of care. For multi-person rooms, for example, patients with infectious diseases and same sex in one room-rules may complicate the room assignments. Basically there are two types of decisions in this type of problem: (1) shifting admission dates and (2) transferring patients between wards according to medical preferences.

Below, we highlight literature on patient admission scheduling, bed assignment and admission control. Sets of benchmark instances for the offline optimization of bed assignments[2] and the patient admission scheduling problem[3] are available online.

Mixed Integer programming

MIP models are incorporated in online decision support systems to optimize bed assignments. For example, [196] and [242] determine the optimal ward and/or bed assignment for each patient with respect to the bed censuses for all wards, the adequate level of care for as many patients as possible, and the number of transfers required during treatment. The authors in [24] also apply an MIP approach to assign patients to beds. Elective patients request a time window in which they require treatment, whereas for emergency patients this window starts at the current time and is equal to the LoS.

Bed assignment decisions may also be optimized in an offline setting, in which all patients scheduled for admission are assigned to beds in an optimal fashion. To solve an operational offline patient scheduling and bed assignment problem for a weekday ward, the authors of [41] use an MIP model to optimize a schedule over all medically preferred patient access times. The authors of [101] present a heuristic based on an MIP model to satisfy as many bed-assignment constraints as possible in an offline

[2] https://people.cs.kuleuven.be/ wim.vancroonenburg/pas/
[3] http://satt.diegm.uniud.it/index.php?page=pasu

optimization model, while taking into account that some patients may require care from multiple medical specialties.

Heuristics

The patient admission scheduling problem including all constraints on bed-assignments and patient access times has been proven by [241] to be NP-hard. Therefore, recent literature is more frequently applies heuristics to improve the admission schedule. The authors in [130], for example, apply a "great deluge" algorithm to optimize admission dates and bed assignments in an offline setting. These authors compare several heuristics and conclude that their great deluge algorithm can compete with more familiar heuristics such as simulated annealing.

Queueing theory

The authors of [98] investigate an operational admission control for an ICU using an Erlang loss queue with both elective and emergency arrivals. In analyzing the bed census they show that both the arrival streams and service rates can be combined into a single queue with multiple servers (e.g. an $M/G/c/c$ queue). The authors analyze the system by controlling the elective patient admission dates based on the bed census using Euler's method to analyze the loss queue with time-dependent arrival rates. Using historical data, they show that is possible to estimate the most probable level of bed-occupancy several days in advance, given the bed occupancy on the current day. In addition, the model is able to predict the expected split between emergency and elective patients over the coming days. Based on the expected bed occupancy in the near future, staffing levels can be adjusted.

Markov decision theory

Optimizing patient scheduling decisions using a Markov decision approach typically results in complicated scheduling policies that are difficult to implement in practice. For example, the authors of [19] model patients with a stochastic LoS for multiple hospital resources (e.g. beds and operating rooms) such that emergency patients can always be admitted and elective patients are delayed or deferred. Even the approximate dynamic programming model was not solvable within the set time limits for realistically-sized instances, and the authors evaluate some heuristics based upon the results for small instances.

A Markov decision process approach is also used by [256] to decide which surgeries have to be rescheduled so that the ICU capacity is not exceeded. The authors base a heuristic solution approach on the obtained optimal policy and apply it to data on cardiothoracic ICU surgery requests. It appears that the heuristic policy significantly outperforms the current admission policy.

Markov decision models are also used to determine an optimal bed assignment policy in an online setting. For example, in [220] a hospital is considered in which patients should be admitted to a bed in their medically preferred ward or one of the predetermined alternative wards. The authors in [64] approach a similar problem, and

use approximate dynamic programming to optimize assignment of patients to their medically preferred ward or to the "second-best" ward.

Transferring patients during their stay may optimize the bed assignments and shorten the time between admission and bed assignment. These transfer decisions are often optimized together with the assignment of newly admitted patients. Transferring patients during their stay could be optimal from a bed census perspective. Other factors (e.g quality of care, patient condition and staff workload) should also be considered when implementing such decision rules.

Simulation

Simulation models are used to investigate a number of bed assignment policies for specific hospital case-studies. For example, in [42] policies are studied that reserve beds for patients who are about to be brought to the OT. The authors of [144] evaluate policies for reserving beds for patients admitted to the hospital through the ED.

In Section 3.7.1 we describe a simulation model to investigate how many beds should be reserved for high-care patients, which implies that patients with lower care-requirements should be admitted to a ward that is not their medically-preferred one.

3.6.5 Nurse-to-patient ratio

The physical beds in a ward are often not the limiting factor for the number of patients that can be accommodated. Instead, the number and type of nurses and the specific patients present in the ward determine whether there is capacity for new admissions. The nurse-to-patient ratio indicates how many 'average' patients one nurse can take care of; if there are five patients with high care demands a ward can be full, whereas a ward with fifteen patients with low care demands may still have available capacity. An acceptable workload is important for the well-being of nurses and the quality of care.

Integer programming

Linear programming approaches are the primary methods used in the literature, to balance the workload fairly among nurses. For example, in [2] patient acuity scores and travel distances for the nurses are considered in optimizing nurse-patient assignments. The authors of [41] also consider the continuity of care, education and patient or nurse preferences in the optimization, while [189] applies a goal programming to optimize nurse-patient assignments, extending an MIP approach from [213]. In the model described in [213], patients have a nurse-dependent acuity, motivated by differences in experience,or training or by the preferences of the nurses.

3.6.6 Length of stay and readmission forecast

The hospital LoS is typically not known with precision before the patient is admitted, and sometimes it is not known exactly even the day before the patient may be discharged. Moreover, when a patient is discharged, there is always a possibility that the patient may be readmitted for further treatment. Knowing the time at which a patient

is medically ready to be discharged and the patient's readmission probability are useful in the patient scheduling process, as these determine how many new patients may be admitted. In recent literature, we see queueing theory, simulation, machine learning and regression approaches to this problem, of which we provide examples below.

Heuristics

An example of a machine learning approach (random forest model) is used to forecast the LoS of obstetric patients, using information from a patient's medical history drawn from electronic medical records ([84]). The authors of [201] forecast the readmission probability for a cardiac ICU, comparing approaches to this problem that use a support vector machine, decision trees, and a logistic regression. The results of these studies may be implemented in a decision support tool, and provide guidelines to practitioners on which clinical measurements indicate a relatively high risk of prolonged LoS or an increased readmission probability.

Queueing theory

Queueing theory is used, for example, to investigate the effects of different discharge policies at an ICU ([161]). These authors investigate the practical implications of the best policies using a simulation. When a patient needs to be admitted to the ICU at a time when all beds are occupied, typically the "most healthy" patient is discharged to a ward with a lower level of care; optimizing such decisions may improve the quality of care significantly.

For a general ward, the authors of [53] develop an infinite server queue in which a server may only be released after an inspection, which mimics the final doctor visit before a patient may be discharged. The results of the queueing analysis indicate that inspections should be at equal time intervals and additional inspections have decreasing marginal rewards.

Simulation

An example of a discrete event simulation model is given by [62], which analyzes the effects of different discharge strategies on the readmission rate and ED crowding for a complete hospital. The authors conclude that a more "aggressive" discharge policy that discharges patients as early as possible increases the readmission rate significantly.

3.6.7 Conclusion

We have provided a broad overview of planning problems for which OR analyses can benefit patients and staff in hospital wards. Obviously, each situation may require a different modeling approach, but as we have demonstrated above, many models are applicable for analyzing capacity decisions of these types. Queueing models can generate estimates quickly and are often applied to develop a first indicator of, for example, the required capacity. Typically, Markov chain models are less 'broadly applicable' than queueing models, as they are more difficult to re-apply to other wards, but they are easier than queueing models to model transient behavior and ward-specific patient

admissions, discharges, and transfers. Similar to Markov chain analyses, MIP approaches are relatively case-study specific. Although, MIP approaches may be used to optimized processes and schedules, a frequent complicating factor is often the stochasticity in healthcare processes. Machine learning and regression approaches are useful for analyzing large amounts of hospital data and are increasingly used to assist medical decision-making in wards.

3.7 OR for Wards Illustrated Cases

In this section we, together with colleagues from CHOIR, present three illustrated cases that are conducted at partnering hospitals. In all projects, the hospital has implemented the results of the research. In these illustrative cases, we focus on the practical approach that was taken to implemented results and insights from OR models and reflect on the implementation. Each case study gives unique insights into success factors, pitfalls, and lessons learned. Although, these illustrations does not have great methodological contributions to literature, we hope to challenge other OR researchers to discuss implementations in their research. By doing this, we hope other OR researchers use these insights for their implementations and so improve the impact of OR in hospitals. In the following chapters, we will present more insights in our OR modeling approaches.

3.7.1 Case study I: Balancing bed census[4]

Both over- and under-capacity are a problem for wards. In a ward with under-capacity, patients cannot always be accommodated in the medically preferred ward. As a consequence, patients needing treatment in one medical specialty are placed in many different wards and doctors spend much time visiting their patients. Having over-capacity is a problem for hospital staff as many patients from other medical specialties are likely to be placed in the ward. As a consequence, nurses from the ward have to care for patients for with conditions for which they were not fully trained, and may experience a high workload if they do not feel qualified to treat patients from other medical specialties. In both scenarios, patients do not always receive the best possible care, which increases the willingness of all stakeholders to solve this problem.

Two medical wards of the Jeroen Bosch Hospital (JBH) located in Den Bosch, The Netherlands, experienced unbalanced bed occupancies during 2012 and the first months of 2013. In the neurology department, patients' LoS had been reduced significantly, resulting in over-capacity. At the same time, the department of internal medicine experienced increasing numbers of patients, resulting in crowded wards and many patients being deferred to other wards (e.g. under-capacity).

Project organization

In accordance with the list of factors in [226], at the start of this project we commissioned a steering group consisting of all stakeholders in this problem: a neurologist, an

[4]This case study was conducted, among others, by CHOIR-colleague N.M.(Maartje) van de Vrugt.

internal medicine specialist, an administrator from the patient admission scheduling office, and all involved ward managers. The hospital management made this steering group responsible for finding a solution for the over- and under-capacity of the wards, and included one organizational advisor and a healthcare logistics advisor in the steering group. A representative of the highest management level below the JBZ board of directors was made chair of the steering group. The neurologist and internist were selected based upon the trust and goodwill they had from their peers. These representatives were not necessarily the heads of departments; since it was required that these be doctors who spend time on the wards and experience the problems on a daily basis.

The first meeting of the steering group started with all members of the group getting to know one another; although all stakeholders work on closely related topics, typically they do not often meet or talk to each other. The group discussed the extent to which they experience a problem in the ward or during patient scheduling. Supporting this discussion, the logistical advisor presented the results of a data-analysis with information on (1) the bed occupancy of all hospital wards, (2) the bed requirement per medical specialty, and (3) the number of patients per medical specialty who were not treated in their medically preferred ward. The data that was used for this analysis was routinely collected hospital data on admittance and discharge date, medical specialty and ward. The data analysis objectified the discussion significantly. For example, the neurology ward nurses experienced a high workload, and the hospital data confirmed that the nurses had to take care of a relatively lager number of patients from other medical specialties, which increased the experienced workload.

Analysis of possible interventions

The result of the first session was that the steering group wanted to investigate two possible interventions:

1. Opening an AMU.
2. Reassigning medical specialties to wards.

Using an $M/G/s/s$ queue, the required bed capacity to achieve at most a 5% blocking probability was determined for each specialty. This analysis confirmed the belief of the steering group that the distribution of beds among the specialties was not adequate but it was not necessary to add overall capacity to the system. For intervention 2, each of the possible scenarios required serious rebuilding of units or splitting of medical specialties among multiple wards. Re-building several wards would have been costly and would have taken several months. Therefore, the steering group decided to discard this intervention option.

The effects of intervention 1 were analyzed for several scenarios using an $M(t)/M(t)/s/s$ queue [226]. The conclusion of this analysis was that the AMU would not be beneficial for the hospital's case-mix, and the steering group discarded this intervention as well.

At this point in the project, the steering group was looking for new interventions, and decided to investigate the possibility of creating an overflow ward for internal medicine within the neurology ward. In the analysis of intervention 1, each doctor

had to determine a list of diagnoses for their specialty that were eligible to be treated at an acute medical unit. This list consisted of diagnoses that required a relatively low care-level, and each acute patient with a diagnoses from the list would be admitted to the AMU. With minor moderations made by the internist, the list was adequate for identifying the eligible overflow patients.

Since the admission data had been anonymized, an exact analysis of the overflow ward was not possible. Financial hospital data revealed what fraction of all internal medicine patients had diagnoses from the list, and this was used to estimate the total overflow bed requirements. This number of beds was sufficient to alleviate the pressure on the internal medicine ward, and was low enough to be accommodated in the neurology ward.

Choosing an intervention

Based on this promising result, the organizational advisor helped doctors and nurses from internal medicine and neurology to investigate what would be required to implement the intervention, for example in terms of skills, education, and doctors' rounds at the wards. The most important decisions at this level were how often the internists would visit the overflow patients, which medical decisions were the neurologists would be allowed to make, and when an internist should be called for assistance. Based on these discussions, nurses and doctors were confident that the quality of the care provided to the overflow patients would be good.

Additionally, the logistical advisor conducted a simulation study in which historical data was used to determine the best policy to start and stop the overflow of patients. In this simulation, each patient was randomly eligible for the overflow ward. The steering group requested this additional research, as the neurology ward manager feared that, due to the overflow patients, not enough beds would be available for neurology patients. Several overflow policies and their effects were presented to the steering group.

Based on all gathered results the steering group decided to implement the overflow ward using the following policy: patients would be sent to the overflow ward only if both (1) three or fewer beds are available at the internal medicine ward, and (2) two or more beds were available at the neurology ward. In September 2013, the intervention was implemented.

After intervention

In January 2014, data analysis and interviews with the staff showed that the intervention had the desired effect: the neurology ward accommodated more internal medicine patients (on average 2.5 beds) and fewer patients from other specialties. Both effects were statistically significant. Additionally, the internists reported a reduction of the time required for their rounds, and the neurology nurses experienced a reduction in the fluctuations in the workload and were confident of their ability to deliver a high quality of care. An apparent downside of the implementation was a higher workload at the internal medicine ward, as many of the "easier" patients were admitted to the neurology ward.

Lessons learned

The success of this case study was a result of involving clinical leaders, proposing interventions, and objectifying the effects of those interventions prospectively using data. Based on these analyses, the steering group was able to choose the most promising intervention to implement. The higher management let the steering group choose what interventions to investigate, but had set a clear target to find a solution for the problem. This autonomy was greatly appreciated by the steering group.

Another important part of the project was checking all assumptions and data-analyses with nurses and doctors working on the wards. The goal of many data validation discussions was to come to agreement that the data indeed reflected what happened in reality on the wards. Lengthy discussion about data or assumptions during steering group meetings would be undesirable, as these would lead the group away from finding a solution.

During the project there was an emphasis on finding a solution that all steering group members and involved staff would consider a clear improvement over the current situation. To this end, in addition to data-analysis, a thorough risk analysis was done for every intervention the steering group suggested. It was important not to ignore any of the concerns of the steering group, as this would decrease members' willingness to cooperate in implementing the intervention. For each of the concerns raised, data analysis was performed, if possible, and the steering group took time to discuss all concerns thoroughly, until either the issue was alleviated or the corresponding intervention was discarded. One example is the neurologists' concern that the internal medicine patients would displace neurology patients. This concern was alleviated by a simulation analysis with multiple scenarios, which eventually lead to a decision rule for the patient admission planners.

Before the project started, higher management had emphasized with the steering group members that the project had been initiated to improve both quality of care and employee satisfaction. When discussions within the steering group focused on competing interests of members, the chair of the meeting reminded everyone to stay focused on the quality of care and employee satisfaction. In all discussions this reminder sufficed to lead memebers to find a common goal and, eventually, a solution to the given problem. The autonomy of the steering group and the iterative process of testing possible interventions resulted in an intervention supported by all involved staff. This support was the primary key to the success of the intervention. Moreover, the intervention proved to be effective in reality, which was the ultimate goal of the project.

3.7.2 Case study II: Dimensioning wards

The LUMC dealt with multiple logistical problems in its wards. These problems related to small wards in terms of the number of beds and a medically illogical distribution of patient groups among wards. These problems resulted in rising numbers of refusals at the ED and growing waiting lists. Small wards have more difficulties coping with variability such as the number of arrivals and LoS as it has relatively more impact. Therefore small wards will often have over- or under-capacity.

Project organization

In 2014 the hospital board of directors decided to redistribute patient groups among wards and re-dimension wards. A project was initiated with a steering group consisting of a management director (project lead), a project manager, all care managers, an organizational consultant, a change management consultant, a human resource consultant and a healthcare logistics consultant. Multiple topics for further analysis were defined, and one working group for each topic was formed. To overcome the earlier mentioned logistical problems, the project had to implement the following interventions:

- Redistribute patient groups among wards for long stay patients (e.g. a LoS of at least 5 days).
- Introduce a ward for short stay patients (e.g. LOS of less than 5 days).
- Introduce a ward for day treatments (e.g. LoS of at most 8 hours).
- Introduce a ward for acute admissions (e.g. an AMU).
- Merge ward staff and management that have medical affinity so that beds are interchangeable at these wards (e.g. orthopedic surgery with traumatology or nephrology with endocrinology wards).
- Introduce a new management consisting of a physician and nurse manager.

The hospital management decided that the total capacity should not be increased, so all interventions should be achieved without increase in the number of beds and nurses.

Analysis of possible interventions

As boarders are a risk to the quality of care, the hospital wanted to minimize the probability of refusals in the medically preferred ward. As presented in Section 3.6, queueing models dominate strategic and tactical analyses. We therefore chose to model multiple scenarios of assignments of patient groups among wards as an $M/G/s/s$ queue. Based on [69], we analyzed each scenario (e.g. the patient load from selected medical specialties in a ward or merged wards) on two performance measures: (1)the blocking probability given an occupancy rate of 85% and (2) the occupancy rate given a blocking probability of 5%. Based on these measures we redistributed the patient groups. Furthermore, we developed a simulation model to analyze the flow of acute admissions via the AMU. Patients stay at the AMU for at most 2 days. If the need further treatment, they are transferred to the inpatient ward of their medical specialty. From this analysis it appeared that solely introducing an AMU would not solve the problem of emergency admission refusals. We performed an analysis to determine the number of allocated beds in each ward to minimize the number of refusals. We also showed the effect of bed shortage in wards (e.g. finally resulting in an overcrowded AMU and ED). To prevent flow congestion, we analyzed scenarios with different numbers of beds dedicated for these transfers in each inpatient ward. We used two heuristics to determine the best number of beds for each ward or for merged wards. This simulation study was executed by a healthcare logistics consultant and ward management (nurse manager and medical manager of the AMU).

Choosing an intervention

The actual re-dimensioning of wards and re-distribution of medical specialties to these wards was based entirely on the queuing model results. Significant effort was required to convince the stakeholders of the reliability of the model outcomes. We used pseudonymized admission data from 2013 and 2014 as input for the model and invested several weeks in discussing the assumptions of the model, results, and data. This is an important step when using queueing models in practice. Although queueing models are based on straightforward formulas, it can be challenging for stakeholders to interpret them. To ensure acceptance, the model should be thoroughly discussed and not used as a "black box". We planned sessions with each medical specialty to discuss the data, showing first the current patient load for each ward. We showed individual patients records to medical staff in terms of admission and discharge dates. Next, we discussed the model assumptions the input used, and KPIs. Lastly, we showed the proposed redistribution of patient groups and the effects on bed usage. Taking time to present the model and answering questions from all stakeholders in both plenary and individual settings, we finally convinced most stakeholders of the proposed benefits of the redistribution and re-dimensioning.

After intervention

After the interventions had been completed, the hospital still struggled to create a schedule for patients for the weekday ward so that this ward could be closed on the weekends. Furthermore, at a later point in time some medical specialties were again redistributed among wards based on new insights into organizing wards according to a new hospital strategy that focused on more thematic care (e.g. oncology care and transplantation care). Again, a queueing model was used for this new redistribution of medical specialties.

Lessons learned

A pitfall in this type of analysis is the requests for more up-to-date data. Given the size of the project, we needed 6 months to discuss the analyses with all stakeholders. During 6 months, many things can change, and therefore some stakeholders requested a new analysis with up-to-date input data such that the model would be more reliable (e.g. real time hype). These requests delayed the project by a year. In the end, we did not update the data and reasoned with stakeholders that we used highly aggregated data over a long time horizon, which means that mainly trends and strategic decisions would be detectable in the results. We also showed the added value of merging wards given the performance measures. In practice merging wards in university hospitals has major implications for nursing staff. As nursing staff are highly trained for specific treatments and specialties, merging wards requires that they be trained in other fields of medical expertise as well, as more patient types can be placed in merged wards.

Using a simulation model for the introduction of the AMU, we were able to show stakeholders how the emergency admissions process would evolved over time ([204]). This visual representation and the implementation of tailored process characteristics significantly contributed to convince stakeholders. Therefore, achieving consensus was

easier compared to the scenario in which queueing models had been used for the re-dimensioning and re-distribution of wards and was also reflected in the time needed to convince all stakeholders (2 months). The simulation model was used as a tactical tool in the planning and control cycle; in the model we updated the distribution of dedicated beds each quarter using data with a rolling horizon (i.e. adding the last quarter and deleting the first quarter of the data).

The success of this project was a result of a clear sense throughout the organization of the urgency of becoming future-proof. Care providers dealt with the consequences of the suboptimal distribution of beds and small units on a daily basis. The determination of the board of directors and the persuasiveness of the management convinced all stakeholders to let go of strict bed allocation policies, resulting in larger units. Additionally, the use of OR models gave the management a safe environment in which to experiment with new bed distribution and their KPIs.

3.7.3 Case study III: Bed assignment optimization [5]

Massachusetts General Hospital (MGH, USA) deals with operational bed occupancies between 95% and 100%. As a consequence, patients generally have a long wait upon admission or transfer before an inpatient bed is available. This results in flow congestion at the post-anesthesia care unit (PACU) and the ED. Particularly for emergency patients long wait times increase risks. The state of Massachusetts therefore has a "Code Help" policy, requiring hospitals to move all admitted inpatients out of the ED within a 30-minute period after the ED's maximum occupancy – adjusted for the number of patients present and their acuity – is reached or exceeded. Activating Code Help causes the hospital to prioritize moving patients out of the ED, which results in delaying bed assignments for patients from other areas of the hospital, potentially requiring cancellation of elective surgeries and other activities. The consequences of Code Help require significant management attention and can affect hospital operations for several days. In 2015, notifications that the hospital was approaching or had reached Code Help frequently occurred multiple times per week.

Project organization

The continuous lock down gave rise to a hospital-wide redesign of admission scheduling. Under supervision of the CEO, a project was initiated with a team that consisted of the head of the perioperative department, the head of and a bed manager from the admitting department, the nurse managers and resource nurses from several clinical units, a professor, a postdoctoral fellow, a graduate student in healthcare OR, and two advisors from the department of process improvement at MGH.

Analysis of possible interventions

Before the intervention, elective surgical same-day admits were pre-assigned to beds that were occupied with patients who were to be discharged on that day. This was done to guarantee a continual patient flow from the PACU to inpatient units; by reserving

[5] This case study was conducted by, among others, our CHOIR-colleague Aleida Braaksma ([42]).

a bed for a surgical patient, the bed could not be taken by a patient from the ED or another department. However, the exact timing of discharges was unknown and uncertain. As a consequence, patients were frequently waiting for their pre-assigned beds while simultaneously other beds were waiting for their pre-assigned patients. Data analysis showed that the average time patients waited for a bed ranged between 2.5 and 26.9 hours, while a subset of four surgical inpatient units (129 beds) experienced a total bed-waiting-for-patient time of 11,2 hours, or 466 bed-days, in 2015.

To alleviate this problem, a simulation model was developed to investigate the effects of multiple interventions, among which was a just-in-time (JIT) bed assignment strategy. This strategy assigns patients to empty beds just before the patients are medically ready, and therefore beds cannot be pre-assigned to patients. A second intervention investigated was virtually pooling the capacity of two surgical wards, as they were clinically similar. This intervention requires that patients not be pre-assigned to wards, but that the ward be selected the moment a patient is assigned to a bed.

The input for the simulation model was 1 year of hospital data, including timestamps for admission and discharge. From the data, empirical distributions were determined for, for example, bed cleaning duration and patient transportation time. The model was made more realistic by implementing bed closures based on the hospital data (for example due to staffing shortages) and the hospital's policy with respect to gender and infection precautions in semiprivate rooms. Additionally, the model improved patient cohorting by occasionally swapping a patient from one room to another, mimicking the policy that was used in practice.

After intervention

Based on the simulation results and two earlier projects by graduate students in healthcare OR, the hospital implemented the JIT bes assignment policy and pooling policy in 12 surgical inpatient units. In the 5 months post-implementation, the average patient wait time for bed decreased by 18.1% for ED-to-floor transfers ($P < 0.001$), by 30.5% for PACU-to-floor transfers ($P < 0.001$), and by 10.0% for ICU-to-floor transfers ($P < 0.05$). As a consequence, patients receive their required care earlier, which improved the quality of care. Additionally, the intervention resulted in a smoothed workload for nurses and bed cleaners, and less congestion in the ED and PACU. Another positive side-effect was an increased focus on patient flow: because the JIT bed assignment policy, nurses wonder why a bed is empty for a long time, which may speed up, for example, hand-offs and transportation.

Lessons learned

For physicians and nurses, simulation is relatively easy to understand compared to mathematical modeling. Therefore the project team was convinced the intervention would have a positive effect in practice. The determination of involved clinical leadership (e.g. the head of the perioperative department) was also key to success. In the first days of implementation, nurses were sometimes skeptical about the intervention. The project team leaders showed empathy for the struggles related to the new situation while simultaneously encouraging nurses to stay put. Daily short evaluation

meetings were the opportunities to quickly react to unforeseen side-effects of or negative sentiments about the intervention, before these could evolve into larger problems. The new policies resulted in more stable admission and discharge rates throughout the day. These more stable rates were also noticed by the bed cleaning department, as their peaks in workload were reduced. This was considered a positive but unforeseen result of the project.

3.8 Impact in Practice

In this section we provide our views on how OR researchers can increase the likelihood of results being implemented in practice based on the literature and the case studies in Section 3.7. There exist many studies in which OR models have been used to analyze wards. However, actual use of these models in practice seems scare; only a few of the articles reviewed for this chapter reflect on actual implementation results or the use of the models in practice. A widely used quote is: "the final test of a theory is its capacity to solve the problems which originated it" [65]. In this section we report on the problems faced while implementing research results, and the lessons learned from the implemented research included in this chapter.

The most important contributions to a successful implementation, is the involvement of clinical leadership, who are important medical stakeholders in the process and have the respect of their colleagues. These leader should be able to speak on behalf of his/her colleagues, and should discuss the project often with peers. Researchers should earn the trust of these clinical leaders so that they are convinced of the soundness of the the methodological approach and proposed interventions. Ultimately these clinical leaders can (and should) convince other colleagues.

Model input determines, to a large extent, the outcome and the acceptance of the results. On several occasions the already available hospital data appeared to be insufficient to provide all necessary input for the models, or the database was incomplete [141, 145]. Hospital data is often inconsistent or partly missing across different databases; financial data does not always match raw admission and discharge data. Depending on the goals of the research, different databases may be used. Even in times of increasing use of technology, we cannot trust the data to reflect reality completely. The entry of admission and discharge data, for example, is in many hospitals still a manual task, often performed when nurses have relatively low workload or at the end of a shift. Additionally, it is important to realize that all hospital data is the *realized* process and most hospitals do not register deferred or denied patients, so actual patient demand is often difficult to obtain. Knowing the ins and outs of the healthcare process is also essential in reading the data; for example, for an ICU, the LoS is affected by the bed occupancy since intensivists often transfer the healthiest patient to make a bed available for a new patient when all beds are occupied. A careful sensitivity analysis should be performed to ensure that the best possible scenario for implementation is included in the analysis.

In any mathematical model, assumptions are necessary for tractability. Some assumptions may be too unrealistic to be of practical relevance. Therefore, in all our projects we start with one or several observation rounds, in which we study the pro-

cess in reality, become familiar with the practitioners and their decisions, to determine which assumptions it is important to maintain, and learn what flexibility and stochasticity are present. Letting practitioners draw a typical patient process is often not sufficiently accurate to provide all modeling assumptions. Moreover, seeing the outcomes of the process in the data does not mean that the preferred medical process was followed. The time that is invested in making the assumptions and relations in the model more realistic will significantly reduce the time spent on data-analysis. Additionally, making the model more realistic will increase the likelihood it will be adopted in practice.

To further increase the likelihood of implementation, a researcher should be able to convey to healthcare practitioners how a model works, and thereby earning the trust of practitioners. Expectation management is very important; practitioners should know what the model can and cannot do. Often, when the mathematics behind the modeling approach becomes less complex, practitioners find the results easier to grasp and trust. A major advantage of simulation models is their ability provide a visualization of the analyzed process. Visualization leads to an understanding of the contributions, understanding leads to commitment by decision makers (e.g. clinical leadership) and commitment leads to implementation. After the project is completed for the practitioners, a researcher can continue thoroughly investigating the problem or extending the model to make the approach interesting enough to publish in an OR journal. Alternatively, or perhaps simultaneously, publishing together with the practitioners in medical journals may be considered.

For adoption, an iterative process will be more effective; first mimic the current process and let practitioners check it (and repeat this if necessary), and second iteratively investigate scenarios and discuss them with practitioners. The most promising predictor of implementation is evidence that the stakeholders are actively participating in the iterative process by proposing the interventions to be investigated. Additionally, presenting the results in insightful graphics will increase their impact; checking the results with the involved clinical leader before presenting them to all practitioners allows the researcher to adapt the presentation to the audience. One risk of this iterative process is that the project never ends as more and more scenarios are investigated. This risk can be avoided by setting clear performance targets early in the project, and by keeping to strict project schedule.

Discussing possible interventions can be challenging because desired outcomes are often based on extreme incidents. Exploring interventions mathematically and thus rationally often simplifies the discussions significantly. Conveying the chosen intervention to colleagues becomes easier for the clinical leaders, as the decision was based on rational arguments.

In summary, the stakeholders play a significant role in increasing the likelihood of implementation. Additionally, researchers should be thorough in their data collection, sensitivity and robustness analyses, and implementation support. Additional information on project life cycles for general healthcare applications is found in [108].

Part II

Integral Capacity Planning in Hospitals

CHAPTER 4

Allocating Emergency Beds Improves the Emergency Admission Flow[1]

4.1 Introduction

To avoid overcrowded emergency departments (EDs), congestion and fluctuations in downstream resources, sophisticated (process) analysis is required [114, 146]. Demographic changes and improved patient survival rates have contributed to the increasing number of hospitalizations of complex and(or) chronic patients [10]. In addition, society demands cost-effective healthcare delivery, which puts pressure on available resources. This results in bed occupancy rates above 85 % in inpatient wards [47], leaving marginal slack for admission flow fluctuations and resulting in refused patients [15, 260]. A major side effect of a decreasing number of available beds is the increasing number of so-called boarders. Boarders are emergency patients waiting for admission or emergency patients placed outside of their designated specialty ward due to bed unavailability [172]. In general, boarders have a significantly longer length of stay (LOS), experience a decreased quality of care, are less satisfied, have increased mortality rates and are associated with patient safety issues [27, 59, 172, 260].

To improve the emergency admission flow, some hospitals use acute medical units (AMUs) [151]. "An AMU is a designated hospital ward specifically staffed and equipped to receive medical inpatients presenting with acute medical illness from EDs and outpatient clinics for expedited multidisciplinary and medical specialist assessment, care and treatment for up to a designated period (typically between 24 and 72 hours) prior to discharge or transfer to medical wards" [206]. From an Operations Management perspective an AMU operates as a buffer. A buffer can operate in two different configurations: (1) as an inflow buffer and (2) as an outflow buffer. An inflow buffer transforms a (highly) variable inflow into manageable outflow by accommodating all arrivals in the buffer before they are transferred further downstream. With the second configuration, the buffer is used only if a downstream inpatient ward is fully occupied.

[1]This chapter is based on A.J. Schneider, P.L. Besselink, M.E. Zonderland, R.J. Boucherie, W.B. van den Hout, J. Kievit, P. Bilars, A.J. Fogteloo and T.J Rabelink. Allocating Emergency Beds Improves the Emergency Admission Flow. *INFORMS Journal on Applied Analytics*, 48:4:384-394, 2018.

In this study we analyze an AMU operating as an inflow buffer, where the timing of transfers can be managed (between 24 and 72 hours) so that inpatient wards have time to make capacity available. As a result, downstream inpatient wards can attain higher bed utilization without increasing the number of refused patients. An AMU initially reduces pressure on the ED utilization. However, in the case of structural lack of coordination between AMU and downstream hospital wards, the AMU cannot transfer patients to the downstream hospital wards [1], again resulting in an overcrowded AMU and ED [206]. Ultimately, this increases the number of the aforementioned boarders, and will contribute to the downward spiral of more emergency admission refusals.

As discussed in Section 3.5, bed capacity management focuses on efficiently allocating beds (and thus staff) between patient types (i.e., emergency versus elective patients between patients from different specialties) and OR can offer useful managerial insights into trade-offs for capacity management, such as the relations between the probability of refusals, bed occupancy and throughput. Using OR, possible interventions can be safely evaluated and so reduce the risk of implementing an intervention that may turn out to be counter-productive.

The remainder of this chapter is organized as follows. Section 4.2 explains the objectives of this study. Section 4.3 presents the process of the emergency admission flow in more detail. In Section 4.4, we discuss our modeling approach and key performance indicators. The data analysis required for the input of the model is discussed in the section 4.5, followed by a description of the model in Section 4.6. The results are presented in Section 4.7 section and we discuss the implementation of our results in practice in Section 4.8. Finally, further managerial implications, limitations, and potential extensions of our study are discussed in the Section 4.9.

4.2 Objectives

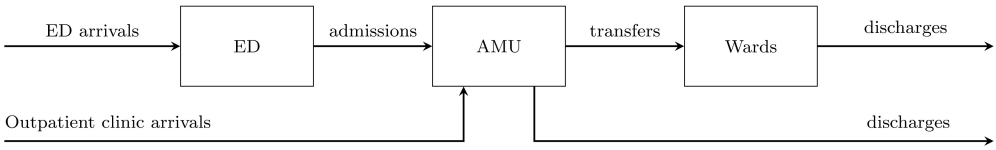
The partnering hospital, Leiden University Medical Center, introduced an AMU in 2014. However, management of both the AMU and inpatient wards were still spending significant time to transfer patients from the AMU to downstream inpatient wards, due to a lack of beds and organizational guidelines and/or protocols. To simplify the transfer process, the concept of allocated emergency beds was introduced, meaning that each inpatient ward allocates a part of its bed capacity to accommodate patient transfers from the AMU. The first objective of this study is thus to evaluate the effect of allocating inpatient bed capacity for patients of the emergency admission flow. At the same time, the board of directors of the partnering hospital also decided to restructure its inpatient wards into care units, based on liaison specialties (i.e., pooling specialties which cooperate with each other such as nephrology and endocrinology). Therefore, we formulated a second objective for this study: evaluate the effect of pooling wards on the required number of allocated emergency beds. The concept of pooling resources is extensively studied [50, 162] and more specific in a hospital setting by [69] and [238]. Ultimately we want to structurally improve the emergency admission flow by implementing the model into the partnering hospital's planning and control cycle.

4.3 Process Description

Figure 4.1 shows the basic patient flow we are analyzing. In this process there are three types of hospital beds: (1) beds at the ED, (2) beds at the AMU and (3) beds at the various downstream inpatient wards. There are two arrival streams: (1) from outside the hospital to the ED and (2) from the hospital outpatient clinic circumventing the ED at the AMU. After a stay at the AMU, patients can either leave the hospital (discharged) or transferred to an inpatient ward. After a stay at the inpatient ward, patients leave the hospital.

Here, we describe this basic process in more detail, including its logistical charac-

Figure 4.1: A basic Flow Chart of the Emergency Admission Flow.



teristics and decision moments. Patients who need immediate medical care will visit the ED. Therefore ED arrivals arrive from outside the hospital at unpredictable rates. Arriving patients are seen and provisionally assessed, identifying those in need of a bed (while those who do not need a bed are treated and leave the hospital). If a bed at the ED is available, the patient is assigned a bed and a priority status is determined by triage. Based on this priority status the patient is seen by an emergency-room physician either immediately or at a later time. The attending physician determines an initial diagnosis, and the ensuing main medical specialty that will be responsible for the patient. If no bed is available at the ED, the patient will be refused hospital admission and referred to a nearby hospital. ED arrivals who need an overnight stay are admitted to the AMU and later, if needed, transferred to an inpatient ward. In this study we focus on admitted emergency patients; so we only address admitted ED arrivals.

After a random LOS at the ED, patients are admitted to the AMU once a bed is available. Otherwise, patients wait in a bed at the ED. Admissions at the AMU can also originate from the hospital's outpatient clinics, thus circumventing the ED, when a physician at the outpatient clinic indicates that immediate hospitalization is required. Also, patients from the outpatient clinics are refused and referred to nearby hospitals when the AMU is fully occupied. At the AMU patients will be observed, further diagnosed and, if necessary, given treatment. The LOS at the AMU is limited to 48 hours. This limit is a management decision of the partnering hospital and is applied with some flexibility. For instance, a patient who has already spent 48 hours at the AMU and is expected to be discharged within 24 hours, will remain at the AMU. A transfer is time consuming for staff and stressful for patients, and therefore these patients stay at the AMU. Additionally, patients will not be transferred between 9 pm and 9 am, since staffing levels at the downstream inpatient wards are minimal during these hours. If further treatment is necessary, patients are transferred

to other inpatient wards depending on their treatment specialty and stay there for a certain random LOS, also depending on the treatment specialty, after which they are discharged. Patients can only transfer to their destination inpatient ward if an allocated emergency bed is available at that ward. We assume beds which are allocated for emergency admissions cannot be used by elective patients and vice versa.

The ED and AMU solely dedicate beds for emergency patients, while wards also deal with elective patients. Given the objectives of our study, one could be tempted to focus on the effect of the number of allocated emergency beds at inpatient wards and model only the wards. This does not capture the patient flow through the AMU. In addition, for explanation to the stakeholders, we want to show the effects of the allocated emergency beds on all departments involved. We therefore include the departments types: ED, AMU and wards.

4.4 Methods

In this section we explain our modeling approach and key performance indicators (KPIs) for analyzing scenarios.

4.4.1 Model Approach

We use discrete event simulation (DES) to analyze the emergency admission flow, since analytical modeling of the non-homogeneous inter arrival times, the different LOS per ward and specialty and the time interval in which patients can be transferred from the AMU is analytically intractable. Further, DES provides a visual representation of the process for implementation purposes. DES is widely used for decision support and planning in healthcare; see for example the online reference database described in [117] and the systematic reviews of [102, 118, 184].

4.4.2 Performance Indicators

We formulate the following KPIs as output for our simulation model: (1) the relative LOS at the AMU, (2) the fraction of refused arrivals and (3) the utilization of the beds allocated at the inpatient wards. The first KPI is an accurate parameter to measure the level of throughput [257]. When patients cannot be transferred to an inpatient ward, the LOS will increase immediately. The relative LOS is defined as the ratio between the average LOS at the AMU divided by the average LOS at the AMU in the case of unlimited capacity. We define the term relative LOS, so we can directly interpret the factor that causes the scenario to improve or worsen and compare it to unlimited capacity (which has only marginal waiting time when patients completed their LOS between 9 pm and 9 am and therefore must wait until 9 am to be transferred). The second KPI is the fraction of refused patients (related to the total number of arrived patients) and is an accurate measure of a full system (i.e., no beds are available). The third KPI is the average utilization of the allocated beds for each inpatient ward and the beds at the AMU and gives information about potential bottlenecks.

4.5 Data

The model requires the following input data:

- arrival rates at the ED and AMU per hour;
- LOS at the ED, AMU and hospital wards based on specialty;
- distribution of number of admissions per specialty;
- transfer rates per specialty from the AMU to the inpatient wards or for patients leaving the system;
- number of allocated emergency beds at the ED, AMU and wards (of course at the ED and AMU all bed capacity is available for emergency admissions).

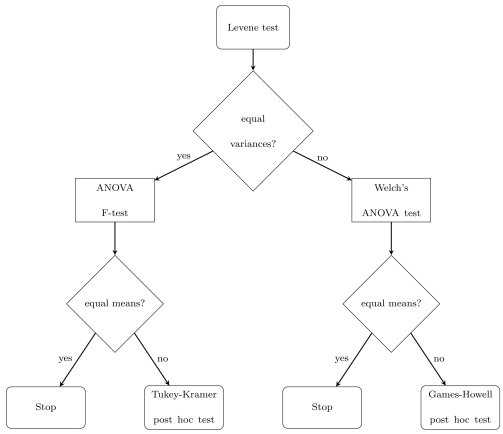
We obtained patient data from the hospital's management information system. The data set consists of 4,446 admissions at the AMU between 2012 and 2014. To overcome the high diversity in specialties and patient flows, only the top 99 percent of admissions were taken into account excluding the remaining 1 percent atypical cases in terms of specialty and/or ward. This resulted in 7 hospital wards (67 percent reduction) and 6 specialties (50 percent reduction), simplifying calculations significantly.

4.5.1 Data Analysis

The data analysis serves the following purposes: (1)finding the distribution of specialties and patient flows (to which ward patients are transferred from the AMU), (2) clustering of patient groups, (3) fitting clustered patient groups to probability distributions for modeling the LOS and (4) determineing the arrival patterns. The distribution of specialties and patients flows are based on the historical data using frequency tables.

Patients from the same specialty can be transferred from the AMU to different wards. This could increase the complexity of our model by means of specialty and ward options. To keep our model as simple as possible in terms of options and increase the statistical power of the samples for fitting the probability distributions, we cluster patient groups from two perspectives: (1) the LOS of patients whose care is within the same medical specialty on different wards and (2) the LOS of patients on the same ward with different specialties. Patients cared for by the same specialty could have a similar LOS because of the similar nature of their disease or injury, and(or) because they are treated by the same staff. Our clustering process is based on the logic in Figure 4.2. First, a Levene test [149] is executed to identify differences between sample variances. If the results of the Levene test show unequal variances between samples, an ANOVA F-test [39] with a Welch test statistic [249] is required; otherwise a normal ANOVA F-test is performed to determine unequal means between samples. If the ANOVA F-test shows significant unequal means, the final step in the clustering process is a post hoc test to analyze which sample(s) is (are) significantly different compared to the samples that share the same mean and variance. The post hoc used test depends on the results of the Levene test. With unequal variances and unequal means between samples a Games-Howell test [83] is required to determine significant differences between samples. For samples with equal variances but unequal means a Tukey range test [222] is performed. All tests are performed with SPSS, an IBM statistical software package. The results of this clustering process showed that two

Figure 4.2: Statistical Clustering Process of Patient Groups.

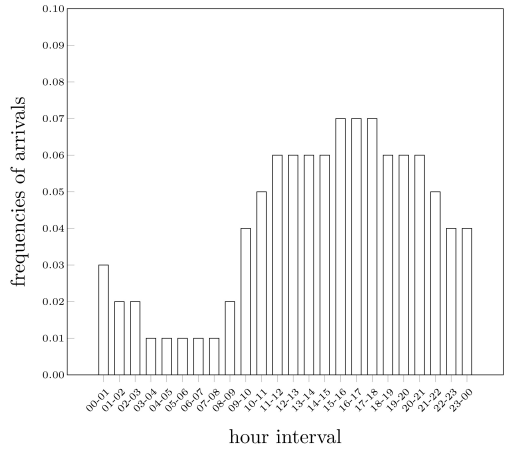


hospital wards have equal LOS independent of the specialties on these wards and one specialty has the same LOS independent of its designated wards. All other wards and specialties have significant different means.

Using the outcomes of the clustering process, we fit probability distributions to each clustered patient group. For this we use Rockwell's ARENA Input Analyzer (version 11) based on goodness-of-fit tests (i.e., the Chi-squared test). The following probability distributions are used for fitting: Gamma, Erlang, Exponential and Lognormal. Due to the limit, the historical data of the LOS at the AMU displays a specific gradient, where the probability mass is centered around 24 hours and 48 hours. Therefore, we used the empirical distribution derived from the historical data to model the LOS at the AMU.

Since emergency patients arrive unscheduled we want to find the arrival patterns. One can see from Figure 4.1 that the process has two arrival streams: (1) arrival at the ED and (2) arrival at the AMU from the hospital's outpatient clinics. Data analysis shows that patients arrive according to a non-homogeneous process. For instance, peak hours are between 3 pm and 8 pm. We therefore determine hourly arrival rates based on the historical data and assume that arrivals occur according to a (non-homogeneous) Poisson process, as is common practice in modeling unscheduled patient arrivals [218]. Seasonality or differences in weekdays and weekends are not taken into account in the arrival rates. As an example, the daily ED arrival frequencies are given in Figure 4.3.

Figure 4.3: Frequencies of Emergency Department Arrivals per Hour.



4.6 Model Implementation

The DES is implemented in Tecnomatix Plant Simulation Sofware (Siemens, version 9). The model has a generic setup, therefore, various configurations (e.g., the number of wards and beds, medical specialties, patients flows, and LOS) can be analyzed without changing the core design of the model.

Arrivals at the ED or AMU are assigned a specialty and a certain destination (e.g., after a stay at the AMU the patient will be discharged or transferred to a hospital ward for further treatment) according to a single random Bernoulli trial, using probabilities derived from the frequency tables mentioned in the *Data Analysis* section. The LOS at each department is based on the department's medical specialties. When new patients arrive at the ED or AMU and all beds are occupied, they are refused and leave the system. When patients are ready to transfer from the ED to the AMU or from the AMU to an inpatient ward and the destination is occupied, they wait at their current department.

4.6.1 Simulation Initialization

To obtain (statistically) reliable results from our simulation we need to initialize our model. We start initializing the simulation with all parameters derived from the data analysis: frequency tables for specialties and destinations, the probability distributions for the LOS and the arrival patterns. We also need to dimension the ED, AMU and wards according to the hospital's practice. The ED and AMU have 8 and 24 allocated beds, respectively and a patient can be transferred to one of seven wards (the number of allocated beds at wards will vary for each scenario).

Since the simulation starts with an empty system (i.e., no patients are present), a warm-up period is required to reach steady state. We therefore exclude all results from the warm-up period. The length of the warm-up period is determined by means

of the Welch method. This method plots moving averages of the means from the i_{th} observation for a number of replications and for an arbitrary long run length per KPI. The mean of multiple replications of the i_{th} observation smooths the variability over individual observations and therefore gives insight into the dependency on the initial state. We then use moving averages over these means to smooth out high-frequency oscillations. The warm-up period is determined through graphical interpretation of the plotted moving averages per KPI resulting in a length of 365 days.

We determine the run length using the convergence method of [199]. This method implies convergence of the cumulative means of KPIs over multiple replications as the run length increases. The convergence level is measured as the ratio of the positive difference between the maximum and the minimum of the cumulative mean of all replications until day t, divided by the same maximum of the cumulative mean of all replications until day t. We set the convergence level to 0.01 resulting in a run length of 3,250 days. We then rounded up the run length to 10 years (i.e., 3,650 days).

The final step is to determine the required number of replications. We use a stopping criterion on the relative error of the aforementioned KPIs. The relative error bound was set to 1 percent and five replications proved to be sufficient.

4.6.2 Heuristics

With the simulation model we analyze various distributions of allocated emergency beds at inpatient wards. To locate a feasible solution, we developed one heuristic per objective respectively (see Tables 4.1 and 4.2).

Table 4.1: Heuristic Locating Feasible Allocations of Emergency Beds at Inpatient Wards

1.	Initialization phase	Set allocated emergency bed capacity at 100 for each ward (approximating unlimited capacity)
2.	Base phase	Set capacity to average occupied beds per ward from step 1
3.	Optimization phase	Increase capacity of ward with highest utilization rate;
4.	Iteration phase	Repeat step 3 until outcomes of Initialization phase are approached sufficiently (arbitrarily maximum deviation of 3% from the relative LOS at AMU).

In the first heuristic we start with an unlimited bed capacity (the Initialization phase of the heuristic and scenario Init in Table 4.3). As we mentioned above, this scenario shows the best performance because patients only have marginal waiting time because transfers to inpatient wards cannot take place between 9 pm and 9 am. With this scenario we have found an upper bound of our solution space. The distribution of beds among inpatient wards for the next phase of the heuristic (the base phase) is based on the average utilization of the initialization phase. The base phase provides a lower bound for our solution space. Using these averages we do not take into account the stochasticity of the process and therefore this scenario (Avg in Table 4.3) is characterized by underperformance. In the following phase of the heuristic (the optimization phase) we consequently analyze which ward is the bottleneck (i.e., the

ward with the highest utilization rate) and increase the number of allocated beds in this ward by one. The heuristic stops iterating when the stopping criterion is met. We arbitrarily choose a maximum deviation of 0.03 from the relative LOS at the AMU of the initialization phase as stopping criterion.

The second heuristic (see Table 4.2) starts again with an initialization phase, where we use the solution of the first heuristic as input. Since we suspect that pooling resources will improve performance (e.g., the required bed capacity will be lower), we want to know which care unit has the lowest utilization rate. The heuristic now decreases the capacity of the care unit with the lowest utilization rate (optimization phase) and iterates until the stopping criterion is met (iteration phase).

Table 4.2: Heuristic 2 Locating a Feasible Allocation of Emergency Beds at Care Units (i.e., pooled inpatient wards)

1.	Initialization phase	Set allocated emergency bed capacity of care units equal to capacity of pooled wards (see table 4.3)
2.	Optimization phase	Decrease capacity (i.e., number emergency beds) of care unit with lowest utilization rate
3.	Iteration phase	Repeat step 2 until outcomes of separate wards are approached sufficiently (arbitrary percentage refused patients < 0.01)

4.7 Results

The first objective of this study is to evaluate the effect of allocating beds within inpatient wards for patients of the emergency admissions flow. Using the simulation model and heuristic for this objective we have analyzed 14 scenarios (see Table 4.3). Table 4.3 lists these scenarios with their input parameters and output values for the KPIs. Per scenario (i.e., a row in the table) the bed capacity of inpatient ward i with the highest utilization is increased by one bed, graphically shown in bold numbers.

The last row in Table 4.3 shows that in total 33 allocated emergency beds are required to achieve similar performance as in the Init scenario. However, the bed utilization per ward does not exceed 70 percent, which is quite low.

To evaluate the second objective we pooled the wards and applied the second heuristic. We pooled the wards according to the configuration of the partnering hospital. The heuristic starts with the initialization phase (scenario Init in Table 4.4). Per scenario (i.e., a row in the table) the bed capacity of the care unit with the lowest utilization rate is decreased by one bed, graphically shown in bold numbers and stops when the stopping criterion is met. Care Unit 1 consists of the pooled wards 1 and 2; Care Unit 2 of wards 3 and 4; and Care Unit 3 of wards 5, 6, and 7.

The results given in Table 4.4 show that pooling resources in terms of allocated beds between wards further improves outcomes. In the best performing scenario using the first heuristic for separate inpatient wards, 33 allocated emergency beds are required. When pooling inpatient wards, the required number of emergency beds decreases to 24 without significant decrease in performance (based on the KPIs).

Table 4.3: Input Parameters and Results for Each Scenario for Allocated Emergency Beds Within Individual Wards.

Scenario	B	AMU		W1		W2		W3		W4		W5		W6		W7		Total Beds
		rel LOS	ρ	beds	ρ	beds	ρ	beds	ρ	beds	ρ	beds	ρ	beds	ρ	beds	ρ	
Init	0.00	1.00	0.35	100	0.09	100	0.04	100	0.02	100	0.02	100	0.01	100	0.02	100	0.01	700
Avg	0.16	2.58	0.95	9	0.81	4	0.78	2	0.97	2	0.85	1	0.72	2	0.75	1	0.80	21
A	0.15	2.30	0.89	9	0.86	4	0.82	**3**	0.68	2	0.89	1	0.76	2	0.79	1	0.85	22
B	0.12	2.17	0.87	9	0.88	4	0.84	3	0.70	**3**	0.61	1	0.78	2	0.81	1	0.87	23
C	0.09	2.11	0.85	**10**	0.80	4	0.86	3	0.70	3	0.61	1	0.78	2	0.83	1	0.88	24
D	0.07	1.97	0.80	10	0.81	4	0.88	3	0.71	3	0.61	1	0.79	2	0.84	**2**	0.45	25
E	0.05	1.77	0.76	10	0.83	**5**	0.71	3	0.73	3	0.62	1	0.80	2	0.86	2	0.45	26
F	0.01	1.48	0.64	10	0.85	5	0.74	3	0.75	3	0.65	1	0.84	**3**	0.60	2	0.48	27
G	0.01	1.19	0.61	**11**	0.80	5	0.71	3	0.77	3	0.67	1	0.85	3	0.67	2	0.49	28
H	0.01	1.09	0.54	11	0.80	5	0.71	3	0.78	3	0.68	**2**	0.43	3	0.67	2	0.50	29
I	0.00	1.07	0.52	**12**	0.74	5	0.71	3	0.78	3	0.68	2	0.43	3	0.67	2	0.68	30
J	0.00	1.05	0.49	12	0.74	5	0.71	**4**	0.59	3	0.68	2	0.43	3	0.67	2	0.50	31
K	0.00	1.04	0.48	**13**	0.68	5	0.71	4	0.59	3	0.68	2	0.43	3	0.67	2	0.50	32
L	0.00	1.03	0.46	13	0.68	**6**	0.59	4	0.59	3	0.68	2	0.43	3	0.67	2	0.50	33

B = percentage of refused patients, ρ = bed utilization, rel LOS = relative LOS at AMU, W_i = ward i

Table 4.4: Input Parameters and Results for Each Scenario for Allocated Emergency Beds Within Care Units.

Scenario	B	AMU		CU 1		CU 2		CU 3		Total beds
		rel LOS	ρ	Beds	ρ	Beds	ρ	Beds	ρ	
Init	0.00	0.79	0.36	19	0.66	7	0.59	7	0.51	33
I	0.00	0.79	0.36	19	0.66	7	0.59	**6**	0.60	32
II	0.00	0.81	0.37	19	0.66	**6**	0.69	6	0.60	31
III	0.00	0.85	0.39	19	0.66	6	0.69	**5**	0.72	30
IV	0.00	0.86	0.39	**18**	0.69	6	0.69	5	0.72	29
V	0.00	0.87	0.40	**17**	0.73	6	0.69	5	0.72	28
VI	0.00	0.98	0.45	17	0.73	**5**	0.84	5	0.72	27
VII	0.01	1.26	0.57	17	0.72	5	0.83	**4**	0.89	26
VIII	0.01	1.27	0.58	**16**	0.77	5	0.83	4	0.89	25
IX	0.01	1.30	0.59	**15**	0.82	5	0.82	4	0.89	24

B = percentage of refused patients, ρ = bed utilization, rel LOS = relative LOS at AMU, CU_i = care unit i

4.8 Implementation in Practice

This research started with a request from AMU management to analyze the bottlenecks in the AMU's patient flows. Since these flows are multidepartmental, we involved the management of the other departments as well (e.g., ED and inpatient wards). After we reached consensus about the problem and potential solutions, we started constructing the simulation model. In every step of the simulation study the management was involved. We discussed the results with management and the board of directors. We also discussed the structure, related to tactical decision making using the outcomes of the simulation model, which is now embedded in the planning and control cycle of the partnering hospital. This means that at the beginning of every quarter the distribution of the allocated emergency beds at the wards is re-evaluated. The outcomes of this re-evaluation are implemented at the start of the next quarter and meaning that ward managers adjust the number of allocated emergency beds (at the expense of beds available for elective patients). This allows management to adjust other resources (mainly staffing levels) and adjust the planning for elective patients to the new situation. For the evaluation process we use a one-year rolling horizon of data. Our experience shows that an adjustment of zero to three beds per inpatient ward is required every quarter.

As we mentioned above, ward management must work with and manage multiple stakeholders. In practice management could prioritize elective patients and therefore not completely adjust the number of allocated emergency beds as suggested by our model. Overall, this model resulted in a 70 percent decrease in the number of patients refused admission, while elective admissions also increased.

4.9 Discussion

This study shows the positive impact of allocating emergency beds on the emergency admission flow in terms of emergency patients who are refused admission, AMUs LOS

and utilization levels of these beds. This allocation strategy eliminates boarders and therefore a positive contribution to the quality of care can be expected [212]. It may also improve patient satisfaction, since crowding is associated with lower patient satisfaction [190]. One of the primary objectives during the development phase was to create a generalizable model. The model can easily be adopted to different settings, such as extra wards or specialties, varying numbers of beds and different LOS. The model could therefore be useful for other hospitals facing similar problems. It is a user-friendly planning tool, granting (medical) management the power to determine the optimal number of allocated emergency beds with respect to flow dynamics and resource utilization. It provides immediate input for inter-departmental alignment and the tool allows for evaluation of capacity decisions on the patient flow, simplifying the real-life tactical capacity decisions management must make. Therefore this DES model is a universal and powerful tool supporting the planning and control cycle. The partnering hospital uses the tool on a regular basis for tactical decision making and has completely integrated it into the hospital's practices.

Allocating a shared resource (e.g., beds) for specific populations could result in suboptimal utilization since flexibility is reduced. This study addresses this problem by analyzing the effect of pooling capacity (e.g., the pooling of inpatient wards). The results show a significant improvement in bed utilization without decreased performance in the fraction of refused patients and the LOS at the AMU. This supports earlier research, showing that hospital wards can improve performance in terms of bed utilization and refused patients by pooling [69, 238].

Including more case studies is necessary to determine the correlation between allocated emergency beds, flow congestion and boarders. A limitation of the model is the empirical distribution the LOS of the AMU. Substituting beds among emergency patients and elective patients is complex since factors such as the length of the waiting list and the seriousness of conditions of elective patients are important as well. Other factors that influence performance such as staffing levels are currently also not included. To improve both the accuracy of the results and the model validation, further research should be done on the arrival patterns analyzing differences between days and hours and predictors for the AMU LOS. Also, external outflow blockage (e.g., transfers to a nursing home) is a likely cause of significant longer LOS at inpatient wards resulting in increased bed usage, with the potential risk of refusals at the ED and AMU. Finally, different configurations of pooled wards could be analyzed since we looked solely at the configuration of the partnering hospital.

Our model could also facilitate in capacity allocation decisions for emergency patients from a regional perspective. In addition, the visuality of the simulation model adds to the intuition of the flow dynamics when dedicating beds for emergency patients and so increases the likelihood of successful implementation.

This research shows not only that allocating beds for emergency patients at hospital wards improves the emergency admission flow, but its implementation into a tool also helps elucidate the pros and cons of this allocation and thus facilitates implementation.

CHAPTER 5

Scheduling Surgery Groups Considering Multiple Downstream Resources[1]

5.1 Introduction

The operating theater (OT) is one of the most expensive resources [100] and a central hub in hospital patient flow. Therefore, the OT gets a lot of attention to improve productivity. By focusing on OT improvements, other resources get out of sight and are therefore easily forgotten. After surgery, patients are transferred to downstream departments such as the intensive care unit and inpatient wards (hereafter referred to as wards). Therefore, the performance of these downstream departments is directly influenced by the OT [81]. Focusing solely on OT improvements results in large fluctuations in downstream resources, and therefore, requires overcapacity. To optimize all resources involved in the flow of surgical patients, a holistic approach is required. In other words, while improving the productivity of the OT (e.g. optimizing surgery planning), it is crucial to also consider the effect on downstream departments.

According to the organizational decision hierarchy described in [106] and [81], surgery planning consists of three stages: (1) the strategic case mix planning, (2) the tactical master surgery scheduling (MSS) and (3) the operational surgery planning. In the first stage of planning, OT capacity is roughly divided among surgical specialties via blocks (e.g. a day or half a day). Then, the assigned OT blocks are scheduled in a cyclic schedule, which means that the schedule is repeated (bi)weekly, and this results in the tactical MSS. Finally, on an operational level, patients are scheduled within the OT blocks of their surgical specialty.

In this chapter, we discuss the tactical MSS problem while optimizing the effect on the downstream inpatient resources (e.g. bed usage in wards and the ICU). Although counter-intuitive, we observe, from hospital data, that the fluctuations in bed occupancy are mostly caused by artificial (e.g. self induced) variation, and are therefore a result of planning. For this reason, we focus on elective surgery planning. We propose a single step model where bed usage variation is minimized and the OT utilization is

[1]This chapter is based on A.J. Schneider, J.T. van Essen, M. Carlier and E.W. Hans. Scheduling surgery groups considering multiple downstream resources. *European Journal of Operational Research*, 280.2:741-752, 2020.

maximized. We distinguish three concepts related to OT scheduling: (1) OT blocks consisting of (half) a day in which a surgical specialty can perform surgeries in an OT, (2) surgery groups are clusters of surgery types sharing comparable characteristics and (3) surgery types that define a specific surgical procedure. Different from recent research where surgical specialties are assigned to OT blocks in the MSS, we schedule surgery groups within these OT blocks. Surgery groups are clusters of surgery types that share comparable characteristics (e.g. duration, specialty, and/or expertise of surgeons). As a result of the wide variety of surgery types, some surgery types are not performed (bi-)weekly. Therefore, these surgery types cannot be taken into account individually, which makes is necessary to cluster several surgery types within a surgery group. By scheduling surgery groups instead of OT blocks, we want to bridge the gap between the tactical and operational level. Scheduling OT blocks on a tactical level leaves many options for scheduling different surgery types with different expected durations on the operational level, which increases the probability of variation in OT utilization and bed usage. Therefore, we show that scheduling surgery groups reduces the probability of overtime and variation in bed usage.

In the remainder of this chapter, we start with an overview of available literature on OT scheduling and position our research (Section 5.2). In Section 5.3, we discuss three elements of our model: (1) the constraints, (2) the probability distributions of bed usage in the downstream departments for a given cyclic schedule of surgery groups, and (3) the objective function. Section 5.4 describes our global and local search approach, and Section 5.5 discusses the results of both approaches. We analyze several variants of our model in Section 5.6. Finally, we discuss the implications of our approach in Section 5.7.

5.2 Literature review & research positioning

OT planning and scheduling literature is broadly available. For an overview on general OT scheduling literature, we refer to the systematic reviews of [49] and [100]. Here, we solely consider OT planning and scheduling literature that take downstream resources into account.

Two approaches are used by [22] to model bed occupancy of a single ward while creating an MSS: (1) a mixed integer programming (MIP) based approach, linear as well as quadratic, and (2) simulated annealing (SA). In [23], this model is extended with multiple wards. Furthermore, [22] assume the number of patients per OT block to be deterministically dependent on the type of surgery and fixed for each surgeon, while [23] assume a multinomial distribution function for this. Two hierarchical goal programming approaches are developed by [23], that both consist of two goal programming models that are solved successively.

A MIP was developed by [202] using average values for the LoS (Length of Stay: the sojourn time at wards). Their model has two objectives: maximizing daily bed utilization and maximizing throughput and mix of patients. A mixed integer linear programming (MILP) model by [254] levels the daily beds and nurse workload, while considering surgeons preferences.

Elective surgical types that are frequently performed are cyclical scheduled in[233]

. The solution approach consists of two steps: (1) an integer linear program (ILP) which ignores the required number of beds and that is solved by an implicit column generation approach and (2) a MILP with the objective to minimize the required number of beds. They incorporate three types of beds which can be prioritized. Also surgeries are assigned to a day in the cyclic schedule by [3] , as in [233]. However, [3] use a stochastic LoS that outperforms a deterministic LoS. The extension of [233] in [4] also accounts for emergency patients. They use simulation to create an operational schedule based on the obtained tactical schedule with emergency patients.

Two other studies, [240] and [239], assign OT time to specialties, just as [22], by computing the ward occupancy distributions, the patient admission/discharge distributions, and the distributions for the ongoing interventions/treatments required by recovering patients. In [240], they swap OT blocks and surgical specialty assignments to find a good solution. This model is extended in other literature. The analytical approach of [240] is used to determine the number of required beds [231]. Two solution methods are used: (1) ILP and (2) SA. To be able to use an ILP, the objective function is replaced by the maximum of the expected number of required beds. An extension of the approach of [240] is made by taking multiple wards and the ICU into account and consider several heuristic solution methods [81]. In [79], this is even further extended by including multiple ICUs and outpatient flows in downstream resources. Another extension by [80] includes outpatients and emergency surgeries during the weekends.

Simulation is used to investigate a stochastic surgery scheduling problem while considering ICU beds [173]. Surgery durations and LoS on the ICU are assumed stochastic with known distributions. Monte Carlo simulation is combined with a MIP to predict the impact of an MSS on bed occupancy [56]. The simulation model predicts the daily demand of beds and the MIP (based on [23]) optimizes the bed occupancy by scheduling surgery blocks and patient types within each block. Also, a MIP model to find an MSS was proposed by [17]. Their objective is to maximize the number of surgeries planned while minimizing the violation of due dates. Next to the MIP model, they also simulate the MIP solution for robustness.

The impact of variability in admissions and LoS on the required amount of bed capacity with an approximation method is analyzed by [21]. Given an admission pattern, their quadratic programming model determines the mean bed occupancy of each day. The Markov Decision Process (MDP) model in [12] provides scheduling policies for all surgeries, given an MSS, that minimize the time a patient spends on the waiting list, OT overtime and ward congestion. They use approximate dynamic programming to solve the MDP of a realistic problem.

We extend the previous work of [233] and [81] by scheduling surgery groups within OT blocks and by developing an single step solution method instead of decomposition approaches. Scheduling surgery groups complicates modeling the overtime constraint and utilization of the OT, because there are multiple options for scheduling surgery groups within an OT block. To cluster surgical procedure types into surgery groups, we use techniques from data mining. Furthermore, we linearize the overtime constraint by a piecewise linear function and the objective function by using the expected variation in bed occupancy.

5.3 Problem formulation

In this section, we formulate our problem of creating a schedule that specifies which surgery groups should be scheduled in an OT block. First, we explain our clustering approach for defining the surgery groups in Section 5.3.1. From Section 5.3.2 to 5.3.4, we explain the mathematical model.

5.3.1 Clustering surgery types into surgery groups

As mentioned in the introduction, not every individual surgery type can be considered on the tactical level as some are not performed weekly. Thus, we need to cluster surgery types into groups to model all surgery types. We cluster surgery types into surgery groups using data mining techniques. Data mining improves the understanding of the relations between predictor and response variables, underlying structures and/or distributions of the input data. Therefore, data mining potentially improves the results of the considered model. Data mining techniques can be split into two main categories: supervised learning and unsupervised learning. Supervised learning makes use of labeled training and predicts a response variable with predictor variables [227]. Unsupervised learning only uses unlabeled (e.g. predictor) variables and analyzes the underlying structure or distribution of the data (e.g. clustering or association). For our model, we want to use the predictor variables *specialty* and *surgery* type to predict the response variables *surgery duration* (for OT utilization) and *LoS* (for bed usage) as is done in supervised learning. The response variable could then be split into certain classes such as short LoS and short surgery duration. However, these labels are dependent on the classification we would like to make, and are therefore not available. The other category, unsupervised learning, assumes unlabeled data and does not split the variables into response and predictor variables.

Clustering algorithms examine the data to find groups of similar instances. We would like instances with the same specialty and surgery type to be in one cluster, so they account for surgeon specialization. However, most clustering algorithms (unsupervised learning) assume independent instances. Moreover, in most clustering algorithms we cannot specify what type of clusters we want. This means that one cluster could contain instances where the dispersion of LoS is small and the dispersion of surgery duration is large, and vice versa in another cluster. Therefore, we combine supervised and unsupervised learning techniques in our approach: first, we divide the surgery types of a specialty into short and long stay clusters based on the *median* LoS of the surgery type (e.g LoS groups). This means that the cut-off point between short and long stay clusters depends on the specialty. The cut-off point is determined by maximizing the precision, based on all instances, of both clusters. Precision is an evaluation measure of the confusion matrix and is defined as the fraction of correct positive predictions among the total number of positive predictions [227]. In our study, this equals the number of instances in a cluster that were indeed lower (for short stay) or higher (for long stay) than the cut-off point among all instances of a surgery type. We tested the precision both on the median and mean of each surgery type and results show that for clustering the LoS the median results in higher precision. Next, we further divide each short and long stay cluster into three sub clusters based on the surgery

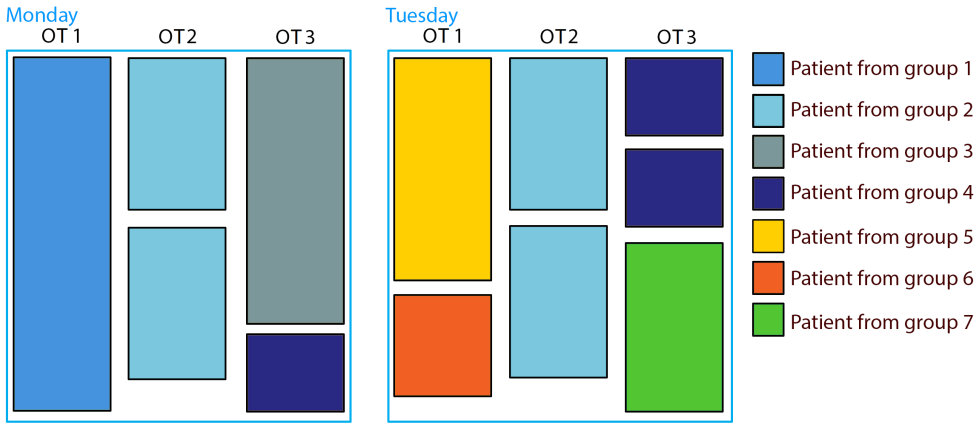
duration. The clustering for the surgery duration is similar as for the LoS, although now we take the *mean* of each surgery type. The cut-off points are again determined by maximizing the precision of each cluster. This means that our clustering approach results in six groups per surgical specialty. Here, we also tested the precision bases on both the median and mean. Results show that for surgery durations, the means results in a higher precision.

To ensure that the sizes of the resulting surgery groups do not become too small, we set the cut-off point such that the number of instances assigned to a group is at least 20% of the number of instances that can be divided. In addition, we use a two-sample t-test with a 5% significance level to determine whether two groups are significantly different. When the two groups fail this test, i.e., when they are not significantly different, we decrease the number of groups.

5.3.2 Conceptual model

In our approach, we assign surgery groups to OT blocks instead of surgical specialties. See Figure 5.1 for a graphical example of scheduling surgery groups. Assigning a surgery group to an OT block allows for a single surgery type of that group to be scheduled during the next planning stage. Multiple surgery groups can be assigned multiple times to the same OT on the same day as long as the surgery groups belong to the same specialty of the allocated OT block. The order in which individual patients of the surgery groups are scheduled on the operational level is undefined. For example, our MSS specifies that surgery group X and Y are scheduled on the same day and OT, but does not specify if surgery group X must be scheduled before or after surgery group Y during that day. Hence, a variable amount of surgery groups can be scheduled in an OT block of the MSS. The objective of our model is to find an optimal schedule of surgery groups that maximizes OT utilization while minimizing the variance of bed usage at the wards.

Figure 5.1: OT Schedule Example With Surgery Groups



5.3.3 Constraints

Multiple constraints are taken into account for our model, e.g., the need for specific OTs (e.g. thoracic surgeries require specific operating tables) , the need for specific equipment, the total available OT time during opening hours and the number of scheduled surgery groups. Let $\mathcal{O}$ be the set of given OTs and $\mathcal{K}$ the set of days in the MSS. Then, an OT block (o, k) is defined as a combination of day $k \in \mathcal{K}$ of the MSS and OT $o \in \mathcal{O}$. The set of given surgery groups is denoted by set $\mathcal{J}$.

The integer decision variable z_{okj} specifies the number of surgeries from surgery group $j \in \mathcal{J}$ that are scheduled in OT block (o, k). To ensure equitable access for each surgery group, we set a lower bound β_j on the number of scheduled surgeries per surgery group $j \in \mathcal{J}$ and assume that waiting lists are inexhaustible. The following constraints ensure that all groups $j \in \mathcal{J}$ are scheduled a minimum of β_j times.

$$\sum_{o \in \mathcal{O}, k \in \mathcal{K}} z_{okj} \geq \beta_j, \qquad \forall j \in \mathcal{J}. \tag{5.1}$$

Let $\mathcal{S}$ be the set of specialties and $\mathcal{J}_s \subseteq \mathcal{J}$ the set of surgery groups belonging to specialty $s \in \mathcal{S}$. We introduce binary parameters ϵ_{oks} that are one when specialty $s \in \mathcal{S}$ can be allocated to OT block (o, k) in the MSS, and zero otherwise. Furthermore, we introduce binary decision variables u_{oks} which are one when a surgery group of specialty $s \in \mathcal{S}$ is scheduled in OT block (o, k) and zero otherwise. Now we can ensure that only surgery groups of the specialty that is allocated to OT block (o, k) can be scheduled:

$$u_{oks} \leq \epsilon_{oks}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}, s \in \mathcal{S}, \epsilon = 1. \tag{5.2}$$

The relation between z_{okj} and u_{oks} is given by constraints (5.3), where M_s is the maximum number of surgeries of a specialty $s \in \mathcal{S}$ that fit in one OT block:

$$\sum_{j \in \mathcal{J}_s} z_{okj} \leq M_s \cdot u_{oks}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}, s \in \mathcal{S}. \tag{5.3}$$

To ensure that only one specialty $s \in \mathcal{S}$ can be assigned to each OT block, we introduce the following constraints and binary parameters χ_{ok} which are one when OT $o \in \mathcal{O}$ is open on day $k \in \mathcal{K}$ and zero otherwise:

$$\sum_{s \in \mathcal{S}} u_{oks} \leq \chi_{ok}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}. \tag{5.4}$$

The total surgery duration of the surgery groups we assign to an OT block is limited by the opening hours of the OT. The surgery duration ζ_j of surgery group $j \in \mathcal{J}$ is a stochastic variable with mean μ_j and variance σ_j^2. Let g_{ok} denote the stochastic variable representing the total duration of the surgery groups that are scheduled in OT block (o, k). The available OT time on day $k \in \mathcal{K}$ in OT $o \in \mathcal{O}$ is denoted by τ_{ok}. We introduce constraints (5.5) to ensure that the probability of overtime is below α with $0 \leq \alpha \leq 1$. Overtime occurs when the total sum of the duration of the scheduled groups exceeds the available time of that OT block:

$$P(g_{ok} \leq \tau_{ok}) \geq 1 - \alpha, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}. \tag{5.5}$$

Some surgery groups require specific equipment that is not available in every OT, and therefore, have to be scheduled in specific OTs, while other surgery groups can be scheduled in every OT. To model this, we define a set of OT types $\mathcal{R}$ and we denote the subset of surgery groups that can be performed in OT type $r \in \mathcal{R}$ by $J_r \subseteq \mathcal{J}$. Binary parameters v_{okr} are one when OT $o \in \mathcal{O}$ on day $k \in \mathcal{K}$ is of type $r \in \mathcal{R}$ and zero otherwise. This leads to the following constraint:

$$\sum_{j \in \mathcal{J}_r} z_{okj} \leq N_r v_{okr}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}, r \in \mathcal{R}, \tag{5.6}$$

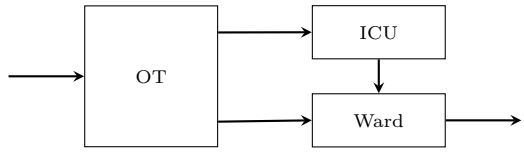
where N_r is the maximum number of surgeries belonging to OT type r in one OT block.

5.3.4 Bed usage distributions

Next, we want to determine the bed usage distributions of the wards in three steps: (1) we calculate the bed usage distribution for the wards per surgery group, (2) we calculate the bed usage distribution for overlapping cycles, and (3) we calculate the bed usage distribution for an entire OT block. This final step needs to be repeated for every cyclic schedule. The first two steps can be done beforehand. Based on these inputs, we will describe the objective function.

As mentioned before, we further extent the work of [239] and [81] by assuming that patients from the same surgery group can be admitted at different wards (e.g. different ICUs or different wards belonging to the same surgical specialty). This extension is based on practice, where hospitals can have specific ICUs and wards for specific patient types en thus surgery types such as thoracic surgery. Therefore, we take into account all wards where patients of a certain surgical specialty can be admitted.

Figure 5.2: Main Hospital Flows for Surgical Patients



We assume that patients can take two paths after surgery: (1) directly to a ward or (2) first to the ICU followed by a transfer to a ward (see Figure 5.2). Finally, patients are discharged and leave the system. Let set $\mathcal{I}$ denote all ICUs and let set $\mathcal{W}$ denote all wards. For all $j \in \mathcal{J}$, we define subsets $\mathcal{J}_i \subseteq \mathcal{J}$ and $\mathcal{J}_w \subseteq \mathcal{J}$ for the surgery groups that are transferred to the ICU $i \in \mathcal{I}$ and ward $w \in \mathcal{W}$, respectively. The LoS (in days) in the ICU $i \in \mathcal{I}$ or ward $w \in \mathcal{W}$ of each surgery group is modeled by discrete empirical distributions based on historical data. The empirical distribution of the LoS is determined per surgery group, regardless of the ward they are transferred to. The following input parameters are required for every surgery group $j \in \mathcal{J}$:

- a_{ij} represents the probability that a patient of surgery group $j \in \mathcal{J}$ is transferred to ICU $i \in \mathcal{I}$ after surgery.

- b_{wj} represents the probability that a patient from surgery group $j \in \mathcal{J}$ is transferred to ward $w \in \mathcal{W}$ after surgery or ICU.
- c^I_{jn} represents the probability that a patient from surgery group $j \in \mathcal{J}$ stays exactly n days in the ICU after surgery.
- c^{WS}_{jn} represents the probability that a patient from surgery group $j \in \mathcal{J}$ stays exactly n days in the ward after surgery.
- c^{WI}_{jn} represents the probability that a patient from surgery group $j \in \mathcal{J}$ stays exactly n days in the ward after a stay in the ICU.

The probability that a patient from surgery group $j \in \mathcal{J}$ is transferred to the ICU is given by $\sum_{i \in I} a_{ij}$ and a transfer to the ward is given by $1 - \sum_{i \in I} a_{ij}$. The probabilities c^I_{jn}, c^{WS}_{jn} and c^{WI}_{jn} are not given separately for every ward or ICU, because for every surgery group $j \in \mathcal{J}$, the probability of a patient staying exactly n days is independent of the ward or ICU. We also assume a bed is occupied a whole day if a patient is discharged on that day.

Single surgery group

The first step of our approach is similar to the approach presented in [81]. As [81], we start by calculating conditional probabilities d^I_{jn+1} that a patient from surgery group $j \in \mathcal{J}$ is transferred from the ICU to a ward on day $n+1$ (which is n days after surgery). In a similar way, the conditional probabilities d^{WS}_{jn+1} that a patient from surgery group $j \in \mathcal{J}$, who is in the ward on day n, is discharged on day n can be determined. Conditional probabilities d^{WI}_{jn+1} represent the probability that a patient from surgery group $j \in \mathcal{J}$, who is in the ward on day n after being transferred from the ICU, is discharged on day n, where we assume that the patient is transferred from the ICU on day 1.

As [81], we can now calculate probabilities e^I_{jn} that a patient from surgery group $j \in \mathcal{J}$, who had surgery on day 1, is still occupying a bed on day n. For $n = 1$ and the ICU, this is simply the probability that the patient is transferred to the ICU after surgery. We assume a patient stays at least one day in the ICU, otherwise, a patient is transferred directly to the ward. Therefore, for $n = 2$, we have the same probability as for $n = 1$. For $n \in \{3, ..., N^I_j + 1\}$, where N^I_j is the maximum number of days that a patient from surgery group $j \in \mathcal{J}$ stays in the ICU after surgery, this is the probability that the patient was not transferred to the ward the day before, i.e., day $n-1$, multiplied by the probability that the patient was still in the ICU the day before.

Similarly, probabilities e^{WS}_{jn} and e^{WI}_{jnm} are determined, i.e., the probabilities that a patient from surgery group $j \in \mathcal{J}$ who had surgery on day 1, is still occupying a bed in the ward on day n and the probability that after an ICU stay of m days, a patient from surgery group $j \in \mathcal{J}$ is still in the ward on day n, respectively. Probabilities e^{WS}_{jn} and e^{WI}_{jmn} are combined to calculate the probability e^W_{jn} that a patient of surgery group $j \in \mathcal{J}$ is in the ward on day n.

Different from [81], we consider multiple ICUs. Therefore, we also need to calculate the probability that a patient from surgery group $j \in \mathcal{J}$ is in ICU $i \in \mathcal{I}$, given that this patient is in the ICU. $\sum_{i \in I} a_{ij}$ is the probability that a patient of surgery group $j \in \mathcal{J}$ is in the ICU. So, for all $j \in \mathcal{J}$ and $i \in \mathcal{I}$, we have conditional probability $\hat{a}_{ij} = \frac{a_{ij}}{\sum_{i \in I} a_{ij}}$ that a patient of surgery group $j \in \mathcal{J}$ is in ICU $i \in \mathcal{I}$, given that this

patient is in the ICU. For the wards, this probability is given by b_{wj}. We do not need to normalize this probability, since every patient in our model is transferred to the ward. Patients who do not stay at the ward are represented using a ward LoS of zero days.

The probability distributions of the number of patients from surgery group $j \in \mathcal{J}$ in ICU $i \in \mathcal{I}$ or ward $w \in \mathcal{W}$ on day n are denoted by f^I_{ijn} and f^W_{wjn}. The discrete stochastic variables that are associated with these probability distributions are given by f^I_{ijn} and f^W_{wjn}, respectively. Since different from [81], we schedule surgery groups instead of OT blocks, the number of patients in an ICU or ward can only equal zero or one. So, the probability that there is one patient in the ward or ICU is calculated by multiplying the probability that a patient from surgery group $j \in \mathcal{J}$ goes to ICU $i \in \mathcal{I}$ (ward $w \in \mathcal{W}$), given this patient is in the ICU (ward), with the probability that this patient is in the ICU (ward) on day n. The probability that there are zero patients is equal to one minus the probability that there is one patient.

$$P(\mathrm{f}^I_{ijn} = 0) = 1 - \hat{a}_{ij} e^I_{jn}, \qquad i \in \mathcal{I}, j \in \mathcal{J}, n \in \{1, ..., N^I_j\}; \tag{5.7}$$

$$P(\mathrm{f}^I_{ijn} = 1) = \hat{a}_{ij} e^I_{jn}, \qquad i \in \mathcal{I}, j \in \mathcal{J}, n \in \{1, ..., N^I_j\}; \tag{5.8}$$

$$P(\mathrm{f}^W_{wjn} = 0) = 1 - b_{wj} e^W_{jn}, \qquad w \in \mathcal{W}, j \in \mathcal{J}, n \in \{1, ..., N^W_j\}; \tag{5.9}$$

$$P(\mathrm{f}^W_{wjn} = 1) = b_{wj} e^W_{jn}, \qquad w \in \mathcal{W}, j \in \mathcal{J}, n \in \{1, ..., N^W_j\}. \tag{5.10}$$

Cyclical surgery group

Now that we have all probabilities for single surgery groups, we can calculate the bed usage distribution for overlapping cycles using the approach of [239], since the maximum LoS of a patient can exceed the cycle length. The distribution of the number of patients in overlapping cycles is denoted by F^I_{ijl} and F^W_{wjl} for surgery group $j \in \mathcal{J}$ in ICU $i \in \mathcal{I}$ and ward $w \in \mathcal{W}$, respectively, on the lth day of a cycle, when the surgery group is scheduled on day one of the cycle. The number of overlapping cycles depends on the maximum LoS in the ICU and wards, N^I_j and N^W_j, respectively, and on the cycle length L, which is the number of elements in $\mathcal{L}$. Depending on the day l in the cycle, we have $\lfloor (N^I_j - l)/L \rfloor + 1$ overlapping cycles for the ICU and $\lfloor (N^W_j - l)/L \rfloor + 1$ overlapping cycles for the ward.

Cyclic schedule

We now have all the elements for the final step: calculating the bed usage distributions for a cyclic surgery group schedule. The calculations in this step differ from [239] and [81], since we schedule surgery groups instead of OT blocks. This means that we consider a variable amount of surgery groups and number of the same surgery group within an OT block, where [81, 239] consider an average number of surgeries and durations for an OT block. A cyclic schedule is given by the integer decision variables z_{okj}, which represent the total number of surgeries from surgery group $j \in \mathcal{J}$ that are scheduled in OT $o \in \mathcal{O}$ on day $k \in \mathcal{K}$. Let $\mathbf{1}_{z_{okj}}$ be an indicator function that is equal to one if z_{okj} is greater than zero and equal to zero if z_{okj} is zero. The bed usage distribution in ICU $i \in \mathcal{I}$ and ward $w \in \mathcal{W}$ on day l of a cyclic schedule when

scheduling surgery group $j \in \mathcal{J}$ once in OT block (o,k) is given by G^I_{iokjl} and G^W_{wokjl}, respectively.

Next, we shift both distributions F^I_{ijl} and F^W_{wjl} to the day on which the surgery group is scheduled. Here, l is the day for which we are determining the bed usage distribution and k is the day on which the surgery group is scheduled in the cyclic schedule. If $l \geq k$, we shift F^I_{ijl} and F^W_{wjl} by $k-1$ days. If $l < k$, the bed usage distribution on day l results only from surgery groups scheduled on day k of previous cycles. Thus, we shift by $k-1-L$ days. We multiply these distributions by $\mathbf{1}_{z_{okj}}$, which is only non-zero if the surgery group $j \in \mathcal{J}$ is assigned to OT block (o,k).

$$G^I_{iokjl} = \begin{cases} F^I_{ijl-k+1}\mathbf{1}_{z_{okj}}, & l \geq k \\ F^I_{ijl-k+1+L}\mathbf{1}_{z_{okj}}, & \text{otherwise.} \end{cases} \tag{5.11}$$

$$G^W_{wokjl} = \begin{cases} F^W_{wjl-k+1}\mathbf{1}_{z_{okj}}, & l \geq k \\ F^W_{wjl-k+1+L}\mathbf{1}_{z_{okj}}, & \text{otherwise.} \end{cases} \tag{5.12}$$

Next, we obtain the bed usage distributions for an OT block. We use the indicator function $\mathbf{1}_{z_{okj}}$ to indicate that a surgery group $j \in \mathcal{J}$ is assigned at least once to OT block (o,k). However, a surgery group might be assigned multiple times to one OT block. To obtain the distribution of patients from an entire OT block, we need the convolution of all distributions G^I_{iokjl} and G^W_{wokjl} of the surgery groups scheduled in that OT block. If a surgery group is assigned n times to one OT block, we need to convolute the distribution n times with itself, before convolving it with the distributions of other surgery groups assigned to that OT block. Therefore, we use the convolution power, which is defined as the n-fold iteration of the convolution with itself. For h, a function $\mathbb{Z} \to \mathbb{R}$ and $n \in \mathbb{N}_{>0}$, we have:

$$h^{*n} = \underbrace{h * h * ... * h * h}_{n}, \qquad h^{*0} = \delta_0, \tag{5.13}$$

where δ_0 is Dirac's delta function. Dirac's delta function focuses the mass of a function around zero. When we convolve a distribution zero times, the probability of being zero is equal to one.

The bed usage distribution in ICU $i \in \mathcal{I}$ and ward $w \in \mathcal{W}$ on day l of the cyclic schedule per surgery group $j \in \mathcal{J}$ in OT block (o,k) is given by $\hat{G}^I_{iokjl}$ and $\hat{G}^W_{wokjl}$:

$$\hat{G}^I_{iokjl} = {G^I_{iokjl}}^{*z_{okj}}, \qquad i \in \mathcal{I}, o \in \mathcal{O}, k \in \mathcal{K}, j \in \mathcal{J}_i, l \in \mathcal{L}. \tag{5.14}$$

$$\hat{G}^W_{wokjl} = {G^W_{wokjl}}^{*z_{okj}}, \qquad w \in \mathcal{W}, o \in \mathcal{O}, k \in \mathcal{K}, j \in \mathcal{J}_w, l \in \mathcal{L}. \tag{5.15}$$

Now we can define distributions H^I_{iokl} and H^W_{wokl}, which represent the bed usage distributions on day l at ICU $i \in \mathcal{I}$ and ward $w \in \mathcal{W}$, resulting from all surgery groups $j_1, j_2, ..., j_{\max} \in \mathcal{J}_i$ and $j_1, j_2, ..., j_{\max} \in \mathcal{J}_w$, respectively.

$$H^I_{iokl} = \hat{G}^I_{iokj_1l} * \hat{G}^I_{iokj_2l} * ... * \hat{G}^I_{iokj_{\max}l}, \qquad i \in \mathcal{I}, o \in \mathcal{O}, k \in \mathcal{K}, l \in \mathcal{L}, \tag{5.16}$$

$$H^W_{wokl} = \hat{G}^W_{wokj_1l} * \hat{G}^W_{wokj_2l} * ... * \hat{G}^W_{wokj_{\max}l}, \quad w \in \mathcal{W}, o \in \mathcal{O}, k \in \mathcal{K}, l \in \mathcal{L}. \tag{5.17}$$

Following the approach of [239] and [81], we convolve the distributions of all the OT blocks in the cyclic schedule to obtain the bed usage distributions resulting from the complete cyclic schedule. $\hat{H}^I_{il}$ denotes the distribution of patients in ICU $i \in \mathcal{I}$ on day l of the cyclic schedule and $\hat{H}^W_{wl}$ denotes the distribution of recovering patients in ward $w \in \mathcal{W}$ on day l of the cyclic schedule. The last OT and the last day in the cyclic schedule on which surgeries take place are denoted by $\max\{\mathcal{O}\}$ and $\max\{\mathcal{K}\}$ respectively.

$$\hat{H}^I_{il} = H^I_{i11l} * H^I_{i12l} * ... * H^I_{i1\max\{\mathcal{K}\}l} * H^I_{i21l} * H^I_{i22l} * ... * H^I_{i\max\{\mathcal{O}\}\max\{\mathcal{K}\}l}, \quad i \in \mathcal{I}, l \in \mathcal{L}, \tag{5.18}$$

$$\hat{H}^W_{wl} = H^W_{w11l} * H^W_{w12l} * ... * H^W_{w1\max\{\mathcal{K}\}l} * H^W_{w21l} * H^W_{w22l} * ... * H^W_{w\max\{\mathcal{O}\}\max\{\mathcal{K}\}l}, \quad w \in \mathcal{W}, l \in \mathcal{L}. \tag{5.19}$$

We define the probability of having n patients in ICU $i \in \mathcal{I}$ or ward $w \in \mathcal{W}$ on day l by $\hat{H}^I_{il}[n]$ and $\hat{H}^W_{wl}[n]$.

For a given cyclic schedule ψ, we want to determine the variation in bed occupancy. This means that we calculate for each day l and with probability p that there are at most n patients, thus n required beds, by summing over the probabilities that there are at most n patients in the ICU or ward. The required number of beds $\gamma_{il}(\psi)$ on day l in ICU $i \in \mathcal{I}$ for a given solution $\psi \in \Psi$ is then given by:

$$\gamma_{il}(\psi) = \min\left\{ n \middle| \sum_{m=0}^{n} \hat{H}^I_{il}[m] \geq p \right\}. \tag{5.20}$$

The required number of beds $\gamma_{wl}(\psi)$ on day l in ward $w \in \mathcal{W}$ for a given solution ψ is given similarly by:

$$\gamma_{wl}(\psi) = \min\left\{ n \middle| \sum_{m=0}^{n} \hat{H}^W_{wl}[m] \geq p \right\}. \tag{5.21}$$

Peaks in bed occupancy occur during weekdays since new patients arrive to undergo scheduled surgeries. These peaks may cause surgery cancellations, because not enough beds are available. Therefore, we are interested in minimizing the variation in bed occupancy during weekdays. As no surgeries are scheduled during the weekends, the bed occupancy is lower. The variation in bed occupancy, denoted by $\gamma_i(\psi)$ and $\gamma_w(\psi)$, in ICU $i \in \mathcal{I}$ and ward $w \in \mathcal{W}$ is given by the difference between the maximum and minimum number of required beds during the week and are given by:

$$\gamma_i(\psi) = \max_{l \in \mathcal{K}} \gamma_{il}(\psi) - \min_{l \in \mathcal{K}} \gamma_{il}(\psi), \tag{5.22}$$

$$\gamma_w(\psi) = \max_{l \in \mathcal{K}} \gamma_{wl}(\psi) - \min_{l \in \mathcal{K}} \gamma_{wl}(\psi), \tag{5.23}$$

where $\mathcal{K}$ is the set of all workdays as defined in Section 5.3.3.

Objective function

Our model has two main goals: (1) to maximize the OT utilization and (2) to minimize the variation in bed occupancy. Because the available OT time is determined at the strategical level, it is constant in our model. Hence, maximizing the OT utilization is equal to maximizing the time allocated for scheduled surgery groups. The utilized OT time is the sum of the mean surgery durations μ_j of the scheduled surgery groups. Furthermore, we want to minimize the variation in bed occupancy, γ_i and γ_w. Finally, we include weights θ_i and θ_w, so we can manage the balance between the variation in bed occupancy and the OT utilization. The objective function is now given by:

$$\max \sum_{o \in \mathcal{O}} \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \mu_j \cdot z_{okj} - \sum_{i \in I} \theta_i \cdot \gamma_i(\psi) - \sum_{w \in W} \theta_w \cdot \gamma_w(\psi), \tag{5.24}$$

where the objective function value for a given schedule ψ is denoted by $OB(\psi)$.

5.4 Solution methods

The majority of calculations in Section 5.3.4 can be performed beforehand. However, the calculations for a cyclic schedule still involve the convolution of several probability distributions and have generated within the model. Moreover, the constraints in (5.5) are nonlinear which makes the model nonlinear. Therefore, we use two different approaches to solve our problem: (1) approximate our model using linearizations in a MILP and (2) use an approximation approach, for this we use simulated annealing (SA), to run our model. MILP and SA are widely used for solving the MSS problem and are also compared on the trade-off between the objective function value and computational performances by [49].

5.4.1 Global approach

Our global approach uses an approximation of the objective function and a linearized version of nonlinear constraints (5.5) in order to formulate a MILP which we can solve with a commercial solver. In Section 5.4.1, we linearize the overtime constraints (5.5). Because there is no direct relation between a given OT-schedule and the number of required beds, we also linearize the objective function in Section 5.4.1.

Linearization of the surgery duration constraint

In the problem formulation introduced in Section 5.3.3, we have nonlinear constraints that make the surgery schedule more robust against overtime. We linearize the overtime constraint using the same approach as [36] and shown in equation 5.5.

The 3-parameter lognormal distribution is the best fit for surgery duration distributions [169]. However, since there is no known exact result for the distribution of the sum of 3-parameter lognormal distributed stochastic variables we approximate the distribution of the sum of the surgery durations with a normal distribution as is done by [107] and [233]. For this, we assume that the total duration of OT block (o, k) is

normally distributed with mean μ_{ok} and variance σ^2_{ok}. Thus, $g_{ok}(x) \sim \mathcal{N}(\mu_{ok}, \sigma_{ok})$. Then, the overtime constraints can be written as:

$$P(g_{ok} \leq \tau_{ok}) = \Phi\left(\frac{\tau_{ok} - \mu_{ok}}{\sigma_{ok}}\right) \geq 1 - \alpha, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}. \tag{5.25}$$

Where Φ is the z-score of the normal distribution. Rewriting equation (5.25) gives:

$$\mu_{ok} + \Phi^{-1}(1-\alpha)\sigma_{ok} \leq \tau_{ok}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}. \tag{5.26}$$

The mean and variance of the total surgery duration g_{ok} of OT block (o, k) can be written as:

$$\mu_{ok} = \sum_{j \in \mathcal{J}} z_{okj}\mu_j \quad \text{and} \quad \sigma^2_{ok} = \sum_{j \in \mathcal{J}} z_{okj}\sigma^2_j. \tag{5.27}$$

Substituting the latter two expressions into the overtime constraints (5.26) gives:

$$\sum_{j \in \mathcal{J}} z_{okj}\mu_j + \Phi^{-1}(1-\alpha)\sqrt{\sum_{j \in \mathcal{J}} z_{okj}\sigma^2_j} \leq \tau_{ok}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}. \tag{5.28}$$

To linearize this constraint, we approximate the square root function $f(x) = \sqrt{x}$ by a piecewise linear function. The square root function needs to be approximated on the interval $[x_{\min}, x_{\max}]$. We do not want to underestimate the function $f(x)$, so the approximation function must be greater than or equal to $f(x)$ for all $x \in [x_{\min}, x_{\max}]$. The intervals of the piecewise linear functions are determined by breakpoints $n \in N$, where $N = \{0, 1, ..., m\}$. Here, x_n is the value on the x-axis of breakpoint $n \in N$. We define x_0 as the first x-value and x_m as the last x-value for which we approximate the square root function. The other x-values, x_n for $n = \{1, ..., m-1\}$, are intersection points of the linear approximations. Let y_n be the function value of the linear approximation function at breakpoint n, so $y_n = \sqrt{x_n}$. See for more details the appendix in the supplementary materials.

Once the breakpoints are known, we can use the λ-formulation by [32] to model piecewise linear functions together. The function value of any point between two breakpoints is the weighted sum of the function values of these two breakpoints. Let λ_{okn} denote n nonnegative weights for each OT block (o, k) such that their sum equals one. Then, the piecewise linear approximation of the overtime constraint can be written as:

$$\sum_{j \in \mathcal{J}} z_{okj}\mu_j + \Phi^{-1}(1-\alpha) \sum_{n \in N} \lambda_{okn} y_n \leq \tau_{ok}, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}, \tag{5.29}$$

$$\sum_{n \in N} \lambda_{okn} x_n = \sum_{j \in \mathcal{J}} z_{okj}\sigma^2_j, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}, \tag{5.30}$$

$$\sum_{n \in N} \lambda_{okn} = 1, \qquad \forall o \in \mathcal{O}, k \in \mathcal{K}. \tag{5.31}$$

Considering overtime constraints (5.28), we show that when scheduling surgery groups instead of surgical specialties, we can at least assign the same number of surgeries to

one OT block. Assuming that there exists a surgery group $j \in \mathcal{J}_s$ with $\mu_j \leq \mu_s$ and $\sigma_j^2 \leq \sigma_s^2$, where μ_s and σ_s^2 represent the mean and variance of the surgery duration for surgical specialty $s \in \mathcal{S}$, we have that

$$z\mu_j + \Phi^{-1}(1-\alpha)\sqrt{z\sigma_j^2} \leq z\mu_s + \Phi^{-1}(1-\alpha)\sqrt{z\sigma_s^2} \leq \tau \tag{5.32}$$

where z denotes the number of assigned surgeries to a given OT block. This means that the cyclic schedule obtained when scheduling surgery groups instead of surgical specialties allows us to schedule at least the same number of surgeries and possibly more.

Linearization of the objective function

Our approach for linearizing the objective function is an extension of the approach of [22]. Instead of using γ_i and γ_w, we use the expected number of beds at ward $w \in \mathcal{W}$ and ICU $i \in \mathcal{I}$ on day l of the cycle. For a solution ψ, this is given by $\bar{\gamma}_{wl}(\psi)$ and $\bar{\gamma}_{il}(\psi)$, respectively. We use the expected value of the distribution functions $\hat{H}_{il}^I$ and $\hat{H}_{wl}^W$, which are defined as the probability distributions of the bed usage in the ICU and ward, respectively. The expected value of $\hat{H}_{il}^I$ is given by:

$$\begin{aligned}\bar{\gamma}_{il} &= \mathbb{E}\left(\hat{\mathcal{H}}_{il}^I\right) \\ &= \sum_{o\in O}\sum_{\substack{k\in\mathcal{K}\\ l\geq k}}\sum_{j\in\mathcal{J}_i}\sum_{n=0}^{\lfloor D_{jkl}^I/L\rfloor} \hat{a}_{ij} e^I_{j(l-k+1+nL)}\cdot z_{okj} \\ &+ \sum_{o\in O}\sum_{\substack{k\in\mathcal{K}\\ l<k}}\sum_{j\in\mathcal{J}_i}\sum_{n=1}^{\lfloor (D_{jkl}^I-L)/L\rfloor+1} \hat{a}_{ij} e^I_{j(l-k+1+nL)}\cdot z_{okj}\end{aligned} \tag{5.33}$$

with $\lfloor D_{jkl}^L/L\rfloor = \lfloor (N_j^I - (l-k+1))/L\rfloor$ for the number of overlapping cycles on day $l \in \mathcal{L}$ when a surgery group is scheduled on day k and $l \geq k$ and $\lfloor (D_{jkl}^I - L)/L\rfloor + 1 = \lfloor (N_j^I - (l-k+1+L))/L\rfloor + 1$ the number of overlapping cycles on day $l \in \mathcal{L}$ when $l < k$. The expected number of required beds on day l is given by the sum over all surgery groups of the probability that a patient from surgery group $j \in \mathcal{J}$ is in an ICU on day l, accounting for all cycles, multiplied by the number of times this surgery group is scheduled in all OT blocks (o,k). Similarly, we obtain:

$$\begin{aligned}\bar{\gamma}_{wl} &= \mathbb{E}\left(\hat{H}_{wl}^W\right) \\ &= \sum_{o\in O}\sum_{\substack{k\in\mathcal{K}\\ l\geq k}}\sum_{j\in\mathcal{J}_w}\sum_{n=0}^{\lfloor D_{jkl}^W/L\rfloor} b_{wj} e^W_{j(l-k+1+nL)}\cdot z_{okj} \\ &+ \sum_{o\in O}\sum_{\substack{k\in\mathcal{K}\\ l<k}}\sum_{j\in\mathcal{J}_w}\sum_{n=1}^{\lfloor (D_{jkl}^W-L)/L\rfloor+1} b_{wj} e^W_{j(l-k+1+nL)}\cdot z_{okj}.\end{aligned} \tag{5.34}$$

Since $\sum_n \hat{a}_{ij} e^I_{j(l-k+1+nL)}$ and $\sum_n b_{wj} e^W_{j(l-k+1+nL)}$ are constant, the new objective function is linear in the decision variables z_{okj}. Again, we want to obtain the maximum and minimum of both $\bar{\gamma}_{il}(\psi)$ and $\bar{\gamma}_{wl}(\psi)$ to determine the variation in bed occupancy during the week. The maximum and minimum operator are not linear. Therefore, we add the following constraints:

$$\bar{\gamma}_i^{\max} \geq \bar{\gamma}_{il}, \qquad \forall i \in \mathcal{I}, l \in \mathcal{L}, \tag{5.35}$$
$$\bar{\gamma}_w^{\max} \geq \bar{\gamma}_{wl}, \qquad \forall w \in \mathcal{W}, l \in \mathcal{L}, \tag{5.36}$$
$$\bar{\gamma}_i^{\min} \geq -\bar{\gamma}_{il}, \qquad \forall i \in \mathcal{I}, l \in \mathcal{L}, \tag{5.37}$$
$$\bar{\gamma}_w^{\min} \geq -\bar{\gamma}_{wl}, \qquad \forall w \in \mathcal{W}, l \in \mathcal{L}. \tag{5.38}$$

Additionally, let

$$\hat{\gamma}_i = \bar{\gamma}_i^{\max} + \bar{\gamma}_i^{\min}, \qquad \forall i \in \mathcal{I}, \tag{5.39}$$
$$\hat{\gamma}_w = \bar{\gamma}_w^{\max} + \bar{\gamma}_w^{\min}, \qquad \forall w \in \mathcal{W}. \tag{5.40}$$

The resulting MILP model is now given by:

$$\max \sum_{o \in \mathcal{O}} \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \mu_j \cdot z_{okj} - \sum_{i \in \mathcal{I}} \theta_i \hat{\gamma}_i - \sum_{w \in \mathcal{W}} \theta_w \hat{\gamma}_w \tag{5.41}$$
$$\text{s.t. } (5.1) - (5.3), (5.4) - (5.6), (5.29) - (5.40)$$

We refer to this problem as the linear OT schedule problem, which is NP-hard as proven by [231]. Note that a solution obtained by solving the linear OT schedule problem will still be evaluated by using the original objective function (5.24).

5.4.2 Local search approach

Similarly to [23], [107], [22], and [231], we use SA as local search approach. First, we explain how we define neighbor solutions, and then, we describe how we determine the cooling scheme. To obtain feasible neighbor solutions, we set a generator function that uses the current solution as input and produces a new solution. We consider four strategies to generate a neighbor solution:

- *Removing a surgery group*
  We find a neighbor solution by removing one surgery group from a certain OT-day. To find a feasible new solution, it is important to only remove a surgery group if it is scheduled more often than the required minimum amount.
- *Adding a surgery group*
  Similarly, adding one surgery group to a certain OT-day leads also to a neighbor solution. To find a feasible new solution, it is important to only add surgery groups from the specialty assigned to the selected OT-day and to check if adding this surgery group does not violate the overtime constraint.
- *Swap two OT blocks*
  Similar to [23], [22], and [231], we define neighbor solutions by swapping two OT blocks including their surgery groups. This can only be done if they have the same available time for surgeries and the same specialty can operate in the OTs.

We do not swap two OT blocks that take place on the same day, because this leads to a symmetric solution.

- *Swap two groups*
  Similar to [107], we define neighbor solutions by swapping two surgery groups that have been scheduled in the current solution. They can only be swapped if either the OT or the day on which they are scheduled is different. Furthermore, the new solution is only feasible when surgery groups from the same specialty are swapped and the overtime constraint is not violated.

Per iteration, one strategy is selected with equal probability $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ to find the next neighbor solution.

We follow a similar approach as [231] to select appropriate values for the initial temperature T_{in} and final temperature T_f. In our preliminary tests, we used $\theta = \theta_w = \theta_i$, so each ward was given the same weight. The maximum possible decrease of the objective function is given by $\max_{j \in \mathcal{J}} \mu_j + \theta$, which depends on the parameters θ and the surgery groups $\mathcal{J}$. At the start of the procedure, we want to accept this maximum decrease with probability 0.5. Thus, the initial temperature is given by:

$$T_{in} = \frac{-\left(\max_{j \in \mathcal{J}} \mu_j + \theta\right)}{\ln(0.5)}. \tag{5.42}$$

We determine the final temperature using the same approach. Near the end of the procedure, we want to accept negative changes in the objective function with a low probability. This way, the procedure converges to a local minimum. Our minimum negative change is given by removing the surgery group with the shortest surgery duration, while not influencing the variation in bed occupancy. We set the probability of accepting this change to 0.001 and this gives:

$$T_f = \frac{-\min_{j \in \mathcal{J}} \mu_j}{\ln(0.001)}. \tag{5.43}$$

Next to the initial and final temperature values, we also need to set the reduction factor, the number of iterations per temperature and the maximum number of accepted solutions per temperature. We used sensitivity analysis to determine the best combination of parameters considering both computational time and solution quality.

5.5 Computational results

In this section, we present the results of our two approaches. To compare the performance of the global approach and the SA approach, we use a real-life data set. This data includes a master surgery schedule where each OT block is assigned to a specialty. The cycle length is 14 days with 13 OTs where 9 surgical specialties operate. We have 11 wards and one ICU. Data was gathered from interviews with surgeons involved with planning, OT management and the hospital data warehouse. As a result of missing time stamps, 75% of the data set is used. For each surgery group obtained from the data, the mean and variance in surgery duration and LoS are determined. Furthermore, we determine the probability of patients from a surgery group going to ICU $i \in I$ and ward $w \in \mathcal{W}$. With the model described in Section 5.3.4, we determine the bed usage distribution resulting from scheduling the surgery groups. The changeover time between surgeries is set to 15 minutes.

In the global approach, we calculate the objective function value differently from the SA approach. The objective function of the MILP is an approximation of the original objective function and only depends on the expected number of beds, while the SA approach considers the original objective function. In order to make a fair comparison, we also determine the original objective function value for the solution given by the MILP. We also combine both approaches by starting with the global approach and then try to improve that solution with SA. Finally, we compare the performance of the best solution of both approaches with the performance of the real-life data set in Section 5.5.6. For analysis, we also consider computation time as a performance indicator.

We start this section with the results of our clustering approach and parameter settings for both the global and local approach. In Section 5.5.3, we compare the results of both approaches and try to further improve the value of the objective function by combining both approaches. In Section 5.5.5, we compare the result of our approach with the commonly used block scheduling approach. Finally, we validate our model using historical data in Section 5.5.6.

Solving the MILP model is done by using version 4.2.3 of AIMMS. For our MILP model, we use CPLEX version 12.6.3. The SA procedure is implemented in MATLAB R2016b. All computational experiments are performed on a PC with an Intel Core i7 6700K 4.20 GHz with 16 GB RAM.

5.5.1 Clustering

For each specialty, we use the clustering approach as described in Section 5.3.1. First, we determine the threshold between the short stay group and the long stay group per surgical specialty. The procedures with a median LoS of less than the threshold are denoted as short stay, while the procedures with a median LoS higher than the threshold are in the long stay group. Next, each LoS group is divided into three surgery groups based on the surgery duration of the surgery types. Two thresholds are determined and procedures are put into a short, medium or long surgery duration group depending on the mean surgery duration. However, some LoS groups did not contain enough procedures to be split into three significantly different surgery duration

groups. In these cases, only two surgery duration groups are defined. This approach leads to a total of 62 different surgery groups. Four medium surgery duration groups have a precision of less than 0.6 and all belong to different specialties. For these groups, the interval between the two thresholds defining the three surgery duration groups is small (less than 30 minutes). Therefore, the mean surgery duration of certain procedures may fall into the interval between the two thresholds, but many realized instances are outside these bounds, which leads to a low precision. However, the three surgery duration groups have significantly different means, and therefore, our method does define three groups instead of two. Defining less thresholds would increase cluster variance, and therefore, we decided not to adjust our clustering approach for groups with low precision.

5.5.2 Parameter settings

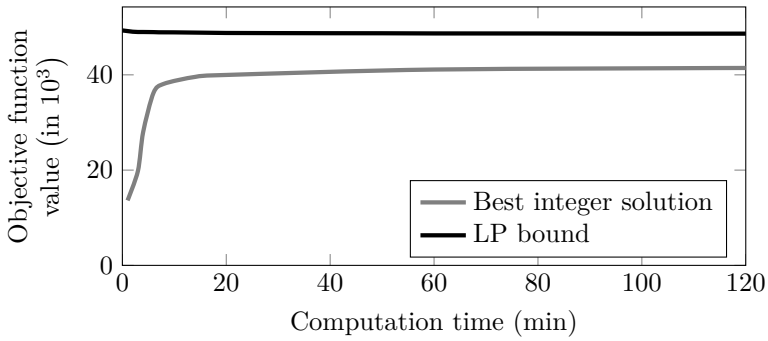
In this section, we discuss the input parameters for both the global and SA approach.

Global approach

In our MILP model, we only have to define the input parameters based on managerial decisions. These consist of parameter α that denotes the overtime probability and parameters θ_w and θ_i to balance the OT utilization and the variation in bed usage at the wards. Preliminary results indicate that setting $\theta_w = \theta_i$ provides the best trade-off between the variation in required number of beds and OT utilization. This means that we would remove scheduled surgery groups with a total OT time of 500 minutes if this would reduce the variation in required number of beds by one. The overtime probability α is set to 0.3.

In Figure 5.3, the LP bound and current best solution are shown when increasing the computation time. We see that the solution improves with longer computation times, however, the speed of improvement decreases rapidly after 20 minutes. Since we are creating a tactical schedule, which in theory should only be calculated a couple of times per year, we decided to set the computation time to 90 minutes.

Figure 5.3: The Value of the LP Bound and Best Integer Solution With Increasing Computation Time



SA approach

As initial solution for SA, we use the incumbent solution obtained after solving our MILP for 60 seconds. Furthermore, we have to set the following parameters: the initial temperature, reduction factor, final temperature and the maximum number of iterations within one temperature. As in Section 5.5.2, we use $\alpha = 0.3$ and $\theta_w = \theta_i = 500$. Table 5.1 gives an overview of the parameter settings for the SA approach. Using our data, $T_{in} \approx 1000$ and $T_f \approx 5$. However, preliminary results showed that we should set the stopping temperature to $T_f < 1$ to make sure SA converges to a local optimum. Furthermore, the preliminary results showed that we should set the number of iterations for one temperature, given by ω, to 450 and the maximum number of new solutions accepted for one temperature, denoted by ω_{new}, to 150 to obtain acceptable solutions.

Table 5.1: Parameter Setting for SA Approach

Symbol	Value	Description
T_{in}	1000	Initial temperature
T_f	1	Final temperature
ρ	0.97	Reduction factor
ω	450	Number of iterations for one temperature
ω_{new}	150	Maximum number of new solutions accepted for one temperature

5.5.3 Comparing the global and local approach

We compare the best solutions of both approaches to determine which approach performs best, using five key performance indicators (KPIs): (1) objective value, (2) OT utilization, (3) total number of used beds, (4) total difference in used beds during the cycle, and (5) computation time. The objective function values for both the MILP and SA are calculated using the 90-percentile and 85-percentile of the probability distribution of the number of required beds.

Given the parametrization used in our SA procedure, SA is slower than the MILP approach. The best obtained solution is shown in Table 5.2 and required seven hours to compute. This can be explained by the large amount of convolutions needed to calculate the objective function value. Recall that we set the computation time of the MILP to 90 minutes.

Table 5.2: Results for The Best Solution of MILP and SA Procedure With 90-percentile

KPI	MILP	SA
Objective value	41778	38699
OT utilization	0.839	0.855
Number of beds	152	159
Difference in beds	12	20
Computation time (hr)	1.5	7

The MILP also performs best compared to the SA approach for the 85-percentile (see Table 5.3).

Table 5.3: Results for The Best Solution of MILP and SA With 85-percentile

KPI	MILP	SA
Objective value	41278	36518
OT utilization	0.839	0.843
Number of beds	146	149
Difference in beds	13	23
Computation time (hr)	1.5	6

5.5.4 Improving MILP solution with SA

We also test whether the MILP solution can be improved by the SA procedure. With the initial temperature at $T_{in} = 1000$, we did not obtain better solutions. Therefore, we analyzed different initial temperatures. Results improve slightly for $T_{in} = 10$: the OT utilization improves with 0.66 percentage point to 84.57% and the variation in required number of beds decreases by 1 bed to 11 beds. We need an additional 30 minutes of computation time to obtain this solution.

5.5.5 Scheduling surgical specialties instead of surgery groups

In our introduction, we state that scheduling surgery groups instead of surgical specialties reduces the OT overtime probability and variation in bed usage. In Section 5.4.1, we have already shown that under the assumption that there exists a surgery group $j \in \mathcal{J}_s$ with $\mu_j \leq \mu_s$ and $\sigma_j^2 \leq \sigma_s^2$, we can schedule at least the same number of surgeries in one OT block when compared to scheduling surgical specialties. This assumption holds for our data.

In addition, our results show that we can even schedule more surgeries when scheduling surgery groups instead of surgical specialties. If we evaluate the solution obtained by scheduling surgery groups on data on surgical specialty level, we see that the OT utilization increases from 84.57% to 100.71%, which means that the obtained solution is not feasible when aggregating the data on surgical specialty level. In addition, we see that the bed variation increases from 12 to 36 beds and the maximum number of required beds increases from 152 to 192. This means that by scheduling the same number and type of surgeries, we need to reserve more OT and bed capacity when aggregating the data on surgical specialty.

Next to this, if we schedule surgical specialties, we see that we cannot meet the restriction on the minimum number of surgeries that should be scheduled per surgical specialty. By relaxing this constraint, i.e., by setting $\beta_s := 0.75\beta_s$, we do obtain a feasible solution with an OT utilization of 61.53%, bed variation of 23 and maximum number of required beds equal to 122. This means that this solution performs worse in terms of maximizing OT utilization and minimizing bed variation and that it is not feasible according to our original constraints.

5.5.6 Historical versus model performance

The average OT utilization registered in the data set was 71%. The weekly variation in bed occupancy over all wards was 53 beds. When we compare this with our best solution, the variation in bed occupancy can be improved by 42 beds, while the OT utilization can be improved to 84.57%. Given that the available OT capacity has not changed, these results show that more surgeries can be performed while the variation in the number of required beds decreases. In Table 5.4, we see the historical mean bed variation and the bed variation resulting from our best solution for each ward and the ICU. We also see that for each ward the variation in bed occupancy decreases, or in case of long stay ward 8, stay the same.

Table 5.4: Comparison Between The Historical Mean Bed Variation and The Bed Variation Results

Ward	Bed variation historical	Bed variation model
Day treatment	9	0
Weekday ward	13	2
Long stay 1	5	0
Long stay 2	4	2
Long stay 3	4	1
Long stay 4	5	2
Long stay 5	2	1
Long stay 6	1	0
Long stay 7	2	0
Long stay 8	4	4
ICU	4	1

5.6 Problem and model variants

To show the robustness and potential of our model, we analyze several variants of the model using minor modifications. In the first variant of our model, discussed in Section 5.6.1, we try to avoid occupied beds during the weekends at the weekday ward, as this ward closes during the weekends. In Section 5.6.2, we describe and test a variant of our model that minimizes the total number of required beds instead of the variation in number of required beds. In Section 5.6.3, we analyze a relaxation of the MILP. Finally, we apply our MILP model to data from another hospital in Section 5.6.4.

5.6.1 Closure during the weekends

The weekday ward (WDW) is intended to be only used during weekdays. When there are still patients admitted at this ward when the weekend starts, these patients have to be transferred to other wards. This setting has not yet been included in our current model. However, we could schedule the surgery groups in such a way that no patients

are admitted at the weekday ward during the weekend. We do so by adding a penalty Q for each patient admitted at the weekday ward during the weekend. We introduce variable r which denotes the number of patients admitted at the weekday ward during the weekend. The weekday ward is abbreviated by WDW and the Saturdays and Sundays in the planning horizon are given by set $\mathcal{L}_r \subset \mathcal{L}$. The resulting MILP model is given by

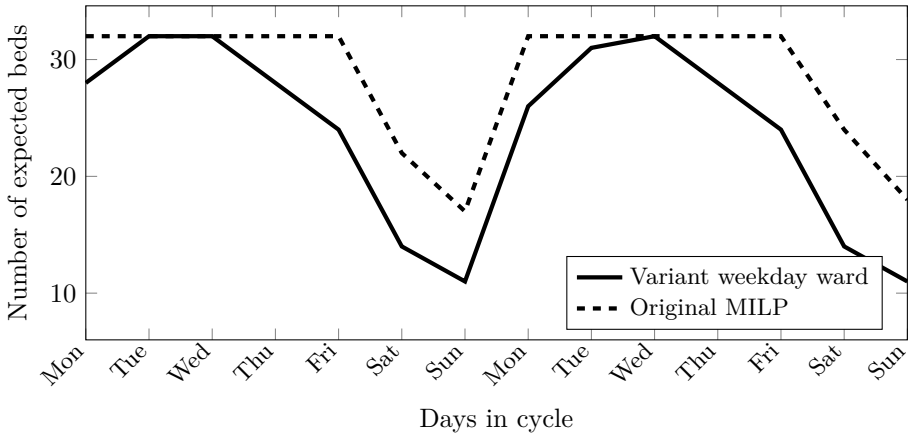
$$\max \sum_{o \in \mathcal{O}} \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \mu_j \cdot z_{o,k,j} - \sum_{i \in \mathcal{I}} \theta_i \hat{\gamma}_i - \sum_{w \in \mathcal{W}} \theta_w \hat{\gamma}_w - Q \cdot r \tag{5.44}$$

$$\text{s.t. } (5.1) - (5.3), (5.4) - (5.6), (5.29) - (5.40)$$

$$r \geq \bar{\gamma}_{WDW,l}, \qquad \forall l \in \mathcal{L}_r \tag{5.45}$$

In Figure 5.4, the results of the model with $Q = 10000$ and a computation time of 90 minutes is compared to the best solution obtained by the initial MILP model. We see that the expected number of beds during the weekend is reduced, but does not reach zero. It also affects the OT utilization, which decreases by 7.5 percentage point and the difference in beds, which increases by 12 beds. For higher values of Q, the results for the weekday ward do not improve. These results can be explained by the fact that each surgery group has to be scheduled a minimum number of times. In the solution provided when $Q = 10000$, each surgery group for which patients are admitted to the weekday ward are scheduled the minimum number of times. However, the used surgery groups are not the best predictor for the ward the patients need to be admitted, because every surgery group has some probability that a patient will be admitted at the weekday ward. Therefore, always some patients will be admitted at the weekday ward during the weekend given the used surgery groups.

Figure 5.4: The Expected Number of Beds at The Weekday Ward



5.6.2 Minimize the number of beds

In our model, we minimize the variation in the number of required beds. However, personnel to keep the beds open is expensive. Therefore, instead of minimizing the variation in bed usage, we can also minimize the number of required beds. Even though there might be more variation in bed usage, the number of required beds may decrease.

To minimize the number of required beds, we modify our linear model described in Section 5.4.1. In the modified model, we use the maximum values of $\bar{\gamma}_{i,l}$ and $\bar{\gamma}_{w,l}$ instead of using the difference between the maximum and minimum values of $\bar{\gamma}_{i,l}$ and $\bar{\gamma}_{w,l}$. The resulting MILP is:

$$\begin{aligned}
&\max \sum_{o\in\mathcal{O}}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} \mu_j z_{o,k,j} - \sum_{i\in\mathcal{I}} \theta_i \hat{\gamma}_i - \sum_{w\in\mathcal{W}} \theta_w \hat{\gamma}_w && (5.46)\\
&\text{s.t. } (5.1)-(5.3), (5.4)-(5.6), (5.29)-(5.34)\\
&\quad \hat{\gamma}_i \geq \bar{\gamma}_{i,l}, \qquad \forall i \in \mathcal{I}, l \in \mathcal{L}\\
&\quad \hat{\gamma}_w \geq \bar{\gamma}_{w,l}, \qquad \forall w \in \mathcal{W}, l \in \mathcal{L}.
\end{aligned}$$

The results for this variant of the model are shown in Table 5.5. We cannot compare objective function values, since different objective functions are used for both methods. The variation in number of required beds is a lot higher, as was to be expected. However, the OT utilization is also higher while less beds are needed in total.

Table 5.5: Results for Two Variants of The Model: (1) Minimizing Variation in Bed Utilization and (2) Minimizing The Number of Beds

KPI	Minimize variation	Minimize required beds
OT utilization	83.9%	85.2%
Number of beds	152	147
Difference in beds	12	26

5.6.3 Scheduling without blocks

Our model uses the OT blocks of the MSS as input. This means that surgery groups can only be scheduled within the OT blocks that are allocated to this surgical specialty. In this variant of the model, we relax our model by excluding the MSS, meaning that every surgery group can be scheduled at any day within the cycle. In Table 5.6, we see that the complexity of the problem increases when not using an MSS. After 90 minutes of computation time, the optimality gap is still 46%. The solution has a lower objective function value than the solution of the basic model, which has an objective function value of 41778, an OT utilization of 83.9%, and a variation in number of required beds of 12. After six hours, the found solution has a higher objective function value than the solution of the basic model. The OT utilization has improved, but the variation in the number of required beds has increased.

Scheduling without OT blocks shows advantages, but ignores many other factors that affect the MSS, e.g. schedules of surgeons, staff and equipment availability, and therefore, implementation is challenging.

Table 5.6: Results of Telaxation Variant Disregarding The Blocks in the MSS

KPI	90 min	360 min
Objective value	40561	41847
OT utilization	87.0%	87.5%
Number of beds	163	162
Difference in beds	18	16
Optimality gap	45.8%	39.3%

5.6.4 Applying the model to different instances

To analyze whether our approach also works for other real-life instances, we obtained a data set from another hospital. This data set contains 43 surgery groups for which the minimum, mean and standard deviation and LoS probability distributions per surgery group are given. Furthermore, one ward is taken into account. We use the 95-percentile to calculate the required number of beds. The results can be found in Table 5.7 for different computation times. The same data set and overtime probability are used as in [36]. Their decomposition approach consists of: (1) maximizing the OT utilization and (2) minimizing the required number of beds. The best solution found in [36] yields an OT utilization of 91% that needs 45 beds in total. Our solution has 1.4 percentage point lower OT utilization, however, the required number of beds decreases by 6 beds. In [36], the OT blocks are formed beforehand, so there is no flexibility in assigning surgery groups to OT blocks when minimizing the required number of beds.

Table 5.7: Results From Other Hospital Data Set

KPI	10 min	90 min
OT utilization	86.7%	89.6%
Number of beds	40	39
Difference in beds	3	3
Optimality gap	15.4%	12.2%

5.7 Discussion

In this chapter we show the positive impact of the holistic perspective on surgery scheduling. We introduce two single step approaches for scheduling surgery groups while taking into account the overtime constraint and maximizing the OT utilization and minimizing variation of bed usage. Scheduling surgery groups instead of OT blocks leaves fewer options on an operational level to schedule surgeries, and therefore, the probability of overtime and variation in bed usage as a result of surgery scheduling on an operational level decreases. We also added weights to the objective function, θ_i and θ_w, to balance the managerial trade-off between variation of bed usage and OT utilization.

We compare two approaches for finding a good feasible solution for large real-life instances. Both on computational results and on computation time, the MILP outperforms the SA approach. We also combine both approaches where we first optimize the MSS with the MILP, and then try to further improve the objective function value with SA. This combination leads to slightly better results. The MILP shows good results for large real-life instances without long computation times and is therefore suitable for practical applications. Comparing the results of the model with the historical performance derived from the data set, the variation in beds is improved from 53 beds to 11 beds and the OT utilization can be improved from 71% to 85%.

With the use of the MILP, we also analyze some model variants that can give more managerial insights. The first variant focuses on closing the weekday ward during the weekend, as in practice, weekday wards are only opened during weekdays. Weekday wards often struggle with patients that are still admitted during weekends. Without changing the surgery planning, the ward management has two options to solve this problem: (1) extend the opening hours of the weekday ward or (2) transfer these patients to other wards on Friday. So, in the first variant of our model, we extended the model by including a penalty in the objective function for patients being admitted on a weekday ward during the weekend. The computational results show that it is difficult to close such wards during weekends. The next variant of the model minimizes the usage of beds instead of the variation of bed usage which can be achieved by modifying the objective function. Results show that with this approach, the number of required beds can be further reduced and OT utilization increased. However, this also results in an increase in the variation of bed usage.

Furthermore, we relaxed our model such that every surgery group can be scheduled in any OT block in the cycle. The results show that OT utilization can be improved at the cost of an increase in bed variation and required number of beds. To analyze the robustness of our model, we compare our model with another solution approach and data set. Results show that our model has 1.4 percentage point lower OT utilization, however, the required number of beds decreases by 6 beds.

An important step in our approach is the clustering of surgery types into surgery groups. Our clustering approach has a major effect on the group variation, in terms of surgery duration and length of stay, and possible destination wards. With the surgery groups used for our model, we were not able to close the weekday ward during the weekend, because too many surgery groups may use the weekday ward after surgery. Therefore, we conclude that the groups at hand are still too aggregated for this model variant. Further research on applying data mining on such instances could increase the predictive value of clusters (in our case surgery groups), and therefore, improve the robustness of planning.

When clusters are only based on specialty, we obtain a model which schedules OT blocks similar as previous work ([81, 231, 239]). As shown here, our clustering approach results in more precise predictions for surgery duration and LoS, and therefore, results in a higher OT utilization and lower variation in bed usage. However, smaller clusters (e.g. clustered surgery types versus clusters based on specialty) require more data to attain similar precision levels. Therefore, our approach assumes no limitations in data availability. Since most hospitals nowadays have advanced electronic health record systems, this would be a fair assumption to make. Furthermore, this type of data is

transient and thus the data and the model should be analyzed repeatedly (e.g every year). Next to possible data limitations, our model assumes that the clusters also account for the biweekly number of realizations of surgery types (e.g. at least once every two weeks) such that each block can be filled with surgeries of that type on the operational level. Furthermore, the dispersion of durations within surgery groups should be limited. When this is not the case, it could result in under- or overutilization of resources, since on the operational level, surgery types with significant shorter or longer individual durations could be scheduled than was accounted for when scheduling surgery groups on the tactical level.

The model can also be extended to optimize the schedule of surgeons. To achieve this, the model should not only take the OT and its downstream resources into account, but also its upstream resources such as the outpatient clinic given that surgeons also work there. To realize this potential in practice is a difficult task, as this requires discipline from specialists concerning their schedule. Another potential direction for further research is optimizing break-in moments for OT cleaning.

Overall, this research provides a way to bridge the gap between tactical and operational planning of surgeries. It reduces the variation in bed usage and improves the robustness of the schedules. The use of surgery groups makes it possible to easily implement our model into practice, and for operational planners, it is instantly clear where to schedule what type of surgery. With only minor model modifications, we show that a broad range of variants on OT scheduling can be analyzed to obtain valuable managerial insights.

CHAPTER 6

The Hospital Online Multi-Appointment Scheduling Problem[1]

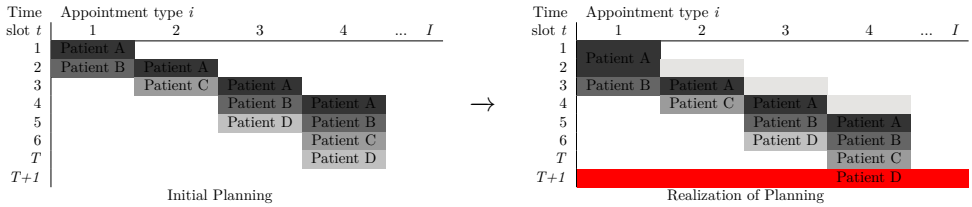
6.1 Introduction

The value-based healthcare (VBHC) paradigm has received increasing attention [95] since the early 2000s [194]. Worldwide, healthcare organizations aim to implement VBHC to attain better outcomes for their patients, with value defined as patient health outcomes (e.g. mortality, or patient safety) per dollar spent [191]. This is envisaged as a shift from volume to value and requires shared accountability among involved healthcare providers. Process measurements and improvements are important conditions for creating value. To operationalize the VBHC concept, integrated practice units (IPUs) were proposed among other things, in [192]. In that study, an IPU is defined as: "a dedicated team made up of both clinical and nonclinical personnel providing the full care cycle for the patient's condition". Implementing IPUs requires redesign of the organizational structure, shifting from a focus on specialties or interventions to an emphasis on total care pathways that encompass all services and activities that jointly determine success in meeting a set of patient needs. The usual minimal coordination between different schedules involved in the same care pathway results in large fluctuations in downstream resources, last-minute adjustments or even cancellations. Hospitals struggle to implement IPUs and therefore often fail to deliver and measure value [193]. Designing and organizing IPUs and their care pathways is challenging, as all appointments at different resources should be optimally scheduled for all pathways to guarantee equitable access and waiting times for all patients. To overcome these problems, IPUs may organize integral coordination of appointment planning (i.e. multi-appointment planning), to optimally align each step of the healthcare delivery process. Therefore, a single administrative and scheduling organization for each IPU should be in place [192]. In this chapter we analyze the scheduling problem of IPUs from an online multi-appointment scheduling perspective.

Appointment planning balances trade-offs between efficient resource utilization and waiting time [155, 258]. Here, we describe the different phases of appointment plan-

[1]This chapter is based on A.J. Schneider, J.W.M. Otten, M.E. Zonderland, R.J. Boucherie and M.J. Schalij. The Hospital Online Multi-Appointment Scheduling Problem. *Working paper*.

Figure 6.1: Illustration of the Lack of Robustness of Multi-Appointment Scheduling: First Appointment of Patient A is Extended and Affects All Subsequent Appointments and Schedules.



ning based on the top-down organizational decision hierarchy illustrated in [106]: (1) the strategic case-mix planning, (2) tactical master scheduling and (3) operational appointment scheduling. In strategic case-mix planning, hospitals determine the required capacity levels for appointments (e.g. outpatient clinics and diagnostic facilities) based on the desired patient case-mix. Furthermore, capacity is roughly divided among medical specialties, including physicians, on this level. Next, capacity is cyclically scheduled, meaning a schedule is repeated after a cycle length (e.g. weeks or months) and results in a master schedule. In a master schedule the relations between different types of capacity, as a result of patient flows, should be considered to balance the load for each capacity. Therefore, multi-appointment scheduling is considered on the tactical level. Third, operational appointment scheduling involves scheduling of patients to appointment sessions. Sessions define the horizon in which an appointment can be scheduled and typically have a length of a day or half a day. This final stage is further divided into the offline (appointments are scheduled at a later point in time) and online scheduling (appointment requests are instantly scheduled).

In this chapter we discuss a specific instance of the operational online multi- appointment scheduling problem in hospitals. During the day, all sorts of events may occur that have an impact on the realization of appointment schedules. In our day-to-day work, we observe that hospitals lack online coordination, which increases the possibility of overtime. This lack of coordination has even greater impact on multi-appointment schedules and therefore the possibility of overtime increases. Here, we give an example of this impact on multi-appointment scheduling: a patient's first appointment takes longer than expected and thus the session is likely to run into overtime. In a multi-appointment setting, this could also have an impact on the patient' subsequent appointments in other sessions. This means that other sessions may also run into overtime as this patient will not arrive on time for the next appointments. Thus, the delay of a single appointment may result in overtime for all other sessions related to this patient. The robustness of multi-appointment schedules is therefore fragile. We graphically show this phenomenon in Figure 6.1. We assume appointments are optimally scheduled offline such that arrivals are equally distributed during the day and available resources and demand are balanced. We analyze the problem of online multi-appointment scheduling and rescheduling given the actual status of the system and expected future arrivals during an appointment session. Incorporating such system dynamics (see example in Figure 6.1) in our strategies and in line with

other appointment scheduling problems, we aim to minimize both the patient' sojourn time and resource overtime. As mentioned, patients have a number of appointments scheduled, and appointments can be arranged in a different order during the session. To overcome model intractability, we propose a decomposition approach for two decisions: (1) acceptance of arrivals and (2) scheduling of patients to appointments. Given the lack of robustness of multi-appointment schedules, there there is a possibility that new arrivals will not finish all appointments within the session and will have to come back for a later session. We assume that it is preferable to reschedule all appointments of these arrivals and therefore take this acceptance decision into account. For the acceptance decision, we developed a finite-horizon discrete-time Markov decision process (MDP) model to analyze optimal online policies minimizing costs for rejecting arrivals and for the "unfinished" patient at the end of the horizon. For the scheduling decision, we developed an ILP that schedules patients one time slot ahead and therefore has maximum flexibility in sequencing the appointments of a patient. This means that we will not consider the order of appointments scheduled offline. We combine both decisions as follows: at each decision epoch we determine the optimal policy for acceptance or rejection of arrivals and the ILP determines at each decision epoch the optimal schedule for all accepted patients one decision epoch ahead.

We start this research by providing an overview of available literature on online multi-appointment scheduling in healthcare and we position our research (Section 6.2). In Section 6.3, we present the formal problem formulation of our approach. In Section 6.4, we present an implementation of our model for a real-life instances based on a case study. We also show, in Section 6.4.4, a full implementation in practice of our solution and analyze its impact. Finally, we discuss the implications of our approach in Section 6.5.

6.2 Literature Review & Research Positioning

Literature on appointment scheduling problems in healthcare are broadly available. Systematic reviews are given in [6, 25, 51, 103, 118]. Here, we solely consider research analyzing online multi-appointment schedules. For this, we use the recently available systematic reviews of [147] and [166]. We categorize the available literature based on modeling approach, namely: markov decision process, integer linear programming and simulation.

6.2.1 Markov Decision Process

To optimize online multi-appointment scheduling for a cardiac diagnostic testing center, [67] developed a finite-horizon, discrete-time MDP. The authors use a heuristics approach to develop a real-time decision support system as the MDP became intractable. The authors of [203] formulated a discounted infinite-horizon MDP for scheduling cancer treatments in radiation therapy units. The authors approximate the optimal policy with approximate dynamic programming (ADP) and solve an equivalent LP model with column generation. Another application of MDP is used to schedule multidisciplinary, multistage appointments at a bariatric clinic in [71]. They also use

an ADP approach to deal with intractability. Allocating capacity to patients at the moment of their arrival at a rehabilitation clinic, to maximize the total number of requests booked within their corresponding access time targets, is an approach analyzed in [31]. The authors also use ADP to analyze real-life instances. In another study [30], the same authors use a decomposition approach to develop a tractable MDP model to analyze a radiation department appointment scheduling problem. In the first phase of their decomposition approach, appointment dates and linacs are assigned to incoming patients on the day they arrive (for treatment), taking into account future arrivals using an MDP model. In the second phase, specific appointment times are assigned to the patients on a weekly basis, taking into account time constraints and patient time preferences using an ILP model.

Clearly, most of the MDP models discussed here become intractable for real-life instances and therefore the authors use approximations. Our approach is in line with that is used in [30]. However, we analyze decisions within the same session. Furthermore, instead of scheduling all patients over the whole scheduling horizon, we schedule patients only one time slot ahead.

6.2.2 Integer Linear Programming

A MILP model is developed in [14], to optimize multi-appointment scheduling for pathology laboratory tests, where arrival times are unknown and appointments have partial precedence constraints. As the model becomes intractable, the authors develop a genetic algorithm to analyze their problem. To optimize online multi-appointment scheduling requests of a nuclear medicine clinic, the authors in [188] use an integer programming method in combination with scheduling algorithms. Two scheduling algorithms are assessed: (1) scheduling requests on arrival and (2) scheduling requests on arrival taking into account possible future requests. They use simulation to compare the performances of the two algorithms. The authors of [41] use an ILP approach to optimize online multi-appointment scheduling of a rehabilitation clinic and they use simulation to analyze solution performances.

The ILP models from the presented literature also become quickly intractable for analyzing real-life instances and therefore approximations are used. Furthermore, the ability to model uncertainty of arrivals and durations is limited using an ILP approach compared to an MDP approach.

6.2.3 Simulation

Simulation is by far the most used modeling approach for analyzing online multi-appointment scheduling problems. Multi-appointment schedules are complex. As a result, analytical models quickly become intractable. This is one of the reasons simulation is favored in this literature. The authors in [125] simulate different multi-appointment scheduling rules and load smoothing strategies to minimize the patient sojourn time at a breast cancer center. The authors of [48] use appointment type sequencing in care pathways for an orthopedic consultation suite and test its impact on patient waiting time and the percentage of consultations performed in overtime. For a nuclear medicine department, the authors of [187] analyze several multi-appointment

scheduling algorithms. In [70], researchers analyze multiple block schedules for multi-appointment schedules where blocks between different schedules are aligned (i.e. do not overlap). To analyze current bottlenecks in a radiotherapy department practice, the authors in [127] use simulation to show that the number of linacs and physician availability have a great impact on flow congestion in this practice. Another simulation study on radiation therapy planning conducted in [250] shows that decreasing radiation therapist capacity has the largest impact on extending delays, while improving the oncologist processing time has the largest positive impact on reducing delays. The authors of [153] and [168] both use simulation to analyze online multi-appointment scheduling for oncology centers.

Agent-based approaches are also often applied to analyze online multi-appointment scheduling systems. The authors of [185] use agent based modeling in which patient agents minimize their sojourn time and resource agents maximize utilization. In [245], a semi dynamic agent-based model is implemented in which appointments are initially scheduled without complete information about all plans that have to be scheduled, and a simulation model is used for online dynamic rescheduling to analyze plans with all information available. The authors in [116] use timed Petri nets to define care pathways and use simulation to find feasible schedules for these care pathways. Finally, in [174] the authors analyze the outpatient clinic of an eye care hospital with open access policy (solely unscheduled arrivals). They develop a hybrid ant agent algorithm to find good feasible paths for an appointment in the care pathway.

Simulation results are difficult to generalize as endless features can be modelled and therefore results present the performance of the specific instance that is analyzed. However, we will use simulation to evaluate the performance of our approach compared to a heuristic.

6.2.4 Conclusion

Online multi-appointment scheduling is a relatively new research topic and has received increasing attention. As mentioned earlier, models become quickly intractable when considering multiple features of real-life instances. For this reason, approximation approaches are widely used in the literature. Most studies consider online appointment systems in which requests for future sessions are instantly served. In this research, we consider the instant handling of arrivals and appointments that are scheduled within the same session. We use an MDP approach as it enables us to model the uncertainty of future arrivals and durations. Although the technical contributions of this research are limited, we present a novel decomposition approach for analyzing a novel online multi-appointment scheduling type of problem using MDP. We test this approach with real-life instances and compare our approach with a heuristic using discrete event simulation.

6.3 Model Formulation

For the problem at hand, we aim to determine how to optimally respond to random arrivals of new patients. Deriving such a policy involves sequential decision-making

under uncertainty, and an MDP is particularly suited for this. We discretize the scheduling horizon in equal time slots of 1 minute, where appointment durations have a length of multiple time slots and a maximum of 1 patient can arrive during the length of a time slot. In this section we formulate the problem as an MDP and propose a method that enables us to determine the optimal policy efficiently.

6.3.1 Model Formulation

We consider the acceptance and scheduling of patients during a single session, hence we use a finite horizon MDP. The session is divided into equal-length time slots. At the start of each time slot two decisions are made: (1) accept or reject a newly arrived patient and (2) which patient is assigned to which tests. All patients are given a type, which is defined by the remaining tests a patient needs to visit. After completion of a test, the patient will reenter the system as a different type of patient. As an example, suppose that a patient who needs to visit tests A and B is assigned to test A. After completion of this test, the patient reenters the system as a patient of the type that needs to visit test B. Patients leave the system once all the required tests have been performed. Costs are accrued for patients who are rejected and for patients who have not completed all required tests at the end of the session. Tn both cases, the costs increase with the number of required tests. We now formally describe the MDP.

Decision Epochs. Decisions are made at the start of each time slot t, $t = 1, \ldots, T$.

State Space. The state space is denoted by

$$\begin{aligned} S = \{&(x_1, \ldots, x_N, y_1, \ldots, y_M, z) | 0 \leq x_i \leq C_i, i = 1, \ldots n, \\ &0 \leq y_j \leq N, j = 1, \ldots, M, 1 \leq z \leq N\}, \end{aligned} \tag{6.1}$$

where x_i denotes the number of incumbent patients of type i, for $i = 1, \ldots, N$; y_j denotes the type of patient currently receiving test j, for $j = 1, \ldots, M$; and z denotes the type of patient arriving in the system.

Action Space. $A(s)$ is the set of all possible actions when the system is in state $s \in S$. As mentioned earlier, an action consists of two parts: (1) a decision is made whether or not to accept an arrival and (2) a decision is made as to which of the patients in the system are assigned to which tests. Whenever a test is available, a patient of the appropriate type can be assigned to it.

Transition Probabilities. Between successive decision epochs $t-1$ and t, the state of the system can change in three different ways. First, with probability $p_{t,i}$, a new patient of type i arrives. Second, a patient can complete a test, and then either leave the system or remain in the system if there are remaining tests to be visited. We assume the service time of test j to be geometrically distributed with mean $\frac{1}{q_j}$. Thus q_j is the probability that a patient finishes test j between decision epochs $t-1$ and t. Third, there are transitions as result of the chosen actions, for example, an arriving patient is accepted and added to the current patients and a patient already in the system is assigned to a vacant test.

Costs. The immediate cost of rejecting a new patient is defined as $r_t(s, \text{Reject}) = \sum_{j=1}^{M} k_r l_j z_j$, where k_r is a constant cost for rejecting a patient and l_j is a cost based on the service time of test j defined by the number of time slots. After the final decision epoch, $t = T$, a cost is accrued for any patient who did not complete service. This is denoted by $r_T(s) = \sum_{i=1}^{N} k_e x_i + \sum_{j=1}^{M} k_e \mathbb{1}(y_j)$, where k_e denotes the cost for a patient's not completing a test and $\mathbb{1}(y_j)$ is 1 if patient of type j needs to visit more tests after finishing the current test. The costs for not finishing a patient are independent of the patient type, as we assume rescheduling is not preferable for patients regardless of the number of appointments that need to be rescheduled.

Policy. A policy describes the decision that should be taken in state s at time t and is denoted by $\pi_t(s)$.

Value Function. The value function is denoted by $V_t^{\pi}(s)$ and gives the expected total reward, from time t onward, when the system is in state s and policy π is followed.

The problem reduces to finding an optimal policy π^* such that the following optimality equations are satisfied:

$$\begin{aligned} V_t^*(s) &= \max_{a \in A(s)} \left\{ r(s,a) + \sum_{s' \in S} P(s'|s) V_{t+1}^*(s') \right\} \quad \forall s \in S, \quad t = 1, \ldots, T-1, \\ V_T^*(s) &= r_T(s), \qquad \forall s \in S, \end{aligned} \tag{6.2}$$

where V^* is the optimal value function [195].

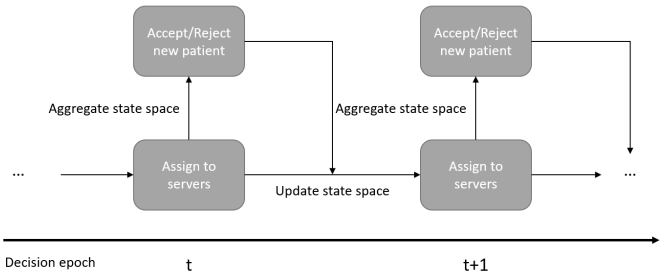
6.3.2 Solution Approach

To solve the described MDP we need to keep track of the number of patients in the system for each patient type. As a result the problem becomes intractable. This is caused by both the modeling approach (e.g MDP) and the definition of the patient types. The set of types consists of every possible subset of tests that are available. As an illustration, if there are six tests and we assume that at any given moment there are no more than five patients per type in the system, the number of states already exceeds 10^{50}. In this section, we describe a method to overcome this curse of dimensionality. To this end we decouple the two decision-making processes of the MDP: namely, the acceptance process of accepting and rejecting arriving patients, and the allocation process of assigning patients already in the system to the different tests. Our approach resembles that of [30], who also split a large scheduling problem into two subproblems, whith the output of one is part of the input of the other subproblem. Our approach differs in that we reject arriving patients and do not put them on a waiting list, and our allocation model is less complex. Instead of scheduling all patients over the whole time horizon we only assign patients to tests one time slot ahead. Alternatively, we could use ADP approach. However, applying standard ADP methods to the described MDP would not incorporate the specific two-decision-processes structure of our MDP. By decoupling a large decision process into an into acceptance and allocation models we aim to exploit this structure to obtain a good approximation of the MDP.

Combined Decision Process

In the next sections, we describe the two separate models for the acceptance and allocation of patients in the system. In this section we describe how these two models interact when applied in practice. During the whole decision process the complete state of the system, including the number of patients per type present in the system, the type of patient currently in service for each test and the type of patient requesting entry, should be tracked. At each decision epoch, this information is used as input for the allocation model. This model determines which patients should be assigned to which tests. Subsequently, the state space is aggregated by counting the number of patients who require service of a certain test. The acceptance model provides an optimal policy to either accept or reject an arriving patient based on this aggregated state. This optimal decision for the reduced model, together with the possible arrival and departure of patients leads to a new segregated state of the system for which the whole decision cycle starts over. The interaction of the two models is depicted in Figure 6.2.

Figure 6.2: Combination of the Acceptance and Allocation Decision Processes.



Acceptance Model

The goal of the acceptance part of the decision process is to determine an optimal policy to accept and reject patients, given the current state of the system and probabilistic knowledge about the future. For this we adjust the MDP described above in the following way. First, we decouple the allocation process from the MDP and second we aggregate the state space by keeping track of the number of patients requiring a certain test instead of keeping track of the specific type of patients. The modified MDP is described below.

Decision Epochs. Decisions are made at the start of each time slot t, $t = 1, \ldots, T$.

State Space. The state space is denoted by

$$S = \{(y_1, \ldots y_M, z_1, \ldots, z_M) | 0 \leq y_j \leq C_j, z_j \in \{0, 1\}, j = 0, \ldots, M\}, \tag{6.3}$$

where y_j denotes the number of patients in the system who require test j and z_j is a binary variable denoting whether an arriving patient requires test j or not.

Action space. The patient that requests access can be accepted or rejected. However, when accepting would lead to a state with more patients than the capacity C_j for a certain appointment, the patient must be rejected. Therefore, the action space is

$$A(s) = \begin{cases} \{\text{Accept, Reject}\}, & \text{if } \max_j(y_j + z_j) \leq C_j, \\ \{\text{Reject}\}, & \text{otherwise.} \end{cases}$$

Transition Probabilities. System transitions consist of three factors. First, we assume patients arrive according to a Poisson arrival process. As mentioned earlier, the length of a time slots t is small (e.g 1 minute), thus we can assume that at time t a new patient of type i can request admission with probability $p_{t,i}$. We assume that there are N patient types, $i = 1, \ldots, N$, and we characterize elements of vector $\bar{z}^i$ by the appointments required by a patient of type i, thus $\bar{z}^i = (\bar{z}_1, \ldots, \bar{z}_M)$. Second, a patient may finish service at test j, with probability q_j. Third, if a patient is accepted into the system, y_j is increased by one for each of the tests j the new patient requires.

$$\begin{aligned} &(y_1, \ldots y_M, z_1, \ldots, z_M) \to (\hat{y}_1 + z_1, \ldots \hat{y}_M + z_M, \tilde{z}_1, \ldots, \tilde{z}_M), && \text{if } a = \text{ Accept}, \\ &(y_1, \ldots y_M, z_1, \ldots, z_M) \to (\hat{y}_1, \ldots \hat{y}_M, \tilde{z}_1, \ldots, \tilde{z}_M), && \text{if } a = \text{ Reject}, \end{aligned}$$

where

$$\hat{y}_j = \begin{cases} y_j - 1, & \text{with probability } q_j, \\ y_j, & \text{with probability } 1 - q_j, \end{cases}$$

and $\tilde{z}_j$ denotes the requirements of the next arriving patient.

Costs. The immediate cost of rejecting a new patient is defined as $r_t(s, \text{Reject}) = \sum_{j=1}^{M} k_r l_j z_j$, where k_r is a constant cost for rejecting a patient and l_j is a cost based on the service time of test j defined by the number of time slots. After the final decision epoch, $t = T$, a cost is accrued for any patient who did not complete service. This is denoted by $r_T(s) = \sum_{j=1}^{M} k_e y_j$, where k_e denotes the cost for not completing a test.

Value Function. Now, inserting the described transition probabilities in Equation 6.2 reduces the problem to one to finding an optimal policy π^* such that the following optimality equations are satisfied:

$$V_t^*(s) = \begin{cases} \max \left\{ \sum_{j=1}^{M} k_r l_j z_j + \sum_{i=1}^{N} \lambda p_{t,i} V_{t+1}^*(\hat{y}_1, \ldots \hat{y}_M, \bar{z}^i), \right. \\ \quad \left. \sum_{i=1}^{N} \lambda p_{t,i} V_{t+1}^*(\hat{y}_1 + z_1, \ldots \hat{y}_M + z_M, \bar{z}^i) \right\} \text{ if } \max_j(y_j + z_j) \leq C_j, \\ \sum_{j=1}^{M} k_r l_j z_j + \sum_{i=1}^{N} \lambda p_{t,i} V_{t+1}^*(\hat{y}_1, \ldots \hat{y}_M, \bar{z}^i), \\ \quad \text{otherwise, } t = 1 \ldots, T-1, \end{cases} \tag{6.4}$$

$$V_T^*(s) = r_T(s), \qquad \forall s \in S, \tag{6.5}$$

where V^* is the optimal value function and $\lambda \leq 1$ is a discount factor. We discount the future costs by a factor λ in order to incorporate the effect of the allocation model on

the acceptance model. At the beginning of a session there are more options to schedule appointments than near the end. Furthermore, patients who require many tests are more difficult to schedule than patients who require only a few tests. However, after receiving partial service these difficult-to-schedule patients reduce to easier to schedule patients making it easier to reshuffle patients when the time until the end of the session is longer. These two observations together form the rationale behind the use of the discount factor in this MDP. We emphasize that it is not a standard discount factor and therefore we use the term scheduling factor.

Aggregating the state space reduces the number of states considerably. For instance, for six different appointments and at most five patients per server at any given time the number of states of the original MDP exceeds 10^{50}. For the aggregated MDP, this is reduced to 10^6, for which the MDP can be solved in reasonable time.

Allocation Model

The allocation part of the decision process consists of assigning patients to available servers. Since we decoupled the allocation and admission processes, the arrival process no longer plays a role in the allocation process. Therefore, the allocation problem reduces to a scheduling problem. At each decision epoch, a given set of patients, with appointments, are present in the system. To service as many patients as possible and to minimize the number of patients who did not finish all appointments, we need to maximize the occupancy of the servers and prioritize patients with more appointments to be completed. Although there may be approaches that are more efficient for this type of problem, we use the following ILP approach for allocating patients:

$$\begin{aligned} \max \quad & \sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_i u_{ij}, & \\ \text{s.t.} \quad & \sum_{j=1}^{M} u_{ij} \leq \beta_{ij} x_i, & 1 \leq i \leq N, \\ & u_{ij} \in \{0,1\}, & \end{aligned}$$

where $u_{ij} = 1$ if a patient of type i is assigned to server j and is 0 otherwise, α_i is a cost parameter for patients of type i, that increases with the number of appointments for a patient type, and β_{ij} is a parameter that is one if patients of type i require appointment j and zero otherwise. Parameter β_{ij} can also be used for precedence constraints between appointments (e.g. appointment A have to take place before appointment B).

Policy Structure

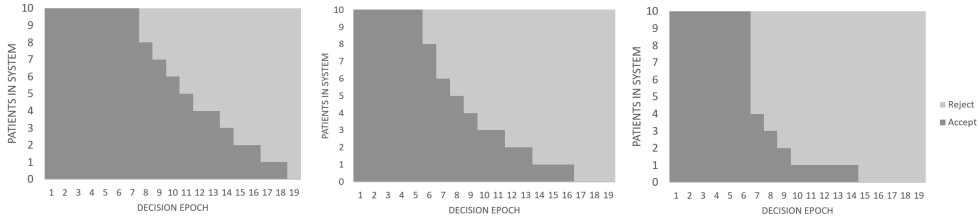
To show the structure of the optimal policy of the admission model we introduce a small example. Suppose there are three tests (A, B and C) and four types of patients. Table 6.1 lists the required tests per patient type. We assume that arriving patients are of either type with equal probability, that the expected duration of test A, B and C are two, three and eight time slots, respectively; and that the cost factors are $k_r = 1$

Table 6.1: Required Tests per Patient Type.

Patient type	Servers required
I	A
II	B
III	C
IV	A,C

and $k_e = 4$. Furthermore, the session is divided into 20 time slots and the scheduling factor is set at $\lambda = 0.9$. Figure 6.3 shows the optimal policy for patients of types I,

Figure 6.3: Optimal Acceptance Decision for Patients of Type I (left), type II (middle) and type III(right).



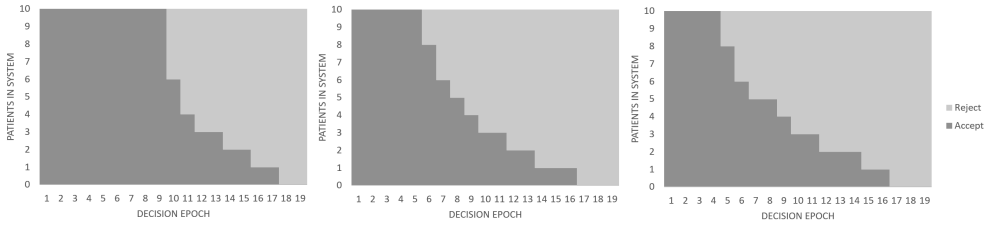
II and III. We see that the policy is more restrictive if the expected workload of an arriving patient is higher and that the threshold for accepting a patient decreases as the decision process progresses, that is, toward the end of the session, more patients are rejected than at the start. The cause for this monotone behavior of the optimal policy is that toward the end of a session there is less capacity available to test all present patients, making it more beneficial to reject a new patient.

Since an MDP considers the immediate rewards and the expected future rewards the optimal policy will, to a large extent, depend on the cost factors k_r and k_e, and in particular on the ratio of these factors. In Figure 6.4 the optimal policy for patients of type I is shown for three cases of the cost factors. We see that an increase of the ratio of k_r and k_e result in more rejections. Clearly, the increasing cost for not completing a service makes it more profitable to reject an arriving patient.

6.4 Case study

In this section we demonstrate how we applied our approach to a real-life case. The cardiology outpatient clinic at the LUMC is a preeminent IPU with many care pathways. The clinic consists of two units: (1) the heart laboratory for heart function tests and (2) the outpatient clinic for cardiologist consultations. Since there is typically a lunch break between sessions, the cardiology outpatient clinic has independent morning and afternoon appointment sessions. The department was confronted with increasing waiting times, resulting in overcrowded waiting rooms and unsatisfied patients.

Figure 6.4: Optimal Acceptance Decision for Patients of Type I with $\frac{k_e}{k_r} = 2$ (left), $\frac{k_e}{k_r} = 3$ (middle) and $\frac{k_e}{k_r} = 4$ (right).



Patients visiting the clinic have, on average, two appointments; in some cases, they have up to five appointments. The clinic aims to schedule all appointments within the same session. A regular clinic visit starts with the function tests, followed by the consultation with the cardiologist, so that the results of the tests can be discussed. Unfortunately, the clinic cannot always schedule all appointments of a patient in a single session, and therefore some patients will have multiple visits distributed over multiple appointment sessions. Afternoon sessions start quickly after the end of the morning sessions, so for the morning sessions there is limited time to run in overtime. This also applies for afternoon sessions, as staff work in strict shifts and working in overtime is costly. Furthermore, cardiologists are scheduled for one session in the outpatient clinic per day and will do other tasks during the other sessions of that day. As the association between patient and cardiologist is strict, patients cannot be rescheduled to a latter session on the same day. Furthermore, the department did not have an online management system in place. As a result, waiting times increased and the number of patients who had not completed all of their appointments by the time a session ended increased. Given the overtime limitations, these appointments had to be canceled and rescheduled to later sessions, resulting in more visits to the clinic. More visits, as a result of cancellations, will lead to more dissatisfied patients. We started analyzing the problem and decided to first implement a heuristic, to quickly observe the impact in practice of an online multi-appointment scheduling approach (for results see Section 6.4.4. In parallel we also started analyzing the problem with our decomposition approach. The used heuristic is simple to understand and therefore suitable for rapid implementation. The heuristic is described in Section 6.4.3.

6.4.1 Model Input

Here we define all input parameters to analyze the case study with our MDP model.

Scheduling Horizon As explained earlier, the cardiology outpatient clinic has independent morning and afternoon appointment sessions. Therefore, we set the model horizon to half a day (i.e. 4 hours), which also reduces the solution space. Each slot has a duration of 1 minutes, so the scheduling horizon consists of 240 slots.

Table 6.2: Appointment Clusters and Explanations for each Cluster.

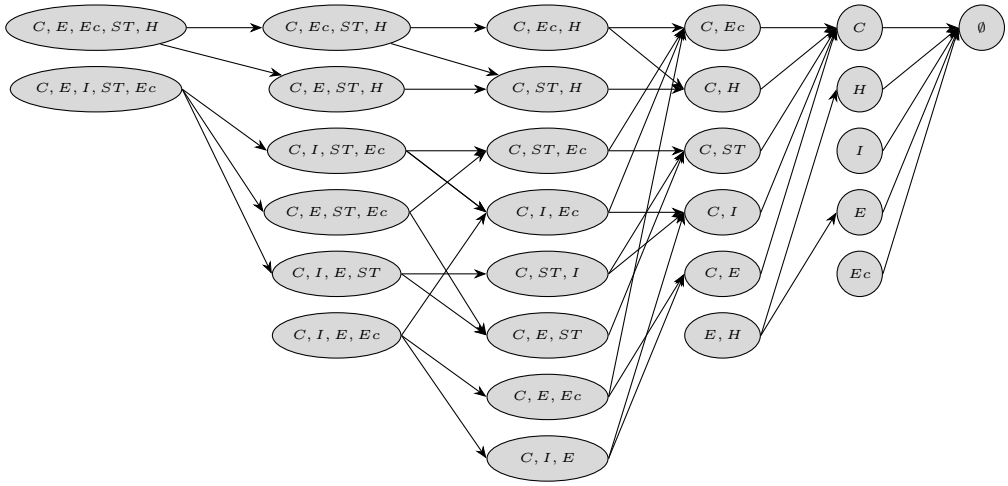
Appointment cluster	Explanation of cluster
Consultation (C)	Consultation with cardiologist
Echocardiogram (E)	Sonogram of the heart
Electrocardiogram(Ec)	Measuring the electrical activity of the heart
Cardiac stress test (ST)	Testing the heart's ability to respond to external stress
Holter check (H)	Check of a mobile ECG monitoring device
ICD check (I)	Check of an implantable cardiovascular device (ICD)

Table 6.3: Frequencies of Arriving Patient types at the Cardiology Outpatient Clinic of the LUMC.

Patient type	Frequency
C,Ec	0.350
E	0.143
I	0.127
C,E,Ec	0.119
C,I,Ec	0.063
Ec	0.046
H	0.042
C,E,Ec,I	0.031
E,H	0.030
C,E,Ec,H,ST	0.022
C,E,Ec,I,ST	0.021

Patient Types and Arrivals We define patient types according to the combination of appointments that need to be scheduled within the same session for a patient. Furthermore, we need to determine the arrival rate per patient type. Currently, the cardiology outpatient clinic has over 30 different appointment types. To reduce the number of possible appointment combinations and thus patient types, we cluster, working with the department's management team, the appointment types. See Table 6.2 for all clusters. Based on appointment data from 2018, we derive all possible patient types and their frequencies. For computational tractability, we take the top 90% of patient types into account, resulting in 11 patient types, and we normalized these frequencies. For example, 35% of the patients arriving in our model will have a combination of appointments consisting of an electrocardiogram and a consultation with the cardiologist. See Table 6.3 for all frequencies of arriving patient types. Next, we define all the transitions between patient types (see Figure 6.5). Here, we consider the precedence constraints that are also used during offline scheduling ($E \rightarrow ST \rightarrow H$), meaning that, for example, the echo-cardiogram always must take place before the cardiac stress test. The arrival rate is also based on historical data and we assume a static arrival rate of 0.45 per slot (e.g. 0.45 per minute). With this arrival rate, we cover 85% of the realized cases in 2018, where the mean arrival rate was 87 patients per session.

Figure 6.5: Transition Diagram of Patient Types When Starting a New Appointment.



E = Echocardiogram, Ec = Electrocardiogram, H = Holter test, ST = Exercise Stress Test, I = ICD check and C = Cardiologist consultation

Table 6.4: Clusters of Appointments, Expected Durations in Minutes, Number of Slots and Number of Testing Stations.

Appointment cluster	Expected Duration (min.)	Expected No. of Slots	No. of Testing stations
Consultation (C)	15	15	5
Echocardiogram (E)	35	35	5
Electrocardiogram(Ec)	5	5	4
Cardiac stress test (ST)	30	30	1
Holter check (H)	15	15	1
ICD check (I)	30	30	2

Appointment Durations We assume appointment durations are geometrically distributed and therefore define the durations by the expected number of slots (see Table 6.4).

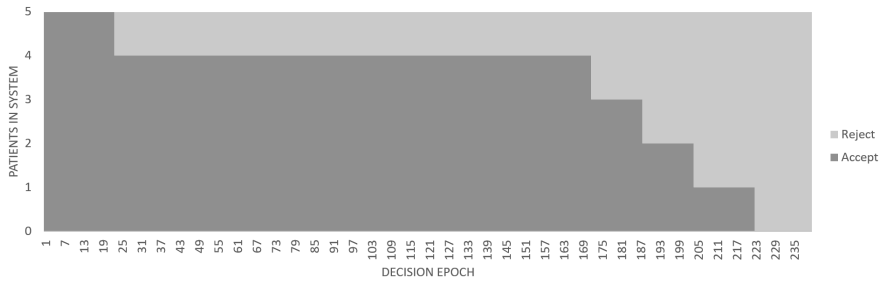
Costs As described in Section 6.3, the cost for rejecting a patient increases with the number of appointments (e.g. tests and consultation) a patient has scheduled in a session. In addition, we want the cost to depend on the duration of each appointment and thus the rejection cost also increases with the duration of an appointment. Otherwise, our MDP would prioritize appointments with short durations to maximize the number of patients who finish all appointments. Hence, the cost of rejecting a certain patient of type i is defined as the total capacity that a patient would need, multiplied by a constant cost factor k_r. Furthermore, we define the end cost ($r_T(s)$) as penalty costs for the patients who have not completed their appointment(s) at time T. This penalty is defined as the total number of tests not finished multiplied by a constant cost factor k_e. By giving a penalty for patients who still have appointments

at time T, we want to find an optimal policy where a maximum number of patients have finished all appointments at the end of the planning horizon. It follows that the cost factor k_r should be lower than the cost factor k_e. If that were not the case, it immediately follows that it would be optimal to reject all patients. The decision of k_r and k_e mainly depend on how undesirable it is that a patient be unable to receive all tests during a session compared to rejecting a patient. This trade-off is a managerial decision. For this case study, we arbitrarily use the ratio $\frac{k_e}{k_r} = 3.6$.

6.4.2 Results

Both the example and the case are implemented in MATLAB R2019b. In Section 6.3.2 we showed that for a small example the optimal policy has a monotone structure. This is also the case for our case study. In Figure 6.6 the optimal policy for a newly arriving patient requiring a Holter check is given. During the first 22 minutes of the session these patients will be accepted whenever the system is not full. After that, the system becomes increasingly more likely to reject patients even while there is capacity. In the last 17 minutes, no patients will be accepted even when the system is empty.

Figure 6.6: Optimal Decision for Acceptance of Patients Requiring a Holter Check.



6.4.3 Simulation

To demonstrate the performance, we evaluate our approach against the implemented heuristic. For this, we use a DES model. The simulation was carried out using Python v3.8 and the Simulus v1.2.1 package. The simple heuristic implemented, is based on the *join shortest queue (JSQ)* policy [104]. Under JSQ, *available patients* (e.g. new arrivals or patients who just finished an appointment and still have other appointments scheduled) are routed to the next appointment with the least number of patients waiting for this appointment. Furthermore, the queuing discipline *first come first served (FCFS)* is used. This differs from our approach for the allocation decision (see Section 6.3.2), which is similar to a single queue with *longest remaining processing time first (LRPT)* discipline. This means that waiting patients with more appointments will be served first. Furthermore in the JSQ scenario no patients are blocked. We have summarized the different settings in Table 6.5. For the scenario using the MDP policy, the simulation model checks for every arrival the optimal decision for the corresponding the state of the model. To evaluate the performance of both

Table 6.5: Summary of System Settings per Simulation Scenario.

System setting	MDP	JSQ heuristic
Number of queues	Single queue	Different queues per test
Queuing discipline	Longest remaining processing time	First come first serve
Entry policy	Blocking	None

approaches, we defined the following KPIs:

- Number of patients who are not completely served (e.g. blocked and unfinished patients).
- Average total waiting time of finished patients.
- Utilization of tests and consultation schedules.

Clearly, the simulation model is terminating as the start and end state are defined by the outpatient clinic session length of one-half of a day (e.g. 240 minutes). It starts with an empty system as at the start of the session no patients have arrived. Based on historical data, we assume a static patient arrival rate defined by a Poisson distribution with an arrival rate of 0.45 patients per minute, and time-dependent frequencies for the types of patients arriving: every 30 minutes, different frequencies are used to determine the patient type of arrivals. Based on data analysis, we fit all service times to log-normal distributions as was done in [52]. For statistically accurate results, we determine the required number of replications (i.e. number of appointment sessions) using the convergence method [199] with $\alpha = .05$. Results show that 1000 replications proved to be sufficient.

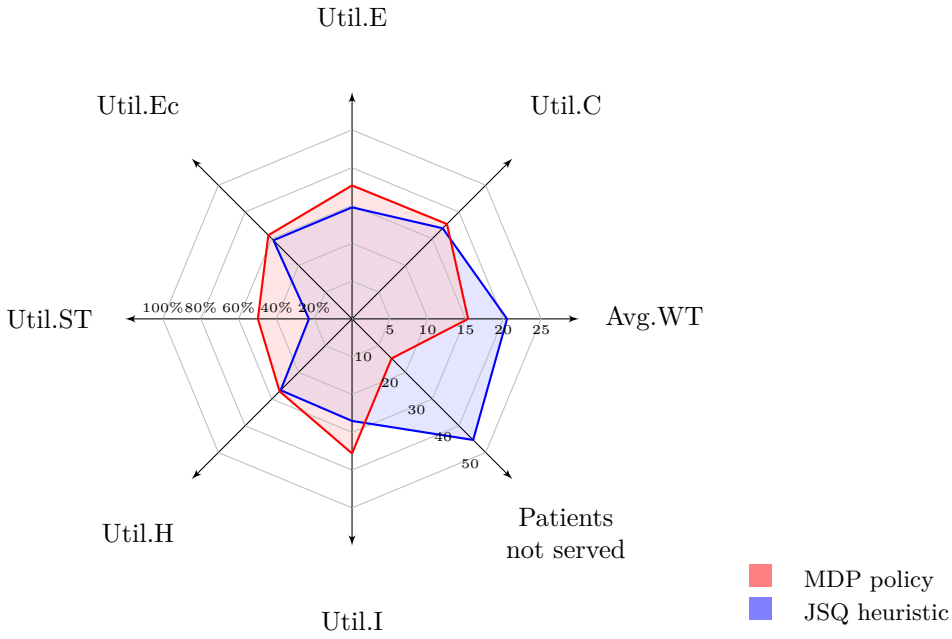
Simulation Results

In Figure 6.7, we summarize the results of our simulation study to determine the performance of the scheduling algorithms. Clearly, for this experiment, the MDP policy outperforms the JSQ heuristic. We expect the timing of the decision on when to allocate a patient to the next appointment may lead to these results. The JSQ heuristic will allocate each patient instantly (i.e. on arrival or after completing an appointment), while the MDP extends this decision by having a single queue and new schedules will be made based on all waiting patients. This may explain both a lower number of patients not served and a lower average waiting time for the MDP scenario. Clearly, when serving more patients, the utilization increases for the MDP scenario.

6.4.4 Implementation in Practice

In this section, we describe the benefits for practice from using an approach for online multi-appointment scheduling by evaluating the performance before and after implementation. The results of the MDP policy were not available at the time of implementation in practice. And as the problems of the department required rapid improvement, the JSQ heuristic was implemented in a decision support tool. The tool generated interesting data for use in analyzing the effects in practice of using an online multi-appointment scheduling approach. In the online setting of our problem, such a system should be able to continuously monitor system dynamics. Using solutions

Figure 6.7: Kiviat Diagram of Simulation Results.



from the aviation sector, the implementation embraced an advanced self-service system. Patients use self-service kiosks upon arrival to register themselves. At the end of this registration process, the decision support tool decides what the first appointment is and patients sit in the waiting room until further notification. Using a real-time data connection with the hospital's electronic medical record system, the decision support tool derives the registering patient's patient type (i.e. combination of scheduled appointments) and shows the patient what the first appointment will be and the expected waiting time. Patients are further notified via monitors in the waiting room. After each appointment patients are notified what the next appointment is and return to the waiting room, or they leave the clinic if they have finished all appointments.

More LUMC departments expressed their interest in the decision support tool. We therefore analyzed the benefits in practice based on the following KPIs:

- Patient sojourn time
- Number of patients waiting in waiting room
- Number of patients waiting at registration desk
- Patient satisfaction with waiting time
- Patient satisfaction with the waiting time information

These KPIs were measured before and after implementation during comparably busy weeks and gave us the opportunity to analyze the data and test for significant differences between the samples using *t-tests* (see Figure 6.8). The data shows an 18% improvement in patient sojourn time (i.e. the time between the arrival and departure of a patient) ($t(326) = -5,34;\ p < 0.01$). Starting with an average of 90 minutes, this means a reduction of almost 20 minutes per visit. As expected, a comparable reduction (17%) is seen in the number of patients waiting ($t(68) = -2,56;\ p < 0.01$), and this changes by the hour, ranging from a 40% reduction (e.g. > 66 patients) to an increase of 3%. The data analysis similarly shows a 66% reduction ($t(64) = -2,44;\ p < 0.01$) in the number of patients waiting at the registration desk. Based on these results, the management of the outpatient clinic decided to reduce the staff at the registration desk by 50%. We also asked patients their opinion on the perceived waiting time and on the provision of waiting time information, asking them to indicate their responses via a 5-point Likert scale (i.e. from very dissatisfied to very satisfied). Although the data analysis showed a significant reduction in the number of patients waiting, the waiting time satisfaction of patients was not improved. We did find a significant improvement in the patient satisfaction on the provision of waiting time information (i.e. showing patients their expected waiting time).

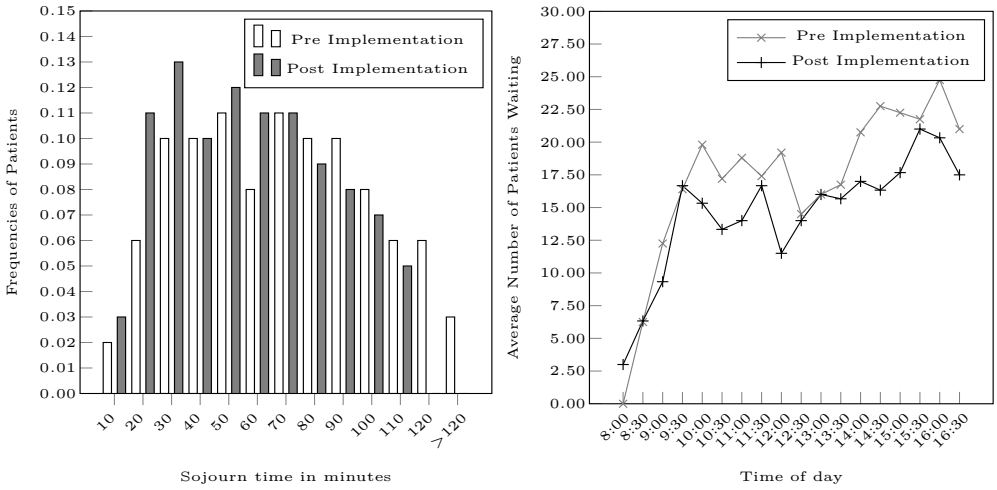
The improved sojourn time is likely a result of waiting time reduction as the number of waiting patients also was significantly reduced and appointment durations were not adjusted during the project. Furthermore, we have analyzed data from comparable weeks and are confident the improvements are a result of the implemented decision support system. These conclusive results led to the hospital-wide implementation of the self-service kiosks including the decision support tool.

6.5 Discussion

This research was inspired by the practical problem observed in the cardiology outpatient clinic of the LUMC. We have shown the positive impact of online multi-appointment scheduling incorporating real-time system dynamics. While most literature analyzes online multi-appointment scheduling for later sessions, we derive optimal decisions for the same session.

We formulated the scheduling problem as an MDP model. By decomposing the acceptance and scheduling decisions, we simplified the state space and were able to define a tractable MDP for real-life instances. The acceptance decision, is inspired by the conviction that it is better to not serve a patient when it is expected that the patient cannot finish all appointments, rather than finishing a subset of the scheduled appointments and reschedule the others. A major assumption influencing the result-

Figure 6.8: Histograms and Line Charts of The Patient Sojourn Time and The Average Number of Patients Waiting Before and After Implementation of The Decision Support Tool.



ing policy derived from our MDP, are the geometrically distributed service times. For tractability, memoryless distributions are required. We therefore expect that our MDP model is over-fitting the expected durations of service times, resulting in an increasing number of blocked patients as it is expected these arrivals will not finish their appointments within the session horizon. However, the MDP approach still outperforms the JSQ heuristic on this KPI.

Using our model, healthcare management can easily balance the trade-off between resource utilization, waiting time and blocking probability for complex appointments systems incorporating multi-appointment scheduling. As shown in Section 6.4.4, implementing a relatively simple heuristic already results in significant improvements in waiting time and thus in overtime. Implementation of sophisticated algorithms, requires advanced digitization of decision support systems. Although it is often said that healthcare it is not suitable for digital transformations as it requires face-to-face contact, solutions as presented here result in process information and efficiency gain for both patients and healthcare professionals. In this way, such solutions could contribute to keeping healthcare accessible. As mentioned in Section 6.1, another important development is the increasing complexity of appointment systems incorporating multi-appointment scheduling. These complex appointment systems balance many conflicting objectives, and optimizing such scheduling problems requires sophisticated analytics, as demonstrated here.

Our approach is also suitable for other types of appointment systems (e.g. advanced access systems [175]), incorporating both scheduled and unscheduled arrivals, as we optimize the decision of accepting or rejecting to ultimately balance demand and supply. Furthermore, our approach can also be used to optimize specific appointment schedules of future sessions. As all possible states defined are analyzed by our MDP, a decision support system can decide based on the progression of that session and

expected future arrivals whether it is still optimal to accept an arrival or to reschedule the arrival given the managerial priorities as translated into the cost function of our model. Future research could focus on multi-departmental or multi-location online multi-appointment scheduling. This problem increases the state space significantly as more patient types (e.g. combination of appointments) are possible and travelling distances between appointment location could be analyzed.

An unique feature of this research was the opportunity to analyze the impact of one of the algorithms in practice. We are therefore grateful for the trust and patience of the management of the cardiology department. The department invested multiple resources (both staff and monetary) to successfully implement the solution at hand and analyze the impact. Whereas clinical innovations are tested extensively before exposed to patients, the effects of organizational and process innovations are rarely evaluated in healthcare, while the effects on both patients and healthcare professionals can be tremendous. Of course, to analyze the impact of organizational innovations is methodological and ethical challenging. Another reason for this lack of evaluation could be that designing this kind of research is time consuming or that researchers are not familiar with these types of research design. We hope to encourage researchers to also focus on implementation and to analyze the impact in practice.

This research contributes to furthering online multi-appointment scheduling analysis and implementation. From a methodological perspective, our research contributes to analyzing complex appointment systems such as multi-appointment scheduling and ultimately keeping healthcare accessible.

Bibliography

[1] Joanna Abraham and Madhu C Reddy. Challenges to inter-departmental coordination of patient transfers: a workflow perspective. *International journal of medical informatics*, 79(2):112–22, mar 2010. ISSN 1872-8243. doi: 10.1016/j.ijmedinf.2009.11.001.

[2] Ilgin Acar and Steven E. Butt. Modeling nurse-patient assignments considering patient acuity and travel distance metrics. *Journal of Biomedical Informatics*, 64:192–206, 2016. ISSN 15320464. doi: 10.1016/j.jbi.2016.10.006.

[3] I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X.T. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009. ISSN 1386-9620.

[4] 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. ISSN 0377-2217.

[5] Larsson Agneta and Fredriksson Anna. Tactical capacity planning in hospital departments. *International Journal of Health Care Quality Assurance*, 32(1): 233–245, jan 2019. ISSN 0952-6862. doi: 10.1108/IJHCQA-11-2017-0218.

[6] Amir Ahmadi-Javid, Zahra Jalali, and Kenneth J Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3–34, 2017. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2016.06.064.

[7] E Akcali, JC Murray, and C Lin. A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science*, 9(4):391–404, 2006.

[8] R Akkerman and M Knip. Reallocation of beds to reduce waiting time for cardiac surgery. *Health Care Management Science*, 7(2):119–126, 2004.

[9] Berhanu Alemayehu and Kenneth E. Warner. The lifetime distribution of health care costs. *Health Services Research*, 39(3):627–642, 2004. doi: 10.1111/j.1475-6773.2004.00248.x.

[10] Faranak Aminzadeh and William Burd Dalziel. Older adults in the emergency department: A systematic review of patterns of use, adverse outcomes, and effectiveness of interventions. *Annals of Emergency Medicine*, 39(3):238–247, mar 2002. ISSN 01960644. doi: 10.1067/mem.2002.121523.

[11] Robert N. Anthony. *Planning and control systems; a framework for analysis*. Division of Research, Graduate School of Business Administration, Harvard University Boston, 1965.

[12] D. Astaraky and J. Patrick. A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling. *European Journal of Operational Research*, 245(1):309–319, 2015. ISSN 0377-2217.

[13] Stefan Auener, Danielle Kroon, Erik Wackers, Simone van Dulmen, and Patrick Jeurissen. Covid-19: A window of opportunity for positive healthcare reforms. *International Journal of Health Policy and Management*, 2020. ISSN 2322-5939. in press, forthcoming.

[14] Ali Azadeh, Milad Baghersad, Mehdi Hosseinabadi Farahani, and Mansour Zarrin. Semi-online patient scheduling in pathology laboratories. *Artificial Intelligence in Medicine*, 64(3):217–226, jul 2015. ISSN 18732860. doi: 10.1016/j.artmed.2015.05.001.

[15] A. Bagust, M. Place, and J. W Posnett. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *BMJ*, 319(7203):155–158, 1999. ISSN 0959-8138. doi: 10.1136/bmj.319.7203.155.

[16] Norman 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. ISSN 00359246. doi: 10.1111/j.2517-6161.1952.tb00112.x.

[17] C. Banditori, P. Cappanera, and F. Visintin. A combined optimization-simulation approach to the master surgical scheduling problem. *Ima Journal of Management Mathematics*, 24(2):155–187, 2013. ISSN 1471-678X.

[18] Henri Barki and Alain Pinsonneault. A model of organizational integration, implementation effort, and performance. *Organization Science*, 16(2):165–179, 2005. doi: 10.1287/orsc.1050.0118.

[19] Christiane Barz and Kumar Rajaram. Elective Patient Admission and Scheduling under Multiple Resource Constraints. *Production and Operations Management*, 24(12):1907–1930, 2015. ISSN 19375956. doi: 10.1111/poms.12395.

[20] 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. ISSN 02545330. doi: 10.1007/s10479-009-0570-z.

[21] René Bekker and Paulien M. Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):237–249, 2011. ISSN 13869620. doi: 10.1007/s10729-011-9163-x.

[22] 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. ISSN 0377-2217.

[23] 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. ISSN 1094-6136.

[24] Rym Ben Bachouch, Alain Guinet, and Sonia Hajri-Gabouj. An integer linear model for hospital bed planning. *International Journal of Production Economics*, 140(2):833–843, 2012. ISSN 09255273. doi: 10.1016/j.ijpe.2012.07.023.

[25] Bjorn Berg and Brian T. Denton. Appointment planning and scheduling in outpatient procedure centers. In *International Series in Operations Research and Management Science*, pages 131–154. Springer New York LLC, jan 2012. doi: 10.1007/978-1-4614-1734-7_6.

[26] Mathilde A. Berghout, Isabelle N. Fabbricotti, Martina Buljac-Samardžić, and Carina G. J. M. Hilders. Medical leaders or masters?—a systematic review of medical leadership in hospital settings. *PLOS ONE*, 12(9):1–24, 09 2017. doi: 10.1371/journal.pone.0184522.

[27] Steven L. Bernstein, Dominik Aronsky, Reena Duseja, Stephen Epstein, Dan Handel, Ula Hwang, Melissa McCarthy, and et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10, jan 2009. ISSN 10696563. doi: 10.1111/j.1553-2712.2008.00295.x.

[28] Donald M. Berwick and Andrew D. Hackbarth. Eliminating Waste in US Health Care. *JAMA*, 307(14):1513–1516, 04 2012. ISSN 0098-7484. doi: 10.1001/jama.2012.362.

[29] TJ Best, B Sandikci, D Eisenstein, and D Meltzer. Managing hospital bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management*, 17(2):157–176, May 2015.

[30] Ingeborg A. Bikker. *Organizing timely treatment in multi-disciplinary care*. PhD thesis, University of Twente, Netherlands, November 2018.

[31] Ingeborg A. Bikker, Martijn R.K. Mes, Antoine Sauré, and Richard J. Boucherie. Online capacity planning for rehabilitation treatments: An approximate dynamic programming approach. *Probability in the Engineering and Informational Sciences*, page 1–25, 2018. doi: 10.1017/S0269964818000402.

[32] J. Bisschop. *AIMMS optimizating modelling*, 2016.

[33] EL Blair and CE Lawrence. 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.

[34] Richard M.J. Bohmer. Leading clinicians and clinicians leading. *New England Journal of Medicine*, 368(16):1468–1470, 2013. doi: 10.1056/NEJMp1301814. PMID: 23594000.

[35] Nick Bosanquet. Human: solving the global workforce crisis in healthcare. *British Journal of Healthcare Management*, 25(5):206–206, 2019. doi: 10.12968/bjhc.2019.25.5.206.

[36] J. Bosch. Better utilisation of the or with less beds. Master's thesis, Industrial Engineering & Management, University of Twente, 2011.

[37] T Bountourelis, L Luangkesorn, A Schaefer, L Maillart, SG Nabors, and G Clermont. Development and validation of a large scale icu simulation model with blocking. In S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, editors, *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pages 1143–1153. IEEE, 2011.

[38] T Bountourelis, MY Ulukus, JP Kharoufeh, and SG Nabors. The modeling, analysis, and management of intensive care units. In Brian T. Denton, editor, *Handbook of Healthcare Operations Management*, volume 184 of *International Series in Operations Research & Management Science*, book section 6, pages 153–182. Springer New York, 2013.

[39] GEP Box. Non-normality and tests on variances. *Biometrika*, 40(3):318–335, 1953.

[40] Kenneth K. Boyer and Peter Pronovost. What medicine can teach operations: What operations can teach medicine. *Journal of Operations Management*, 28 (5):367 – 371, 2010. ISSN 0272-6963. doi: https://doi.org/10.1016/j.jom.2010.08.002.

[41] A Braaksma, J Deglise-Hawkinson, B T Denton, M P Van Oyen, R J Boucherie, and M R K Mes. Online appointment scheduling with different urgencies and appointment lengths. Forthcoming, 2014.

[42] Aleida Braaksma, Elizabeth Ugarph, Retsef Levi, Ana Cecilia Zenteno, Bethany J Daily, Benjamin Orcutt, and Peter F Dunn. Just-in-time Bed Assignment Improves Surgical Patient Flow. Forthcoming, 2018.

[43] 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. doi: 10.1057/jos.2009.10.

[44] Mark Britnell. *Human: solving the global workforce crisis in healthcare*. Oxford University Press, 2019.

[45] James R. Broyles, Jeffery K. Cochran, and Douglas C. Montgomery. A statistical Markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657, 2010. ISSN 03772217. doi: 10.1016/j.ejor.2010.06.021.

[46] Timothy W. Butler, G.Keong Leong, and Linda N. Everett. The operations management role in hospital strategic planning. *Journal of Operations Management*, 14(2):137–156, 1996. doi: 10.1016/0272-6963(95)00041-0.

[47] Simon Capewell. The continuing rise in emergency admissions. *BMJ*, 312(7037): 991–992, 1996. doi: 10.1136/bmj.312.7037.991.

[48] Brecht Cardoen and Erik Demeulemeester. Capacity of clinical pathways - A strategic multi-level evaluation tool. *Journal of Medical Systems*, 32(6):443–452, dec 2008. ISSN 01485598. doi: 10.1007/s10916-008-9150-z.

[49] Brecht Cardoen, Erik Demeulemeester, and Jeroen Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201(3):921 – 932, 2010. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2009.04.011.

[50] Kyle Cattani and Glen M. Schmidt. The pooling principle. *INFORMS Transactions on Education*, 5(2):17–24, 2005. doi: 10.1287/ited.5.2.17.

[51] Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, jan 2010. ISSN 10591478. doi: 10.1111/j.1937-5956.2003.tb00218.x.

[52] Tugba Cayirli, Emre Veral, and Harry Rosen. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1): 47–58, 2006. ISSN 1572-9389. doi: 10.1007/s10729-006-6279-5.

[53] Carri W. Chan, Jing Dong, and Linda V. Green. Queues with Time-Varying Arrivals and Inspections with Applications to Hospital Discharge Policies. *Operations Research*, 65(2):469–495, 2016. ISSN 0030-364X. doi: 10.1287/opre.2016.1536.

[54] CW Chan, VF Farias, N Bambos, and GJ Escobar. Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research*, 60(6): 1323–1341, 2012.

[55] Chandra Charu and Kumar Sameer. Enterprise architectural framework for supply-chain integration. *Industrial Management & Data Systems*, 101(6):290–304, jan 2001. ISSN 0263-5577. doi: 10.1108/EUM0000000005578.

[56] V.S. Chow, M.L. Puterman, N. Salehirad, W.H. Huang, and D. Atkins. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management*, 20(3):418–430, 2011. ISSN 1059-1478.

[57] JK Cochran and A Bharti. Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*, 9(1):31–45, 2006.

[58] D Conforti, F Guerriero, R Guido, MM Cerinic, and ML Conforti. An optimal decision making model for supporting week hospital management. *Health Care Management Science*, 14(1):74–88, 2011.

[59] R Conway, D O'Riordan, and B Silke. Long-term outcome of an AMAU–a decade's experience. *QJM : monthly journal of the Association of Physicians*, 107(1):43–9, jan 2014. ISSN 1460-2393. doi: 10.1093/qjmed/hct199.

[60] MW Cooke, J Higgins, and P Kidd. Use of emergency observation and assessment wards: a systematic literature review. *Emergency Medicine Journal*, 20 (2):138–142, 2003.

[61] AX Costa, SA Ridley, AK Shahani, PR Harper, V De Senna, and MS Nielsen. Mathematical modelling and simulation for planning critical care capacity*. *Anaesthesia*, 58(4):320–327, 2003.

[62] Elizabeth A. Crawford, Pratik J. Parikh, Nan Kong, and Charuhas V. Thakar. Analyzing discharge strategies during acute care: A discrete-event simulation study. *Medical Decision Making*, 34(2):231–241, 2014. ISSN 0272989X. doi: 10.1177/0272989X13503500.

[63] Frans Cruijssen, Wout Dullaert, and Hein Fleuren. Horizontal cooperation in transport and logistics: A literature review. *Transportation Journal*, 46(3):22–39, 2007. ISSN 00411612, 2157328X.

[64] J. Dai and Pengyi Shi. Inpatient Overflow: An Approximate Dynamic Programming Approach. *Ssrn*, Available, 2017. doi: 10.2139/ssrn.2924208.

[65] George B. Dantzig. *Linear programming and extensions*. Princeton university press, Princeton, New Jersey, 1963.

[66] R Davies. Simulation for planning services for patients with coronary artery disease. *European Journal of Operational Research*, 72(2):323–332, 1994.

[67] Robert W. Day, Matthew D. Dean, Robert Garfinkel, and Steven Thompson. Improving patient flow in a hospital through dynamic allocation of cardiac diagnostic testing time slots. *Decision Support Systems*, 49(4):463–473, nov 2010. ISSN 01679236. doi: 10.1016/j.dss.2010.05.007.

[68] Theodore Eugene Day, Albert Chi, Matthew Harris Rutberg, Ashley J. Zahm, Victoria M. Otarola, Jeffrey M. Feldman, and Caroline A. Pasquariello. Addressing the variation of post-surgical inpatient census with computer simulation. *Pediatric Surgery International*, 30(4):449–456, 2014. ISSN 14379813. doi: 10.1007/s00383-014-3475-0.

[69] A. M. de Bruin, R. Bekker, L. van Zanten, and G. M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178 (1):23–43, 2010. ISSN 02545330. doi: 10.1007/s10479-009-0647-8.

[70] Niimish Dharmadhikari and Jun Zhang. Simulation optimization of blocking appointment scheduling policies for multi-clinic appointments in centralized scheduling systems. *International Journal of Engineering and Innovative Technology*, 2(11):196–201, 2013.

[71] Adam Diamant, Joseph Milner, and Fayez Quereshy. Dynamic Patient Scheduling for Multi-Appointment Health Care Programs. *Production and Operations Management*, 27(1):58–79, 2018. doi: 10.1111/poms.12783.

[72] G Dobson, Hi Lee, and E Pinker. A model of icu bumping. *Operations Research*, 58(6):1564–1576, 2010.

[73] S W Douma and H Schreuder. *Economic approaches to organizations*. Pearson Education Limited, 6th edition, 2017. ISBN 9781292175720 1292175729.

[74] David Dreyfus, Anand Nair, and Claudia Rosales. The impact of planning and communication on unplanned costs in surgical episodes of care: Implications for reducing waste in hospital operating rooms. *Journal of Operations Management*, 66(1-2):91–111, 2020. doi: 10.1002/joom.1070.

[75] MB Dumas. Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels. *Health Services Research*, 20(1): 43–61, 1985.

[76] The Economist. The world's most valuable resource is no longer oil, but data, May 2017. URL `https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data`.

[77] 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.

[78] RB. Ferreira, FC. Coelli, WC . Pereira, and RMVR Almeida. Optimizing patient flow in a large hospital surgical centre by means of discrete-event computer simulation models. *Journal of Evaluation in Clinical Practice*, 14(6):1031–1037, 2008.

[79] A. Fügener. An integrated strategic and tactical master surgery scheduling approach with stochastic resource demand. *Journal of Business Logistics*, 36(4): 374–387, 2015. ISSN 0735-3766.

[80] A. Fügener, G.M. Edenharter, P. Kiefer, U. Mayr, J. Schiele, F. Steiner, R. Kolisch, and M. Blobner. Improving intensive care unit and ward utilization by adapting master surgery schedules. *A A Case Rep*, 6(6):172–180, 2016.

[81] Andreas Fügener, Erwin W. Hans, Rainer Kolisch, Nikky Kortbeek, and Peter T. Vanberkel. Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, 239(1):227–236, 2014. ISSN 03772217. doi: 10.1016/j.ejor.2014.05.009.

[82] Steve Gallivan and Martin Utley. A technical note concerning emergency bed demand. *Health Care Management Science*, 14(3):250–252, 2011. ISSN 13869620. doi: 10.1007/s10729-011-9158-7.

[83] Paul A Games and John F Howell. Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study. *Journal of Educational and Behavioral Statistics*, 1(2):113–125, jun 1976. doi: 10.3102/10769986001002113.

[84] Cheng Gao, Abel N Kho, Catherine Ivory, Sarah Osmundson, Bradley A Malin, and You Chen. Predicting Length of Stay for Obstetric Patients via Electronic Medical Records. *Studies in health technology and informatics*, 245:1019–1023, 2017. ISSN 1879-8365.

[85] L Garg, S McClean, M Barton, B Meenan, and K Fullerton. Forecasting hospital bed requirements and cost of care using phase type survival trees. In *Intelligent Systems (IS), 2010 5th IEEE International Conference*, pages 185–190, Bed Occupancy; Cardiology, 2010.

[86] GM Garrison and JL Pecina. Using the m/g/inf queueing model to predict inpatient family medicine service census and resident workload. *Health Informatics Journal*, 2015.

[87] Daniel Gartner and Rainer Kolisch. Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research*, 233(3):689–699, 2014. ISSN 03772217. doi: 10.1016/j.ejor.2013.08.026.

[88] John Glaser. It's time for a new kind of electronic health record, 2020.

[89] Sholom Glouberman and Henry Mintzberg. Managing the care of health and the cure of disease - Part I: Differentiation. *Health Care Management Review*, 26(1):56–89, 2001a. ISSN 0361-6274. doi: 10.1097/00004010-200101000-00006.

[90] Sholom Glouberman and Henry Mintzberg. Managing the Care of Health and the Cure of Disease—Part II:Integration. *Health Care Management Review*, 26 (1):70–84, 2001b. ISSN 0361-6274. doi: 10.1097/00004010-200101000-00007.

[91] F Gorunescu, SI McClean, and PH Millard. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24, 2002.

[92] F Gorunescu, SI McClean, and PH 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.

[93] M Gorunescu, F Gorunescu, and A Prodan. Continuous-time markov model for geriatric patients behavior. optimization of the bed occupancy and computer simulation. *Korean Journal of Computational & Applied Mathematics*, 9(1): 185–195, 2002.

[94] S C Graves. Manufacturing planning and control systems. In Resende M Pardalos P., editor, *Handbook of Applied Optimization*, volume 1, pages 728–746. Oxford University Press, New York, US, 2002. doi: 10.1080/09537289008919307.

[95] Muir Gray. Value: Operations Research and the new health care paradigm. *Operations Research for Health Care*, 1(1):20–21, 2012. ISSN 2211-6923. doi: https://doi.org/10.1016/j.orhc.2012.01.002.

[96] L Green and V Nguyen. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research*, 36:421–442, 2001.

[97] 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.

[98] J. D. Griffiths, V. Knight, and I. Komenda. Bed management in a Critical Care Unit. *IMA Journal of Management Mathematics*, 24(2):137–153, 2013. ISSN 14716798. doi: 10.1093/imaman/dpr028.

[99] JD Griffiths, JE Williams, and RM Wood. Modelling activities at a neurological rehabilitation unit. *European Journal of Operational Research*, 226(2):301–312, 2013.

[100] Francesca Guerriero and Rosita Guido. Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, 14(1):89–114, 2011. ISSN 1572-9389. doi: 10.1007/s10729-010-9143-6.

[101] Rosita Guido, Maria Carmela Groccia, and Domenico Conforti. An efficient matheuristic for offline patient-to-bed assignment problems. *European Journal of Operational Research*, 268(2):486–503, 2018. ISSN 03772217. doi: 10.1016/j.ejor.2018.02.007.

[102] M. M. Günal and M. Pidd. Discrete event simulation for performance modelling in health care: A review of the literature. *Journal of Simulation*, 4(1):42–51, 2010. ISSN 17477778. doi: 10.1057/jos.2009.25.

[103] Diwakar Gupta and Brian Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions (Institute of Industrial Engineers)*, 40(9):800–819, jul 2008. ISSN 0740817X. doi: 10.1080/07408170802165880.

[104] Varun Gupta, Mor Harchol Balter, Karl Sigman, and Ward Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007. ISSN 0166-5316. doi: https://doi.org/10.1016/j.peva.2007.06.012.

[105] Chris Ham. Improving the performance of health services: the role of clinical leadership. *The Lancet*, 361(9373):1978–1980, jun 2003. ISSN 0140-6736. doi: 10.1016/S0140-6736(03)13593-3.

[106] Erwin W. Hans, Mark van Houdenhoven, and Peter J. H." Hulshof. A framework for healthcare planning and control. In Randolph Hall, editor, *Handbook of Healthcare System Scheduling*, pages 303–320. Springer US, Boston, MA, 2012. ISBN 978-1-4614-1734-7. doi: 10.1007/978-1-4614-1734-7_12.

[107] E.W. Hans, G. Wullink, M. van Houdenhoven, and G. Kazemier. Robust surgery loading. *European Journal of Operational Research*, 185(3):1038 – 1050, 2008. ISSN 0377-2217.

[108] PR Harper and MA Pitt. On the challenges of healthcare modelling and a proposed project life cycle for successful implementation. *Journal of the Operational Research Society*, 55(6):657–661, 2004.

[109] PR Harper, VA Knight, and AH Marshall. Discrete conditional phase-type models utilising classification trees: Application to modelling health service capacities. *European Journal of Operational Research*, 219(3):522–530, 2012.

[110] RA Harris. Hospital bed requirements planning. *European Journal of Operational Research*, 25(1):121–126, 1986.

[111] GW Harrison, A Shafer, and M Macky. Modelling variability in hospital bed occupancy. *Health Care Management Science*, 8(4):325–334, 2005.

[112] Jonathan Helm and Mark P. Van Oyen. Design and Optimization Methods for Elective Hospital Admissions. *Ssrn*, 62(6):1265–1282, 2014. ISSN 0030-364X. doi: 10.2139/ssrn.2437936.

[113] Lene Berge Holm, Hilde Lurås, and Fredrik A. Dahl. Improving hospital bed utilisation through simulation and optimisation. With application to a 40% increase in patient volume in a Norwegian general hospital. *International Journal of Medical Informatics*, 82(2):80–89, 2013. ISSN 13865056. doi: 10.1016/j.ijmedinf.2012.05.006.

[114] Nathan R Hoot, Larry J LeBlanc, Ian Jones, Scott R Levin, Chuan Zhou, Cynthia S Gadd, and Dominik Aronsky. Forecasting emergency department crowding: a discrete event simulation. *Annals of emergency medicine*, 52(2):116–25, aug 2008. ISSN 1097-6760. doi: 10.1016/j.annemergmed.2007.12.011.

[115] Grant Howard. No single approach will solve healthcare's problems—we need operations management. *BMJ*, 368, 2020. doi: 10.1136/bmj.m1114.

[116] Fu Shiung Hsieh. A hybrid and scalable multi-agent approach for patient scheduling based on Petri net models. *Applied Intelligence*, 47(4):1068–1086, dec 2017. ISSN 15737497. doi: 10.1007/s10489-017-0935-y.

[117] Peter J. H. Hulshof, Richard J. Boucherie, J. Theresia van Essen, Erwin W. Hans, Johann L. Hurink, Nikky Kortbeek, Nelly Litvak, Peter T. Vanberkel, Egbert van der Veen, Bart Veltman, Ingrid M. H. Vliegen, and Maartje E. Zonderland. ORchestra: an online reference database of OR/MS literature in health care. *Health Care Management Science*, 14(4):383–384, nov 2011. ISSN 1386-9620. doi: 10.1007/s10729-011-9169-4.

[118] Peter J H Hulshof, Nikky Kortbeek, Richard J Boucherie, Erwin W Hans, and Piet 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. ISSN 2047-6965. doi: 10.1057/hs.2012.18.

[119] Peter J H Hulshof, Richard J. Boucherie, Erwin W. Hans, and Johann L. Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science*, 16(2):152–166, jun 2013. ISSN 13869620. doi: 10.1007/s10729-012-9219-6.

[120] Peter J.H. Hulshof, Martijn R.K. Mes, Richard J. Boucherie, and Erwin W. Hans. Patient admission planning using Approximate Dynamic Programming. *Flexible Services and Manufacturing Journal*, 28(1-2):30–61, 2016. ISSN 19366590. doi: 10.1007/s10696-015-9219-1.

[121] Mark W. Isken, Timothy J. Ward, and Steven J. Littig. An open source software project for obstetrical procedure scheduling and occupancy analysis. *Health Care Management Science*, 14(1):56–73, 2011. ISSN 13869620. doi: 10.1007/s10729-010-9141-8.

[122] David Johnston, Adam Diamant, and Fayez Quereshy. Why do surgeons schedule their own surgeries? *Journal of Operations Management*, 65(3):262–281, 2019. doi: 10.1002/joom.1012.

[123] Maheshkumar P Joshi, Ravi Kathuria, and Stephen J Porth. Alignment of strategic priorities and performance: an integration of operations and strategic management perspectives. *Journal of Operations Management*, 21(3):353 – 369, 2003. ISSN 0272-6963. doi: https://doi.org/10.1016/S0272-6963(03)00003-2.

[124] Drupsteen Justin, van der Vaart Taco, and Pieter van Donk Dirk. Integrative practices in hospitals and their impact on patient flow. *International Journal of Operations & Production Management*, 33(7):912–933, jan 2013. ISSN 0144-3577. doi: 10.1108/IJOPM-12-2011-0487.

[125] Alan J Kalton, Medini R Singh, David A August, Christopher M Parin, and Elizabeth J Othman. Using simulation to improve the operational efficiency of a multi-disciplinary clinic. *Journal of Social Health Systems*, 5(3):43–62, 1997. ISSN 1043-1721.

[126] Erin M. Kane, James J. Scheulen, Adrian Püttgen, Diego Martinez, Scott Levin, Bree A. Bush, Linda Huffman, Mary Margaret Jacobs, Hetal Rupani, and David T. Efron. Use of systems engineering to design a hospital command center. *The Joint Commission Journal on Quality and Patient Safety*, 45(5):370–379, 2019. ISSN 1553-7250. doi: https://doi.org/10.1016/j.jcjq.2018.11.006.

[127] T Kapamara, K Sheibani, D. Petrovic, O. Hass, and C. Reeves. A simulation of a radiotherapy treatment system: A case study of a local cancer centre. In *Proceedings of the ORP3 conference*, pages 29–35, 2007.

[128] Kyle Keepers and Gary W. Harrison. Internal flows and frequency of internal overflows in a large teaching hospital. In El-Darzi E Nugent C McClean S. Millard P., editor, *Studies in Computational Intelligence*, volume 189 of *Studies in Computational Intelligence*, pages 185–192. Elsevier, 2009. ISBN 9783642001789. doi: 10.1007/978-3-642-00179-6_11.

[129] Samee U Khan, Albert Y Zomaya, and Assad Abbas. *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Springer, Cham, 2017. ISBN 9783319582801. doi: https://doi.org/10.1007/978-3-319-58280-1.

[130] Saif Kifah and Salwani Abdullah. An adaptive non-linear great deluge algorithm for the patient-admission problem. *Information Sciences*, 295:573–585, 2015. ISSN 00200255. doi: 10.1016/j.ins.2014.10.004.

[131] SC Kim, I Horowitz, KK Young, and TA Buckley. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research*, 115(1):36–46, 1999.

[132] SC Kim, I Horowitz, KK Young, and TA Buckley. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, 18 (4):427–443, 2000.

[133] Yoon Hee Kim, Fabian J. Sting, and Christoph H. Loch. Top-down, bottom-up, or both? toward an integrative perspective on operations strategy formation. *Journal of Operations Management*, 32(7):462 – 474, 2014. ISSN 0272-6963. doi: https://doi.org/10.1016/j.jom.2014.09.005. Special Issue on Implementing Operations Strategy for Competitive Advantage.

[134] 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.

[135] Alexander Kolker. Interdependency of hospital departments and hospital-wide patient flows. In Randolph Hall, editor, *Patient flow*, volume 206 of *International Series in Operations Research & Management Science*, pages 43–63. Springer US, Boston, MA, 2013. ISBN 978-1-4614-9511-6. doi: 10.1007/978-1-4614-9512-3_2.

[136] Cornelia Körber, David Strååt, Jan-Inge Henter, Andreas Ringman-Uggla, and Mandar Dabhilkar. Implementing valuebased health care at the provider level: an operations management view. In *EurOMA conference paper*, 2016.

[137] N. Kortbeek, A. Braaksma, C. A.J. Burger, P. J.M. Bakker, and R. J. Boucherie. Flexible nurse staffing based on hourly bed census predictions. *International Journal of Production Economics*, 161(0):167–180, 2015. ISSN 09255273. doi: 10.1016/j.ijpe.2014.12.007.

[138] N Kortbeek, A Braaksma, FHF Smeenk, PJM Bakker, and RJ Boucherie. Integral resource capacity planning for inpatient care services based on bed census predictions by hour. *Journal of the Operational Research Society*, 66(7):1061–1076, 2015.

[139] A Kumar and J Mo. Models for bed occupancy management of a hospital in singapore. In *Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management.*, pages 1–6. IIEOM & cosponsored by INFORMS, 2010.

[140] S Kumar. Modeling hospital surgical delivery process design using system simulation: optimizing patient flow and bed capacity as an illustration. *Technology and Health Care*, 19(1):1–20, 2011.

[141] RJ Kusters and PMA Groot. Modelling resource availability in general hospitals design and implementation of a decision support model. *European Journal of Operational Research*, 88(3):428–445, 1996.

[142] Hmwe Hmwe Kyu et al. Global, regional, and national disability-adjusted life-years (dalys) for 359 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1859 – 1922, 2018. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(18)32335-3.

[143] P Landa, M Sonnessa, E Tanfani, and A Testi. A discrete event simulation model to support bed management. In *Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), 2014 International Conference on*, pages 901–912, Bed Occupancy; Emergency Service, Hospital;, 2014.

[144] Paolo Landa, Michele Sonnessa, Elena Tànfani, and Angela Testi. A Discrete Event Simulation Model to Support Bed Management. In *Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), 2014 International Conference on*, pages 901–912, 2014. doi: 10.5220/0005161809010912.

[145] SD Lapierre, D Goldsman, R Cochran, and J DuBow. Bed allocation techniques based on census data. *Socio-Economic Planning Sciences*, 33(1):25–38, 1999.

[146] Eva K. Lee, Hany Y. Atallah, Michael D. Wright, Eleanor T. Post, Calvin Thomas, Daniel T. Wu, and Leon L. Haley. Transforming Hospital Emergency Department Workflow and Patient Care. *Interfaces*, 45(1):58–82, feb 2015. ISSN 0092-2102. doi: 10.1287/inte.2014.0788.

[147] A G Leeftink, I A Bikker, I M H Vliegen, and R J Boucherie. Multi-disciplinary planning in health care: a review. *Health Systems*, 0(0):1–24, 2018. doi: 10.1080/20476965.2018.1436909.

[148] Federico Lega and Carlo De Pietro. Converging patterns in hospital organization: beyond the professional bureaucracy. *Health Policy*, 74(3):261 – 281, 2005. ISSN 0168-8510. doi: https://doi.org/10.1016/j.healthpol.2005.01.010.

[149] H. Levene. Robust tests for equality of variances. In I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann, editors, *Contributions to Probability and Statistics*, pages 278–292. Stanford Univ. Press, Stanford, 1960.

[150] Marcel Levi. Evidencebased beleid. *Medisch Contact*, 43, October 2018. URL `https://www.medischcontact.nl/opinie/blogs-columns/column/evidencebased-beleid.htm`. in Dutch.

[151] Jordan Y Z Li, Tuck Y Yong, Denise M Bennett, Lauri T O'Brien, Susan Roberts, Paul Hakendorf, David I Ben-Tovim, Paddy a Phillips, and Campbell H Thompson. Outcomes of establishing an acute assessment unit in the general medical service of a tertiary teaching hospital. *The Medical journal of Australia*, 192(7):384–7, apr 2010. ISSN 0025-729X.

[152] X. Li, P. Beullens, D. Jones, and M. Tamiz. An integrated queuing and multiobjective bed allocation model with application to a hospital in China. *Journal of the Operational Research Society*, 60(3):330–338, 2009. ISSN 01605682. doi: 10.1057/palgrave.jors.2602565.

[153] Bohui Liang, Ayten Turkcan, Mehmet Erkan Ceyhan, and Keith Stuart. Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. *International Journal of Production Research*, 53(24):7177–7190, dec 2015. ISSN 1366588X. doi: 10.1080/00207543.2014.988891.

[154] Nelly Litvak, Marleen van Rijsbergen, Richard J. Boucherie, and Mark van Houdenhoven. Managing the overflow of intensive care. *European Journal of Operational Research*, 185(3):1–16, 2006. ISSN 03772217. doi: 10.1016/j.ejor.2006.08.021.

[155] Nan Liu, Serhan Ziya, and Vidyadhar G. Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2):347–364, 2010. doi: 10.1287/msom.1090.0272.

[156] Delaney Lori. The challenges of an integrated governance process in healthcare. *Clinical Governance: An International Journal*, 20(2):74–81, jan 2015. ISSN 1477-7274. doi: 10.1108/CGIJ-02-2015-0005.

[157] M Mackay. Practical experience with bed occupancy management and planning systems: an australian view. *Health Care Management Science*, 4(1):47–56, 2001.

[158] M Mackay and M Lee. Choice of models for the analysis and forecasting of hospital beds. *Health Care Management Science*, 8(3):221–230, 2005.

[159] F Mallor and C Azcarate. Combining optimization with simulation to obtain credible models for intensive care units. *Annals of Operations Research*, 221(1):255–271, 2014.

[160] F Mallor, C Azcarate, and J Barado. Control problems and management policies in health systems: application to intensive care units. *Flexible Services and Manufacturing Journal*, pages 1–28, 2014.

[161] Fermín Mallor, Cristina Azcárate, and Julio Barado. Optimal control of ICU patient discharge: from theory to implementation. *Health Care Management Science*, 18(3):234–250, 2015. ISSN 13869620. doi: 10.1007/s10729-015-9320-8.

[162] Avishai Mandelbaum and Martin I. Reiman. On pooling in queueing networks. *Management Science*, 44(7):971–981, 1998. doi: 10.1287/mnsc.44.7.971.

[163] Avishai Mandelbaum, Petar Momčilović, and Yulia Tseytlin. On Fair Routing from Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers. *Management Science*, 58(7):1273–1291, 2012. ISSN 0025-1909. doi: 10.1287/mnsc.1110.1491.

[164] E Marcon, S Kharraja, N Smolski, B Luquet, and JP Viale. Determining the number of beds in the postanesthesia care unit: a computer simulation flow approach. *Anesthesia and Analgesia*, 96(5):1415–1423, 2003.

[165] YN Marmor, TR Rohleder, DJ Cook, TR Huschka, and JE Thompson. Recovery bed planning in cardiovascular surgery: a simulation case study. *Health Care Management Science*, 16(4):314–327, 2013.

[166] Joren Marynissen and Erik Demeulemeester. Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, 272(2):407–419, 2019. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.03.001.

[167] BJ Masterson, TG Mihara, G Miller, SC Randolph, ME Forkner, and AL 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.

[168] Marie E. Matta and Sarah Stock Patterson. Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. *Health Care Management Science*, 10(2):173–194, may 2007. ISSN 13869620. doi: 10.1007/s10729-007-9010-2.

[169] Jerrold H. May, David P. Strum, and Luis G. Vargas. Fitting the lognormal distribution to surgical procedure times*. *Decision Sciences*, 31(1):129–148, 2000. doi: 10.1111/j.1540-5915.2000.tb00927.x.

[170] JO McClain. A model for regional obstetric bed planning. *Health Services Research*, 13(4):378–394, 1978.

[171] ML McManus, MC Long, A Cooper, and E Litvak. Queuing theory accurately models the need for critical care resources. *Anesthesiology*, 100(5):1271–1276, 2004.

[172] Marion E T McMurdo and Miles D Witham. Unnecessary ward moves. *Age and ageing*, 42(5):555–6, sep 2013. ISSN 1468-2834. doi: 10.1093/ageing/aft079.

[173] D.K. Min and Y. Yih. Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, 206(3):642–652, 2010. ISSN 0377-2217.

[174] Jyoti R. Munavalli, Shyam Vasudeva Rao, Aravind Srinivasan, and G. G. van Merode. Integral patient scheduling in outpatient clinics under demand uncertainty to minimize patient waiting times. *Health Informatics Journal*, 2020. ISSN 17412811. doi: 10.1177/1460458219832044. in press, forthcoming.

[175] Mark Murray and Donald M. Berwick. Advanced Access Reducing Waiting and Delays in Primary Care. *JAMA*, 289(8):1035–1040, 02 2003. ISSN 0098-7484. doi: 10.1001/jama.289.8.1035.

[176] Navonil Mustafee, Terry Lyons, Paul Rees, Lee Davies, Mark Ramsey, and Michael D. Williams. Planning of bed capacities in specialized and integrated care units: Incorporating bed blockers in a simulation of surgical throughput. *Proceedings - Winter Simulation Conference*, pages 1–12, 2012. ISSN 08917736. doi: 10.1109/WSC.2012.6465102.

[177] JM 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.

[178] JP Oddoye, MA Yaghoobi, M Tamiz, DF 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, 2007.

[179] 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, feb 2009. ISSN 03772217. doi: 10.1016/j.ejor.2007.10.029.

[180] OECD. Health care resources. stats.oecd.org, May 2020.

[181] Jan Olhager, Martin Rudberg, and Joakim Wikner. Long-term capacity management: Linking the perspectives from manufacturing strategy and sales and operations planning. *International Journal of Production Economics*, 69(2):215 – 225, 2001. ISSN 0925-5273. doi: https://doi.org/10.1016/S0925-5273(99)00098-5. Strategic Planning for Production Systems.

[182] Irene Papanicolas, Liana R. Woskie, and Ashish K. Jha. Health Care Spending in the United States and Other High-Income Countries. *JAMA*, 319(10):1024–1039, 03 2018. ISSN 0098-7484. doi: 10.1001/jama.2018.1150.

[183] Lillrank Paul. Integration and coordination in healthcare: an operations management view. *Journal of Integrated Care*, 20(1):6–12, jan 2012. ISSN 1476-9018. doi: 10.1108/14769011211202247.

[184] SA Paul, MC Reddy, and CJ DeFlitch. A systematic review of simulation studies investigating emergency department overcrowding. *Simulation*, 2010.

[185] T. O. Paulussen, N. R. Jennings, K. S. Decker, and A Heinzl. Distributed patient scheduling in hospitals. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 1224–1229, 2003.

[186] Canan Pehlivan, Vincent Augusto, Xiaolan Xie, and Catherine Crenn-Hebert. Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach. In *IEEE International Conference on Automation Science and Engineering*, pages 137–142, 2012. ISBN 9781467304283. doi: 10.1109/CoASE.2012.6386385.

[187] Eduardo Pérez, Lewis Ntaimo, Wilbert E. Wilhelm, Carla Bailey, and Peter McCormack. Patient and resource scheduling of multi-step medical procedures in nuclear medicine. *IIE Transactions on Healthcare Systems Engineering*, 1(3): 168–184, jul 2011. ISSN 19488319. doi: 10.1080/19488300.2011.617718.

[188] Eduardo Pérez, Lewis Ntaimo, César O. Malavé, Carla Bailey, and Peter McCormack. Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. *Health Care Management Science*, 16(4):281–299, dec 2013. ISSN 13869620. doi: 10.1007/s10729-013-9224-4.

[189] Gilles Pesant. Balancing nursing workload by constraint programming. In Claude-Guy Quimper, editor, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9676, pages 294–302. Springer International Publishing, 2016. ISBN 9783319339535. doi: 10.1007/978-3-319-33954-2_21.

[190] Jesse M Pines, Robert J Batt, Joshua a Hilton, and Christian Terwiesch. The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Annals of emergency medicine*, 58(4):331–40, oct 2011. ISSN 1097-6760. doi: 10.1016/j.annemergmed.2011.03.004.

[191] Michael E Porter. What Is Value in Health Care? *New England Journal of Medicine*, 363(26):2477–2481, 2010. doi: 10.1056/NEJMp1011024.

[192] Michael E Porter and Thomas H Lee. The strategy that will fix health care. *Harvard Business Review*, 91(10):1–19, 2013.

[193] Michael E. Porter, Stefan Larsson, and Thomas H. Lee. Standardizing patient outcomes measurement. *New England Journal of Medicine*, 374(6):504–506, 2016. doi: 10.1056/NEJMp1511701. PMID: 26863351.

[194] Micheal E. Porter and Elizabeth O. Teisberg. *Redefining Health Care: Creating Value-Based Competition on Results*. Harvard Business School Press, Boston, 2006.

[195] Martin Puterman. Finite-horizon markov decision processes. In *Markov Decision Processes*, chapter 4, pages 74–118. John Wiley & Sons, Ltd, 2008. ISBN 9780470316887. doi: 10.1002/9780470316887.ch4.

[196] Schmidt R., Geisler S., and Spreckelsen C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. *BMC medical informatics and decision making*, 13(1):3, 2013. ISSN 14726947. doi: Article.

[197] M Ramakrishnan, D Sier, and PG Taylor. A two-time-scale model for hospital patient flow. *IMA Journal of Management Mathematics*, 16(3):197–215, 2005.

[198] JC Ridge, SK Jones, MS Nielsen, and AK Shahani. Capacity planning for intensive care units. *European Journal of Operational Research*, 105(2):346–355, 1998.

[199] Stewart Robinson. *Simulation: the practice of model development and use*, volume 50. Wiley Chichester, 2004.

[200] Aleda V. ROTH and Roland van DIERDONCK. Hospital resource planning: Concepts, feasibility, and framework. *Production and Operations Management*, 4(1):2–29, 1995. doi: 10.1111/j.1937-5956.1995.tb00038.x.

[201] Yazan F. Roumani, Yaman Roumani, Joseph K. Nwankpa, and Mohan Tanniru. Classifying readmissions to a cardiac intensive care unit. *Annals of Operations Research*, 263(1-2):429–451, 2018. ISSN 15729338. doi: 10.1007/s10479-016-2350-x.

[202] P. Santibanez, 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. ISSN 1386-9620.

[203] Antoine Sauré, Jonathan Patrick, Scott Tyldesley, and Martin L Puterman. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research*, 223(2):573–584, 2012. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2012.06.046.

[204] A. J. (Thomas) Schneider, P. Luuk Besselink, Maartje E. Zonderland, Richard J. Boucherie, Wilbert B. Van Den Hout, Job Kievit, Paul Bilars, A. Jaap Fogteloo, and Ton J. Rabelink. Allocating emergency beds improves the emergency admission flow. *Journal of Applied Analytics*, 48(4):384–394, 2018. ISSN 1526551X. doi: 10.1287/inte.2018.0951.

[205] Gerard Scholten, Linda Muijsers-Creemers, Jan Moen, and Roland Bal. Structuring ambiguity in hospital governance. *The International Journal of Health Planning and Management*, 34(1):443–457, 2019. doi: 10.1002/hpm.2693.

[206] Ian Scott, Louella Vaughan, and Derek Bell. Effectiveness of acute medical units in hospitals: A systematic review. *International Journal for Quality in Health Care*, 21(6):397–407, 2009. ISSN 13534505. doi: 10.1093/intqhc/mzp045.

[207] Antonio Sebastiano, Valeria Belvedere, Alberto Grando, and Antonio Giangreco. The effect of capacity management strategies on employees' well-being: A quantitative investigation into the long-term healthcare industry. *European Management Journal*, 35(4):563–573, 2017. ISSN 0263-2373. doi: https://doi.org/10.1016/j.emj.2016.12.001.

[208] Claire Senot, Aravind Chandrasekaran, and Peter T. Ward. Role of bottom-up decision processes in improving the quality of health care delivery: A contingency perspective. *Production and Operations Management*, 25(3):458–476, 2016. doi: 10.1111/poms.12404.

[209] AK Shahani, SA Ridley, and MS Nielsen. Modelling patient flows as an aid to decision making for critical care capacities and organisation. *Anaesthesia*, 63 (10):1074–1080, 2008.

[210] A Shmueli, CL Sprung, and EH Kaplan. Optimizing admissions to an intensive care unit. *Health Care Management Science*, 6(3):131–136, 2003.

[211] W Shonick and JR Jackson. An improved stochastic model for occupancy-related random variables in general-acute hospitals. *Operations Research*, 21(4): 952–965, 1973.

[212] Adam J Singer, Henry C Thode, Peter Viccellio, and Jesse M Pines. The association between length of emergency department boarding and mortality. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 18(12):1324–9, dec 2011. ISSN 1553-2712. doi: 10.1111/j.1553-2712.2011.01236.x.

[213] Mustafa Y. Sir, Bayram Dundar, Linsey M. Barker Steege, and Kalyan S. Pasupathy. Nurse-patient assignment models considering patient acuity metrics and nurses' perceived workload. *Journal of Biomedical Informatics*, 55:237–248, 2015. ISSN 15320464. doi: 10.1016/j.jbi.2015.04.005.

[214] AA Sissouras and B Moores. The optimum number of beds in a coronary care unit. *Omega*, 4(1):59–65, 1976.

[215] Nigel Slack, Stuart Chambers, and Robert Johnston. *Operations management*. Pearson Education, 2010.

[216] Vicki L. Smith-Daniels, Sharon B. Schweikhart, and Dwight E. Smith-Daniels. Capacity management in health care services: Review and future research directions*. *Decision Sciences*, 19(4):889–919, 1988. doi: 10.1111/j.1540-5915.1988.tb00310.x.

[217] RW Swain, KE Kilpatrick, and JJ Marsh. Implementation of a model for census prediction and control. *Health Services Research*, 12(4):380–395, 1977.

[218] Gordon Swartzman. The Patient Arrival Process in Hospitals. *Health services research*, 5:320–329, 1970.

[219] GJ Taylor, SI McClean, and PH 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.

[220] Steven Thompson, Manuel Nunez, Robert Garfinkel, and Matthew D. Dean. OR Practice—Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges. *Operations Research*, 57(2):261–273, 2009. ISSN 0030-364X. doi: 10.1287/opre.1080.0584.

[221] PM Troy and L Rosenberg. Using simulation to determine the need for icu beds for surgery patients. *Surgery*, 146(4):608–617, 2009.

[222] John W Tukey. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99–114, jun 1949. ISSN 0006341X. doi: 10.2307/3001913.

[223] Martin Utley, Steve Gallivan, Katie Davis, Patricia Daniel, Paula Reeves, and Jennifer Worrall. Estimating bed requirements for an intermediate care facility. *European Journal of Operational Research*, 150(1):92–100, 2003. ISSN 03772217. doi: 10.1016/S0377-2217(02)00788-9.

[224] Martin Utley, Steve Gallivan, Tom Treasure, and Oswaldo 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. ISSN 13869620. doi: 10.1023/A:1023333002675.

[225] Martin Utley, Steve Gallivan, and Mark Jit. How to take variability into account when planning the capacity for a new hospital unit. In *Health Operations Management: Patient Flow Logistics in Health Care*, Routledge health management series, pages 146–161. Routledge Taylor & Francis Group, 2005. ISBN 0203356799. doi: 10.4324/9780203356791.

[226] N. M. (Maartje) van de Vrugt, A. J. (Thomas) Schneider, Maartje E. Zonderland, David A. Stanford, and Richard J. Boucherie. Operations research for occupancy modeling at hospital wards and its integration into practice. In Cengiz Kahraman and Y Ilker Topcu, editors, *Operations Research Applications in Health Care Management*, volume 262, pages 101–137. Springer US, Boston, MA, 2018. doi: 10.1007/978-3-319-65455-3_5.

[227] Wil M. P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, Berlin, Heidelberg, Germany, 2011. ISBN 978-3-642-19345-3. doi: 10.1007/978-3-642-19345-3.

[228] Annelies van der Ham, Henri Boersma, Arno van Raak, Dirk Ruwaard, and Frits van Merode. Identifying logistical parameters in hospitals: Does literature reflect integration in hospitals? a scoping study. *Health Services Management Research*, 32(3):158–165, 2019. doi: 10.1177/0951484818813488. PMID: 30463453.

[229] N. M. van Dijk and N. Kortbeek. Erlang loss bounds for OT-ICU systems. *Queueing Systems*, 63(1):253–280, 2009. ISSN 02570130. doi: 10.1007/s11134-009-9149-2.

[230] J. Theresia van Essen, Mark van Houdenhoven, and Johann L. Hurink. Clustering clinical departments for wards to achieve a prespecified blocking probability. *OR Spectrum*, 37(1):243–271, 2015. ISSN 14366304. doi: 10.1007/s00291-014-0368-5.

[231] J.T. van Essen, J.M. Bosch, E.W. Hans, M. van Houdenhoven, and J.L. Hurink. Reducing the number of required beds by rearranging the or-schedule. *Or Spectrum*, 36(3):585–605, 2014. ISSN 0171-6468.

[232] Liza van Lonkhuyzen. 'is het niet een gevaar als artsen níet op dit soort technologie leunen?', January 2020. URL `https://www.nrc.nl/nieuws/2020/01/28/het-algoritme-maakt-de-dokter-beter-a3988495`. in Dutch.

[233] 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 operating room departments. *OR Spectrum*, 30(2):355–374, 2008. ISSN 0171-6468.

[234] Catharina J. Van Oostveen, Aleida Braaksma, and Hester Vermeulen. Developing and testing a computerized decision support system for nurse-to-patient assignment: A multimethod study. *CIN - Computers Informatics Nursing*, 32 (6):276–285, 2014. ISSN 15389774. doi: 10.1097/CIN.0000000000000056.

[235] Nederlandse Vereniging van Ziekenhuizen. Zorg in de toekomst: standpunten [in dutch]. Booklet available at www.nvz-ziekenhuizen.nl, Utrecht, 2013.

[236] Peter T VanBerkel and John T Blake. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health care management science*, 10(4):373–85, 2007. ISSN 1386-9620. doi: 10.1007/s10729-007-9035-6.

[237] Peter T Vanberkel, Richard J Boucherie, Erwin W Hans, Johann L Hurink, and Nelly Litvak. A Survey of Health Care Models that Encompass Multiple Departments. *International Journal of Health Management and Information*, 1 (1):37–69, 2010.

[238] Peter T. Vanberkel, Richard J. Boucherie, Erwin W. Hans, Johann L. Hurink, and Nelly Litvak. Efficiency evaluation for pooling resources in health care. *OR Spectrum*, 34(2):371–390, 2012. ISSN 1436-6304. doi: 10.1007/s00291-010-0228-x.

[239] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, W. 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. ISSN 0160-5682.

[240] 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 and Analgesia*, 112(6):1472–1479, 2011. ISSN 0003-2999.

[241] Wim Vancroonenburg, Federico Della Croce, Dries Goossens, and Frits C R Spieksma. The Red-Blue transportation problem. *European Journal of Operational Research*, 237(3):814–823, 2014. ISSN 03772217. doi: 10.1016/j.ejor.2014.02.055.

[242] Wim Vancroonenburg, Patrick De Causmaecker, and Greet Vanden Berghe. A study of decision support models for online patient-to-room assignment planning. *Annals of Operations Research*, 239(1):253–271, 2016. ISSN 15729338. doi: 10.1007/s10479-013-1478-1.

[243] C Vasilakis and E El-Darzi. A simulation study of the winter bed crisis. *Health Care Management Science*, 4(1):31–36, 2001.

[244] C Vasilakis, E El-Darzi, and P Chountas. A decision support system for measuring and modelling the multi-phase nature of patient flow in hospitals. In Panagiotis Chountas, Ilias Petrounias, and Janusz Kacprzyk, editors, *Intelligent Techniques and Tools for Novel System Architectures*, volume 109 of *Studies in Computational Intelligence*, book section 12, pages 201–217. Springer Berlin Heidelberg, 2008.

[245] Ivan Vermeulen, Sander Bohte, Koye Somefun, and Han La Poutré. Multi-agent Pareto appointment exchanging in hospital patient scheduling. *Service Oriented Computing and Applications*, 1(3):185–196, oct 2007. ISSN 18632386. doi: 10.1007/s11761-007-0012-1.

[246] J. Vissers and R. Beech. *Health Operations Management: Patient flow logistics in health care*. Routledge, New York, 2005.

[247] 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. doi: 10.1080/09537280175039771 6.

[248] J. Volland, A. Fügener, J. Schoenfelder, and J.O. Brunner. Material logistics in hospitals: A literature review. *Omega (United Kingdom)*, 69:82–101, 2017. doi: 10.1016/j.omega.2016.08.004. cited By 43.

[249] B L Welch. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, jan 1947. doi: 10.1093/biomet/34.1-2.28.

[250] Greg Werker, Antoine Sauré, John French, and Steven Shechter. The use of discrete-event simulation modelling to improve radiation therapy planning processes. *Radiotherapy and Oncology*, 92(1):76–82, jul 2009. ISSN 01678140. doi: 10.1016/j.radonc.2009.03.012.

[251] Jeffrey L Whitten, Lonnie D Bentley, and Kevin C Dittman. *Systems Analysis and Design Methods 5e*. McGraw-Hill Higher Education, 2000.

[252] J. Williams, S. Dumont, J. Parry-Jones, I. Komenda, J. Griffiths, and V. Knight. Mathematical modelling of patient flows to predict critical care capacity required following the merger of two district general hospitals into one. *Anaesthesia*, 70 (1):32–40, 2015. ISSN 13652044. doi: 10.1111/anae.12839.

[253] Wayne L Winston and Jeffrey B Goldberg. *Operations research: applications and algorithms*, volume 3. Thomson Brooks/Cole Belmont, 2004.

[254] Z. Yahia, A.B. Eltawil, and N.A. Harraz. The operating room case-mix problem under uncertainty and nurses capacity constraints. *Health Care Management Science*, 19(4):383–394, 2016. ISSN 1572-9389.

[255] M Yang, MJ Fry, J Raikhelkar, C Chin, A Anyanwu, J Brand, and C Scurlock. A model to create an efficient and equitable admission policy for patients arriving to the cardiothoracic icu. *Critical Care Medicine*, 41(2):414–422, 2013.

[256] Muer Yang, Michael J. Fry, and Corey Scurlock. The ICU will see you now: Efficient-equitable admission control policies for a surgical ICU with batch arrivals. *IIE Transactions (Institute of Industrial Engineers)*, 47(6):586–599, 2015. ISSN 15458830. doi: 10.1080/0740817X.2014.955151.

[257] Philip Yoon, Ivan Steiner, and Gilles Reinhardt. Analysis of factors influencing length of stay in the emergency department. *CJEM : Canadian journal of emergency medical care = JCMU : journal canadien de soins médicaux d'urgence*, 5 (3):155–61, 2003. ISSN 1481-8035.

[258] Christos Zacharias and Michael Pinedo. Managing Customer Arrivals in Service Systems with Multiple Identical Servers. *Manufacturing & Service Operations Management*, 19(4):639–656, 2017. ISSN 1523-4614. doi: 10.1287/msom.2017.0629.

[259] Zhu Zhecheng. An online short-term bed occupancy rate prediction procedure based on discrete event simulation. *Journal of Hospital Administration*, 3(4): p37, 2014. ISSN 1927-6990. doi: 10.5430/jha.v3n4p37.

[260] Jian-Cang Zhou, Kong-Han Pan, Dao-Yang Zhou, San-Wei Zheng, Jian-Qing Zhu, Qiu-Ping Xu, and Chang-Liang Wang. High hospital occupancy is associated with increased risk for patients boarding in the emergency department. *The American journal of medicine*, 125(4):416.e1–7, apr 2012. ISSN 1555-7162. doi: 10.1016/j.amjmed.2011.07.030.

[261] Z Zhu. Impact of different discharge patterns on bed occupancy rate and bed waiting time: a simulation approach. *Journal of Medical Engineering & Technology*, 35(6-7):338–343, 2011.

[262] Z Zhu. An online short-term bed occupancy rate prediction procedure based on discrete event simulation. *Journal of Hospital Administration*, 3(4):p37, 2014.

[263] W H M Zijm. Towards intelligent manufacturing planning and control systems. *OR-Spektrum*, 22(3):313–345, 2000. ISSN 1436-6304. doi: 10.1007/s002919900032.

[264] Maartje E. Zonderland and Richard J. Boucherie. Queuing networks in health care systems. In Randolph Hall, editor, *Handbook of Healthcare System Scheduling*, pages 201–243. Springer US, Boston, MA, 2012. ISBN 978-1-4614-1734-7. doi: 10.1007/978-1-4614-1734-7_9.

[265] Maartje E Zonderland, Richard J Boucherie, Michael W Carter, and David A Stanford. Operations Research for Health Care Modeling the effect of short stay units on patient admissions. *Operations Research for Health Care*, 5:21–27, 2015. ISSN 2211-6923. doi: 10.1016/j.orhc.2015.04.001.

Acronyms

ADP	Approximate Dynamic Programming
AMU	Acute Medical Unit
C	Consultation
CEO	Chief Executive Officer
CM	Capacity Management
DES	Discrete Event Simulation
E	Echocardiogram
Ec	Electrocardiogram
ED	Emergency Department
EMR	Electronic Medical Record
FCFS	First Come First Serve
GEN	General ward
H	Holter test
I	ICD check
ICD	Implantable Cardiovascular Device
ICM	Integral Capacity Management
ICU	Intensive Care Unit
ILP	Integer Linear Programming
IPU	Integrated Practice Unit
JBH	Jeroen Bosch Hospital
JIT	Just-in-Time
JSQ	Join the Shortest Queue
KPI	Key Performance Indicator
LoS	Length of Stay
LRPT	Longest Remaining Processing Time
LUMC	Leiden University Medical Center
MDP	Markov Decision Process
MGH	Massachusetts General Hospital
MILP	Mixed Integer Linear Programming
MIP	Mixed Integer Programming
MSS	Master Surgery Schedule
NP	Non-deterministic Polynomial-time
OBS	Obstetric Ward
OR	Operations Research

OT	Operating Theater
PACU	Post-Anesthesia Care Unit
RCT	Randomized Controlled Trial
rel LOS	relative Length of Stay
SA	Simulated Annealing
ST	Exercise Stress Test
VBHC	Value Based Health Care
WDW	Weekday Ward

Summary

Introduction

Care pathways in hospitals usually encompass multiple resources and healthcare professionals. This makes managing hospital processes and capacities challenging. To prevent myopic optimization, process improvements should consider multiple steps in care pathways. This dissertation aims to improve complex decision-making that integrally manages capacity for care pathways. Operations research may play a crucial role by analyzing such capacity decisions in a safe environment before actual implementation. However, despite the vast amount of available research and its potential, it appears that the actual implementation of operations research models and results in healthcare practice is rarely described in the literature. This is surprising, as implementation is the ultimate step in realizing improvement. We try to improve this final step by distinguishing two approaches: (1) organizing the timing and alignment of the optimal decisions among related capacities and (2) analyzing (near) optimal capacity decisions considering multiple capacities.

Part I Integral Capacity Management in Hospitals

We start this thesis analyzing the organization of capacity decisions in hospitals in *Chapter 2*. We observe that current capacity management (CM) in hospitals organizes departments as silos, or even as single cost centers, with their own operations management systems and a top-down deployment of decision-making processes. We aim to realize this potential by breaking through the siloed system, by optimizing flow rather than myopically optimizing utilization. We do this by aligning capacity in care pathways. We propose integral capacity management (ICM) as the successor to CM. This is the first theoretical introduction of ICM. We distinguish three dimensions for organizational integration: hierarchical, patient-centeredness, and domain. We discuss alignments on and between these dimensions to integrally organize capacity decisions. Hierarchical integration concerns top-down and bottom-up decision-making processes, in which higher levels set boundaries, targets and planning objectives (i.e. increasingly disaggregated information) for lower levels and lower levels provide input for improvement of decision-making on higher levels. Patient-centeredness concerns the coordination and alignment of capacity across departments and organizations to optimize care pathways. Domain integration encompasses alignment of managerial domains: clinical, financial and nonrenewable resources. This study is a first step

for theoretical development of ICM. We therefore derive multiple directions for future research.

In *Chapter 3* we review operations research (OR) literature applied to hospital wards. Based on logistical characteristics and patient flow problems, we distinguish the following particular ward types: intensive care, acute medical units, obstetric wards, weekday wards, and general wards. We analyze typical trade-offs of performance indicators for each ward type, review OR models commonly applied to it, and discuss typical capacity management and planning decisions. Additionally, we provide three illustrative cases, discuss both theoretical and practical challenges, and provide directions for future research. With this review we aim to guide both researchers and healthcare professionals dealing with hospital bed capacity and on which OR models best suits each specific capacity decision and the type of ward.

Part II Integral Capacity Planning in Hospitals

In *Chapter 4* we analyze the process of emergency admissions. The increasing number of admissions to hospital emergency departments (EDs) during the past decade has resulted in overcrowded EDs and decreased quality of care. The emergency admission flow that we discuss in this study relates to three types of hospital departments: EDs, acute medical unit (AMUs), and inpatient wards. The study in this chapter has two objectives: (1) to evaluate the impact of allocating beds in inpatient wards to accommodate emergency admissions and (2) to analyze the impact of pooling the number of beds allocated for emergency admissions in inpatient wards. To analyze the impact of various allocations of emergency beds, we develop a discrete event simulation model. We evaluate the bed allocation scenarios using three performance indicators: (1) the length of stay in the AMU, (2) the fraction of patients refused admission, and (3) the utilization of allocated beds. We develop two heuristics to allocate beds to wards and show that pooling beds improves performance. The partnering hospital has embedded a decision support tool based on our simulation model into its planning and control cycle. The hospital uses it every quarter and updates it with data on a 1-year rolling horizon. This strategy has substantially reduced the number of patients who are refused emergency admission.

Chapter 5 analyzes optimal surgery schedules considering multiple resources. Surgery groups are clustered surgery procedure types that share comparable characteristics (e.g. expected duration). Scheduling operating theater (OT) blocks leaves many options for operational surgery scheduling and this increases the variation in usage of both the OT and downstream beds. Therefore, we schedule surgery groups to reduce the options for operational scheduling, ultimately bridging the gap between tactical and operational scheduling. We propose a single step mixed integer linear programming (MILP) approach that approximates the bed and OR usage along with a simulated annealing approach. Both approaches are compared on a real-life data set and results show that the MILP performs best in terms of solution quality and computation time. Furthermore, the results show that our model may improve the OR utilization from 71% to 85% and decrease the bed usage variation from 53 beds to 11 beds compared to

historical data. To show the potential and robustness of our model, we discuss several variants of the model requiring minor modifications. The use of surgery groups makes it easier to implement our model in practice as, for operational planners, it is instantly clear where to schedule different types of surgery.

Chapter 6 presents an innovative methodology to overcome Markov decision process (MDP) intractability for online multi-appointment scheduling problems. As a result of increasing treatment options and far-reaching specialization, the number of appointments for patients has increased. This makes appointment schedules fragile as dependencies between schedules increase. Decomposing the decisions (e.g. accept/reject and allocation decisions) allows us to analytically solve practical online multi-appointment scheduling problems. We use an MDP to derive optimal decisions for accepting or rejecting new arrivals based on capacity availability and future arrivals. Once accepted, we developed an ILP to allocate patients to their next appointment. Based on a case study at the Leiden University Medical Center cardiology outpatient clinic, we then present the implementation of our model for a real-life instance. We compare the performance of our approach with a heuristic and show that our approach outperforms that of the heuristic. Furthermore, we show a full implementation and analyze the impact in practice.

Future developments for Hospital Capacity Management & Planning

We see multiple future directions for ICM and planning in hospitals. In Western countries, most hospitals have emerged from the digitization era and are now discovering the value of the newly available information. This will explode the number of research opportunities for all types of analytics (i.e. descriptive, predictive and prescriptive). To increase their impact in practice, researchers should embrace data-driven optimization. For example, both descriptive and predictive analytics may be used to improve the input data for prescriptive analytics and therefore improve the results of prescriptive analytics. Prescriptive analytics quickly become intractable when the number of decisions increases. Therefore, the number of decisions resulting from descriptive and predictive analytics should be balanced to ensure tractable prescriptive analytics. This may result in decreasing quality of both descriptive and predictive analytics. Analyzing trade-offs between quality measures of different types of analytics could be an interesting research topic. Capacity planning and management software systems that in real-time automate all steps of capacity planning and management may be developed using data-driven optimization. Ultimately, this may further reduce the waste caused by failures in care coordination.

Currently, regulation of data is organized well. As the aforementioned scalability of innovations is finally taking place, the exchange of data and information has proven to be difficult as there are still compatibility and data ownership problems. *"Data is the new oil"* [76]. First and foremost, health data will increase the quality of care as information availability improves. For operations research this will raise opportunities as more data about the total care pathways of patients becomes, to some extent, available through, for example, wearables.

In the coming years, clinical practice will become more technical as a result of the

digitization era, technical medicine and the introduction of health analytics. The first small steps are currently being taken. For example, predictive medicine already supports medical decision-making on antibiotic dosage in sepsis for intensive care patients [232]. This also has an impact on capacity planning and management. Therefore, both analytical and medical scholars should further embrace each other's field of expertise. Together, they can integrally shape the hospital of the future.

Samenvatting

Introductie

Zorgpaden in ziekenhuizen bestaan in toenemende mate uit verschillende capaciteiten (zoals bedden, diagnostiek en operatiekamers) en zorgprofessionals. Dit maakt het managen van ziekenhuisprocessen en -capaciteiten uitdagend. Om eenzijdige optimalisatie van zorgpaden te voorkomen (bijvoorbeeld uitsluitend het optimaliseren van de operatieplanning), moet er in procesverbeteringen rekening gehouden worden met meerdere stappen (integrale aansturing) in een zorgpad. Dit proefschrift is gericht op het verbeteren van complexe besluitvorming op integrale capaciteitsplanning en -management. Operations research kan een cruciale rol spelen bij het analyseren van capaciteitsbeslissingen in een veilige omgeving voordat deze worden geimplementeerd in de praktijk. Ondanks de enorme hoeveelheid aan onderzoek en het potentieel ervan, lijkt het erop dat de daadwerkelijke implementatie van operations research modellen en/of resultaten in de gezondheidszorgpraktijk zelden wordt beschreven in de literatuur. Dit is verrassend, omdat implementatie van uitkomsten de ultieme stap is om procesverbetering te realiseren. We proberen deze laatste stap te verbeteren door twee benaderingen onderscheiden: (1) het organiseren van besluitvorming en afstemming van de optimale beslissingen tussen gerelateerde capaciteiten en (2) het analyseren van optimale capaciteitsbeslissingen rekening houdend met verschillende capaciteiten.

Deel I Integraal Capaciteitsmanagement in Ziekenhuizen

We starten dit proefschrift met het analyseren van de organisatie rondom capaciteitsbeslissingen in ziekenhuizen in *Hoofdstuk 2*. We observeren dat capaciteitsmanagement (CM) in ziekenhuizen momenteel georganiseerd is in silo's, of zelfs als afzonderlijke resultaatverantwoordelijke eenheden, met hun eigen managementsystemen. Optimalisatie van patiëntdoorstroming vraagt de verschillende capaciteiten in zorgpaden op elkaar af te stemmen en daardoor dient het silo-systeem doorbroken te worden. Wij stellen daarom dat integraal capaciteitsmanagement (ICM) de opvolger van CM is. In dit proefschrift wordt ICM voor het eerst theoretisch onderbouwd. We onderscheiden drie dimensies voor integratie op capaciteitsbesluitvorming: hiërarchisch niveau, patiëntgerichtheid en managementdomein. Hierbij richten wij ons op de afstemmingen binnen en tussen deze dimensies om capaciteitsbeslissingen integraal te organiseren. Hiërarchische integratie heeft betrekking op zowel top-down als bottom-up besluitvormingsprocessen, waarbij hogere niveaus kaders, doelen en planningsdoelen stellen

voor lagere niveaus (bijvoorbeeld het detailniveau van informatie neemt toe) en lagere niveaus input leveren voor het verbeteren van besluitvorming op hogere niveaus. Hierin onderscheiden we de volgende niveaus: strategisch, tactisch en operationeel. Patiëntgerichtheid betreft de coördinatie en afstemming van capaciteiten tussen afdelingen en organisaties om zorgpaden te optimaliseren. Met andere woorden, het creëren van flow. Domeinintegratie omvat afstemming van managementdomeinen: klinisch, financieel en supply chain. Deze studie is een eerste stap voor verdere theoretische ontwikkeling van ICM. We hebben daarom meerdere richtingen beschreven voor toekomstig onderzoek.

In *Hoofdstuk 3* beschrijven we systematisch OR-literatuur waarvan modellen worden toegepast op ziekenhuisafdelingen. Op basis van logistieke kenmerken en problemen in patiëntenstromen onderscheiden we de volgende specifieke typen verpleegafdelingen: intensive care afdeling, acute opname afdeling, verloskunde afdeling, kortverblijfafdeling en algemene verpleegafdeling. We analyseren typische afwegingen van prestatieindicatoren voor elk type verpleegafdeling en de OR-modellen die erop worden toegepast en bespreken typische capaciteits- en planningsbeslissingen. Daarnaast presenteren we drie casestudy's, bespreken we zowel theoretische als praktische uitdagingen en geven we aanwijzingen voor toekomstig onderzoek. Met deze review willen we zowel onderzoekers als zorgprofessionals die zich richten op verpleegafdelingen ondersteunen met capaciteitsvraagstukken en hoe en welke OR-technieken / modellen toepasbaar zijn.

Part II Integraal Capaciteitsplanning in Ziekenhuizen

In *Hoofdstuk 4* analyseren we het proces van acute opnames. Het toenemende aantal opnames vanaf spoedeisende hulpafdelingen (SEH's) van ziekenhuizen in het afgelopen decennium heeft geresulteerd in overvolle SEH's en verminderde kwaliteit van de zorg. Het proces van spoedopnames in dit onderzoek heeft betrekking op drie soorten ziekenhuisafdelingen: SEH, acute opname afdeling (AOA) en verpleegafdelingen. Deze studie heeft twee doelstellingen: (1) het evalueren van de impact van het toewijzen van bedden op verpleegafdelingen voor spoedopnames en (2) het analyseren van de impact van het bundelen van het aantal bedden dat is toegewezen voor spoedopnames op verpleegafdelingen. Om de impact van verschillende toewijzingen van spoedbedden te analyseren hebben we een simulatiemodel ontwikkeld. We evalueren de scenario's voor bedtoewijzing met behulp van drie prestatie-indicatoren: (1) de verblijfsduur op de AOA, (2) het aantal geweigerde opnames en (3) de bezettingsgraad van toegewezen spoedbedden. We hebben twee heuristieken ontwikkeld om spoedbedden aan afdelingen toe te wijzen en prestaties te verbeteren door het bundelen van spoedbedden. Het LUMC heeft het simulatiemodel gebruikt binnen de planning- en controlcyclus door elk kwartaal met voortschrijdende horizon van 1 jaar opnieuw te bekijken welke verdeling van bedden wenselijk is. Deze strategie heeft het aantal geweigerde spoedopnames aanzienlijk verminderd.

Hoofdstuk 5 worden optimale operatieschema's geanalyseerd rekening houdend met meerdere capaciteiten. Operatiegroepen zijn geclusterde typen operatieve ingrepen met vergelijkbare kenmerken (bijvoorbeeld de verwachte duur van de ingreep). Het

plannen van OK-blokken laat veel opties over voor het operationeel plannen van operaties en dit vergroot de variatie in gebruik van zowel de OK als bedden. Daarom plannen we operatiegroepen om de mogelijkheden voor operationele planning te verkleinen en uiteindelijk de kloof tussen tactische en operationele planning te overbruggen. We stellen een mixed integer linear programming (MILP) -benadering voor die zowel OK- als bedbezetting optimaliseert. Daarnaast vergelijken we deze benadering met een simulated annealing-benadering. Beide benaderingen worden vergeleken op een dataset vanuit het LUMC en de resultaten tonen aan dat de MILP het beste presteert in termen van oplossingskwaliteit en rekentijd. Bovendien laten de resultaten zien dat ons model het gebruik van de operatiekamer kan verbeteren van 71% tot 85% en de variatie in bedgebruik kan verminderen van 53 bedden tot 11 bedden in vergelijking met de historische data. Om het potentieel en de robuustheid van ons model te laten zien, bespreken we daarnaast verschillende varianten van het model die kleine aanpassingen vereisen. Het gebruik van operatiegroepen maakt het eenvoudiger om ons model in de praktijk toe te passen, omdat het voor operationele planners duidelijk is waar verschillende soorten operaties moeten worden gepland.

Hoofdstuk 6 presenteert een innovatieve methodologie om voor online planningsproblemen met meerdere afspraken een oplosbaar Markov beslissingsproces (MDP) model te realiseren. Door toenemende behandelmogelijkheden en verregaande specialisatie neemt het aantal afspraken voor patiënten toe. Dit maakt afspraakschema's kwetsbaar naarmate de afhankelijkheden tussen schema's toenemen. Door de beslissingen te ontleden (bijvoorbeeld acceptatie / afwijzing en planningsbeslissing) kunnen we praktische online planningsproblemen met meerdere afspraken analytisch oplossen. We gebruiken een MDP om optimale beslissingen te analyseren voor het accepteren of weigeren van nieuwe patiënten op basis van capaciteitsbeschikbaarheid en toekomstige aankomsten. Na acceptatie hebben we een ILP ontwikkeld om patiënten toe te wijzen aan hun volgende afspraak. Op basis van een casestudy op de polikliniek Cardiologie van het LUMC presenteren we vervolgens de implementatie van ons model op deze situatie. We vergelijken de prestaties van onze aanpak met een heuristiek en laten zien dat onze aanpak beter presteert. Verder tonen we een volledige implementatie in de praktijk en analyseren we de impact ervan.

Toekomst van Integraal Capaciteitsmanagement & -planning

We zien meerdere toekomstige onderzoeksrichtingen voor ICM en integrale planning in ziekenhuizen. In westerse landen zijn de meeste ziekenhuizen uit het digitaliseringstijdperk gekomen en ontdekken nu de waarde van de nieuw beschikbare informatie. Hiermee zal het aantal onderzoeksmogelijkheden voor alle soorten analytics (bijv. beschrijvende, voorspellende en optimalisatie analytics) toenemen. Om de impact in de praktijk te vergroten, moeten onderzoekers datagedreven optimalisatie omarmen. Zo kunnen zowel beschrijvende als voorspellende analyses worden gebruikt als input voor optimalisatie analyses waardoor de resultaten van optimalisatie analyses verbeterd kunnen worden. De uitkomsten van voorspellende analyses worden snel onoplosbaar voor optimalisatie modellen wanneer het aantal beslissingen toeneemt. Daarom moet het aantal beslissingen welke voortvloeien uit beschrijvende en voorspellende analy-

ses in evenwicht zijn om te komen tot oplosbare optimalisatie modellen. Dit kan resulteren in een afnemende kwaliteit van zowel beschrijvende als voorspellende analyses. Het onderzoeken van de kwaliteit van de verschillende soorten analyses wanneer deze elkaar dienen kan een interessant onderzoeksthema zijn. Verder kan software voor capaciteitsplanning en managementsoftware die real-time alle stappen van capaciteitsplanning en -beheer automatiseert worden bereikt met behulp van datagestuurde optimalisatie. Uiteindelijk kan hiermee de verspilling als gevolg van zorgcoördinatie verminderd worden.

Momenteel is de regulering van data goed georganiseerd in wet- en regelgeving. Ondanks dat de eerder beschreven schaalbaarheid van innovaties eindelijk plaatsvindt, blijkt de uitwisseling van data en informatie moeilijk te zijn, aangezien er nog steeds problemen zijn met de compatibiliteit en het eigendom van data en informatie. *"Data is the new oil"* [76], en dit geldt ook voor zorgdata. Gezondheidsdata verhogen in de eerste plaats de kwaliteit van de zorg, aangezien de beschikbaarheid van informatie zal verbeteren. Voor operations research biedt dit ook kansen omdat meer data over de totale zorgpaden van patiënten tot op zekere hoogte beschikbaar komen via bijvoorbeeld wearables.

Door de digitalisering, technische geneeskunde en de introductie van analytics wordt de klinische praktijk de komende jaren technischer. Momenteel worden de eerste kleine stappen gezet. Zo ondersteunen voorspellende analytics al de klinische besluitvorming over de dosering van antibiotica bij sepsis voor patiënten op de intensive care [232]. Zulke toepassingen hebben ook gevolgen voor capaciteitsplanning en -management. Daarom moeten zowel analytische als geneeskundewetenschappers elkaars vakgebied verder omarmen. Samen kunnen ze het ziekenhuis van de toekomst integraal vormgeven.

About The Author

Thomas Schneider was born in Ouderkerk aan Amstel, Ouder Amstel, the Netherlands, on October 17, 1984. He currently lives there with his family. Thomas received his M.Sc. degree in Industrial Engineering and Management at the University of Twente in Enschede in 2011. During his masters program, he specialized in Healthcare Production and Logistics Management. He carried out his master's thesis in the Radiology Department at the Leiden University Medical Center (LUMC) and the Center for Healthcare Operations and Improvement Research (CHOIR) at the University of Twente. The topic of his thesis was waiting list management in radiology. After his thesis, he was appointed as business consultant at Division 2 of the LUMC where he focussed on capacity management. During his career, he has gained experience in a wide variety of capacity related topics and in analytics related to capacity management in hospitals. In 2015, he started part-time Ph.D. studies at the University of Twente on Integral Capacity Management and Planning, supervised by prof.dr.Richard Boucherie and prof.dr.ir. Erwin Hans from CHOIR, Universtity of Twente and prof.dr. Job Kievit from LUMC, and was funded by the LUMC. His research was inspired by, and to a large extent implemented in the LUMC. In 2019, Thomas jointly initiated the LUMC Capacity Center, where he was appointed as manager. Here, he completed the final part of his research. His Ph.D. research culminates with this dissertation.

Reflections

Before the start of this research, I had already been working for several years at the LUMC as internal consultant on patient logistics. This position gave me the opportunity to become familiar with the organization, decision-making processes and the dynamics of working with healthcare professionals. Performing multiple analyses throughout the hospitals also gave me an overview of all patient flows and typical related problems at the LUMC. Furthermore, working as a consultant offered me opportunities to learn skills and gather experience dealing with highly trained professionals to implement results from these analyses. These professionals are needed to carry out the improvements based on the results of operations research models (i.e. to adjust their processes) and thus they are crucial for successful implementation.

This research is performed as a part-time Ph.D. project. Having a track record on improvement projects gave me the opportunity to select practical problems that were interesting for research. Therefore, all research in this thesis is inspired by the practice of the LUMC and most of this research is implemented there. Moreover, this research

is strongly related to practice. Selecting interesting projects for both practice and research is difficult, as research requires generalization and therefore specific details of practice can be lost. Conducting research that is strongly embedded in practice, pushes research to incorporate crucial dynamics from practice and contributes to successful implementation. Furthermore, projects and required models that are interesting for the LUMC may already be available in the literature. Therefore, interests can be in conflict. To solve conflicting interests, requires continuous fine tuning of projects and related research questions. Overall, part-time PhD projects give the unique and challenging opportunity to connect research and practice and to further yourself as both a researcher and professional. I therefore strongly recommend part-time Ph.D. projects in applied sciences.

List Of Publications

A.J. Schneider and E.W.Hans. Integral Capacity Management in Hospitals. *Working paper.*
Basis for Chapter 2.

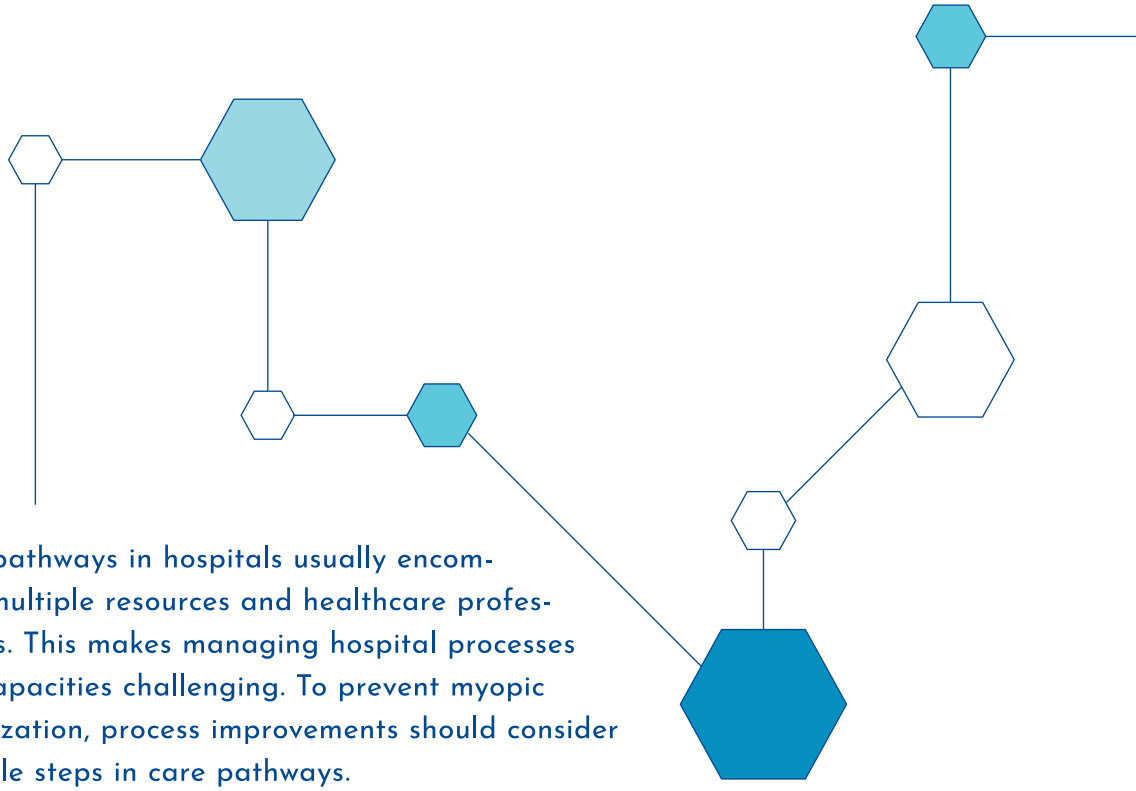
N.M. van de Vrugt, A.J. Schneider, M.E. Zonderland, D.A. Stanford and R.J. Boucherie. Operations Research for Occupancy Modeling at Hospital Wards and Its Integration into Practice In C. Kahraman, Y. Topcu, editors, *Operations Research Applications in Health Care Management*, chapter 5, Springer International Publishing, Cham, United States, 2018. 101-137
Basis for Chapter 3.

A.J. Schneider and N.M. van de Vrugt. Applications of Hospital Bed Optimization. *Submitted.*
Basis for Chapter 3.

A.J. Schneider, P.L. Besselink, M.E. Zonderland, R.J. Boucherie, W.B. van den Hout, J. Kievit, P. Bilars, A.J. Fogteloo and T.J Rabelink. Allocating Emergency Beds Improves the Emergency Admission Flow. *INFORMS Journal on Applied Analytics*, 48:4:384-394, 2018
Basis for Chapter 4.

A.J. Schneider, J.T. van Essen, M. Carlier and E.W. Hans. Scheduling surgery groups considering multiple downstream resources. *European Journal of Operational Research*, 280.2:741-752, 2020
Basis for Chapter 5.

A.J. Schneider, J.W.M. Otten, M.E. Zonderland, R.J. Boucherie and M.J. Schalij. The Hospital Online Multi-Appointment Scheduling Problem. *Working paper.*
Basis for Chapter 6.



Care pathways in hospitals usually encompass multiple resources and healthcare professionals. This makes managing hospital processes and capacities challenging. To prevent myopic optimization, process improvements should consider multiple steps in care pathways.

This dissertation aims to improve complex decision-making that integrally manages capacity for care pathways. We distinguish two approaches to improve integral capacity management and planning: (1) integrally organizing the timing and alignment of the capacity decisions and (2) analyzing (optimal) capacity decisions considering multiple capacities. For this we use both operations management and operations research techniques.

This research creates impact in practice through implementing research outcomes and analyzing the effects of interventions in the Leiden University Medical Center.

Leiden University Medical Center
UNIVERSITY OF TWENTE. | TECHMED CENTRE
Faculty of Electrical Engineering,Mathematics and Computer Science
CHOIR

ISBN: 978-90-365-5034-5