Doctoral theses at NTNU, 2023:401

Thomas Reiten Bovim

Tactical Resource Planning in Surgical Clinics

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Economics and Management
Dept. of Industrial Economics and Technology
Management

NTNU
Norwegian University of
Science and Technology

Thomas Reiten Bovim

Tactical Resource Planning in Surgical Clinics

Thesis for the Degree of Philosophiae Doctor

Trondheim, December 2023

Norwegian University of Science and Technology
Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Economics and Management
Dept. of Industrial Economics and Technology Management

ISBN 978-82-326-7500-5 (printed ver.)
ISBN 978-82-326-7499-2 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:401

Printed by NTNU Grafisk senter

Sammendrag

I mange vestlige land fører den demografiske utviklingen til at behovet for helsetjenester øker, mens andelen i arbeidsfør alder stagnerer eller synker. For å opprettholde dagens helsetilbud, må vi derfor utnytte ressursene bedre i fremtiden. Denne avhandlingen omhandler taktisk operasjonsplanlegging, som er et forskningsfelt innenfor Operasjonsanalyse. I tillegg inneholder avhandlingen et eksempel på bruk av Operasjonsanalyse i ressursplanlegging av akutt- og mottakstjenester i forbindelse med COVID-19.

Taktisk planlegging utgjør overgangen fra strategiske målsetninger og beslutninger, til den operasjonelle planleggingen av enkeltpasienter. Det mest studerte problemet innenfor taktisk operasjonsplanlegging er utviklingen av Kirurgiske Masterplaner (KM). Dette er sykliske blokkplaner hvor et sett av kirurgisek spesialiteter tilordnes til operasjonsrom, hver dag gjennom en planleggingssyklus, med den hensikt å oppnå finansielle målsetninger samt å unngå lange ventetider til utredning og behandling for pasienter.

I den første artikkelen utvikler vi en KM for å understøtte elektiv og akutt operasjonsvirksomhet. I denne situasjonen må vi veie hensynet til effektiv elektiv virksomhet, mot hensynet til å kunne tilby akuttoperasjoner på kort varsel. For å kunne gjøre denne avveiningen, utvikler vi en en tostegs stokastisk optimeringsmodell som lager en KM, og som tar hensyn til den usikre ankomsten av, og sengebehovet for akuttpasienter.

I artiklene to og tre ser vi på samplanlegging av poliklinikk og operasjon. Pasientene mottar behandling i begge enhetene, og kirurgene utfører både konsultasjoner og operasjoner. For å fasilitere en effektiv og koordinert ressursbruk, utvikler vi optimeringsmodeller som lager integrerte masterplaner for poliklinikk og operasjon, hvor både spesialteter og aktivitetstyper tilordnes til rom og dager. Artikkel tre bygger på artikkel to, og vi utvikler et planleggingsrammeverk hvor deler av masterplanen oppdateres jevnlig for å hensynta varierende vantelister.

Den fjerde artikkelen bygger på et prosjekt som ble utført ved St. Olavs hospital i mars 2020, da sykehuset forberedte seg på COVID-19 pandemien. Det ble bestemt at alle pasienter som ankom sykehuset med mistanke om COVID-19-smitte skulle undersøkes i akuttmottaket, og at smittemistenkte pasienter skulle transporteres med ambulanse både til og fra sykehuset. I denne artikkelen utvikler vi en diskret-hendelsessimuleringsmodell for å estimere ressursbehovet ved akuttmottaket og i ambulansetjenesten ved pandemiens topp.

Summary

Due to demographic changes, many western countries experience a growing demand for health care and a stagnation in the working-age population. To maintain the level of care in the future, the health care resources must be better utilized. This thesis considers tactical surgery planning, a subfield of the more general research areas of Resource Management and Operations Research. It also includes a case related to emergency care planning during the COVID-19 pandemic.

Tactical decisions facilitate the transfer of strategic objectives and decisions, to the operational planning of individual patients. The most frequently studied problem within tactical surgery planning, is the development of the Master Surgery Schedule (MSS). The MSS is a blueprint schedule where surgical specialties are assigned to operating room blocks thorough a planning cycle, aiming to achieves financial goals and serve patients in a timely manner.

In the first paper, we consider the problem of constructing an MSS for both planned and emergency surgeries. Here, we face the trade-off between an efficient handling of planned surgeries, while ensuring responsive services for the emergencies. To face this trade-off, we develop a two-stage stochastic optimization model that accounts for the stochastic arrivals and bed loading of emergency patients when constructing the MSS.

The second and third paper consider the integrated planning of the outpatient clinic and the operating theatre. Patients require services in both units, and the surgeons perform both consultations and surgeries. To facilitate efficient and coordinated use of resources, we develop optimization models that construct cross-unit integrated master schedules, where both specialties and different activity types are assigned to operating room blocks. This allows us to adjust the activity type assignments, based on the current waiting lists. The third paper extends on the second paper, and here we propose a planning framework where parts of the master schedule is periodically refined to account for stochastic waiting lists.

The fourth paper is based on a real-life project performed in March 2020, when St. Olav's Hospital was preparing for the COVID-19 pandemic. It was decided that all emergency patients that entered the hospital with a COVID-19 suspicion should be screened in the Emergency Department, and that they should be transported by ambulance both to and from the hospital. In this paper, we develop a discrete-event simulation model to estimate the impact on the Emergency Department and the ambulance services during the peak of the pandemic.

Acknowledgements

During the last four years, I have done research at the Department of Industrial Economics and Technology Management (IØT) at the Norwegian University of Science and Technology, which has resulted in this thesis. Besides doing research, I have been fortunate to have part-time positions, first in SINTEF, and then at St. Olav's Hospital. I am grateful for having had the opportunity to work with many talented people, which has been utmost inspiring to me.

One year prior to starting my PhD, I received my Master's degree from IØT. During the work with my Master's thesis, I was introduced to the field of operating room planning, and I was fortunate to receive brilliant supervision which inspired me to pursue the research. The excellent supervision continued during my entire PhD work, and for that I want to thank Associate Professor Anders Gullhav. First, I want to thank you for the fruitful discussions and guidance you have provided throughout these years, which has been essential to my work. I also want to thank you for the extensive effort you have put into establishing a solid collaboration with the hospital, and for building a strong foundation for high-quality research. Furthermore, I would like to thank my co-supervisor Professor Henrik Andersson for the excellent supervision and valuable discussions you have provided. I truly appreciate the way you have encouraged my work, also in periods when our research made less progress. I am thankful for having you both as my colleagues.

Colleagues at the Clinic of Orthopaedy, Rheumatology and Dermatology have been essential to the work presented in this thesis. During my work, first as a Master's student and eventually as a PhD candidate, I have had the pleasure of working with Dr. Vigleik Jessen, Trude Mittet, Liv Åse Sommervold and Mette Røsbjørgen, who have shared of their knowledge and provided insights in the complex planning problems that they face.

I would like to thank my parents and my two sisters for showing interest in my work, but mostly for drawing my attention away from research. I appreciate all the dinners, evenings, and relaxing weekends we have had together.

Finally, I want to thank my wife, Ingeborg, for the tireless support and encouragement during these four years. Thank you for all the good moments you bring to my life, every day.

Contents

Chapter 1

Introduction

Due to demographic changes, many western countries experience a growing demand for health care, while the number of people in the working-age stagnates. In 2021, there were slightly more than three Europeans of working-age for every European aged 65 and above. This is a 50% higher coverage compered to 2050, when there will be less than two working-age adults for each elderly person(European Commission, 2023). Norway is facing similar chellenges as the EU countries. From 2000 to 2020, the number of people exceeding 80 years increased from 190 000 to 230 000. However, from 2020 to 2040 this population will more than double, exceeding 250 000. During the same period, starting from the mid 2030s, the number of people in the working-age will decrease in absolute numbers (Norges Offentlige Utredninger, 2023).

Currently, disruptions caused by the recent COVID-19 pandemic applies pressure on health care systems globally. From 2019 to 2020, the number of elective surgeries performed across the EU countries fell by 16.5%, generating backlogs of patients on waiting lists (OECD Publishing, 2022). During the same period, the mean waiting times for somatic patients within the specialist health services increased from 61 to 65 days in Norway (Helsedirektoratet, 2021). Many EU countries have taken actions to address the backlogs by providing additional funding to increase supply of surgery. However, the main constraint in increasing the volume of activities has been the health care workforce (OECD Publishing, 2022).

Current and future challenges impose strain on the health care sector. If we want to maintain, or possibly increase, the level of care in the future, we must utilize the resources more efficiently.

The work presented in this thesis is inspired by resource planning problems faced by departments in our collaboratory hospital, St. Olav's Hospital. The first three papers concentrate on surgery planning, encountering problems faced by the Orthopaedic Clinic. All these papers extend on the so called Master Surgery Scheduling Problem (MSSP), where surgical specialties are assigned to operating rooms (OR) on each day of a planning cycle, with the aim to facilitate efficient use of the ORs, surgeons and hospital beds. The resulting blueprint schedule is referred to as the Master Surgery Schedule (MSS). The final paper demonstrates a real-life application of Operations Research applied to the Department of Emergency Medicine and Prehospital Services when preparing for the increased number of COVID-19 related admissions in March 2020.

In Paper I, we face a planning context where the operating theatre (OT) capacities are shared between planned elective surgeries and unplanned emergency surgeries, and we aim to construct an MSS that reserves capacity for both surgery types. To account for the stochastic arrivals and bed loading of emergency patients, we propose a simulation-optimization approach consisting of a two-stage stochastic optimization model and a discrete-event simulation model. The simulation model is used to generate realistic scenarios for the optimization model, while the optimization model generates an MSS that balances efficiency and responsiveness when serving planned elective surgeries and high-priority emergencies.

In Paper II, we consider the integrated planning of the OC and the OT. The patients require services in both units, and the surgeons perform both OC consultations and surgeries. The main objective in the problem is to maximize the number of initial consultations that can be performed, while ensuring that downstream capacities can cope with the derived demand. We develop an optimization model for generating an integrated master schedule, and use discrete-event simulation to evaluate the schedules across different scheduling policies.

Paper III extends on Paper II, and we develop a more sophisticated optimization model that considers the waiting lists to all OC consultation types and surgery types. The optimization model constructs an integrated master schedule, and we aim to provide short waiting times for all activity types. The master schedule can be separated in one cyclic high-level schedule, and one non-cyclic low-level schedule, and we propose a two-level planning framework that utilizes the structure of the master schedule. First, we construct the entire master schedule for the upcoming planning horizon. Then, to account for stochastic waiting lists, we frequently refine the low-level schedule. Finally, we evaluate the planning framework under different planning strategies.

Paper IV demonstrates a real-life application of Operations Research applied to the Department of Emergency Medicine and Prehospital Services when preparing for the increased number of COVID-19 related admissions in March 2020. At this point in time, it was decided that all patients that enter the hospital with a COVID-19 suspicion should be screened in the Emergency Department, and that they should be transported by ambulance both to and from the hospital. In addition, it was proposed that all patients who were not screened upon arrival would require an ambulance when leaving the hospital. Based on these policies, we develop three simulation models to estimate the additional Emergency Department beds required to house patients that are screened, and the number of additional ambulances required to obtain prepandemic response times for the most urgent patients.

1.1 Background

In this section, we present background on topics that are relevant to the thesis. First, a brief introduction of St. Olav's Hospital, and the Orthopaedic Clinic is provided. Then, we present a classification of planning decisions in health care, allowing us to position our work in the larger picture. Following this, we review relevant literature on tactical surgery scheduling, before presenting contributions made by the Operations Research society during the COVID-19 pandemic.

1.1.1 St. Olav's Hospital and the Orthopaedic Clinic

St. Olav's Hospital is a university hospital, located in Trondheim. It is the largest hospital in the Regional Health Authority of central Norway, and it accommodates approximately 1000 beds. In 2022, the hospital's 11 000 employees served more than 60 000 inpatients and performed 660 000 outpatient clinic (OC) consultations (stolav.no, 2023). The hospital has a decentralized organization with 20 clinics, where each clinic covers a specific medical field. In line with the decentralized organization, all clinics that perform surgery manage their own OT, and there is no central OT. The two main surgical clinics at St. Olav's Hospital are the Clinic of Surgery and the Clinic of Orthopaedy, Rheumatology and Dermatology (referred to as the Orthopaedic Clinic). The Orthopaedic Clinic performs activities in different centers and locations in the hospital, but mainly in the Center of Movement (Bevegelsessenteret).

The Department of Orthopaedic Surgery (referred to as the Orthopaedic Department) is organized under the Orthopaedic Clinic, and the orthopaedic OC and the orthopaedic OT are subordinate units of the department. In the Center of Movement, the department manages eight ORs and eight OC rooms. In addition, the department serves two emergency ORs located in the Emergency Department. During the weekdays, the departments operates 59 beds, and 43 beds are available during the weekends. The department performs about 26 000 OC consultations and 6000 surgeries every year. Almost 50% of the surgeries are emergencies.

1.1.2 Classification of planning decisions in health care

Hulshof et al. (2012) develop a taxonomic classification of planning decisions in health care, extending on the framework by Hans et al. (2012). The taxonomy composes of two axes, and can be seen in Figure 1.1. The vertical axis reflects the hierarchical nature of decision making, including strategic, tactical and off- and online operational decisions. On the horizontal axis the authors position major health care services, including ambulatory, emergency, surgical, inpatient, home care and residential care services. The two latter services are not relevant to this thesis and will not be further introduced.

Strategic planning involves structural, long-term decision making, such as dimensioning, and facility layout and location. Tactical planning translates strategic decisions to guidelines that facilitate operational planning decisions. At this level,

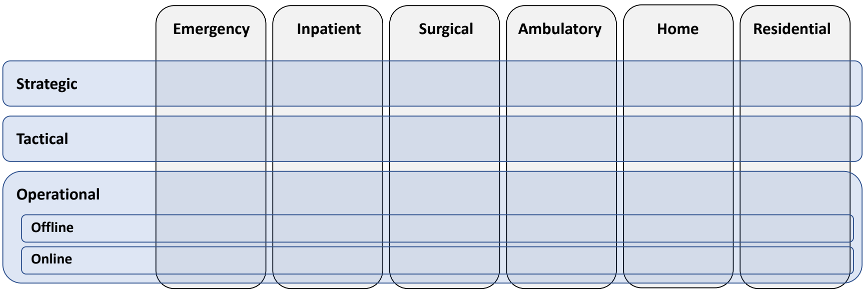


Figure 1.1: The taxonomy for resource capacity planning in health care, proposed by Hulshof et al. (2012)

patient groups are characterized based on disease type, urgency and resource requirements, and the resource capacities settled at the strategic level are divided between the patient groups. Blueprint schedules that assign patient groups to a set of resources are typically developed at the tactical level. Operational planning involves the short-term decision making related to the execution of the services. Following the tactical blueprints, execution plans are designed at the individual patient and resource level. Offline operational planning concerns the advanced planning of operations, while online operational planning involves reacting to unplanned events (Hulshof et al., 2012).

In ambulatory care services, patients are served without requiring a room or a bed. In hospitals, the OCs are the major providers of these services. Emergency care services involve examination and initial treatment of urgent medical problems caused by accidents, traumas or sudden illness. The Emergency Department is the main facility to provide emergency care services in hospitals. Surgical care services, mainly executed in the OTs, provide surgical procedures to patients, while inpatient care services treat hospitalized patients in the intensive care units and general wards (Hulshof et al., 2012).

Considering the taxonomy by Hulshof et al. (2012), tactical planning within surgical services is most relevant to this thesis. In the following, we provide a short introduction to the most relevant planning problems faced at each of the hierarchical decision levels within surgical services.

At the strategic level, the case mix is settled and capacity dimensioning is performed to match the case mix. The case mix involves the number and types of surgical cases that are performed at a facility, and a target case mix is chosen with the objective to optimize net contribution while considering several internal and external factors. Internal factors include the limited resource capacities, while external factors include the expected demand for services in the facility's catchment area and the restricted budgets and service agreements in government-funded systems (Hulshof et al., 2012). In Norway, the financing of public hospitals is split in two: roughly 60% is a basic grant, and the rest (the so called "Innsatsstyrt finansiering" (ISF)) is based on the activities performed by the hospitals. The ISF is founded on a classification system of diagnosis and procedures, the Diagnosis Related Groups (DRG). All treatments performed at a hospital is assigned a DRG depending on the diagnosis of the patient and the performed procedures, and the ISF covers roughly 40% of the average costs related to the DRG (Helsedirektoratet, 2023).

At the tactical level, plans for capacity allocation and admission control are made. The most common way to perform capacity allocation in surgical care services is through block scheduling, and the construction of the MSS. The MSS is a cyclic blueprint schedule where a number of surgical specialties are assigned to operating room blocks on each day of the planning cycle. The schedule is periodically repeated to cover a planning horizon, and it is constructed to facilitate the operational scheduling of patients such that the case mix targets can be achieved. It is common to consider both the surgeon- and bed capacities when constructing the MSS. Admission control is the process of deciding the number of patients from different patient groups to admit for treatment over a period of time. The specialties considered for the MSS are disaggregated into finer groups, and the admissions of different groups are limited by the assignments of specialties in the MSS. Objectives relevant to admission control are related to patient waiting times.

At the offline operational level, surgery case scheduling is performed. This is the process of assigning individual patients to individual resources on specific days. This also involves sequencing the patients during a day, and assigning starting times to the surgical cases. Online operational scheduling involves the scheduling of emergency patients that enter on the day of execution, and the rescheduling of surgeries.

A patient's pathway typically includes several care stages performed by various health care services. Therefore, in the perspective of the presented taxonomy, a strong horizontal interaction can be recognized, in addition to the vertical interaction. As a consequence, planning that concentrates on isolated services tends to be suboptimal, while planning that facilitates flexibility, and integrated decision making for multiple care services show great potential (Hulshof et al., 2012).

1.1.3 Tactical surgery scheduling

The main topic of this thesis is tactical surgery scheduling, and in particular we extend on the Master Surgery Scheduling Problem (MSSP). In this section, we first give a brief overview of the history of Operations Research applied to surgery scheduling, before presenting the state-of-the-art on the MSSP, concentrating on contributions from the past ten years. While the main focus is the MSSP, we also highlight a few relevant contributions on related problems.

A brief overview of the history on surgery scheduling

Surgery scheduling has been a topic of study since the 1950s (Cardoen et al., 2010). While early research was mainly concerned with aspects of operational scheduling, Blake and Carter (1997) suggested that further research should be directed towards solving scheduling issues at strategic and tactical (”administrative”) levels. They also suggested to develop models for integrating operating room scheduling with other hospital operations. In the early 2000s, the body of research on surgery scheduling expanded, and the number of scientific contributions almost doubled in ten years (Cardoen et al., 2010). During the same period, many contributions were made on surgery scheduling at the strategic and tactical levels. Rahimi and Gandomi (2021) reviewed the literature on operating room and surgery scheduling, indicating that the number of publications within the field has been steadily growing for the past ten years. Research in the past decade has mainly focused on elective patients, and most authors consider aspects of uncertainty, where the uncertain surgery duration is most frequently studied. While the research on surgery scheduling has largely expanded over the past 20 years, there are very few contributions that demonstrate real-life implementations of the models. The absence of real-life implementation was addressed in Magerlein and Martin (1978), and the concern has been repeated by others (Samudra et al., 2016; Razali et al., 2022).

Defining the MSS

To the best of our knowledge, the term MSSP was first introduced by Testi and Tànfani (2009). However, the term is not consistently adopted in the literature, and the problem is not properly defined. In this thesis, we use the term MSSP for all problems where the MSS is an outcome of solving the problem. When taking this broad approach, Blake and Donald (2002) and Blake et al. (2002) were the first ones to study the MSSP.

While there is no clear definition of the MSS (Cardoen et al., 2010), it is often referred to as a cyclic schedule where a set of surgical subgroups are assigned to operating room blocks throughout the planning cycle. Razali et al. (2022) characterize the MSS based on three properties: the definition of surgery groups, the length of the planning horizon and the schedule cyclicity. While most authors define the surgery groups based on surgical specialties, it is not uncommon to further divide the groups based on resource consumption related to the expected length of stay and surgery duration. The most commonly considered planning horizons span from seven to 27 days, however there are examples of schedules that cover periods up to a year, and less than a week. The vast majority of contributions construct a cyclic MSS, but a few authors (Agnetis et al., 2012; Santos and Marques, 2022) experiment with constructing a non-cyclic schedule.

There are strong dependencies between the decision levels presented by Hulshof et al. (2012). The tactical level spans wide, and the transitions to the strategic and operational levels are rather vague. While most authors define the MSS as a tactical schedule, others choose different perspectives and decision levels when constructing the schedule (Cardoen et al., 2010). Traditionally, the MSS is strongly linked to the strategic case mix decisions, however it is popular to construct schedules that consider the current waiting lists, or even a set of individual patients to be scheduled. As a consequence, some schedules are primarily governed by budgets, while others are closely linked to the current demand for surgeries.

Both Fügener et al. (2014) and Santos and Marques (2022) assign surgical specialties to each operating room block in the MSS, and they link the demand for each specialty to the case mix settled at the strategic level. Considering the low level of granularity applied by the authors, both offer a relatively high-level decision support. Also Rachuba et al. (2022) ensure that the assignments made in the MSS align with the case mix decisions. However, accounting for a fluctuating demand, the authors assign a fractional number of the different surgical groups each week. At the operational stage, the scheduler can then vary the assignments between weeks, while making sure that it adds up over time. To evaluate the framework, the authors simulate a planning horizon of one year. Schneider et al. (2020) group surgeries that share the same surgical specialty, and similar surgery duration and length of stay. To construct the MSS, the authors assign a number of surgeries from each group to each operating room block. They assume that all waiting lists are inexhaustible and introduce constraints to make sure that a minimum number of surgeries from each group is assigned. By scheduling patients in accordance with the groups in the MSS, the scheduler can ensure efficient use of the resources. In contrast to Rachuba et al. (2022), the authors assign an integer number of surgeries from each group, leaving less flexibility at the operational stage. The lack of flexibility may require a more frequently updated MSS. When testing the model, the authors consider a planning horizon of 14 days.

Several authors construct the MSS based on current waiting lists and thus strengthen the link to the operational scheduling of patients. Both Banditori et al. (2013) and Cappanera et al. (2014) consider the waiting list of patients when constructing the MSS, but they do not consider the individual patients. In line with Schneider et al. (2020) they divide the surgeries into categories based on resource consumption, and assign a number of surgeries from each category to each operating room block. While Banditori et al. (2013) categorize surgeries based only on the expected surgery duration, Cappanera et al. (2014) also consider the patient's expected length of stay when performing the categorization. Banditori et al. (2013) mainly apply a planning horizon of 28 days when testing the model, while Cappanera et al. (2014) consider a planning horizon of 14 days. Some authors (Agnetis et al., 2014; Spratt and Kozan, 2016; Moosavi and Ebrahimnejad, 2020; Mazloumian et al., 2022; Makboul et al., 2022) simultaneously address the MSSP and the Surgical Case Assignment Problem (SCAP), by scheduling individual patients from the waiting lists when constructing the MSS. All authors that integrate the MSSP and SCAP consider a one-week planning horizon.

To summarize the discussion above, we propose Figure 1.2 to categorize the different settings in which the MSS can be constructed. To the left, we illustrate that the tactical level spans from the strategic to the operational decision level. In the middle, we present how the demand is typically defined when we construct the MSS at different decision levels. To the right, we indicate what types of problems that are defined and solved to construct the MSS at the different levels. Here, AC refers to Admission Control.

In Papers I to III, we divide the surgical specialties into finer groups, based

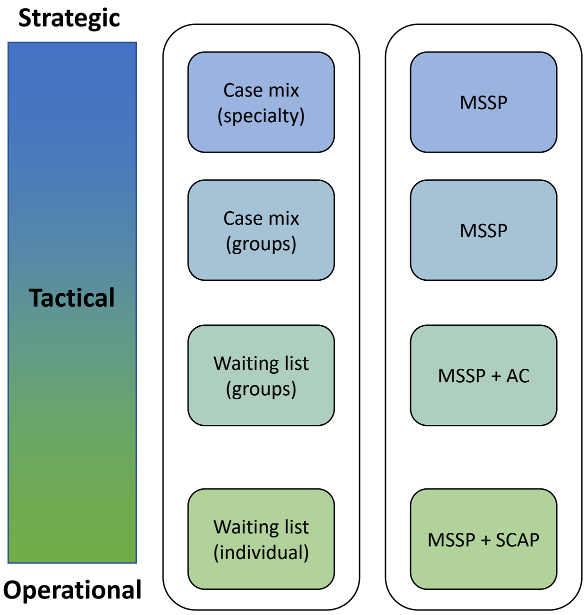


Figure 1.2: The MSS can be constructed at different decision levels. Based on the decision level at which we construct the MSS, we face different definitions of demand and problems to solve.

on resource consumption. In Paper I, we define, for each group, a target number of surgeries to be performed each week. These targets are related to the case mix targets settled on the strategic level. In Papers II and III, demand is directly linked to the waiting lists. The planning horizons considered in Papers I and II are half a year, and we consider static master schedules. In Paper III, we consider a planning horizon of 12 weeks, and we allow to make adjustments in the master schedule as time passes. The following sequence positions Papers I to III from a strategic to an operational connection: Paper I, Paper II, Paper III.

Objectives

There are multiple objectives that have been proposed when constructing the MSS. Many authors apply objective functions related to the downstream ward activities. Schneider et al. (2020), Mazloumian et al. (2022) and Cappanera et al. (2014) minimize the variation in bed occupancy, while Fügener et al. (2014) minimize the costs of operating the downstream wards. Another example is Banditori et al. (2013) who minimize the occurrences of patients resting in beds that are not meant for them. Objectives related to the waiting list management is also frequently considered in the literature (Oliveira et al., 2021; Moosavi and Ebrahimnejad, 2020;

Mazloumian et al., 2022; Banditori et al., 2013). A third recurring objective is related to throughput. Kumar et al. (2018) aim for a high production by maximizing the number of patients scheduled, while minimizing the the expected number of patients that have to be cancelled due to bed shortages. Other authors, like Spratt and Kozan (2016) and Makboul et al. (2022) maximize the gains from assigning surgeries in the schedule. Finally, many authors consider a variety of objectives, either in isolation or combined, when constructing the MSS (Penn et al., 2017; Banditori et al., 2013; Schneider et al., 2020; Mazloumian et al., 2022; Moosavi and Ebrahimnejad, 2020; Britt et al., 2021).

Uncertainty

There are several sources of uncertainty that can be considered when creating the MSS. The three most frequently studied factors of uncertainty are surgery duration, length of stay or bed availability, and emergency arrivals (Razali et al., 2022).

Several authors apply robust optimization to account for uncertain surgery duration (Moosavi and Ebrahimnejad, 2020; Mazloumian et al., 2022; Makboul et al., 2022), while others apply chance constraints to avoid operating room overtime (Rachuba et al., 2022; Schneider et al., 2020). Rachuba et al. (2022) calculate the probability that a surgery of a patient group lasts at least a number of time slots, and use a chance constraint formulation to adjust the proximity of consecutive surgeries. Schneider et al. (2020) approximate the sum of the surgery durations within an operating room with a normal distribution, and use this in combination with a chance constraint logic to adjust the risk of running into overtime. Banditori et al. (2013) investigate the trade-off between efficiency and robustness when creating the MSS. Given an availability of ORs and beds, they aim to obtain a robust schedule by running the optimization model on a set of instances with less resources than what is actually available. The schedules are then implemented in a discrete-event simulation model to guide the scheduling of patients in a system with the original OR and bed capacities. Based on the simulations, the authors evaluate the operational consequences of each MSS related to resource utilization, and the number of executed and cancelled surgeries.

Santos and Marques (2022) state that a cyclic MSS does not imply cyclic bed requirements, and they apply stochastic programming to account for fluctuating bed demands. They formulate a two-stage stochastic model, where the bed requirements from assigning a specialty to a block and a day are represented by the stochastic parameters. In the first stage, the assignments of specialties to blocks are made, and in the second stage the over-consumption of beds is penalized. Kumar et al. (2018) are inspired by stochastic programming when accounting for uncertain bed requirements. Instead of introducing parallel scenarios for the length of stay of patients, they apply a consecutive number of planning cycles where all the patients in each cycle have a randomly sampled length of stay. The authors explicitly impose non-anticipativity constraints by demanding a consistent order of patients to be scheduled and cancelled in each planning cycle. Moosavi and

Ebrahimnejad (2020) present a scenario-based robust formulation to account for the uncertain length of stay of patients, while Makboul et al. (2022) apply a standard robust formulation to model the number of beds available in the ICU each day. Vanberkel et al. (2011) develop a method for calculating the steady-state distribution of the number of patients resting in a ward resulting from an MSS. On a given day following surgery, a patient is present in the ward with a given probability. This behaviour can be described as a Bernoulli trial, and the distribution of the number of patients still present in the ward resulting from assigning a given specialty to a given day is binomially distributed. To calculate the total distribution of the number of patients still present in the ward, the authors apply discrete convolution to add the distributions from all specialties that were assigned on the given day. Finally, to account for patients that have a length of stay that exceeds one planning cycle, the authors again apply discrete convolution to add the days exceeding one cycle. A number of authors have applied, or extended on this framework (Fügener et al., 2014; Fügener, 2015; Schneider et al., 2020; Rachuba et al., 2022).

Both Moosavi and Ebrahimnejad (2020) and Mazloumian et al. (2022) apply a scenario-based robust formulation to account for the emergency demand. Each scenario states the time required to serve emergency patients on each day of the planning horizon. Rachuba et al. (2022) consider the expected number of emergency arrivals on each day of the planning cycle, and assign operating room and bed capacities accordingly.

In Paper I, we consider the uncertain arrivals and bed loading of emergency patients, and we develop a two-stage stochastic model to account for the uncertainties. In Papers II and III, we do not explicitly model uncertainty, but we introduce different planning policies to account for stochastic waiting lists.

Up- and downstream processes

Several authors state that operating room scheduling should not be made in isolation, and call for the inclusion of up- and downstream processes (Blake and Carter, 1997; Cardoen et al., 2010; Hulshof et al., 2012). While there exist contributions on the MSSP that consider the operating rooms in isolation (Agnetis et al., 2014; Spratt and Kozan, 2016), the vast majority of authors consider adjacent processes when constructing the MSS. The most common processes to include are downstream wards, either the Intensive Care Unit (ICU) or the medical wards. Fügener et al. (2014) build on the framework developed by Vanberkel et al. (2011) for calculating the distribution of patients resting in the downstream ward on each day of the planning cycle. However, they extend the model formulation to also include an intermediate stay in an ICU before being transferred to the ward. This framework is adopted by other researchers (Fügener, 2015; Schneider et al., 2020). While the inclusion of downstream wards is most common, some contributions like Moosavi and Ebrahimnejad (2020); Oliveira et al. (2021) also consider that some patients require a stay in a ward before surgery.

Schneider et al. (2020) propose the inclusion of the OC as a direction for future research on the MSSP. The integration of the OC and the OT can be regarded as a multi-appointment system, and be linked to the planning of care processes. The term care process is used for a set of consecutive care stages followed by patients through a hospital. It is the complete path of a patient through the hospital, such as a visit to the OC, a visit to an X-ray, and a revisit to the OC (Hulshof et al., 2016). To the best of our knowledge, the literature on multi-appointment scheduling is developed around outpatient systems, such as the OC or day surgery, and do not consider systems where the patients require a bed following an intervention. Furthermore, the majority of research on multi-appointment scheduling is related to operational scheduling (Marynissen and Demeulemeester, 2019), however a few contributions exist on the tactical level (Hulshof et al., 2013; Bikker et al., 2015; Hulshof et al., 2016).

Hulshof et al. (2013) develop a model for tactical resource allocation and elective patient admission planning in care processes. In the model, the care processes are modelled as a set of consecutive queues, and patients move from one queue to the next with a given probability. To serve a patient in a given queue, a set of resource capacities are required, and since the resource capacities are limited, so is the flow of patients. The main decisions to make is the number of patients to admit from each queue in each time period, and the objective is to minimize the weighted number of patients in each queue, giving higher weights to long waiting times. Hulshof et al. (2016) extend on Hulshof et al. (2013), and introduce stochastic patient arrivals and transmissions between queues. To solve the problem, the authors propose an approximate dynamic programming model. Bikker et al. (2015) study the care process of radiation treatment. All patients that are referred for treatment require a given sequence of examination consultations before the radiation treatment can begin. The aim of the study is to develop a cyclic schedule for the doctors who examine the patients, as a mean to minimize the access time between referral and treatment for all patients. In the schedule, each doctor is assigned to a task and a location on each day he or she is available to work. Instead of modelling queues, the authors calculate the minimum access time of one arriving patient in a system without already scheduled appointments, representing the shortest access time possible for the patient. This calculation is made for each patient type, arriving on each referral day in each location, and the objective is to determine the doctor schedule that minimizes the total access time of all patient types.

In both Papers II and III, we extend on the current literature, constructing integrated master schedules that cover both the OC and the OT. This allows us to level the activities across the two units, and promote a coordinated use of resources.

Flexibility and dynamic planning

Creating more planning flexibility in decision making demonstrates great potential (Hulshof et al., 2012). The topic of planning stability and flexibility is highly relevant in the context of designing an MSS. Here, stability refers to an MSS where all assignments are identical in each planning cycle, and thus offers predictability for the staff. Furthermore, a stable schedule allows for a more predictable pattern in terms of resource consumption. Flexibility concerns the ability to dynamically adapt the plan to the evolution of the waiting lists, allowing for shorter patient waiting times. Stability and flexibility are conflicting, since the former pushes towards having a constant MSS, while the latter seeks variation if necessary. Clinics should aim to find the right trade-off between stability and flexibility (Agnetis et al., 2012). Several authors investigate the value of introducing flexibility in tactical scheduling both related to surgery scheduling (Agnetis et al., 2012; Oliveira et al., 2021) and OC scheduling (Laan et al., 2018). They all find that introducing a very limited degree of flexibility in the master schedule improves resource efficiency and patient waiting times. Furthermore, Agnetis et al. (2012) conclude that small but frequent changes perform better than large but infrequent changes, and Oliveira et al. (2021) find that a static, non-cyclic MSS outperforms its cyclic counterpart.

In Paper II, we consider a static master schedule, but we investigate the value of imposing flexibility in the operational scheduling of patients. In Paper III, we introduce both flexibility and dynamic planning as part of the proposed planning framework. We use the term *dynamic* to describe the frequency by which the schedule can be adjusted, while *flexibility* refers to the amount of change that can be made in each adjustment. Furthermore, we extend the current literature by introducing the term *Agile* planning. By agile, we refer to the delay in time between the time of planning and the time of executing a schedule.

Real-life implementation

According to Samudra et al. (2016), less than 7% of the methods developed for scheduling operating rooms are applied in practice. The authors point to a few examples of models that have been applied in real-life, but these are all tools for operational planning. Regarding the MSSP, no studies exist that demonstrate a real-life use of the developed model (Razali et al., 2022).

1.1.4 Contributions from the OR society during the COVID-19 pandemic

The first outbreak of COVID-19 was observed as a cluster of pneumonia cases in the city of Wuhan, China late in December 2019. On the 20th of January 2020, the first cases of COVID-19 outside China were announced, and in the following weeks the disease spread to many countries. Then, on the 26th of February, the first case of disease was reported in Norway. On the 12th of March, societal restrictions such as closing down public institutions and instructing social distancing, were imposed by the Norwegian government (Tjernshaugen et al., 2023).

In mid March 2020, the management at St. Olav's Hospital decided that all patients that entered the hospital with a COVID-19 suspicion should be screened in the Emergency Department, and that they should be transported by ambulance both to and from the hospital. In addition, it was proposed that all patients who were not screened upon arrival would require an ambulance when leaving the hospital. On the 17th of March, we became part of a team that was established to predict the resource requirements related to the new guidelines for screening and transportation of patients. On the 12th of March 2020, the Norwegian Institute of Public Health (NIPH) had released a *recommended planning scenario* for the evolvement of the COVID-19 pandemic in Norway. The planning scenario was used as a starting point of our analyses, and when an updated scenario was provided on the 24th of March, this was used to update our predictions.

Similar groups were established world-wide to provide decision support for the health services during the pandemic. People from the Operations Research society contributed in many initiatives, and a large body of literature on pandemic-related planning and scheduling emerged. While reviewing the literature on this topic is far beyond the scope of this introduction, we present relevant contributions to serve as a background for positioning our own work.

Currie et al. (2020) discuss how simulation can be used to support decision makers in making informed decisions during the COVID-19 pandemic. The authors identify three main categories of decisions where simulation can provide decision support: decisions affecting disease transmission and interventions, decisions regarding resource management, and decisions about care. Furthermore, they propose a framework to indicate the geographic and time scale over which the decisions are made. The geographic scale includes the global, national, organisational and individual levels. For the time scale, the authors use the disaster operations management framework by Altay and Green (2006) that splits decisions into four phases: mitigation - activities to prevent the onset of disaster or reduce its impact; preparedness - plans to handle an emergency; response - implementation of plans, policies and strategies from the preparedness phase; recovery - long-term planning actions to bring the community back to normality. Decisions regarding resource management, categorized as organisational on the geographical axis, and within preparedness or response on the time scale are most relevant to this thesis. Currie et al. (2020) identify two types of decisions that fall within this category: capacity of inpatient hospital beds and critical care, and staffing.

Several authors apply discrete-event simulation to identify the impact of COVID-19 patients on the ward capacities (Garcia-Vicuña et al., 2020; Le Lay et al., 2020; Wood et al., 2020). Both Garcia-Vicuña et al. (2020) and Wood et al. (2020) concentrate on the COVID-19 patients, disregarding the other patients. Garcia-Vicuña et al. (2020) consider the case where patients are transferred between levels of care depending on the severity of their condition. Simulation is used to model the flow of patients through the system, from hospitalization to discharge, and the results are used to predict the requirements of beds in the upcoming days and weeks. Wood et al. (2020) only consider the patients that require intensive care, and estimate the number of capacity-dependent deaths at a hospital in England under different governmental isolation policies and for different ward capacities. The authors model the intensive care ward as a loss system, implying that patients arriving when no beds are available will be rejected, increasing the probability of death. Le Lay et al. (2020) consider the hospital-wide flow of patients when predicting the bed requirements in the multi-purpose recovery ward for the case hospital. Based on a set of simulation experiments, the authors conclude that a daily arrival rate of five COVID-19 patients will overcrowd the recovery ward, and propose a temporary increase of the ward capacity.

Dai et al. (2022) and Melman et al. (2021) propose new strategies for managing the wards to avoid the interruption of high-priority elective surgeries during the pandemic. Dai et al. (2022) propose a so called *buffered clustered configuration* of beds. Clustered refers to the establishment of a clustered ward where a number of beds from each specialized ward is clustered together to serve low-priority patients, leaving the remaining beds for high-priority patients. The buffer wards are additional beds established to accommodate non-elective patients that must be isolated and tested for COVID-19. The authors develop a mathematical model for assigning elective patients from the waiting list to an operating room day and a ward. Melman et al. (2021) apply discrete-event simulation to evaluate three different resource allocation strategies imposed on the critical care wards of the case hospital: proactive cancellation of elective surgery, reactive cancellation of elective surgery, and ring-fencing OT capacity. In the proactive strategy, the elective OT is closed throughout the pandemic wave, dedicating the critical care capacity to the COVID-19 patients. In the reactive and ring-fencing strategies, the surgery capacities are adjusted dynamically based on the current demand for critical care. According to the ring-fencing strategy, a limited number of elective operating rooms are kept open throughout the pandemic, independently of the critical care demand. This is not the case in the reactive strategy. The authors find that the ring-fencing strategy outperforms the others when balancing the conflicting goals of maximizing the number of elective surgeries performed and minimizing the number of non-elective rejections.

The increased hospital workload during the pandemic causes physical and psychological strain on the health care professionals (Güler and Geçici, 2020). At the case hospital studied by Güler and Geçici (2020), three bed departments are established to handle COVID-19 positive patients, and these require the presence of physicians around-the-clock. Most of the physicians must cover shifts in the new departments in addition to the shifts covered in their regular departments, and the authors develop a mathematical model to solve the physician scheduling problem. To decrease the spread of virus among health care professionals, Kluger et al. (2020) propose to optimize staff scheduling to minimize interactions between the workers and limit the patient pool to which each employee is exposed. The authors find that longer nursing shifts and scheduling designs in which teams of nurses and doctors co-rotate no more frequently than every three days can lead to fewer infections.

The pandemic also affects the outpatient services. To prevent the spread of disease, Otten et al. (2023) aim to reduce the number of patients present simultaneously in the waiting area. The authors consider patient types that require multiple appointments on the same day, and that return to the waiting area between appointments. To avoid crowding in the waiting area, the authors develop an integer linear programming model to construct a blueprint schedule where the different appointment types are assigned to resource blocks through a day. A Monte Carlo simulation model that considers uncertain patient arrival times and appointment durations is developed to evaluate the performance of the blueprint schedule. If the schedule causes excessive crowding, the capacity parameters are decreased before running the optimization model over again. These iterations are repeated until the chances of crowding is below a given threshold.

The Operations Research society has contributed to solve a large variety of problems related to resource planning in hospitals during the pandemic. While it is possible to find similarities between some of the cases presented in the literature, all seem to have distinct characteristics that make them different from the other cases. This is probably due to the urgency under which many projects were established, and the current policies adopted by the different hospitals. The case presented in Paper IV is yet another distinct case that adds to the large body of contributions made by our society.

1.2 Purpose and outline

In this section, we first describe the main purpose of the thesis, and provide the background and motivation behind solving the problems faced in each paper. Then, we outline each of the papers, and present each paper's main contribution.

1.2.1 Purpose of the thesis

The main purpose of the thesis is to make advances in the field of Operations Research on tactical surgery scheduling, while making sure to solve problems that are of relevance to practitioners. Papers I to III contribute to the main purpose of the thesis, and while describing generic problems, they all originate from challenges faced by the Orthopaedic Department at St. Olav's Hospital. Paper IV does not contribute to the main purpose of the thesis, but it demonstrates the application of Operations Research methodology to solve real-life problems at a state of uncertainty.

At St. Olav's Hospital, there are dedicated operating room facilities for emergency surgeries, serving both orthopaedic and surgical emergencies. Surgical emergencies tend to be more urgent than the orthopaedic cases, and in times of high demand, the orthopaedic emergencies experience long waiting times. A main concern faced by the management at the Orthopaedic Department at the time we established contact, was the excessive waiting times of orthopaedic emergencies, and especially the suburgent cases. To reduce the waiting times, the management considered to reserve some capacity in the elective operating rooms to serve the suburgent surgeries. The question was how much capacity to reserve, and this was the starting point and motivation behind Paper I.

Papers II and III originate from another challenge faced by the Orthopaedic Department: to level the activities performed in the OC and the OT, to achieve

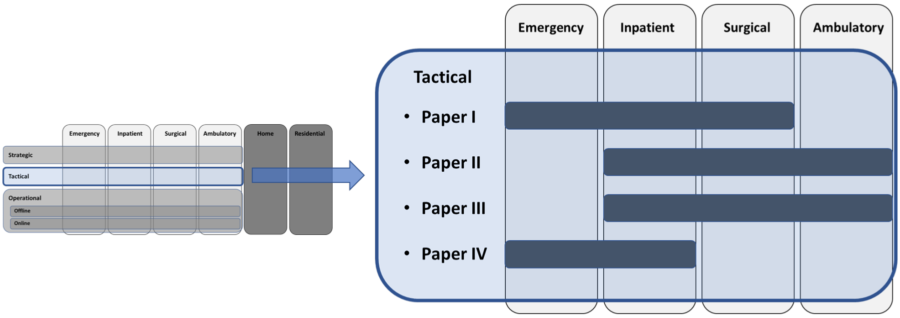


Figure 1.3: Positioning the four papers according to the taxonomy by Hulshof et al. (2012).

stable workloads and short waiting times for patients. The patients require services in both units, and the surgeons perform both consultations and surgeries. To provide specialized services, the surgeons are organized according to orthopaedic specialties, and each specialty accommodates a set of different activity types and corresponding waiting lists. It is challenging to coordinate the use of resources across the two units, and the management state that they experience significant variations in surgery workload. In these papers, we construct master schedules that integrates the planning of the two units.

The problem faced in Paper IV was motivated by the management's desire to estimate the impact of COVID-19 admissions on the Emergency Department, and the ambulance services during early stages of the pandemic.

In Figure 1.3, we position the four papers according to the taxonomy provided by Hulshof et al. (2012). All the papers consider tactical resource planning, covering at least two care services.

1.2.2 Paper I: Stochastic Master Surgery Scheduling

In this paper, we consider the problem of constructing an MSS that reserves capacity for both elective and emergency surgeries. In addition to elective-dedicated operating room slots, flexible slots are assigned to handle the fluctuating demand of emergency patients. If the reserved flexible capacity is insufficient, elective surgeries will be cancelled to free operating room capacity. Most emergency patients require a bed, both while waiting for surgery and while recovering from surgery, and elective surgeries are cancelled if no more beds are available.

To model and solve the MSSP, we propose a two-stage stochastic optimization model, where the stochastic parameters represent the number of emergency patients that must be served each week, and the bed requirements imposed by emergency patients. In the first stage, the MSS is constructed, and in the second stage the scheduling of emergency surgeries and potential elective cancellations are performed. Furthermore, we develop a discrete-event simulation model, and propose a simulation-optimization procedure where the simulation model generates scenarios for the optimization model.

The main contribution of the paper is the two-stage stochastic optimization model that generates an MSS which balances efficiency and responsiveness when serving planned elective surgeries and high-priority emergencies in the same facilities. Furthermore, we demonstrate how a simulation model can be used to generate scenario data that represents alternative planning strategies. Finally, we design a case study to compare the optimized MSS with the existing one, demonstrating the value of our work.

1.2.3 Paper II: Integrated Master Surgery and Outpatient Clinic Scheduling

In this paper, we aim to coordinate the use of resources across the OC and the OT. The inherent demand of the problem is imposed by the new referrals entering the system, as each referral requires an initial consultation in the OC. We know the expected sequence of activities required by patients from different specialties, and this is used to estimate the derived demand for all activity types based on the number of initial consultations scheduled. The objective of the problem is to construct a master schedule that assigns sufficient capacity for initial consultations, while making sure that we can cope with the derived downstream demand for activities.

The main contribution of the paper is an optimization model that generates the integrated master schedule, where a specialty and a number of subordinate activity types are assigned to each unit on each day of the planning cycle. We also develop a discrete-event simulation model to evaluate the operational performance of the master schedule under different operational scheduling policies. We evaluate three scheduling policies, and find that a policy that combines the use of the tactical activity type assignments with some degree of flexibility performs best.

1.2.4 Paper III: A framework for integrated resource planning in surgical clinics

In this paper, we study a similar problem as in Paper II, and it can be seen as an extension of Paper II. In contrast to Paper II, we separate the master schedule in two, one cyclic high-level, and one non-cyclic low level schedule. In the high-level schedule, specialties are assigned to rooms in both units on each day of the planning cycle, referred to as the cycle days. In the low-level schedule, we assign a number of activity types to rooms and days in the planning horizon. Each day in the planning horizon corresponds to a day in the planning cycle, and the assignments made in the low-level schedule are constrained by the high-level schedule assignments. To account for stochastic waiting lists, we propose a two-level planning procedure where the assignments in the low-level schedule are periodically adjusted to handle the current waiting lists.

To generate the master schedule, we develop an optimization model that extends on the model from paper II. The model is more sophisticated, and it explicitly considers all waiting lists in the system. When performing the periodically refinement of the low-level schedule, we solve the model with fixed high-level variables. We also extend on the simulation model from Paper II, and propose a procedure for evaluating the planning framework.

The main contribution of the paper is a two-level planning framework, composing of the optimization model and the two-level planning procedure. Furthermore, we apply the evaluation procedure to study different planning strategies, including a flexible, a dynamic and an agile planning strategy, and we show that combining the strategies yields additive improvements related to patient waiting times.

1.2.5 Paper IV: Simulating emergency patient flow during the COVID-19 pandemic

In mid March 2020, the management at St. Olav's Hospital decided that all patients that entered the hospital with a COVID-19 suspicion should be screened in the Emergency Department, and that they should be transported by ambulance both to and from the hospital. In addition, it was proposed that all patients who were not screened upon arrival would require an ambulance when leaving the hospital. In this paper, we aim to estimate the impact on the Emergency Department and the ambulance services imposed by COVID-19 hospital admissions during the peak of the pandemic.

As a starting point of our analysis, we split the emergency patient population in two: the COVID-19 positive and negative patients. Then, based on the recommended planning scenarios and a set of assumptions regarding the screening policy, we develop a set of scenarios representing the demand of emergency patients to both the Emergency Department and the ambulance services. Two discrete-event simulation models are developed, one for each system, and these are used to analyse the systems across the different scenarios. Finally, the two models are implicitly integrated into a third model, allowing us to study the effects of boarding on the Emergency Department bed requirements.

The main contribution of the paper is to demonstrate the use of discrete-event simulation to provide real-life decision support in a state of uncertainty. The analyses were presented for the hospital management, who established more beds for screening in the Emergency Department, and increased the transportation capacities.

1.3 Contributions

In this section, we present the overarching contributions made in this thesis to the research community and to the industry.

1.3.1 Contributions to the research community

In the following, we discuss the overarching contributions of the thesis to the research community. We have made advances in the field of tactical surgery scheduling, related to two main topics: integrated master scheduling, and the handling of stochastic demand.

Integrated master scheduling

As stated in Section 1.1.3, no contributions exist on the MSSP that integrate the planning of the OC and the OT. In Papers II and III, we develop optimization models that construct integrated master schedules across both units. In Paper II, we aim to level the capacity assigned to initial consultations and to the downstream activity types, such that we serve sufficient patients while making sure that the downstream waiting lists do not increase. In Paper III, we explicitly consider all waiting lists in the system, and we use this information to obtain a high throughput of patients and short waiting times across all waiting lists.

In both papers, we assign activity types in the master schedules to facilitate an efficient use of resources, and strengthen the coordination between the OC and the OT. In Paper II, we compare the outcomes of scheduling patients based on the activity types assigned in the master schedule, to a policy where patients are scheduled purely based on the specialty and a first-come-first-serve policy. The results demonstrate that the first policy suffers from inflexibility, while the second struggles to achieve a coordinated use of resources. However, combining the two policies outperforms the others, especially if downstream resources are scarce.

Handling of stochastic demand

We make several contributions on how modelling can be used to make schedules, and support ways of planning, that account for a stochastic demand for activities.

In Paper I, we construct an MSS that reserves capacity to handle a stochastic demand of suburgent emergency surgeries. Roughly speaking, these patients should be served in the same week as they enter, and we do not know the emergency demand one week ahead. Therefore, as long as our assumptions about demand is not changed, it makes sense to make static capacity reservations for a long period of time. As a contribution to the research society, we propose a two-stage stochastic optimization model that accounts for the stochastic demand of emergency surgeries, and constructs an MSS where the level of robustness can be altered based on the planner's preferences.

In Paper III, we study a problem with stochastic waiting lists of elective patients. In contrast to the emergency patients, the elective waiting lists yield information about the demand for the upcoming weeks and months. Furthermore, given that we do not look too far ahead and that we serve the waiting lists in a first-come-first-serve policy, we can assume a deterministic demand. In this case, it makes sense to periodically (or non-periodically) update the capacity assignments in the master schedules based on the waiting list information. Comparing to the more strategic decisions of reserving capacities to emergency patients, these are tactical decisions that can be made more frequently. As a scientific contribution, we develop a planning framework, composing of a two-level planning procedure and a deterministic optimization model, to handle the stochastic demand of elective patients.

The case of stochastic waiting lists is also encountered in Paper II. Here, we propose a policy for scheduling patients based on the activity type assignments made in the master schedule and the current waiting lists.

Seen in combination, and with regards to the topic of handling a stochastic demand, the contributions made in the three papers align in the vertical axes proposed by Hulshof et al. (2012). This means that they can be implemented independently of each other, or together.

1.3.2 Contributions to the industry

The author of this thesis (referred to as the candidate) has been an employee at St. Olav's Hospital since August 2020, at the Regional Center for Health Services Development (RSHU). RSHU is a multidisciplinary unit, serving the clinical departments in matters of resource planning and logistics, innovation, and health economics. At the time of writing, we are three employees in the team that covers resource planning and logistics, adding up to 1.25 Full Time Equivalents. We all have shared positions between RSHU and the Department of Industrial Economics and Technology Management (IØT), allowing us to create synergies across academia and practice. In addition to conducting projects with the departments, we have arranged a few seminars for clinicians, presenting our methods and research results. The clinicians show an increasing interest in Operations Research, and we are currently working with several departments related to topics such as OC planning, and personnel scheduling.

Our collaboration with the Orthopaedic Department dates back to 2017, when the candidate initiated his Master thesis project. Since then we have conducted multiple projects together, resulting in several Master thesis and peer-reviewed publications. While we have had projects concentrating on operational scheduling, most emphasis has been given to tactical planning. So far, the collaboration has been of an academic character, focusing on the development of theoretical models, and it still remains to demonstrate the real-life potentials of our models. However, we have ambitions to conduct pilot studies based on our work, together with the department. Back in 2019, we applied for funding from St. Olav's Hospital to conduct a pilot study related to Paper I, but unfortunately the project was not funded. In the two following years, the hospital was handling a state of pandemic, causing less capacities to conduct non-clinical activities. Then, in 2022, a new electronic health record system was implemented at St. Olav's Hospital, causing massive amounts of preparations and change management both prior to and following the implementation. We are currently part of a consortium working on a grant application to the Research Council, related to developing and piloting a tool for strategic and tactical surgery scheduling.

The projects referred to in Paper IV made contributions to the hospital management. In contrast to traditional academic projects, the projects referred to in Paper IV were conducted with a sense of urgency. At this point in time, the management had incentives to demonstrate a drive, and they did state a willingness to act upon our advises. Both projects were finished in less than two weeks, and we had project meetings once or twice each day. In hindsight, we see that some of the assumptions made in the analyses can be challenged, and with more time we would have been able to explore more strategies especially related to the ambulance services. Despite this, the results of the models were acted upon by the management. Some months after the projects, new COVID-19 tests were developed that could provide results in less than half the time. When the new tests were to be implemented, we adjusted the simulation model and established a set of updated scenarios to evaluate whether some of the beds established for screening could be brought back to serve the ordinary emergency patients.

The use, and reuse, of a relatively simple simulation model with limited academic interest (in it self) is noteworthy. Before we can implement advanced tools based on complex mathematical models, we must demonstrate the ability to solve the problems currently faced by practitioners and introduce our methods step-by-step. In a dynamic and fluctuating reality, the scientific ethos and rigorous methodologies valued in science can be too tedious: by the time we have properly defined the problem, it may no longer be of interest to solve. On the other hand, we should avoid becoming too involved in firefighting symptoms of underlying problems. Targeting low-hanging fruits and solving problems that are of interest to practitioners should primarily serve as a door opener to targeting the bigger and more important problems.

1.3.3 The author's contributions to each paper in the thesis

In this section we discuss the contributions made by the candidate to each of the papers included in the thesis. The contributions are divided into the following categories: conceptual, implementation, and writing. The conceptual part includes the research idea, the formulation of the problem, the necessary assumptions, the purpose of the paper, and defining the scope of the analysis. Implementation refer to the collection of data, implementing models and solution methods, and analysing and systematising the results. Finally, the writing category includes the structuring and writing of the paper, handling the submission process, and following the revision process. A summary of the candidate's contributions are given in Table 1.1. Here, the contributions are ranked on a scale from 1 ti 3. The rating 1 represents some contribution, 2 means a significant contribution was made, whereas 3 represents a majority of the contributions.

Paper II extends on the Master's thesis of Anita Abdullahu, which was co-supervised by the candidate. Anita implemented the optimization model during her work. While minor improvements were made to the optimization model, the candidate developed the simulation model presented in the paper, conducted the

Table 1.1: The candidate's contribution to each of the papers in the thesis

Paper	**Conceptual**	**Implementation**	**Writing**
Paper I	3	3	3
Paper II	3	3	3
Paper III	3	3	3
Paper IV	3	3	3

analysis, and wrote the paper.

1.4 Concluding remarks and future research

A large body of research has emerged on the topic of tactical surgery scheduling during the past 20 years. While the work presented in this thesis is inspired by real-world problems faced by the Orthopaedic Department at St. Olav's Hospital, it also extends on current knowledge from the literature. Our main contributions to the research society relates to the topics of integrated planning and ways of handling a stochastic demand for surgery. We demonstrate that integrating the OC and the OT when constructing master schedules is useful to mange the waiting lists across both units. Furthermore, we study different planning strategies, both on the tactical and operational level, and we demonstrate how these can be applied handle stochastic waiting lists. While we mainly concentrate on the planning of elective surgeries, we also develop an optimization model that dedicates capacities for emergency patients in the MSS. The model balances the requirements of efficiency and responsiveness when serving elective and emergency surgeries, respectively.

Despite the close collaboration with the Orthopaedic Department, we have yet to demonstrate the real-life potential of our models. However, we have ambitions to conduct pilot studies based on our models in near future. While our work on surgery scheduling has been of an academic character so far, we include a paper that demonstrates how discrete-event simulation were applied to provide decision support to the hospital management during the COVID-19 pandemic. Despite being relatively simple, the model results proved useful to the management in times of uncertainty.

As topics for future research, we propose to compare different planning strategies for the handling of stochastic waiting lists. In the papers included in this thesis, we introduce and demonstrate the use of different planning strategies, both on the tactical and operational level. We believe that some of the policies are rather simple to implement, while others are more complex and will require the involvement of more people to implement. Keeping this in mind, it is interesting to distinguish the values added by adopting more complex strategies and ways of planning. For practical matters there is definitely a trade-off between efficiency and complexity.

Another topic for future research relates to the agile planning strategy introduced in Paper III, where we demonstrate that decreasing the planning delay between the time of planning and executing a schedule improves the waiting list outcomes. However, for reasons related to both the staff and the patients, the planning delay cannot be zero days. Seen from an Operations Research point of view, the sequence of planning and execution in this case seems to suit a stochastic programming framework. Can the use of stochastic programming allow us to add days to the planning delay, without loosing performance?

Finally, in all our master schedules we assign activity types to guide the operational scheduling of patients. As an alternative to this, we can develop an operational optimization model to perform the scheduling of patients, taking the current waiting lists and the master schedule as input. This will allow us to take more detailed information into considerations when performing the scheduling of individual surgeries, which can reduce the chances of overtime etc. Comparing the results from an operational scheduling model to the more naive patterns obtained from the activity type assignments is of interest.

Bibliography

A. Agnetis, A. Coppi, M. Corsini, G. Dellino, C. Meloni, and M. Pranzo. Long term evaluation of operating theater planning policies. *Operations Research for Health Care*, 1(4):95–104, 2012.

A. Agnetis, A. Coppi, M. Corsini, G. Dellino, C. Meloni, and M. Pranzo. A decomposition approach for the combined master surgical schedule and surgical case assignment problems. *Health Care Management Science*, 17:49–59, 2014.

N. Altay and W. G. Green. OR/MS research in disaster operations management. *European Journal of Operational Research*, 175(1):475–493, 2006.

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, 01 2013.

I. A. Bikker, N. Kortbeek, R. M. van Os, and R. J. Boucherie. Reducing access times for radiation treatment by aligning the doctor's schemes. *Operations Research for Health Care*, 7:111–121, 2015.

J.T. Blake and M.W. Carter. Surgical process scheduling: a structured review. *Journal of the Society for Health Systems*, 5(3):17—30, 1997.

J.T. Blake and J. Donald. Using integer programming to allocate operating room time at mount sinai hospital. *Interfaces*, 32(2):63–73, 2002.

J.T. Blake, F. Dexter, and J. Donald. Operating room managers' use of integer programming for assigning block time to surgical groups: A case study. *Anesthesia & Analgesia*, 94(1):143–148, 2002.

J. Britt, M. F. Baki, A. Azab, A. Chaouch, and X. Li. A stochastic hierarchical approach for the master surgical scheduling problem. *Computers & Industrial Engineering*, 158:107385, 2021.

P. Cappanera, F. Visintin, and C. Banditori. Comparing resource balancing criteria in master surgical scheduling: A combined optimisation-simulation approach. *International Journal of Production Economics*, 158:179 – 196, 2014.

B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201 (3):921–932, 2010.

C. S. M. Currie, J. W. Fowler, K. Kotiadis, T. Monks, B. S. Onggo, D. A. Robertson, and A. A. Tako. How simulation modelling can help reduce the impact of COVID-19. *Journal of Simulation*, 14(2):83–97, 2020.

Z. Dai, J. J. Wang, and J. Shi. How does the hospital make a safe and stable elective surgery plan during COVID-19 pandemic? *Computers & Industrial Engineering*, 169:108210, 2022.

European Commission. The impact of demographic change in a changing environment. 2023.

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

A. Fügener, E. W. Hans, R. Kolisch, N. Kortbeek, and P. T. Vanberkel. Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, 239(1):227–236, 2014.

D. Garcia-Vicuña, F. Mallor, and L. Esparza. Planning ward and intensive care unit beds for COVID-19 patients using a discrete event simulation model. In *2020 Winter Simulation Conference (WSC)*, pages 759–770, 2020.

M. G. Güler and E. Geçici. A decision support system for scheduling the shifts of physicians during COVID-19 pandemic. *Computers & Industrial Engineering*, 150:106874, 2020.

E. W. Hans, M. Van Houdenhoven, and P. J. H. Hulshof. *A Framework for Healthcare Planning and Control*, pages 303–320. Springer US, Boston, MA, 2012.

Helsedirektoratet. Ventetider og pasientrettigheter 2021 Norsk pasientregister. 2021.

Helsedirektoratet. Regelverk for innsatsstyrt finansiering 2023 (ISF-regelverket). 2023.

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

P. J. H. Hulshof, M. R. K. Mes, R. J. Boucherie, and E. W. Hans. Patient admission planning using approximate dynamic programming. *Flexible Services and Manufacturing Journal*, 28(1):30–61, 2016.

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

D. M. Kluger, Y. Aizenbud, A. Jaffe, F. Parisi, L. Aizenbud, E. Minsky-Fenick, J. M. Kluger, S. Farhadian, H. M. Kluger, and Y. Kluger. Impact of healthcare worker shift scheduling on workforce preservation during the COVID-19 pandemic. *Infection Control & Hospital Epidemiology*, 41(12):1443–1445, 2020.

A. Kumar, A. M. Costa, M. Fackrell, and P. G. Taylor. A sequential stochastic mixed integer programming model for tactical master surgery scheduling. *European Journal of Operational Research*, 270(2):734–746, 2018.

C. Laan, M. van de Vrugt, J. Olsman, and R. J. Boucherie. Static and dynamic appointment scheduling to improve patient access time. *Health Systems*, 7(2): 148–159, 2018.

J. Le Lay, V. Augusto, X. Xie, E. Alfonso-Lizarazo, B. Bongue, T. Celarier, R. Gonthier, and M. Masmoudi. Impact of COVID-19 epidemics on bed requirements in a healthcare center using data-driven discrete-event simulation. In *2020 Winter Simulation Conference (WSC)*, pages 771–781, 2020.

J.M. Magerlein and J.B. Martin. Surgical demand scheduling: a review. *Health Services Research*, 13(4):418–433, 1978.

S. Makboul, S. Kharraja, A. Abbassi, and A. E. H. Alaoui. A two-stage robust optimization approach for the master surgical schedule problem under uncertainty considering downstream resources. *Health Care Management Science*, 25 (1):63–88, 2022.

J. Marynissen and E. Demeulemeester. Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, 272(2):407–419, 2019.

M. Mazloumian, M. F. Baki, and M. Ahmadi. A robust multiobjective integrated master surgery schedule and surgical case assignment model at a publicly funded hospital. *Computers & Industrial Engineering*, 163:107826, 2022.

G.J. Melman, A.K. Parlikad, and E.A.B. Cameron. Balancing scarce hospital resources during the COVID-19 pandemic using discrete-event simulation. *Health Care Management Science*, 24:356–374, 2021.

A. Moosavi and S. Ebrahimnejad. Robust operating room planning considering upstream and downstream units: A new two-stage heuristic algorithm. *Computers & Industrial Engineering*, 143:106387, 2020.

Norges Offentlige Utredninger. Tid for handling: Personellet i en bærekraftig helse- og omsorgstjeneste. 2023.

OECD Publishing. Health at a glance: Europe 2022: State of health in the eu cycle. 2022. doi: https://doi.org/10.1787/507433b0-en.

M. Oliveira, F. Visintin, D. Santos, and I. Marques. Flexible master surgery scheduling: combining optimization and simulation in a rolling horizon approach. *Flexible Services and Manufacturing Journal*, 34:824–858, 2021.

M. Otten, S. Dijkstra, G. Leeftink, B. Kamphorst, A. O. Meierink, A. Heinen, R. Bijlsma, and R. J. Boucherie. Outpatient clinic scheduling with limited waiting area capacity. *Journal of the Operational Research Society*, 74(2):540–561, 2023.

M.L. Penn, C.N. Potts, and P.R. Harper. Multiple criteria mixed-integer programming for incorporating multiple factors into the development of master operating theatre timetables. *European Journal of Operational Research*, 262(1):194–206, 2017.

S. Rachuba, L. Imhoff, and B. Werners. Tactical blueprints for surgical weeks – an integrated approach for operating rooms and intensive care units. *European Journal of Operational Research*, 298(1):243–260, 2022.

I. Rahimi and A. H. Gandomi. A comprehensive review and analysis of operating room and surgery scheduling. *Archives of Computational Methods in Engineering*, 28(3):1667–1688, 2021.

M. K. Razali, A. H. A. Rahman, M. Ayob, R. Jarmin, F. Qamar, and G. Kendall. Research trends in the optimization of the master surgery scheduling problem. *IEEE Access*, 10:91466–91480, 2022.

M. Samudra, C. Van Riet, E. Demeulemeester, B. Cardoen, N. Vansteenkiste, and F. E. Rademakers. Scheduling operating rooms: achievements, challenges and pitfalls. *Journal of Scheduling*, 19:493–525, 2016.

D. Santos and I. Marques. Designing master surgery schedules with downstream unit integration via stochastic programming. *European Journal of Operational Research*, 299(3):834–852, 2022.

A. J. T. Schneider, J. T. van Essen, M. Carlier, and E. W. Hans. Scheduling surgery groups considering multiple downstream resources. *European Journal of Operational Research*, 282(2):741–752, 2020.

B. Spratt and E. Kozan. Waiting list management through master surgical schedules: A case study. *Operations Research for Health Care*, 10:49–64, 2016.

stolav.no. Nøkkeltall for St. Olavs hospital. 2023. URL `https://stolav.no/om-oss/nokkeltall-for-st-olavs-hospital`.

A. Testi and E. Tànfani. Tactical and operational decisions for operating room planning: Efficiency and welfare implications. *Health Care Management Science*, 12(4):363–373, 2009.

A. Tjernshaugen, H. Hiis, J. F. Bernt, G. S. Braut, V. B. Bahus, and M. M. Simonsen. Koronapandemien. 2023. URL `https://sml.snl.no/koronapandemien`.

P. T. Vanberkel, R. J. Boucherie, E. W. Hans, J. L. Hurink, W. A. M. van Lent, and W. H. van Harten. An exact approach for relating recovering surgical patient workload to the master surgical schedule. *Journal of the Operational Research Society*, 62(10):1851–1860, 2011.

R. M. Wood, C. J. McWilliams, C. P. Thomas, M. J.and Bourdeaux, and C. Vasilakis. COVID-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. *Health Care Management Science*, 23(3):315–324, 2020.

Paper I

T. R. Bovim, M. Christiansen, A. N. Gullhav, T. M. Range, L. Hellemo:

Stochastic Master Surgery Scheduling

Published in European Journal of Operational Research

Chapter 2

Stochastic Master Surgery Scheduling

Abstract

The aim of the Master Surgery Scheduling Problem (MSSP) is to schedule the medical specialties to the different operating rooms available, such that surgeries may be performed efficiently. We consider a MSSP where elective and emergency patients can be treated in the same operating rooms. In addition to elective-dedicated operating room slots, flexible operating room slots are introduced to handle the fluctuating demand of emergency patients.

To solve the MSSP, we propose a simulation-optimization approach consisting of a two-stage stochastic optimization model and a discrete-event simulation model. For the two-stage stochastic optimization model, uncertain arrivals of emergency patients are represented by discrete scenarios. The discrete-event simulation model is developed to address uncertainty related to the surgery duration and the length of stay at the hospital, and to test the Master Surgery Schedule (MSS) developed by the optimization model in a stochastic operational-level environment. In addition, the simulation model is used to generate scenarios for the optimization model.

We present some general advice for surgery scheduling based on testing the optimization model in a numerical study. The simulation-optimization approach is applied to a case study from a hospital department that treats both elective and emergency patients. The optimized MSS outperforms the manually generated MSS, both in terms of emergency waiting time for surgery, and emergency interruptions to the flow of electives.

2.1 Introduction

Demographic changes in Norway and many other countries are increasing the need for hospital services in the years to come. One of the major activities at a hospital is providing surgery to patients. Freeman et al. (2018) state that 60-70% of all patients admitted to a hospital require some surgical intervention, and Essen et al. (2012) state that surgical costs account for approximately 40% of the total hospital costs and that surgeries generate around 67% of hospital revenues. Developing ways to schedule surgery efficiently is key to proper utilization of scarce hospital resources, and a necessity to be able to treat more patients.

Patients are commonly divided into two groups: elective and emergency patients. Elective patients are not experiencing a medical emergency, and their surgery can be scheduled in advance to suit the availability of the surgeon and the patient. Emergency patients, on the other hand, may require surgery within hours or up to a few days. A triage system is often applied to further divide the emergency patients according to the urgency of their condition. Another classification of patients refers to whether the patient may leave the hospital following surgery or not.

One of the major issues within surgery scheduling is how to best balance efficiency and responsiveness when conducting surgeries for scheduled electives and high-priority emergencies. If the OR capacity is shared between electives and emergencies, emergencies can create disruptions to the handling of scheduled surgeries, implying longer elective waiting times, costly resource overtime, cancellations and rescheduling. If some of the OR capacity is dedicated to respond to emergencies and avoid disruptions of electives, there will be times when the dedicated capacity is not utilized as no emergencies are present (Ferrand et al., 2014).

Surgery planning may be divided into three decision stages (Hulshof et al., 2012). At the strategic level, decisions on localization and dimension of the number and size of the operating rooms (ORs) are made. At the tactical level, a Master Surgery Schedule (MSS) is developed to schedule different specialties to the accessible ORs through the week. Finally, at the operational level, the individual patients are scheduled to the ORs covered by the respective specialty.

The main purpose of this paper is to provide tactical decision support for managers of departments that provide surgery to both elective and emergency patients. More specifically, we consider both elective and emergency patients in the Master Surgery Scheduling Problem (MSSP). The goal is to schedule the medical subspecialties to time slots in the ORs such that we ensure a sufficiently high throughput of elective patients while maintaining a high responsiveness for emergency patients. Two types of OR time slots are scheduled: elective slots and flexible slots that are primarily intended to handle sub-urgent emergency patients.

To perform the scheduling, we propose a simulation-optimization approach consisting of a two-stage stochastic optimization model and a discrete-event simulation model (see Figure 2.1). The optimization model generates an MSS, while the simulation model is used to evaluate the MSS and to generate new input scenarios for the optimization model. This allows us to generate a new MSS based

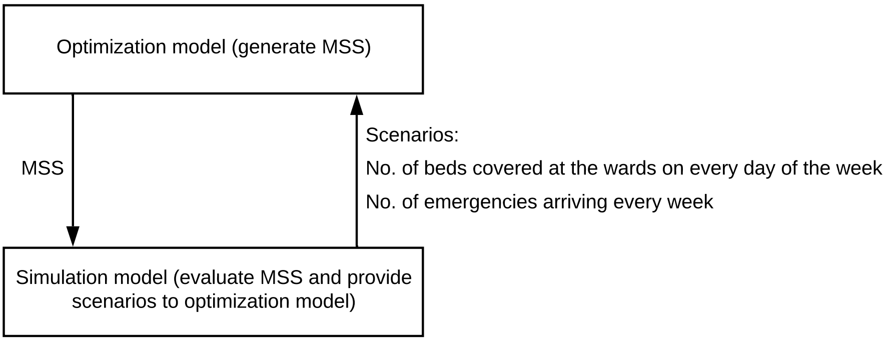

Optimization model (generate MSS)
MSS
Scenarios:
No. of beds covered at the wards on every day of the week
No. of emergencies arriving every week
Simulation model (evaluate MSS and provide scenarios to optimization model)

Figure 2.1: Illustration of our simulation-optimization approach. The optimization model generates an MSS, while the simulation model is used to evaluate the tactical schedule in a dynamic environment, and provides feedback to the optimization model in terms of scenarios.

on the scheduling rules that are applied in the simulation model and the MSS generated in the previous iteration. This is helpful to avoid using historic data that is dependent on the MSS and the scheduling regime that was present when the data was generated.

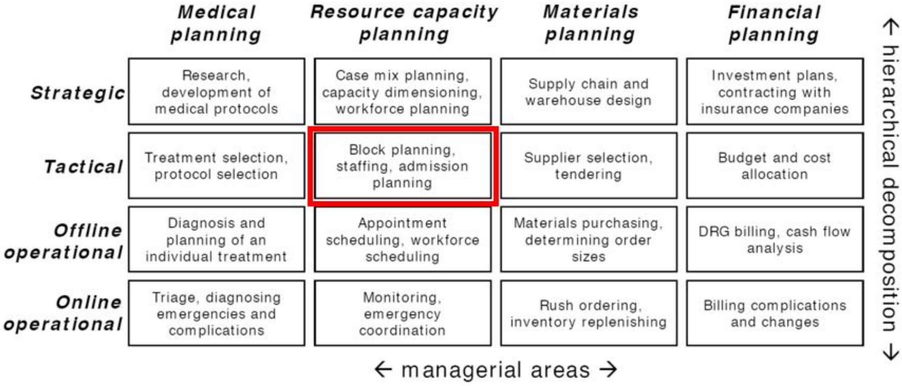
The rest of the paper is outlined as follows: Section 2.2 presents relevant literature. The MSSP is introduced in Section 2.3. In Section 2.4, we present the approach for solving the MSSP. Following this, in Section 2.5, we present the results from a computational study where we investigate the value of applying a stochastic model formulation, managerial insights and a case study from the orthopaedic department at St. Olav's hospital in Trondheim, Norway. Finally, we conclude the paper and suggest topics for further research in Section 2.6.

2.2 Literature review

To provide context, we present relevant literature both on surgery planning and methods for solving the planning problems. First, we present an overview of decision levels within surgery planning.

Hans et al. (2012) propose a holistic planning and control framework for a health care provider, which consists of four managerial areas, combined with a hierarchical decomposition of decision-making levels. Figure 2.2 illustrates the framework and provides examples of planning and control functions for each combination. The MSSP considers the managerial area of resource capacity planning. Furthermore, the MSS is a cyclical block schedule that is repeated for several months, implying that we consider the tactical planning level.

Within the field of OR planning, the majority of publications have considered



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

Figure 2.2: Framework for health care planning and control (Hans et al., 2012)

only the elective patients (Cardoen et al., 2010). The literature on the MSSP is no exception, and most authors argue that the emergency patients are handled with dedicated resources. However, some authors like Freeman et al. (2018), Lamiri et al. (2008), Razmi et al. (2015) and Adan et al. (2011) include emergency patients.

OR capacity is commonly divided into time blocks when solving the MSSP. Many authors consider surgery slots of equal length, while others, like Mannino et al. (2012) include surgery blocks of different lengths. Testi and Tànfani (2008), Mannino et al. (2012) and Adan et al. (2011) consider overtime work at the ORs. Testi and Tànfani (2008) and Mannino et al. (2012) minimize the overtime in the objective function, while Adan et al. (2011) impose hard constraints on the overtime allowed for each OR. Koppka et al. (2018) limit the total OR opening hours available through the week, but allows for the model to decide how the opening hours should be distributed over the different ORs.

The wards are frequently included. However, some authors, like Testi et al. (2007), Li et al. (2017) and Mannino et al. (2012) disregard the wards in their model under the assumption that the access to beds is not imposing a bottleneck on the efficient flow of patients. A few authors, like Li et al. (2017), Fügener et al. (2014) and Adan et al. (2011), include the intensive care unit (ICU) when handling the MSSP. The latter are also among the few that explicitly include the nurses, by imposing restrictions on the amount of nursing hours available at the ICU.

The literature on the MSSP presents numerous objective functions. Testi and Tànfani (2008), Testi et al. (2007) and Penn et al. (2017) include aspects of welfare into their objective functions. Testi and Tànfani (2008) propose an objective function that minimizes loss of welfare among the patients, and the latter two maximize surgeon preferences. A variety of objective functions regarding bed capacity is proposed in the literature. Oostrum et al. (2008) aim to minimize both the number of ORs used and the maximum demand for hospital beds during the planning cycle. Ma and Demeulemeester (2013) minimize the total bed deficit, the maximum daily spare bed volume and the maximum variance of the bed occupancy.

The inclusion of multiple criteria objective functions is used by several authors. In Beliën et al. (2008), the objective function contains three parts: minimization of the total peak mean and variance bed occupancy, minimization of surgeons of the same specialty performing surgery in different rooms and minimization of surgeons not being scheduled to the same room on the same day every week of the planning horizon. Li et al. (2017) aim at minimizing the number of patients not being scheduled, minimizing the underutilization of OR time, minimizing the maximum expected number of patients in the recovery unit and minimizing the expected range of patients in the recovery unit.

Several authors include aspects of uncertainty when handling the MSSP. Oostrum et al. (2008) include a probability distribution for running into overtime in an OR as a function of the number of surgeries scheduled to that OR. Koppka et al. (2018) consider the probability of running into overtime in the ORs that depends on the combination of patients that are scheduled for the OR. Adan et al. (2011), Ma and Demeulemeester (2013) and Li et al. (2017) include probability distributions to account for uncertainty in the patient's length of stay (LOS) following surgery. Fügener et al. (2014) calculate the distribution of patients resting in both the wards and the ICU resulting from a cyclical MSS. These distributions are used in the objective function to minimize the fixed costs, the overcapacity costs, and the staffing costs in both the wards and the ICU when generating the MSS.

According to Higle (2005), stochastic programming is a technique which is well suited when some of the data elements are difficult to predict or estimate. A major framework within stochastic programming is two-stage recourse modeling. A two-stage recourse model consists of a first-stage problem and a second-stage (recourse) problem. The first-stage decisions are determined before knowing the outcome of the stochastic parameters, while the second-stage decisions are made after observing the realization of the stochastic parameters. The goal when applying a two-stage stochastic modelling approach is to identify a first stage solution that performs well in expectation, taking all possible realizations of the stochastic parameters into account. When approximating the continuous probability distributions of the stochastic parameters with discrete scenarios, extra care has to be taken to ensure stability of the solution (see e.g. Kall and Wallace (1994)).

Some authors propose two-stage stochastic models at the tactical level within surgery planning. Koppka et al. (2018) develop a two-stage stochastic model to deal with the varying number of elective patients that require surgery during the planning horizon. However, the authors do not include a recourse option. Kumar et al. (2018) present a method that is inspired by the two-stage stochastic method with recourse. Uncertainty is incorporated by using various LOS scenario realizations, and non-anticipation (see Higle (2005)) is imposed by constraining the model to schedule patients in the same order as their position in the queue. However, in contrast to the traditional two-stage stochastic models, the scenario realizations in this model are chronologically sequential and not parallel. This allows for the model formulation to be deterministic.

Figueira and Almada-Lobo (2014) state that there are three major streams of simulation optimization research : Solution Evaluation (SE), Solution Generation (SG), and Analytical Model Enhancement (AME) approaches. Within these three streams several methods are available. The methods are categorized based on the interaction between simulation and optimization, and on the search algorithm design.

The SE approaches consist of developing a comprehensive simulation model to represent the system and use that model to evaluate performance of various solutions. The results of the simulation model is used to guide the search for new, better solutions. The SG approaches are used when optimization models can be formulated and solved, and their solutions simulated in order to compute realistic values of the variables. The purpose of simulation here is not to verify the advantage of one solution over another, but to compute realistic values for some variables and hence to be part of the whole solution generation procedure. In AME approaches, the optimization model is enhanced by the simulation results, for example by providing better estimates of parameters applied in the optimization model. One of the methods used in the AME approaches is called Stochastic Programming Deterministic Equivalent (SPDE), and here the simulation model is used to generate scenarios for the stochastic optimization model.

Several authors use the SG approach when formulating and solving surgery scheduling problems. Freeman et al. (2018), Ma and Demeulemeester (2013) and Testi et al. (2007) use simulation models to evaluate the tactical schedule produced by the optimization model in an operational setting. This allows them to explore the realistic values of the decision variables when more uncertainty is included. Adan et al. (2011) and Cappanera et al. (2014) use discrete-event simulation to investigate different operational scheduling policies after having first generated a tactical surgery plan. By this they reveal realistic values of the tactical decision variables for different scheduling policies.

In accordance with the AME approach, Lamiri et al. (2008) use Monte Carlo simulations and a sample average approximation (SAA) to solve a stochastic optimization problem with uncertainty related to emergency arrivals. Also Ma and Demeulemeester (2013) apply the AME approach, as the simulated, operational level results are used to alter the parameter for the total bed capacity in the optimization model.

We propose a simulation-optimization approach to solve the MSSP. According to the AME approach (SPDE method), we use a discrete-event simulation model when generating the scenarios for our two-stage stochastic optimization model. We also use the SG approach, as we use the simulation model to evaluate the MSS generated by the optimization model. We introduce different levels of urgency for the emergency patients, and we allow for rescheduling of elective patients to provide capacity for the emergency patients in periods of excessive emergency patient loading. This allows us to investigate a major trade-off faced by the management at many surgical departments, namely the number of electives scheduled for surgery versus the amount of elective rescheduling needed in order to provide surgery for emergency patients.

2.3 Problem description for the MSSP

The MSSP is described in three steps. First, we provide the MSSP with a focus on elective patients. Then, we expand the problem by including emergency patients, and finally we present the concept of flexible slots that are used to handle the flow of emergency patients in periods of high emergency demand.

2.3.1 The elective MSSP

In this MSSP, the aim is to generate a cyclic MSS where the medical subspecialties are scheduled to the available ORs throughout the planning cycle (typically one week). A set of elective patient categories exist that share diagnostic similarities. The patient categories are either in- or outpatients. The surgeons performing the surgeries are divided into different medical subspecialties. Surgeons of a given subspecialty may perform surgery in several patient categories, but each patient category may only receive surgery from surgeons of one subspecialty.

The department has a given number of ORs where surgeries are performed. The ORs are heterogeneous and each OR may only accommodate certain subspecialties. The opening hours of the OR are divided into time slots, and each slot can be scheduled to one subspecialty. The number of slots allowed to schedule for a subspecialty on a given day, and during the cycle is limited. For an OR to be available for scheduling, an anaesthesia resource must be scheduled for the OR. The number of ORs that may be covered by an anesthesia resource each day is limited. Each patient category has an expected surgery duration, and the number of patients that can be scheduled for surgery in an OR is limited by the slot capacity scheduled for a suitable subspecialty.

There are several heterogeneous wards available, where inpatients rest following their surgery. In each ward a given number of beds can be staffed each day. An upper bound on the total number of beds that can be staffed each day. Within the total capacity, we can distribute the number of staffed beds that are available in each ward on a given day. A staffed bed is assigned to each inpatient entering the hospital on the day of arrival, and this bed is occupied by that patient throughout the stay. This scenario is not always true in real life, but it is a fair assumption from a tactical planning perspective.

The subspecialties are scheduled to the ORs according to the cyclical, fixed period MSS. The target throughput of elective patients to be scheduled for each patient category for the cycle is known. For all patient categories, a given share of the target throughput must be scheduled for surgery in each cycle. This minimum throughput should be set such that the average service rate is incrementally higher than the average arrival rate for each patient category, such that we maintain a stable waiting list of elective patients. Because elective patients are (in most cases) scheduled well in advance, short periods of peaks in elective patient demand can be smoothed out by the hospital planners. For that reason, we argue that planning according to the average demand is sufficient at a tactical level.

2.3.2 Considering emergency patients in the MSSP

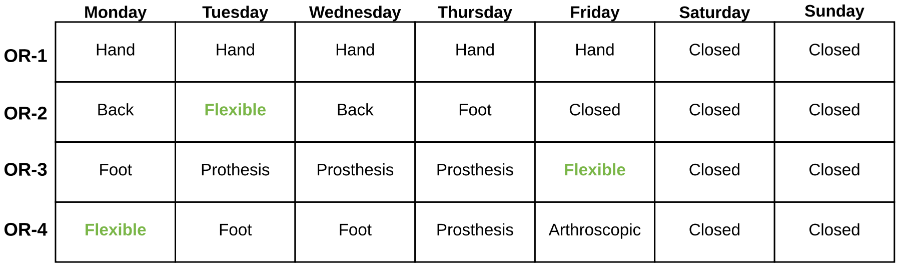
The flow of emergency patients may cause elective cancellations and rescheduling in periods when the emergency OR resources are insufficient to handle the emergency demand for surgery. Therefore, the emergency patients should be considered in the MSSP. The emergency patients represent different urgency categories, and they are grouped according to three emergency scheduling regimes:

- The sub-urgent (SU) emergency scheduling regime applies to the least urgent emergencies. In periods when the emergency OR capacity becomes insufficient to handle all emergency patients, the least urgent emergency patients are scheduled for the elective ORs to free emergency OR capacity.
- The urgent (U) emergency scheduling regime applies to the urgent emergency patients. If no patients from the SU scheduling regime are present, and there is idle capacity in the elective ORs, the urgent emergency patients are scheduled for the elective ORs to free capacity for the most urgent emergency patients in the emergency ORs.
- The critically urgent (CU) emergency scheduling regime applies to the most urgent emergency patients. These patients are always scheduled for the emergency ORs.

Scheduling emergency patients to the elective surgery slots may cause elective cancellations. However, if there is excess capacity in an elective OR slot, after all elective patients have been scheduled, emergency patients might be scheduled to the slot without causing cancellations (given that the idle capacity exceeds the planned surgery duration of the emergency patient).

In periods with a bed shortage in some of the wards, it is possible to let patients rest in wards dedicated to other patient groups. If the number of emergency patients requiring a bed increases, and no more scheduled beds are available, elective inpatients are cancelled to provide additional bed capacity. If the number of emergencies needing beds exceeds the total number of scheduled beds available, more beds need to be staffed. The random arrivals of emergencies cause both the emergency demand for surgery and the number of emergencies occupying beds in the wards each day to differ among cycles.

In periods of high emergency demand, SU emergency patients are displaced from the emergency ORs due to lower priority. These patients are called the excess demand of SU emergencies, and all of these should receive surgery in the elective ORs within the cycle. To handle the SU patients, we schedule flexible slots that are reserved for these patients. In periods of low demand for SU surgeries, the flexible slots can be used for scheduling of U emergency patients. Not all ORs may be accessible for scheduling of flexible slots. Figure 2.3 provides an example of an MSS including both elective-dedicated and flexible slots.



	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
OR-1	Hand	Hand	Hand	Hand	Hand	Closed	Closed
OR-2	Back	Flexible	Back	Foot	Closed	Closed	Closed
OR-3	Foot	Prothesis	Prosthesis	Prosthesis	Flexible	Closed	Closed
OR-4	Flexible	Foot	Foot	Prosthesis	Arthroscopic	Closed	Closed

Figure 2.3: An example of a one-week MSS for four single-slot ORs, where both elective-dedicated and flexible slots are scheduled.

2.3.3 Objectives

Three objectives are relevant for the MSSP. The first is to maximize the number of elective patients scheduled for surgery. Secondly, we aim to minimize both the number of elective cancellations, and the number of patients resting in wards not designated for them.

2.4 Simulation-optimization approach

In this section, we present the simulation-optimization approach used to solve the MSSP. The optimization model is formulated as a two-stage stochastic optimization problem. Scenarios are used to represent the probability distributions of the stochastic parameters, which are the number of emergency patients resting in each ward during each day in the cycle, and the excess demand of SU emergency patients. Figure 2.4 illustrates the main components of the simulation-optimization approach, and their interactions. The simulation-optimization process consists of the following steps:

1. To initiate the procedure, the simulation model is run for an arbitrary MSS to generate input data for the scenarios applied in the optimization model.

2. The scenarios generated from the simulation model is used as input for the optimization model, and the optimization model generates a new MSS based on these.

3. The new MSS is then implemented in the simulation model, and new scenarios are generated. Relevant measures, such as patient waiting time for surgery and the number of elective cancellations are stored.

4. The procedure (2.-3.) is repeated until a stopping criterion is met, and is then terminated.

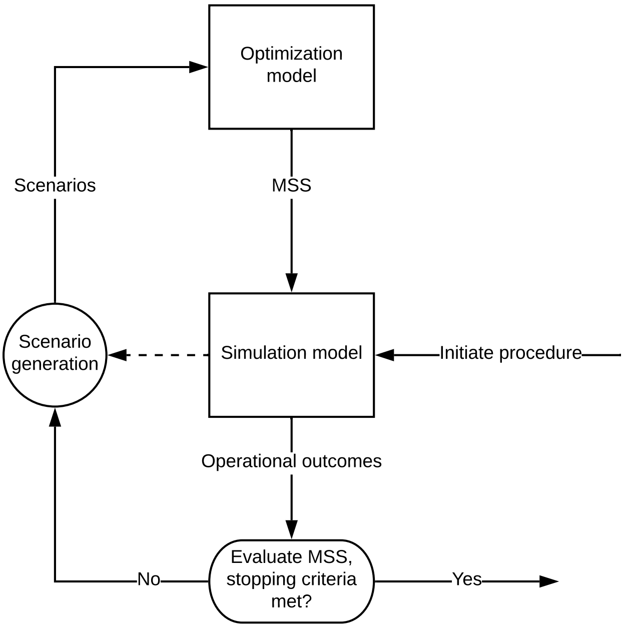


Figure 2.4: The set-up of the simulation-optimization approach, including both the optimization model and the simulation model.

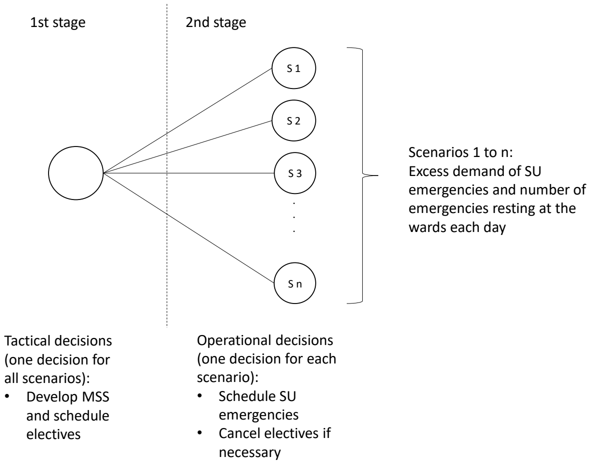

1st stage
2nd stage
S 1
S 2
S 3
S n
Scenarios 1 to n:
Excess demand of SU emergencies and number of emergencies resting at the wards each day
Tactical decisions (one decision for all scenarios):
• Develop MSS and schedule electives
Operational decisions (one decision for each scenario):
• Schedule SU emergencies
• Cancel electives if necessary

Figure 2.5: Illustration of the two-stage decision model

The simulation model takes the MSS generated by the optimization model as input, and evaluates it in an operational environment. In the optimization model, the LOS and surgery duration are deterministic parameters. In the simulation model, these values are drawn from empirical probability distributions for each patient. By generating scenarios from the simulation output, we are able to generate input to the optimization model that is dependent on the scheduling regime applied in the simulation model and the MSS at the previous iteration.

In Sections 2.4.1 and 2.4.2, we present the optimization and simulation models respectively. In Section 2.4.3, we discuss the scenario generation, and in Section 2.4.4, we describe the stopping criterion that terminates the simulation-optimization procedure.

2.4.1 The optimization model

We must decide on the MSS before knowing the exact number of SU emergencies that require surgery in every cycle, or the number of emergencies occupying beds in the wards each day. Therefore, a two-stage modelling framework is suitable for the optimization model, as illustrated in Figure 2.5. The first-stage decisions are tactical-level decisions, and are as follows:

- Assign an anaesthesia resource to the ORs in use.
- Schedule the available OR slots as either flexible or elective, and assign a subspecialty responsible to each slot.
- Schedule the elective patients for surgery in the elective slots.

- Decide on the number of staffed beds in each ward on every day throughout the cycle.

The second stage decisions are operational-level decisions that are made within each cycle:

- Schedule the excess demand of SU patients to the flexible and elective slots (if no more flexible slots are available).
- Cancel elective surgeries if necessary.
- Send inpatients to the wards and let patients rest in wards not designated for them, if necessary due to shortage of bed capacity.
- Staff more beds if necessary to handle all the emergency inpatients.
- If excess capacity is available in the flexible slots, schedule emergency patients belonging to the U emergency scheduling regime to these slots.

The uncertainty included in the optimization model is the excess demand of SU patients that require surgery in the cycle, and the number of emergencies resting in the wards each day through the cycle. These are represented by scenarios, and each scenario contains a complete realization of both aspects for one cycle.

In the following, sets are indicated by calligraphic letters, parameters are given by uppercase letters, and variables are in lowercase Latin or Greek letters. A.1 provides a detailed overview of all indices, sets, parameters and variables used in the model formulation.

The first-stage constraints

The total number of slots available for scheduling is given by constraint (2.1). The variable n_{jkd} represents the number of flexible slots scheduled for subspecialty j in OR k on day d, while y_{jkd} does the same for the elective slots. The parameter M^{CYCLE} represents the number of slots available through the cycle. Constraints (2.2) and (2.3) allocate OR slots to the different subspecialties. Parameter N^D_{jd} gives the number of OR slots available to subspecialty j on day d, while N_j states the number of OR slots available to subspecialty j in the cycle.

$$\sum_{j \in \mathcal{J}} \sum_{d \in \mathcal{D}} \left(\sum_{k \in \mathcal{K}_j \cap \mathcal{K}^F} n_{jkd} + \sum_{k \in \mathcal{K}_j} y_{jkd} \right) \leq M^{CYCLE} \tag{2.1}$$

$$\sum_{k \in \mathcal{K}_j \cap \mathcal{K}^F} n_{jkd} + \sum_{k \in \mathcal{K}_j} y_{jkd} \leq N^D_{jd} \qquad j \in \mathcal{J}, \quad d \in \mathcal{D} \tag{2.2}$$

$$\sum_{d \in \mathcal{D}} \left(\sum_{k \in \mathcal{K}_j \cap \mathcal{K}^F} n_{jkd} + \sum_{k \in \mathcal{K}_j} y_{jkd} \right) \leq N_j \qquad j \in \mathcal{J} \tag{2.3}$$

Scheduling according to the target level for the elective patient categories is ensured by constraints (2.4) and (2.5). The variable x_{ikd} represents the number of elective patients belonging to elective patient category i that are scheduled for OR k on day d, while v_i gives the number of patients belonging to a given elective patient category that are scheduled above the minimum level. The target level of electives belonging to patient category i is given by T_i, and this parameter imposes an upper limit on the number of elective patients belonging to a patient category that may be scheduled for surgery through the cycle. The parameters V_i represents the minimum share of patients belonging to patient category i that should be scheduled for surgery through the cycle.

$$\sum_{k\in\mathcal{K}}\sum_{d\in\mathcal{D}} x_{ikd} \leq T_i \qquad i \in \mathcal{I}^{EL} \tag{2.4}$$

$$\sum_{k\in\mathcal{K}}\sum_{d\in\mathcal{D}} x_{ikd} - v_i = \lceil T_i V_i \rceil \qquad i \in \mathcal{I}^{EL} \tag{2.5}$$

Parameter S_i represents the surgery duration of a patient belonging to patient category i, while B_{kd} is the slot length in OR k on day d. Constraints (2.6) ensure that the scheduled surgery duration at OR k does not exceed the time available for surgery at that OR on day d.

$$\sum_{i\in\mathcal{I}^J_j} S_i x_{ikd} \leq B_{kd} y_{jkd} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D} \tag{2.6}$$

The anaesthesia resources are modelled in constraints (2.7) and (2.8). Binary variable α^A_{kd} indicates whether OR k is covered by an anaesthesia resource on day d, while parameter M^A_d represents the number of ORs that may be covered by an anesthesia resource on day d. The parameter M^{OR}_{kd} represents the maximum number of slots that can be scheduled for OR k on day d. In constraints (2.7) the number of ORs covered by an anaesthesia resource on a given day is restricted by the total amount of anaesthesia resources available on that day. Constraints (2.8) ensure that no more slots, elective or flexible, may be scheduled for an OR on a day than the total number of slots available at the OR on that day. In order to schedule subspecialties to the OR, an anaesthesia resource has to cover the OR.

$$\sum_{k\in\mathcal{K}} \alpha^A_{kd} \leq M^A_d \qquad d \in \mathcal{D} \tag{2.7}$$

$$\sum_{j\in\mathcal{J}} (n_{jk'd} + y_{jkd}) \leq M^{OR}_{kd} \alpha^A_{kd} \qquad k' \in \mathcal{K}^F, \qquad k \in \mathcal{K}, \quad d \in \mathcal{D} \tag{2.8}$$

The first-stage bed constraints are given by (2.9) and (2.10).The number of staffed beds in ward w on day d is given by the variable a_{wd}, while parameters A^{MAX}_w and A_d represent the maximum number of beds available in ward w per day, and the maximum number of staffed beds available on day d, respectively. In constraints (2.9), we require that the number of staffed beds in a ward on a given day cannot exceed the maximum number of beds available in the ward. Constraints (2.10) ensure that we cannot staff more beds in total on a given day than the total number of staffed beds available on that day.

$$a_{wd} \leq A_w^{MAX} \qquad d \in \mathcal{D}, \quad w \in \mathcal{W} \tag{2.9}$$

$$\sum_{w \in \mathcal{W}} a_{wd} \leq A_d \qquad d \in \mathcal{D} \tag{2.10}$$

The second-stage constraints

To model the scheduling of emergency patients and cancellations of elective patients, constraints (2.11) to (2.14) are applied. Each constraint and variable belongs exclusively to a scenario $s \in \mathcal{S}$. Each scenario is defined by the stochastic parameters T_{is}^{SU} and U_{wds}^{EM}. T_{is}^{SU} represents the number of SU emergencies from a given patient category that should be scheduled in the cycle, and U_{wds}^{EM} represents the number of emergencies covering beds at the different wards every day. In Section 2.4.3, we describe how the random parameters are generated.

The variables e_{ijkds} and e_{ijkds}^{EL} are the number of emergency patients that belong to patient category i scheduled for subspecialty j in OR k on day d. The first variable represents the number of emergencies scheduled to flexible slots, while the latter is the number of SU emergencies scheduled to the elective slots. Furthermore, the variable x_{ikds}^{C} gives the number of elective cancellations of elective patient category i in OR k on day d. Constraints (2.11) ensure that all SU emergencies are scheduled for surgery during the cycle. Constraints (2.12) restrict the number of emergency patients scheduled for the flexible slots in an OR by the surgery duration of the patients. In constraints (2.13) the number of electives and SU emergencies scheduled for an OR on a given day are restricted by the slot time scheduled for subspecialties able to handle the patients in that OR on that day. Finally, constraints (2.14) state that we cannot cancel patients who are not scheduled.

$$\sum_{j \in \mathcal{J}} \sum_{d \in \mathcal{D}} \left(\sum_{k' \in \mathcal{K}_j \cap \mathcal{K}^F} e_{ijk'ds} + \sum_{k \in \mathcal{K}} e_{ijkds}^{EL} \right) = T_{is}^{SU} \quad i \in \mathcal{I}^{SU}, s \in \mathcal{S} \tag{2.11}$$

$$\sum_{i \in \mathcal{I}_j^{UJ} \cup \mathcal{I}_j^{SUJ}} S_i e_{ijkds} \leq B_{kd} n_{jkd} \quad j \in \mathcal{J}, k \in \mathcal{K}_j \cap \mathcal{K}^F, d \in \mathcal{D}, s \in \mathcal{S} \tag{2.12}$$

$$\sum_{i \in \mathcal{I}_j^{J}} S_i (x_{ikd} - x_{ikds}^{C}) + \sum_{i \in \mathcal{I}_j^{SUJ}} S_i e_{ijkds}^{EL} \leq B_{kd} y_{jkd} \quad j \in \mathcal{J}, k \in \mathcal{K}_j, d \in \mathcal{D}, s \in \mathcal{S} \tag{2.13}$$

$$x_{ikds}^{C} \leq x_{ikd} \quad i \in \mathcal{I}^{EL}, k \in \mathcal{K}, d \in \mathcal{D}, s \in \mathcal{S} \tag{2.14}$$

The second-stage bed constraints are given by (2.15) to (2.17). Variables u_{iwds} and u_{iwds}^{SU} represent the number of electives and SU-emergencies from patient category i that cover beds in ward w on day d. Variable $b_{ww'ds}$ represents the number of beds occupied in ward w' by patients belonging to ward w on day d in scenario s, and the variable β_{wds} represents the number of additional beds staffed in ward w on day d. The two first bed-constraints count the number of elective and SU emergency patients who rest in the different wards each day, while the last ones ensure that the total bed capacity is respected.

$$\sum_{k \in \mathcal{K}} \sum_{d'=1}^{E_{id}} (x_{ik(d-d'+1)} - x^{C}_{ik(d-d'+1)s}) \leq u_{iwds} \quad w \in \mathcal{W}, i \in \mathcal{I}^{W}_{w}, d \in \mathcal{D}, s \in \mathcal{S} \tag{2.15}$$

$$\sum_{j \in \mathcal{J}} \sum_{d'=1}^{E^{SU}_{id}} \left(\sum_{k' \in \mathcal{K}_j \cap \mathcal{K}^F} e_{ijk'(d-d'+1)s} + \sum_{k \in \mathcal{K}} e^{EL}_{ijk(d-d'+1)s} \right) \leq u^{SU}_{iwds} \quad w \in \mathcal{W}, i \in \mathcal{I}^{SUW}_{w}, d \in \mathcal{D}, s \in \mathcal{S} \tag{2.16}$$

$$\sum_{i \in \mathcal{I}^{W}_{w}} u_{iwds} + \sum_{i \in \mathcal{I}^{SUW}_{w}} u^{SU}_{iwds} + \sum_{w' \in \mathcal{W} | w' \neq w} b_{w'wds} - \sum_{w' \in \mathcal{W} | w' \neq w} b_{ww'ds} \leq a_{wd} + \beta_{wds} - U^{EM}_{wds} \quad w \in \mathcal{W}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{2.17}$$

When implementing constraints (2.15) and (2.16), we link the last day of the cycle to the first day of the cycle. In the implementation, we handle non-positive values in the expression $(d - d' + 1)$ by adding the number of days in the cycle. For a weekly cycle (7 days), day 0 maps to day 7, day -1 maps to day 6, and so on. This way of modelling is possible because the MSS is cyclic, and it allows us to only model the days of the cycle, while at the same time making sure that the bed usage of the whole length of stay is accounted for.

The objective function

The objective function is given by (2.18). Parameter R^{EL}_{i} represents the gain obtained by scheduling more patients belonging to elective patient category i than the lower limit, while P_s gives the probability of ending up in scenario s. Furthermore, C^{C}_{i} represents the penalty of cancelling elective patients belonging to category i, while $C^{W}_{ww'}$ yields the penalty of assigning a patient belonging to ward w to ward w'. C^{SU} indicates the penalty of scheduling SU emergencies for elective ORs, and C^{β} is the penalty related to staffing more beds than scheduled. Finally, R^{U} is the growth obtained when scheduling U emergencies to the flexible slots. The objective function maximizes the gains from scheduling more elective patients than the lower limit in the first stage. In the second stage, we minimize the penalty of cancelling electives, providing beds for patients in wards not originally intended for them, scheduling SU patients to elective slots and staffing more beds than scheduled on the wards. In addition, we maximize the amount of urgent emergency patients scheduled for surgery in the second stage.

$$\begin{aligned} max \sum_{i \in \mathcal{I}^{EL}} R^{EL}_{i} v_i - \sum_{s \in \mathcal{S}} P_s \Bigg[& \sum_{i \in \mathcal{I}^{EL}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} C^{C}_{i} x^{C}_{ikds} + \sum_{w \in \mathcal{W}} \sum_{w' \in \mathcal{W}} \sum_{d \in \mathcal{D}} C^{W}_{ww'} b_{ww'ds} + \\ & \sum_{i \in \mathcal{I}^{SU}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} C^{SU} e^{EL}_{ijkds} + \sum_{w \in \mathcal{W}} \sum_{d \in \mathcal{D}} C^{\beta} \beta_{wds} - \sum_{i \in \mathcal{I}^{U}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} R^{U} e^{U}_{ijkds} \Bigg] \end{aligned} \tag{2.18}$$

Table 2.1: The activities, states and events of the system considered in the simulation model

Activities	States	Events
Preop. stay in wards for inpat.	# of inpat. in each ward, preop.	Arrival of em. pat.
Preop. stay at home for outpat.	# of outpat. at home, preop.	
Postop. stay in wards for inpat.	# of pat. in each ward, postop.	Inpat. leaving ward postop.
Transport of em. pat. to OR		
Surgery (incl. cleaning of OR)	Whether an OR is idle or busy	Completion of surgery

2.4.2 The discrete-event simulation model

The discrete-event simulation model encompasses all elective and emergency patient categories, and it performs the scheduling of these patients either to the elective ORs governed by the MSS or to the emergency ORs. The system is modelled as a queuing network where the emergency patients are the customers waiting in line, and the ORs and the wards are the servers. There are four queues of emergency patients: the CU, the U, the SU outpatient and the SU inpatient queue. The elective patients are also treated by the servers, but they are not waiting in line as they are scheduled for specific time slots and arrive just before surgery. Table 2.1 lists the activities, states and events.

Figure 2.6 illustrates the flow of patients in the model. The emergency patients arrive with an exponentially distributed inter-arrival time. The expected inter-arrival time, $1/\lambda_{i^{EM},\tau}$, is dependent both on the emergency patient category, i^{EM}, and on the time, τ, of the day. Both the U, the CU and the SU inpatients rest in the preoperative wards while waiting in line for surgery. The SU outpatients are sent home to wait. If the number of SU emergencies waiting in line for the emergency ORs is above a threshold value, the SU patients are scheduled for surgery in a flexible surgery slot, preferably within the deadline for surgery. If a flexible surgery slot is not available within a given number of days, elective patients are cancelled in order to provide capacity for the SU emergency patients.

For the emergency ORs, scheduling rules are used to determine which patient should enter next. The first patient in line in the U and CU queues, and the first patient in the SU in- and outpatient queues who have not been scheduled for flexible slots are the first to be chosen for surgery. The scheduling rules applied to choose among the candidates may vary depending on the case department. For the elective ORs, the next patient is the next scheduled patient. This may be either an elective patient or a SU emergency patient. The elective patients that are cancelled are rescheduled to a flexible slot some days ahead. All elective patients are assumed to show up at the scheduled surgery time.

The surgery duration is random, with empirical probability distributions for each patient category. The scheduling of emergency patients is based on the expected surgery duration. We can not schedule for overtime, but overtime may occur as a result of the actual surgery duration. Following each surgery, the OR

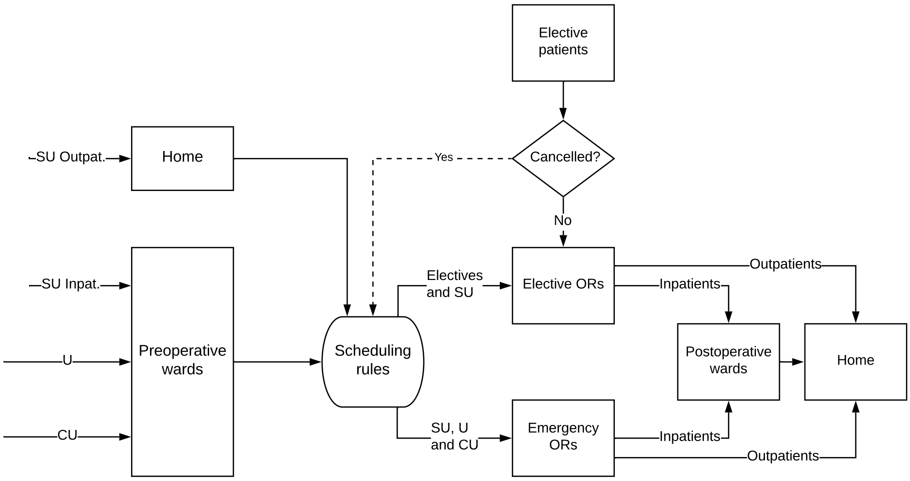


Figure 2.6: A flow chart describing the flow of patients in the simulation model.

must be cleaned, implying that the room is unavailable for some time following surgery.

After surgery, the patients are either sent to the ward or directly home. The preoperative waiting time for emergency patients is calculated as the time from arrival to the system until surgery. The postoperative length of stay is calculated as the time interval from leaving the OR to leaving the postoperative ward. The postoperative length of stay is random, and drawn from empirical probability distributions for each patient category.

Unlike in the optimization model, the wards in the simulation model are assumed to have infinite capacity, implying that no rescheduling is done as a result of the wards being overloaded. The unlimited ward capacity is chosen to allow for scheduling of more beds in periods when many emergencies arrive. This is the normal protocol at many hospitals, as the number of physical beds exceeds the number of staffed beds. Scheduling more beds is not penalized in the simulation model, but rather in the optimization model. All emergency inpatients return to the same ward where they were resting at prior to surgery.

2.4.3 The scenario generation procedure

Each scenario contains data from one cycle in the simulation output. Recall that a cycle corresponds to the number of days considered in the optimization model. In each iteration of the simulation-optimization approach illustrated in Figure 2.4, a number of simulation replications, N^{REP}, are produced, and for each replication, a given number of separated cycles, N^{CYC}, are sampled as the scenarios. In total,

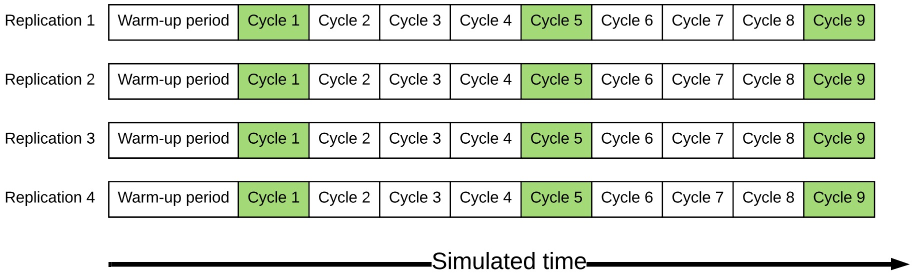


Figure 2.7: An example of how scenarios are chosen in each iteration of the simulation-optimization procedure. Here, the number of simulation replications, $N^{REP} = 4$, and the number of cycles chosen from each replication, $N^{CYC} = 3$, providing a total of 12 scenarios. Note that in each simulation replication there is a warm-up period, and that each cycle is separated by a number of cycles (3 in this example) to ensure reasonable independence between the cycles.

this yields $N^{REP} \cdot N^{CYC}$ scenarios in each iteration. By letting each scenario consist of data gathered from a number of consecutive days from the simulation output, dependency is ensured between the days in each scenario. Furthermore, to keep the scenarios reasonably independent of each other, the selected cycles are separated by a number of cycles. All simulation replications start with an empty system, so to ensure that the scenarios are not affected by this, a warm-up period is implemented in each replication. In Figure, 2.7 an example of the scenario generation procedure is illustrated. Here, $N^{REP} = 4$, and $N^{CYC} = 3$, providing a total of 12 scenarios.

Initially, each scenario contains the number of SU emergency patients that are scheduled for the elective ORs within the cycle and the number of emergency patients resting in each ward every day. Because the SU emergency patients are scheduled in the optimization model, they may be scheduled for other days compared to what the simulation output shows. Because of this, they may end up covering beds on different days as well. Therefore we subtract the postoperative LOS for the SU patients when generating the scenarios. However, we do not subtract the preoperative LOS for the SU patients when generating scenarios. The reason for this is that the optimization model does not explicitly account for the preoperative LOS. Hence, if we had subtracted the preoperative LOS, we would have underestimated the bed loading through the cycle when generating the scenarios.

Equations (2.19) describe how the number of SU emergency patients that require surgery at the elective ORs within the cycle is obtained. SU_{id} is the number of SU emergency patients from patient category i that entered on day d in the cycle and were scheduled for the elective ORs. The set D_s^S represents the days in the drawn cycle that is used to provide scenario s. Equations (2.20) show how we calculate the number of emergency patients resting in the different wards each day of the cycle. E_{wd} state the number of emergency patients resting in ward w on day d, while $E_{wd}^{SU_{post}}$ is the number of SU emergencies that rest in ward w on day d following their surgery.

$$T_{is}^{SU} = \sum_{d \in \mathcal{D}_s^S} SU_{id} \qquad i \in \mathcal{I}^{SU}, \quad s \in \mathcal{S} \tag{2.19}$$

$$U_{wds}^{EM} = E_{wd} - E_{wd}^{SU_{post}} \qquad w \in \mathcal{W}, \quad s \in \mathcal{S}, \quad d \in \mathcal{D}_s^S \tag{2.20}$$

2.4.4 Stopping criterion

The levels of detail in the optimization model and the simulation model differ in their construction. This presents a challenge in terms of the convergence and stability of the solution, as the difference in representation typically leads to different objective values, even for identical MSS. While we do not prove that the model presented is guaranteed to converge, in practice the solution stabilizes after a few iterations of the algorithm illustrated in Figure 2.4.

For the test cases and case study in this paper, we have chosen to use the number of flexible slots scheduled through the cycle as the stopping criterion. At a tactical decision level, determining the share of flexible slots is of high clinical interest for planners at the hospital. If additional stability is required by the user, adding inter-day stabilization would probably be the most useful requirement, but would take more computation time to reach a stable solution.

2.5 Computational study

In the computational study, the optimization model is first run for a set of test instances, providing results that can give both technical and managerial insights. Then we perform a case study, where we apply the simulation-optimization approach on a case department to develop an MSS that handles a fluctuating demand of emergency patients. In all instances, a cycle is set to be one week.

2.5.1 Implementation and setup of study

The optimization model is implemented in the Mosel language and is solved in Xpress 8.3. The simulation model is built in MATLAB. In each simulation-optimization iteration, 20 simulation replications are made ($N^{REP} = 20$), each representing a period of half a year (after half a year warm-up). From each simulation replication, five scenarios are selected ($N^{CYC} = 5$) yielding a total of 100 scenarios. All scenarios are replaced from one iteration to the next.

To ensure independence between the scenarios, we would choose $N^{REP} = 100$ and $N^{CYC} = 1$. However, because of the relatively long warm-up period we decided to use several cycles from each simulation replication to save computational time. The drawn cycles are separated by four weeks to ensure a reasonably degree of independency.

For the test instances, three levels of emergency patient loading are implemented: low, medium and high (EL, EM and EH, respectively). For each of the emergency loading cases, two sets of target throughput of electives - low and high (TL and TH, respectively) - are applied, and for each of these targets we test three bed capacities, resulting in 18 different instances. Because the number of patients is so variable in the instances, we apply four different bed capacities (W1, W2, W3 and W4) to provide three levels of bed capacity for each of the three emergency loading cases. Applying the lowest bed capacity (W1) to the high- and medium emergency (EH and EM) loading cases will yield an unrealistically low bed capacity, while the complete opposite will yield a very large bed capacity.

Note that for the experiments on these test instances, the focus is on the optimization model. The simulation model is only run once for each of the emergency loading levels to generate scenarios for the optimization model. In Appendix A.2 we provide the values assigned to the input parameters of the optimization model in the test instances, and for the case department instance.

2.5.2 The value of the stochastic solution

We will use the value of the stochastic solution (VSS), as described by Birge (1982), to evaluate our model. The VSS measures the value of applying a stochastic, rather than a deterministic, model to solve the problem at hand. It can be interpreted as valuing the flexibility that the stochastic solution provides that is not present in the deterministic solution. The VSS is calculated by first constructing the mean value problem (MVP). In the MVP, the stochastic parameters take their expected value, and the stochastic model is solved deterministically. The values obtained for the first-stage variables when solving the MVP are applied as input parameters for the stochastic model, and we solve the second-stage problems (one problem for each scenario) of the stochastic model. This leaves us with an optimal objective function value referred to as the mean value solution (MVS). The VSS is calculated as the difference between the stochastic solution (SS) obtained from solving the stochastic model (equations (2.1)-(2.18)) and the MVS.

Neither the MVPs nor the stochastic problems are solved to optimality within three hours, so the true VSS may not be calculated. However, the LP gaps are small, and we can provide quite narrow intervals for the VSS. The upper limit is calculated as the difference between the objective value of the upper bound obtained from solving the stochastic model and the MVS. The lower limit is calculated as the difference between the objective value of the IP-solution from solving the stochastic model and the MVS.

When solving the MVP, the values of the first-stage variables (that are fed to the MVS) are the same regardless of the costs associated with the random parameters. Therefore, by increasing these costs, the VSS increases. Furthermore, for this specific problem, the number of flexible slots increases as the costs associated with the random parameters increase. Scheduling flexible slots can be regarded as investing in an insurance against periods of high emergency loading.

Table 2.2: The number of flexible slots scheduled in the SS and the MVS when the cost of cancelling an elective patient is the same as the revenue from scheduling the same patient.

Instance	Flexible slots (SS)	Flexible slots (MVS)
EL-TL-W1	9	9
EL-TL-W2	9	9
EL-TL-W3	9	9
EL-TH-W1	2	2
EL-TH-W2	1	3
EL-TH-W3	0	3
EM-TL-W2	9	9
EM-TL-W3	9	9
EM-TL-W4	9	9
EM-TH-W2	2	5
EM-TH-W3	3	4
EM-TH-W4	2	5
EH-TL-W2	10	10
EH-TL-W3	9	9
EH-TL-W4	10	9
EH-TH-W2	5	8
EH-TH-W3	8	9
EH-TH-W4	7	9

In Table 2.2 we present the number of flexible slots scheduled in both the SS and the MVS, when the cost of cancelling a patient is set to be the same as the revenue from scheduling the patient. As we see, the SSs provide fewer flexible slots than the MVS. This is counterintuitive. The reason for this is that the stochastic model also accommodates scenarios where few emergency patients arrive; the need for flexible slots in these scenarios is less, and the capacity can be utilized to treat more elective patients if they are scheduled in the first stage. As a consequence, more elective patients are scheduled and more electives receive surgery (despite more cancellations in general) in the stochastic solutions.

Table 2.3 includes results from solving both the stochastic model and its deterministic counterpart, when the cost of cancelling elective patients is set to be higher than the revenue from scheduling the same patients. In addition to showing the SS, the MVS and the VSS, the table includes the number of flexible slots scheduled, the mean number of elective cancellations and the mean number of electives treated in solving both the stochastic and the deterministic model for all the instances.

Comparing the instances, the VSS is often higher when the bed capacity is scarce. This has to do with the fact that we penalize rescheduling of beds in the second stage. Proper scheduling of beds in the first stage can decrease the number of elective patients resting in wards not intended for them in the second stage. The benefits of this scheduling are typically higher when the bed capacity is scarce.

Table 2.3: The value of the stochastic solution. For both the stochastic solution (SS) and the mean value solution (MVS) we provide the number of flexible slots scheduled (Flex.), the mean number of cancellations (El. canc.) and the mean number of electives treated (El. treated). Because neither the MVPs or the stochastic problems were solved to optimality, intervals are given for the SS and the VSS.

	# Flex.		# El. canc.		# El. treated		Objective function value		
Instance	SS	MVS	SS	MVS	SS	MVS	SS	MVS	VSS
EL-TL-W1	14	9	2.71	6.20	80.29	84.80	[-9.63, -9.04]	-25.74	[16.11, 16.70]
EL-TL-W2	9	9	0.20	1.55	90.80	89.45	[113.43, 114.96]	100.05	[13.38, 14.91]
EL-TL-W3	9	9	0	0	91.00	91.00	[115.51, 117.14]	115.26	[0.25, 1.88]
EL-TH-W1	8	2	3.29	9.03	93.71	100.97	[-21.35, -19.08]	-46.40	[25.05, 27.32]
EL-TH-W2	4	3	1.22	2.74	106.78	106.26	[107.90, 113.64]	96.24	[11.66, 17.40]
EL-TH-W3	3	3	0.82	0.80	108.18	108.20	[115.26, 120.30]	115.22	[0.04, 5.08]
EM-TL-W2	12	9	3.58	5.98	84.42	85.02	[-278.02, -276.04]	-296.47	[18.45, 20.43]
EM-TL-W3	9	9	0.21	0.57	90.79	90.43	[-19.25, -15.87]	-22.99	[3.74, 7.12]
EM-TL-W4	10	9	0.25	0.27	89.75	90.73	[108.62, 114.22]	109.99	[0, 4.23]
EM-TH-W2	7	5	4.39	7.68	95.61	98.32	[-297.10, -289.21]	-318.15	[21.05, 28.94]
EM-TH-W3	7	4	1.25	2.75	99.75	104.25	[-36.28, -27.83]	-42.76	[6.48, 14.93]
EM-TH-W4	7	5	1.23	2.20	99.77	103.80	[93.76, 102.45]	92.02	[1.74, 10.43]
EH-TL-W2	16	10	6.01	11.86	70.99	78.14	[-6732.91, -6731.50]	-6760.24	[27.33, 28.74]
EH-TL-W3	9	9	1.03	2.91	88.97	88.09	[-1764.45, -1757.14]	-1759.48	[0, 2.34]
EH-TL-W4	11	9	0.33	2.85	87.67	88.15	[-700.23, -691.45]	-719.14	[18.91, 27.69]
EH-TH-W2	12	8	6.65	12.85	79.35	85.15	[-6758.48, -6753.5]	-6786.54	[28.06, 33.04]
EH-TH-W3	11	9	1.20	4.83	89.80	91.17	[-1785.64, -1779.42]	-1813.86	[28.22, 34.44]
EH-TH-W4	11	9	0.62	2.32	90.38	93.68	[-719.23, -712.14]	-728.99	[9.76, 16.85]

2.5.3 Managerial insight

We have destilled some insights from the test instances. Table 2.4 provides the results from running the 18 instances for two levels of cost related to cancelling elective patients. For the low cost cases, the cost of cancelling an elective patient is the same as the revenue from scheduling the same patient. The table includes the number of flexible slots scheduled, the number of electives scheduled for surgery, the mean share of SU surgeries performed in flexible slots, the mean number of electives cancelled, the mean number of electives treated and the mean number of elective inpatients resting in inconvenient wards.

Figures 2.8 and 2.9 illustrate how the number of flexible slots scheduled in the first stage depends on the available OR capacity and the bed capacity, respectively.

Table 2.4: Output from solving the stochastic problem for the 18 instances. Low (L) indicates that the cost of cancelling an elective patient is the same as the revenue from scheduling the same patient, while high (H) represents a higher cost of cancelling electives. We include the number of flexible slots (Flex.) and the number of elective patients scheduled (El. sched.) in the first stage. For the second stage we include the mean number of SU patients scheduled for flexible slots (SU pat. in flex.), the mean number of elective cancellations (El. canc.), the mean number of electives treated (El. treated) and the mean number of patients resting in wards not intended for them (Pat. moved).

	# Flex.		# El. sched.		# SU pat. in flex.		# El. canc.		# El. treated		Pat. moved	
Instance	L	H	L	H	L	H	L	H	L	H	L	H
EL-TL-W1	9	14	91/91	83/91	2.36/2.40	2.40/2.40	5.08	2.71	85.92	80.29	5.76	4.17
EL-TL-W2	9	9	91/91	91/91	2.40/2.40	2.40/2.40	0.27	0.20	90.73	90.80	1.41	1.18
EL-TL-W3	9	9	91/91	91/91	2.40/2.40	2.40/2.40	0	0	91.00	91.00	0	0
EL-TH-W1	2	8	111/115	97/115	1.20/2.40	2.31/2.40	9.89	9.03	101.11	93.71	6.31	4.20
EL-TH-W2	1	4	112/115	108/115	0.60/2.40	1.90/2.40	3.18	2.74	108.82	106.78	2.73	1.93
EL-TH-W3	0	3	114/115	109/115	0/2.40	1.65/2.40	2.62	0.80	111.38	108.18	0.05	0.01
EM-TL-W2	9	12	91/91	88/91	4.80/5.24	5.20/5.24	4.52	3.58	86.48	84.42	6.05	5.42
EM-TL-W3	9	9	91/91	91/91	4.83/5.24	4.78/5.24	0.36	0.21	90.64	90.79	0.49	0.52
EM-TL-W4	9	10	91/91	90/91	4.82/5.24	4.99/5.24	0.47	0.25	90.53	89.75	0.21	0.20
EM-TH-W2	2	7	111/115	101/115	1.83/5.24	4.34/5.24	9.32	4.39	101.68	95.61	10.15	5.24
EM-TH-W3	3	7	109/115	101/115	2.40/5.24	4.37/5.24	3.92	1.25	105.08	99.75	0.43	0.44
EM-TH-W4	2	7	111/115	101/115	1.82/5.24	4.35/5.24	4.68	1.23	106.32	99.77	0.20	0.29
EH-TL-W2	10	16	90/91	77/91	8.86/9.88	9.88/9.88	11.65	6.01	78.35	70.99	9.37	7.39
EH-TL-W3	9	9	91/91	90/91	8.47/9.88	8.50/9.88	1.48	1.03	89.52	88.97	2.55	2.22
EH-TL-W4	10	11	89/91	88/91	9.02/9.88	9.42/9.88	1.01	0.33	87.99	87.67	0.57	0.62
EH-TH-W2	5	12	104/115	85/115	5.15/9.88	9.55/9.88	16.05	6.65	87.95	79.35	9.29	7.17
EH-TH-W3	8	11	97/115	101/115	7.60/9.88	9.41/9.88	3.63	1.20	93.37	89.80	2.12	2.78
EH-TH-W4	7	11	101/115	91/115	7.12/9.88	9.38/9.88	3.98	0.62	97.02	90.38	0.70	0.59

The number of flexible slots increases when the OR capacity is high - in other words, when the target level of elective patients is low. The reason for this is that more OR capacity can be made flexible without having to sacrifice the scheduling of elective patients, and that we assume there is always an U patient who will fill the idle flexible OR capacity.

When the cost of cancelling elective patients is high, the number of flexible slots decreases as the bed capacity increases. There are two reasons for this. First, the flexible slots may be utilized for either SU outpatients or U patients if bed capacity is scarce, while the SU inpatients may be moved to elective slots where they receive surgery without exceeding the bed capacity, providing a kind of option towards scarce bed capacity. Secondly, scheduling flexible slots reduces the number of electives scheduled, implying less demand for beds.

We do not find the same pattern for the low cost instances. In Table 2.4,

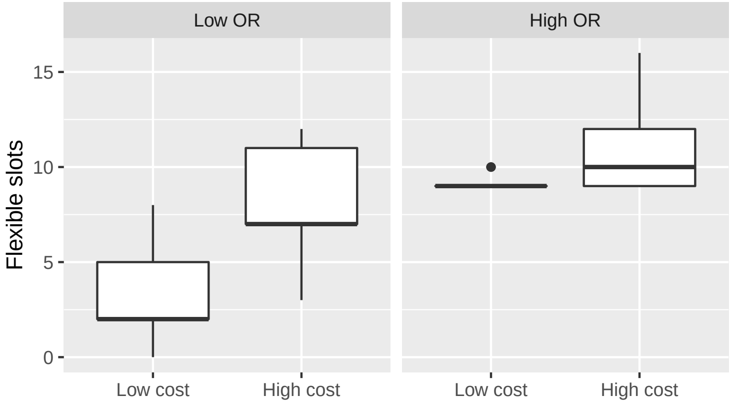


Figure 2.8: How the number of flexible slots scheduled in the first stage depends on the OR capacity when the cost of cancelling elective patients is either low or high.

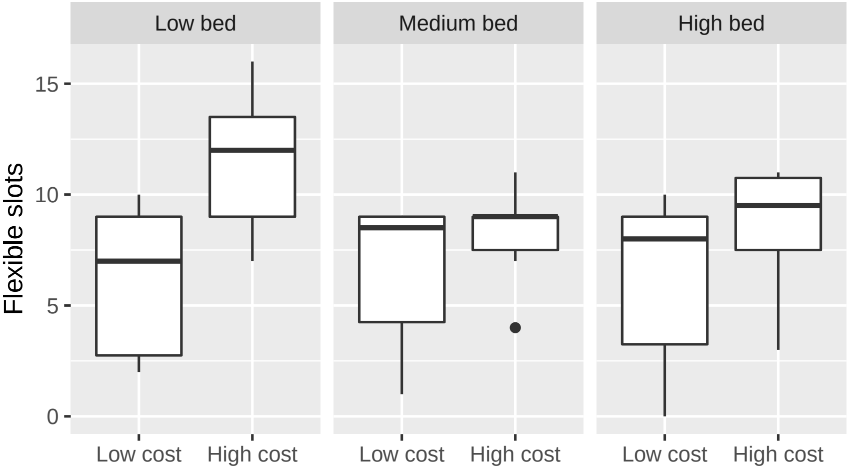


Figure 2.9: How the number of flexible slots scheduled in the first stage depends on the bed capacity when the cost of cancelling elective patients is either low or high.

we observe that going from instance EM-TH-W2 to EM-TH-W3 and from EH-TH-W2 to EH-TH-W3, the number of scheduled flexible slots increases. In the instances EM-TH-W2 and EH-TH-W2, bed capacity is very scarce, resulting in many elective cancellations. Cancelling elective inpatients because of the scarce bed capacity provides ample spare elective OR capacity. This idle OR capacity may be used to schedule SU patients (if they are less demanding in terms of beds than the elective patients that were cancelled), decreasing the need for flexible slots. This happens because the penalty for cancelling elective patients is very low in these instances. For three of the low-target instances, the number of flexible slots increases as we go from the medium to the high bed capacities. This results from the high penalty associated with staffing more beds than scheduled to handle the peaks of emergency bed loading. The chances for avoiding additional staffing are greater for the instances with high bed capacity, and may be avoided if we schedule fewer electives in the first stage. Decreasing the penalty of staffing more beds than scheduled yields solutions that are more aggressive in terms of elective bed loading for the instances with high bed capacity.

Figure 2.10 illustrates the mean share of flexible slots scheduled for the three levels of emergency loading. As expected, the number of scheduled flexible slots increases as emergency loading increases. For the low emergency loading instances, the share of emergency patients is 2.6% for the TL instances and 2.0% for the TH instances. These numbers are 5.4% and 4.4%, and 9.8% and 7.9% for the medium and high emergency loading instances, respectively. For the TH-low cost instances, the mean share of flexible slots is quite similar to the share of emergency patients, implying that the emergency patients receive OR capacity according to the relative size of the group. For the TL-low cost instances, there is less competition for the OR capacity, and the emergency patients receive excessive OR capacity. For the TH-high cost instances, the emergency patients receive a much higher share of the OR capacity compared to the TH-low cost instances, indicating that they gain power in the competition for OR capacity. For the TL instances, the competition from electives is less, and therefore the difference is less between the two cost regimes. In general, if the penalty for cancelling a patient is at least the same as the revenue from scheduling the patient, the share of flexible slots scheduled should be higher than the share of emergency patients.

2.5.4 Case study

For the case study we consider the Master Surgery Scheduling Problem (MSSP) in the orthopaedic department at St. Olav's Hospital. St. Olav's Hospital is a university hospital located in Norway, and it is a relatively large hospital with approximately 1000 beds. The orthopaedic department performs approximately 7000 surgeries every year, and roughly 3000 of these are emergency surgeries. To perform these surgeries, the department has access to 7 elective and 3 emergency ORs. The case department faces many of the issues discussed in this paper.

In this section we apply the simulation-optimization approach to the case department. The aim is to provide an MSS that enables the department to handle

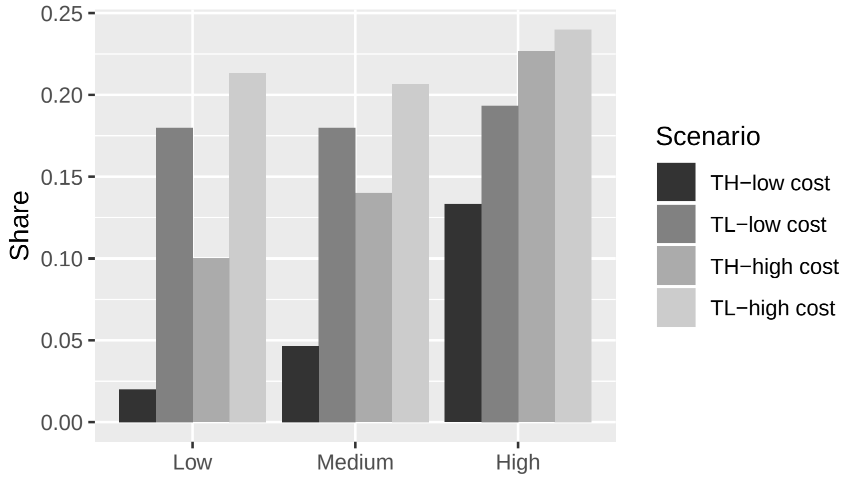


Figure 2.10: How the number of flexible slots scheduled in the first stage depends on the emergency loading when the cost of cancelling elective patients is either low or high.

fluctuating emergency patient demand and at the same time provide a sufficient throughput of elective patients. First, the scheduling rules applied in the simulation model to mimic the case department is presented. Secondly, the new MSS is introduced, and finally we compare outcomes from running both the new MSS and the MSS present at the case department today in the simulation model.

Scheduling rules and the flow of patients in the simulation model

CU patients will always have priority over the U and SU patients. U patients have priority over SU patients except when the next SU patient has exceeded the time limit more than the next U patient. Only CU patients can have surgery during the night, and only the U and CU patients are admitted to surgery on weekends.

All U and CU patients may be summoned for surgery immediately after arrival, while SU patients are scheduled in the morning after arrival. The SU inpatients are primarily lined up for the emergency ORs. If the number of SU inpatients waiting in the queue is above a threshold limit - in this case two - we start scheduling these patients for the flexible slots in the elective ORs. The SU outpatients have to return to the hospital on the day of surgery, so we want to make sure that they do not have their surgery postponed since this would involve too much traveling for the patients. Therefore, the SU outpatients are not lined uo for the emergency ORs, but are scheduled directly into the flexible slots. If no flexible slots are available within a given number of days - in this case eight - elective surgeries must be cancelled to provide capacity for the SU in- and outpatients.

All displaced elective patients will be rescheduled to a flexible slot some days ahead. Rescheduling a patient just before surgery is not preferable, so we introduce a limit on the number of days prior to surgery that rescheduling is not allowed.

Iterative outcomes and the optimization-based MSS

To create an MSS for the case department, we populate the optimization model with the data described in Appendix A.2. On request from the case department, no flexible slots may be scheduled for two of the elective ORs (OR-6 and OR-7), and no elective inpatients should be scheduled on Friday. To initialize the simulation-optimization approach, we apply the MSS present at the orthopaedic department today when running the simulation model in the first iteration. Because no flexible slots are available in the present MSS, all emergency patients are sent to the emergency ORs in the first iteration. However, if the SU emergencies do not receive surgery within a set time, they are rescheduled to the elective ORs, and there will be elective cancellations.

Tables 2.5 and 2.6 present the main outcomes from the optimization model and the simulation model, respectively. The optimality gap in Table 2.5 is calculated as the difference between the objective function of the best integer solution and the upper bound, divided by the upper bound. The MSS generated in the last iteration is illustrated in A.4.

Table 2.5: Outcomes from each iteration with the optimization model when generating the simulation-optimization-based MSS

	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6	Iter. 7	Iter. 8
Obj. func. val.	-2566.701	-75.346	-140.293	-241.623	-44.047	-33.668	-1137.51	58.55
Upper bound	-2563.714	-71.017	-134.99	-230.729	-40.951	-29.512	-1132.74	61.21
Optimality gap	0.12%	5.75%	3.77%	4.51%	7.03%	12.34%	0.42%	4.35%
Flex. slots	12	12	11	14	12	12	12	12
El sched.	71/80	74/80	74/80	70/80	74/80	74/80	74/80	74/80
SU in flex. (avg.)	4.41/4.53	6.47/6.70	6.29/6.58	6.66/6.67	7.26/ 7.60	5.79/6.04	5.91/6.01	7.04/7.16
El. canc. (avg.)	6.95	2.35	3.63	1.94	2.02	2.18	3.45	1.58
U surg. (avg.)	11.22	8.03	6.48	7.94	7.57	9.53	8.76	8.24

Table 2.6: Outcomes from each iteration with the simulation model when generating the simulation-optimization-based MSS

	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 6	Iter. 7	Iter. 8
Canc. per week	4.16	0	0	0	0	0	0	0
Avg. WT (CU)	3.9 h	3.7 h	3.5 h	3.8 h	3.7 h	3.8 h	3.8 h	3.7 h
Avg. WT (U)	31.3 h	20.9 h	20.5 h	20.7 h	18.0 h	20.7 h	21.0 h	18.8 h
Avg. WT (SU)	3.81 d	1.80 d	1.80 d	1.83 d	1.73 d	1.86 d	1.87 h	1.78 d

The number of SU emergencies treated in flexible slots and the number of U emergencies treated reveal information about the flexible slot capacity. Almost all SU emergencies are treated in the flexible slots, and there is ample flexible capacity to treat U emergencies when no more SU emergencies are present. This indicates that the flexible slot capacity is good, and that most of the elective cancellations are due to the shortage of beds.

The simulated results include both the number of elective surgeries that are cancelled each week and the waiting time to receive surgery for the emergency patients. Scheduling flexible slots dramatically decreases the waiting time for both U and SU patients from iteration one to the subsequent iterations. Note that since the bed capacity is treated as unlimited in the simulation model, no electives are cancelled due to the shortage of beds. After the first iteration no electives are cancelled, indicating that the flexible OR capacity is sufficient.

Analysing the performance of the optimization-based MSS

Figure 2.11 illustrates the waiting time to receive surgery for U and SU emergencies when running both the MSS present in the case department today and the optimized MSS 20 times in the simulation model. The optimization-based MSS performs better on average, and the waiting times are less affected by fluctuations in emergency demand for surgery. Figure 2.12 illustrates the queue of SU emergencies at 08.00 through the simulated period when running both the MSS present in the case department today and the optimized MSS once in the simulation model. Note that the minimum number of SU patients waiting in queue for surgery in the MSS present today is almost as large as the maximum number of SU patients waiting for surgery in the optimized MSS. Figure 2.13 shows the number of SU patients treated in both the flexible and elective slots each week from running the two MSSs. For the MSS present today, all SU emergencies are treated in elective slots (as no flexible slots are available), while for the optimized MSS all SU emergencies are treated in flexible slots. As a consequence, there are no elective interruptions when applying the new MSS.

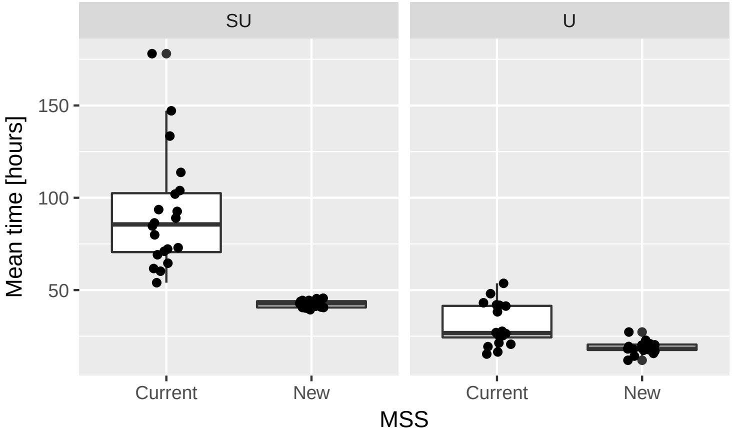


Figure 2.11: The simulation mean waiting time in hours for SU (left) and U (right) emergencies for the current and new optimization-based MSS, results from 20 simulations.

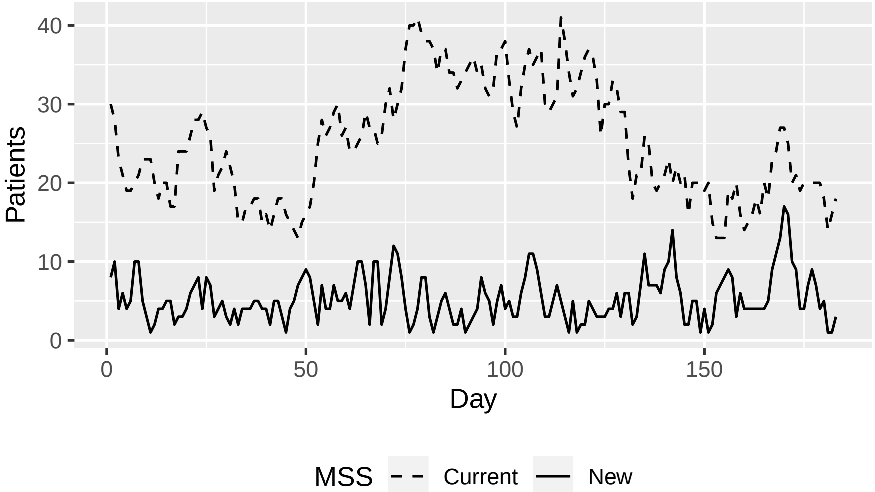


Figure 2.12: The graph illustrates the queue of SU emergencies as 08.00 through the simulated period for both the optimization-based MSS and the MSS present at the case department today.

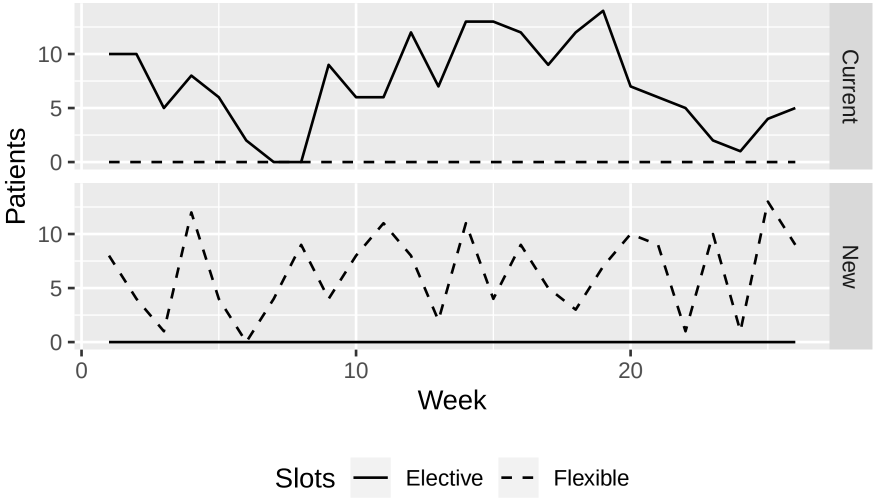


Figure 2.13: The two graphs illustrate the weekly number of SU emergencies treated in flexible and elective slots. The upper graph illustrates the MSS present today, and the lower one illustrates the optimization-based MSS.

2.6 Conclusion

The main purpose of this paper is to present a simulation-optimization approach for developing an MSS and to provide tactical decision support for the management in a department with both elective and emergency patients. Flexible slots are scheduled to the elective ORs to handle the fluctuating demand for emergency surgeries.

A two-stage stochastic optimization model is presented, where uncertainty related to emergency arrivals is included. Furthermore, a discrete-event simulation model is developed to include aspects of uncertainty related to the length of stay of patients following surgery and the surgery duration. The simulation model allows us to evaluate the MSS produced by the optimization model. Also, the simulation model provides scenarios for the optimization model, allowing the model to adapt to different MSSs and scheduling regimes.

The stochastic model outperforms its deterministic counterpart in terms of the Value of Stochastic Solution (VSS), and for realistic cancellation costs the number of flexible slots is higher in the stochastic solution. Furthermore, if the OR capacity is sufficient, or the ward capacity is scarce, a relatively large share of the ORs should be scheduled as flexible.

The simulation-optimization approach is applied to an orthopaedic department at a Norwegian hospital that treats both elective and emergency patients. With the optimized MSS, the emergency waiting time for surgery decreases, and it proves to be able to handle fluctuating emergency surgery demand with less interruptions to the flow of elective patients.

2.7 Acknowledgements

We would like to thank the management in the Orthopaedic Department at St. Olav's Hospital, and in particular Head of Department, Vigleik Jessen, MD for good collaboration throughout the project. We are grateful for the comments and suggestions made by three anonymous reviewers, which helped to improve the paper.

Bibliography

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.

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, 2008.

J. R. Birge. The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical Programming*, 24(1):314–325, 1982.

P. Cappanera, F. Visintin, and C. Banditori. Comparing resource balancing criteria in master surgical scheduling: A combined optimisation-simulation approach. *International Journal of Production Economics*, 158:179 – 196, 2014.

B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201 (3):921 – 932, 2010.

J.T. Van Essen, E.W. Hans, J.L. Hurink, and A. Oversberg. Minimizing the waiting time for emergency surgery. *Operations Research for Health Care*, 1(2): 34 – 44, 2012.

Y. B. Ferrand, M. J. Magazine, and U. S. Rao. Managing operating room efficiency and responsiveness for emergency and elective surgeries—a literature survey. *IIE Transactions on Healthcare Systems Engineering*, 4(1):49–64, 2014.

G. Figueira and B. Almada-Lobo. Hybrid simulation–optimization methods: A taxonomy and discussion. *Simulation Modelling Practice and Theory*, 46:118 – 134, 2014. Simulation-Optimization of Complex Systems: Methods and Applications.

N. Freeman, M. Zhao, and S. Melouk. An iterative approach for case mix planning under uncertainty. *Omega*, 76:160 – 173, 2018.

A. Fügener, E. W. Hans, R. Kolisch, N. Kortbeek, and P. T. Vanberkel. Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, 239(1):227 – 236, 2014.

E. W. Hans, M. Van Houdenhoven, and P. J. H. Hulshof. *A Framework for Healthcare Planning and Control*, pages 303–320. Springer US, Boston, MA, 2012.

J. L. Higle. *Stochastic Programming: Optimization When Uncertainty Matters*, chapter Chapter 2, pages 30–53. 2005.

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

P. Kall and S. W. Wallace. *Stochastic Programming.* John Wiley Sons, 1994.

L. Koppka, L. Wiesche, M. Schacht, and B. Werners. Optimal distribution of operating hours over operating rooms using probabilities. *European Journal of Operational Research*, 267(3):1156 – 1171, 2018.

A. Kumar, A. M. Costa, . Fackrell, and P. G. Taylor. A sequential stochastic mixed integer programming model for tactical master surgery scheduling. *European Journal of Operational Research*, 270(2):734 – 746, 2018.

M. Lamiri, X. Xie, A. Dolgui, and F. Grimaud. A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research*, 185(3):1026 – 1037, 2008.

X. Li, N. Rafaliya, M. F. Baki, and B. A. Chaouch. Scheduling elective surgeries: the tradeoff among bed capacity, waiting patients and operating room utilization using goal programming. *Health Care Management Science*, 20(1):33–54, 2017.

G. Ma and E. Demeulemeester. A multilevel integrative approach to hospital case mix and capacity planning. *Computers Operations Research*, 40(9):2198 – 2207, 2013. Operations research for health care delivery.

C. Mannino, E. J. Nilssen, and T. E. Nordlander. A pattern based, robust approach to cyclic master surgery scheduling. *Journal of Scheduling*, 15(5):553–563, 2012.

M. J. 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.

M. L. Penn, C. N. Potts, and P. R. Harper. Multiple criteria mixed-integer programming for incorporating multiple factors into the development of master operating theatre timetables. *European Journal of Operational Research*, 262 (1):194 – 206, 2017.

J. Razmi, M. Barati, M. S. Yousefi, and J. Heydari. A stochastic model for operating room planning under uncertainty and equipment capacity constraints. *Journal of Industrial Engineering International*, 11(2):269–279, Jun 2015.

A. Testi and E. Tànfani. Tactical and operational decisions for operating room planning: Efficiency and welfare implications. *Health Care Management Science*, 12(4):363, 2008.

A. Testi, E. Tanfani, and G. Torre. A three-phase approach for operating theatre schedules. *Health Care Management Science*, 10(2):163–172, 2007.

Chapter A

Appendices

A.1 The Mathematical Model

In Tables A.1, A.2, and A.3 all the notation used in the mathematical formulation is presented.

Table A.1: Sets used in the mathematical formulation

Set	Description	Indices
$\mathcal{D}$	Days in a cycle	$d \in \mathcal{D}$
$\mathcal{I}$	Patient categories	$i \in \mathcal{I}$
$\mathcal{J}$	Surgical subspecialties	$j \in \mathcal{J}$
$\mathcal{K}$	ORs	$k \in \mathcal{K}$
$\mathcal{K}^F$	ORs that are available for scheduling of flexible slots	$k \in \mathcal{K}^F \subseteq \mathcal{K}$
$\mathcal{W}$	Wards	$w \in \mathcal{W}$
$\mathcal{S}$	Scenarios	$s \in \mathcal{S}$
$\mathcal{I}^{EL}$	El. patient categories	$i \in \mathcal{I}^{EL} \subseteq \mathcal{I}$
$\mathcal{I}^{IN}$	El. patient categories that are inpatients	$i \in \mathcal{I}^{IN} \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^J_j$	El. patient categories that can be treated by subspecialty j	$i \in \mathcal{I}^J_j \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^K_k$	El. patient categories that can be scheduled to OR k	$i \in \mathcal{I}^K_k \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^W_w$	El. patient categories meant for ward w	$i \in \mathcal{I}^W_w \subseteq \mathcal{I}^{IN}$
$\mathcal{I}^{EM}$	Em. patient categories	$i \in \mathcal{I}^{EM} \subseteq \mathcal{I}$
$\mathcal{I}^{SU}$	SU patient categories	$i \in \mathcal{I}^{SU} \subseteq \mathcal{I}^{EM}$
$\mathcal{I}^U$	U patient categories	$i \in \mathcal{I}^U \subseteq \mathcal{I}^{EM}$
$\mathcal{I}^{UJ}_j$	U patient categories that can be treated by subspecialty j	$i \in \mathcal{I}^{UJ}_j \subseteq \mathcal{I}^U$
$\mathcal{I}^{SUIN}$	SU patient categories that are inpatients	$i \in \mathcal{I}^{SUIN} \subseteq \mathcal{I}^{SU}$
$\mathcal{I}^{SUJ}_j$	SU patient categories that can be treated by subspecialty j	$i \in \mathcal{I}^{SUJ}_j \subseteq \mathcal{I}^{SU}$
$\mathcal{I}^{SUW}_w$	SU patient categories meant for ward w	$i \in \mathcal{I}^{SUW}_w \subseteq \mathcal{I}^{SUIN}$
$\mathcal{K}_j$	ORs that can be managed by surgeons with subspecialty j	$k \in \mathcal{K}_j \subseteq \mathcal{K}$

Table A.2: Parameters used in the mathematical formulation

Parameter	Description
A_w^{MAX}	Max. number of beds available in ward w
A_d	Number of staffed beds available on day d
B_{kd}	Time available for surgery in one slot in OR k at day d
C_i^C	Penalty for cancelling an elective patient of category i
C^{SU}	Penalty for scheduling SU patients to elective slots
$C_{ww'}^W$	Penalty for putting a patient belonging to ward w in ward w'
C^β	Penalty for staffing more beds than scheduled
E_{id}	Expected LOS for elective patient category i scheduled for surgery on day d
E_{id}^{SU}	Expected LOS for SU patients of category i scheduled on day d
M_{kd}^{OR}	Max. number of slots that can be assigned in OR k on day d
M_d^A	Max. number of ORs that can be covered by anaesthesia staff on day d
M^{CYCLE}	Total number of slots available through one cycle
N_j	Max. surgeon capacity (slots) of subspecialty j in one cycle
N_{jd}^D	Max. surgeon capacity (slots) of subspecialty j at day d
R_i^{EL}	Reward for scheduling more patients of elective category i than the lower limit
P_s	Probability of ending up in scenario s
R^U	Reward for scheduling a U patient to a flexible slot
S_i	Expected surgery duration of patient category i
T_i	Target throughput of elective patient category i
T_{is}^{SU}	Excess demand of SU patients of category i
U_{wds}^{EM}	Number of emergency patients resting in ward w on day d in scenario s
V_i	Min. share of patient category i that should be scheduled for surgery

Table A.3: Variables used in the mathematical formulation

Variable	Description
a_{wd}	# of staffed beds in ward w on day d
n_{jkd}	# of slots scheduled as flexible for subspecialty j in OR k on day d
v_i	# of elective patients of category i scheduled above the lower limit
x_{ikd}	# of elective patients of category i scheduled to an elective slot in OR k on day d
y_{jkd}	# of elective slots scheduled for subspecialty j in OR k on day d
α_{kd}^A	Indicates whether OR k is covered by anaesthesia staff on day d or not
$b_{ww'ds}$	# of beds occupied in ward w' by patients belonging to ward w on day d and scenario s
e_{ijkds}	# of SU pat. of cat. i scheduled to subspecialty j in a flex. slot in OR k on day d in scenario s
e_{ijkds}^{EL}	# of SU pat. of cat. i scheduled to subspecialty j in an elective slot in OR k on day d in scenario s
u_{iwds}	# of elective patients of category i resting in ward w on day d in scenario s
u_{iwds}^{SU}	# of SU patients of category i resting in ward w on day d in scenario s
x_{ikds}^C	# of elective patients of category i that are cancelled in OR k on day d
β_{wds}	# of additional beds staffed in ward w on day d in scenario s

A.2 Input parameters for the optimization model

Here, we provide the values assigned to the input parameters in the optimization model for both the test instances used in Sections 2.5.2 and 2.5.3, and the large instance uses in Section 2.5.4. In Tables A.4 to A.6, and Tables A.10 and A.11, we present the values applied for the input parameters in the large instance. The number of ORs and wards in the large instance are similar to the case department, and the elective patient categories represent the main patient categories treated at the department. In Tables A.7 to A.9, and Tables A.10 and A.11, we provide the values applied for the test instances.

Table A.4: Values obtained for the subspecialties and the patient categories in the optimization model for the big instance. T_i is the target number of elective patient category i to be scheduled in each cycle, S_i is the expected surgery duration of patient category i , E_{id} is the expected length of stay of patient category i that receive surgery on day d, R_i^{EL} is the gain for scheduling more patients from patient category i than the lower limit, C_i^C is the penalty for cancelling an elective patient of category i, N_{jd}^D is the maximum number of slots available to subspecialty j on day d, and N_j is the maximum number of slots available to subspecialty j in one cycle.

Subspecialty	T_i	S_i	E_{id}	R_i^{EL}	C_i^C	N_{jd}^D	N_j
Elective foot						4	5
Aggregated group	4	143	3	3	3		
Hand						4	7
Aggregated group	8	94	0	2	3		
Carpal tunnel syndrome	3	85	1	3	3		
Plastic						4	14
Aggregated group	15	95	2	3	3		
Plateepitelkarsinom	2	73	1	3	3		
BCC	5	142	1	3	3		
Malingt melanom	4	68	0	2	3		
Cancer mammae	4	97	1	3	3		
Arthroscopic						4	12
Aggregated group	6	123	2	3	3		
ACL	2	186	2	3	3		
Meniscus	3	173	0	2	3		
Back						4	6
Aggregated group	4	295	6	3	3		
Prostheses						4	16
Hip	7	177	4	3	6		
Knee	11	174	4	3	6		
Tumour						2	2
Aggregated group	2	76	1	3	3		
emergency patients							
SU inpatients		192	2				
SU outpatients		131	0				
U patients		165	0				

Table A.5: Values of the parameters related to the ORs in the optimization model for the big instance. M_{kd}^{OR} represents the number of surgery slots available at OR k on day d, and B_{kd} is the time available in each slot in OR k on day d.

OR	M_{kd}^{OR}	B_{kd}	Patient category
1	2	240	Elective foot, hand, plastics, arthroscopic
2	2	240	Elective foot, hand, plastics, arthroscopic
3	2	240	Elective foot, hand, plastics, arthroscopic
4	2	240	Elective foot, hand, plastics, arthroscopic
5	2	240	Back, tumour
6	2	240	Prosthesis
7	2	240	Prosthesis

Table A.6: The ward capacities obtained in the optimization model for the big instance. A_d represents the number of staffed beds available on day d, while A_w^{MAX} is the maximum number of beds available in ward w.

Ward	Name	A_d (week)	A_d (weekend)	A_w^{MAX}	Patient category
1	Trauma			32	El. foot, hand, SU
2	Reconstructive			16	Plastic, tumour
3	Elective			12	Arthroscopic, back
4	Fast-track			16	Prosthesis
5	Hotel-day			5	None, Buffer capacity
		67	44		

Table A.7: Values obtained for the subspecialties and the patient categories in the optimization model for the test instances. T_i is the target number of elective patient category i to be scheduled in each cycle, S_i is the expected surgery duration of patient category i , E_{id} is the expected length of stay of patient category i that receive surgery on day d, R_i^{EL} is the gain for scheduling more patients from patient category i than the lower limit, C_i^C is the penalty for cancelling an elective patient of category i, N_{jd}^D is the maximum number of slots available to subspecialty j on day d, and N_j is the maximum number of slots available to subspecialty j in one cycle.

Subspecialty	T_i^{LOW}	T_i^{HIGH}	S_i	E_{id}	R_i^{EL}	C_i^C	N_{jd}^D	N_j
Subspecialty 1							6	41
Patient category 1	8	12	143	3	3	10		
Subspecialty 2							6	41
Patient category 2	32	38	94	0	2	10		
Patient category 3	18	25	85	1	3	10		
Subspecialty 3							6	41
Patient category 4	21	25	95	2	3	10		
Patient category 5	12	15	73	0	3	10		
emergency patients								
SU inpatients			192	2				
SU outpatients			131	0				
U patients			150	0				

Table A.8: Values of the parameters related to the ORs in the optimization model for the test instances. M_{kd}^{OR} represents the number of surgery slots available at OR k on day d, and B_{kd} is the time available in each slot in OR k on day d.

OR	M_{kd}^{OR}	B_{kd}	Patient category
1	2	240	All
2	2	240	All
3	2	240	All
4	2	240	All
5	2	240	All

Table A.9: The ward capacities obtained in the optimization model for the test instances. The four configurations represent four different bed capacities, A_d represents the number of staffed beds available on day d, while A_w^{MAX} is the maximum number of beds available in ward w.

Configuration	Ward	A_d (week)	A_d (weekend)	A_w^{MAX}	Patient category
W1	1			32	1, 3 and SU
W1	2			25	4
		45	40		
W2	1			32	1, 3 and SU
W2	2			25	4
		55	50		
W3	1			45	1, 3 and SU
W3	2			40	4
		85	50		
W4	1			60	1, 3 and SU
W4	2			50	4
		110	55		

Table A.10: Maximum number of ORs that may be covered by anesthesiologists each day in the optimization model for both the large and the test instances.

Instance size	M_1^A	M_2^A	M_3^A	M_4^A	M_5^A	M_6^A	M_7^A
Large (Section 2.5.4)	7	7	7	7	4	0	0
Small (Sections 2.5.2 and 5.5.4)	5	5	5	5	4	0	0

Table A.11: The values obtained for the other parameters in the optimization model. $C_{ww'}^W$ is the penalty of putting a patient meant for ward w in ward w', C^{SU} is the penalty of scheduling an SU patient to an elective slot, P^U is the gain for scheduling a U patient to a flexible slot, C^β is the penalty of having more patients resting in the wards than the total amount of staffed beds available, V_i is the share of patients belonging to the elective patient category i that needs to be scheduled for surgery, and M^{CYCLE} is the total amount of slots available through the cycle.

Instance size	$C_{ww'}^W$	C^{SU}	P^U	C^β	V_i	M^{CYCLE}
Large (Section 2.5.4)	0.1	2	0.5	1000	0.5	70
Small (Sections 2.5.2 and 5.5.4)	1	2	0.5	1000	0.5	70

A.3 Input for the simulation model

In Tables A.12 and A.13 we provide the input values used in the simulation model when running the model in Section 2.5.4. In Table A.14 we provide the probability distributions that are used to model the stochastic processes in the simulation model. These distributions are based on historical data obtained from the case department.

Table A.12: The wards present in the simulation model

Ward	Patient category	Capacity
1	Elective foot, hand and emergencies	∞
2	Plastic, tumour, emergencies	∞
3	Arthroscopic, back, emergencies	∞
4	Prosthesis, emergencies	∞
5	Emergencies	∞

Table A.13: The ORs present in the simulation model

OR-type	Number of ORs	Opening hours
Elective	7	08.00-16.00 (Monday to Friday)
Emergency	3	Varies dependent on day and OR

Table A.14: Stochastic processes

Process	Probability distribution
Emergency arrivals	Poisson
Surgery duration	Empirical
Length of stay	Empirical

A.4 The case study

Table A.15 provides the optimized MSS.

Table A.15: The MSS generated in the case study. Green letters indicate flexible slots.

OR	Day of week				
	Monday	Tuesday	Wednesday	Thursday	Friday
1	Hand	Hand	El. foot/ Plastic	Plastic/ Arthro.	Plastic/ Plastic
2	Arthro.	Arthro.	El. foot	Plastic	Arthro.
3	Plastic	Plastic/ Arthro.	Arthro.	Hand	Arthro.
4	Plastic	Plastic/-	Plastic	Hand/-	El. foot
5	Tumour/Tumour	Back	Back	Back	-
6	Prosthesis	Prosthesis	Prosthesis/-	Prosthesis	-
7	Prosthesis	Prosthesis	Prosthesis/-	Prosthesis	-

Paper II

T. R. Bovim, A. Abdullahu, H. Andersson, A. N. Gullhav:

Integrated Master Surgery and Outpatient Clinic Scheduling

Published in Operations Research for Health Care

Chapter 3

Integrated Master Surgery and Outpatient Clinic Scheduling

Abstract

In this paper, we study an integrated master surgery and outpatient clinic scheduling problem, motivated by the situation at the Orthopaedic Department at St. Olav's Hospital, Trondheim. During a treatment process, the patients require one or several consultations at the outpatient clinic, and potentially a surgery in one of the operating rooms. The physicians perform both consultations and surgeries, and coordinating the two facilities is challenging. The surgeons are trained to handle different surgical specialties, and they differ in experience. The overall goal is to schedule the specialties, and a number of qualified surgeons, to time slots in the outpatient clinic and operating rooms through the week, to efficiently handle the patient demand. Our main contribution is an optimisation model for solving the integrated master surgery and outpatient clinic scheduling problem. In addition to allocating specialties and a number of surgeons, the model also schedules activity types (surgery categories and outpatient clinic consultation types) to the time slots. These can guide the operational scheduling of individual patients at a later stage. A computational study is performed, demonstrating the use of the optimisation model to provide a set of master schedules, based on a set of different resource capacity cases. We develop a simulation model for evaluating the master schedules in an operational setting, and three different operational scheduling policies are compared. We conclude that scheduling patients to activities governed primarily by the opti-

misation model solution outperforms a FIFO scheduling policy based only on specialty.

3.1 Introduction

Surgical costs account for approximately 40% of the total hospital costs (Van Essen et al., 2012), and Freeman et al. (2018) state that 60-70% of all patients admitted to a hospital require some surgical intervention. However, surgeries are not performed in isolation, and by neglecting aspects of coordination when performing capacity planning, we may arrive at suboptimal solutions. Surgical patients typically require a consultation at the outpatient clinic (OC) both prior to, and following surgery. In addition, many patients require a stay in a hospital ward to recover from surgery. The surgeons serve both the OC rooms and the ORs, and their time must be carefully divided between the two facilities to provide a coordinated and efficient service.

The Master Surgery Scheduling Problem (MSSP) is a frequently studied problem within operations research. The Master Surgery Schedule (MSS), is a cyclic schedule where surgical specialties are assigned to OR slots through the week, such that the demand for surgery is covered. The time horizon considered in the MSSP is typically in the range of months, and the MSS is repeated through the time horizon. An equivalent problem can be formulated for the OC.

In this paper, we study the integrated master surgery and outpatient clinic scheduling problem. The problem is motivated by the situation at the Orthopaedic Department as St. Olav's Hospital, Trondheim. Vik et al. (2022) report poor coordination of key health care activities as one of the major challenges within the department. From our point of view, a lack of coordination between the OC and the ORs can lead to unwanted variations in demand for activities, and we believe that the work presented in this paper can help to increase coordination.

During the treatment process, the patients require one or several consultations in the OC rooms, and potentially a surgery in one of the ORs. The surgeons are trained to handle different surgical specialties, and they differ in experience. The goal is to handle patient demand by allocating specialties, and a number of qualified surgeons, to time slots in the OC rooms and the ORs through the planning horizon. In addition, we also schedule activity types (surgery categories and outpatient clinic consultation types) to the available time slots, which can be used for scheduling individual patients at a later stage. To our knowledge this problem has not been studied in literature. It differs both from multi-appointment scheduling, which mainly considers outpatient services, and from the MSSP, that seldom includes upstream units.

The main contribution in this paper is an optimisation model for solving the integrated master surgery and outpatient clinic scheduling problem. The model produces two cyclic master schedules; one for the OC rooms and one for the ORs. To evaluate the performance of the master schedules, they are implemented in a discrete-event simulation (DES) model, where patient arrivals and the paths of individual patients are modelled as stochastic processes. We compare three different operational scheduling policies. This allows us to investigate the value of scheduling activity types in the master schedules.

A computational study is performed based on data from the Orthopaedic Department at St. Olav's hospital. In this study, five alternative resource capacity cases are considered, each representing a strategy which the department can implement to increase patient throughput. Even though we study the situation at the orthopaedic department, the problem under study is rather generic and can be found in many hospital departments that perform surgical activities. It is common for such departments to serve patients that require multiple services, and where a set of resources are involved in providing more than one service. With some adjustments, the problem under study is relevant to departments that face similar problems.

The rest of the paper is structured as follows. In Section 3.2, a literature review is provided to position our contribution. Then, in Section 3.3 the problem under consideration is described. The mathematical model is presented in Section 3.4, while the simulation model is provided in Section 3.5. In Section 3.6, the computational study is reported on, before presenting managerial insights in Section 3.7. Finally, in Section 3.8, the paper is concluded.

3.2 Literature review

Tactical outpatient and surgery scheduling, that considers multiple resources, constitutes the field of interest. First we present selected literature on surgery scheduling, and the MSSP in particular, before turning to the tactical OC planning.

3.2.1 Tactical OR planning

Within health care planning, the tactical decision level addresses the organisation of the execution of the health care process (Hulshof et al., 2012). At this level, the available resource capacities, settled at the strategic level, are divided among patient groups. Blueprints for the operational planning are created that allocate resources to different tasks, specialties and patient categories. According to Hulshof et al. (2012), the MSSP is considered a tactical planning problem within surgical care services.

Cardoen et al. (2010) review the literature on OR planning and scheduling, and find that about half of the recent contributions limit their scope to an isolated OR. Among the contributions that regard additional facilities, the wards, the Intensive Care Unit and the Post Anesthesia Care Unit are most frequently included. However, the modelling of ORs in interaction with other hospital facilities remains a main topic for further research (Cardoen et al., 2010).

When considering the MSSP, the downstream facilities, and the wards in particular, are frequently considered. Li et al. (2017), Moosavi and Ebrahimnejad (2020) and Adan et al. (2011) include the Intensive Care Unit when analysing the

MSSP, while Schneider et al. (2020) and Fügener et al. (2014) consider multiple downstream units. The upstream activities are seldom regarded in the MSSP literature, and Schneider et al. (2020) propose the inclusion of upstream units, such as the OC, as a topic for future research. Moosavi and Ebrahimnejad (2020) regard the wards as both up- and downstream capacities, acknowledging the fact that patients might need a bed both prior to, and following surgery.

Schneider et al. (2020) cluster surgery types into surgery groups, and instead of scheduling surgery specialties, they schedule surgery groups to the OR blocks. This eases the operational scheduling of individual patients.

Unlike the majority of literature on the MSSP, we consider both up- and downstream activities in our problem. By including the OC we can schedule more of the surgeons' activities, and instead of considering the derived demand for surgery we can include the demand for new referrals. This allows us to control the number of patients that are sent to surgery based on the overall capacity of the system. Furthermore, like Schneider et al. (2020), we schedule surgery groups for the OR slots, and through simulation we evaluate the value of using these when performing the operational scheduling of surgeries.

3.2.2 Tactical OC planning

According to Hulshof et al. (2012), OC planning is categorised as ambulatory care services, and the existing literature is mainly focusing on the operational appointment scheduling. At the tactical level, the allocation of capacity to patient groups is the most frequently studied problem in the OC planning literature (Ahmadi-Javid et al., 2017).

Historically, the majority of research has considered patients requiring a single appointment. However, in recent years, an increasing number of researchers has considered several resources and the fact that patients may require multiple consultations (Marynissen and Demeulemeester, 2019). According to Marynissen and Demeulemeester (2019), the multi-appointment scheduling problem is designed to act as an umbrella for both combination appointments, in which patients require multiple appointments on the same day, and appointment series, in which patients need to revisit the same set of resources several times. Furthermore, the authors define the multi-appointment scheduling problem as an operational problem, but emphasise the importance of reserving capacity at a tactical level. Examples of tactical multi-appointment scheduling problems can be found in Bikker et al. (2015), Nguyen et al. (2015) and Hahn-Goldberg et al. (2014). The first two represent an appointment series problem, while the last is a combination appointment problem for scheduling patients for chemotherapy.

Care processes can be analysed as a multi-appointment scheduling problem, as the patients typically require several visits to the hospital. Hulshof et al. (2013) consider the tactical resource allocation for elective patient admission planning in care processes. The authors analyse a care process comprising of a visit to the OC, followed by surgery and a revisit to the OC, which is similar to the one studied in this paper. In the modelling framework presented by Hulshof et al. (2013), each care process is represented by a set of consecutive queues, and patients are routed between queues to represent the demand for each care process. In each time period, a number of patients are served in each queue, while the patients remain in the queue to the next time period. To serve a patient in a given queue, a set of resources are required throughout the time period. All resources have a given capacity, restricting the flow of patients between queues. The decision variables represent the number of patients treated from each queue, in each time period, and a solution represents the resource capacity devoted to each queue in each time period.

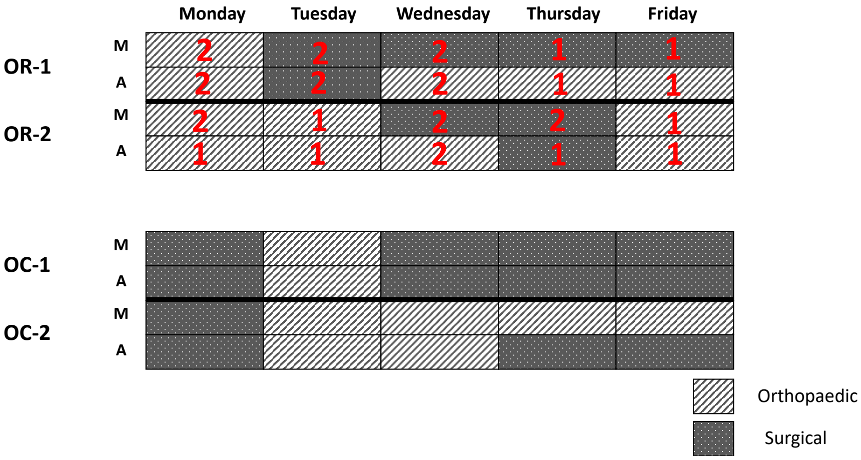
Another branch of hospital planning that relates to multi-appointment scheduling is multi-disciplinary scheduling. Leeftink et al. (2020) define a multi-disciplinary care system as a care system in which multiple interrelated appointments per patient are scheduled, where health care professionals from various facilities, or with different skills are involved. The authors categorise the literature according to a hierarchical planning structure, and within capacity planning the generation of blueprint schedules, patient admission planning and temporary capacity changes are typical outcomes. Like for multi-appointment scheduling, the consultations in multi-disciplinary scheduling can be performed within a day (see Liang et al. (2015)), or as multiple revisits (see Braaksma et al. (2014)).

To the best of our knowledge, the existing papers on tactical multi-appointment scheduling (and similar problems) in hospitals only consider outpatients. Although Hulshof et al. (2013) encounter a setting which has similarities to ours, the problems differ. In Hulshof et al. (2013), the resource requirements to perform an activity (serve patients from a queue) is given. In our problem, surgeons of different experience levels can perform OC consultations, allowing for flexibility. Furthermore, Hulshof et al. (2013) apply discrete time steps in their model, requiring that each activity is started and finished within one time period. If a time step resembles one day, coordination of resources within a day is problematic as surgeons cannot serve one activity in the morning and another in the afternoon. If the time steps represent shorter time periods, coordination is possible. However, requiring that each activity spans one time period makes short time steps problematic. Finally, beds cannot be analysed as resources, as patients cannot wait for a bed following surgery. Therefore, their framework is most suitable for an outpatient setting.

Our model is suitable for handling inpatients, and we consider the demanding task of coordinating and scheduling resources to serve patients that require multiple services.

3.3 Problem description

In the integrated outpatient and master surgery scheduling problem, the aim is to generate two cyclic schedules; one for the OC rooms, and one for the ORs. The time available during the working day is divided into fixed time slots, and we define a slot within a room as a room slot. In both schedules, surgical specialties are assigned



		Monday	Tuesday	Wednesday	Thursday	Friday
OR-1	M	▨ 2	■ 2	■ 2	■ 1	■ 1
	A	▨ 2	■ 2	▨ 2	▨ 1	▨ 1
OR-2	M	▨ 2	▨ 1	■ 2	■ 2	▨ 1
	A	▨ 1	▨ 1	▨ 2	■ 1	▨ 1

OC-1	M	■	▨	■	■	■
	A	■	▨	■	■	■
OC-2	M	■	▨	▨	▨	▨
	A	■	▨	▨	■	■

▨ Orthopaedic

■ Surgical

Figure 3.1: Example of master schedules for the ORs and the OC rooms. Here, the capacity is shared between the orthopaedic and the surgical department. For the ORs, the number of surgeons assigned to each OR is given in addition to the specialty. In this example, two slots are available in each room each day. M is the morning slot, while A is the afternoon slot.

to the available room slots through the planning cycle (typically one week), such that the system can serve at least the expected demand of new referrals during the planning horizon (typically half a year). As an example, Figure 3.1 illustrates two such schedules where two OC rooms and two ORs are shared between the orthopaedic and the surgical department. Here, two slots are available in each room, representing a morning and an afternoon slot. A scheduled slot in the operating theatre comprises both the medical specialty and the number of surgeons assigned to the slot. For the OC, only the specialty is given, as it is always sufficient with one surgeon to perform an OC consultation.

Four activity types can be performed; surgery and three different types of OC consultations. In an *initial consultation* (IC), the patient is examined and the surgeon decides on whether further intervention is needed, and if so, what kind of intervention. If a non-surgical intervention is required, the patient receives a *treatment consultation* (TC) in the OC. Finally, following either surgery or a treatment consultation, the patient is summoned for one or more *follow-up consultations* (FU) in the OC. The length of an OC consultation depends on the specialty of the patient, and what type of consultation that is performed. Surgeries are categorised depending on what procedures that are done during surgery. A surgery category is characterised by the planned surgery duration, the minimum number of surgeons that must be present during surgery, and the planned length of stay (LOS) in a ward following surgery.

There exists a set of surgical specialties, and patients are categorised based on the specialty they belong to. The activity types required by a patient depends on what specialty the patient belongs to. Following the initial consultation, a share of patients belonging to a specialty requires a treatment consultation, while another share requires a surgery. The remaining patients leave the system. Following a treatment consultation or surgery, the patients require one or multiple follow-up consultations. A number of patients within each specialty is referred straight for surgery, and the remaining paths of these patients are identical to the other patients within the same specialty.

Figure 3.2 illustrates the activity types considered in the problem, and how they relate to each other for a specialty with two surgery categories available. In this example, x_{IC} initial consultations are scheduled. The expected demand for treatment consultations and surgery categories one and two are calculated as the probabilities that patients require these activities following an initial consultation multiplied by x_{IC}. Similar logic is used to calculate the expected demand of follow-up consultations based on the number of surgeries and treatment consultations that are scheduled.

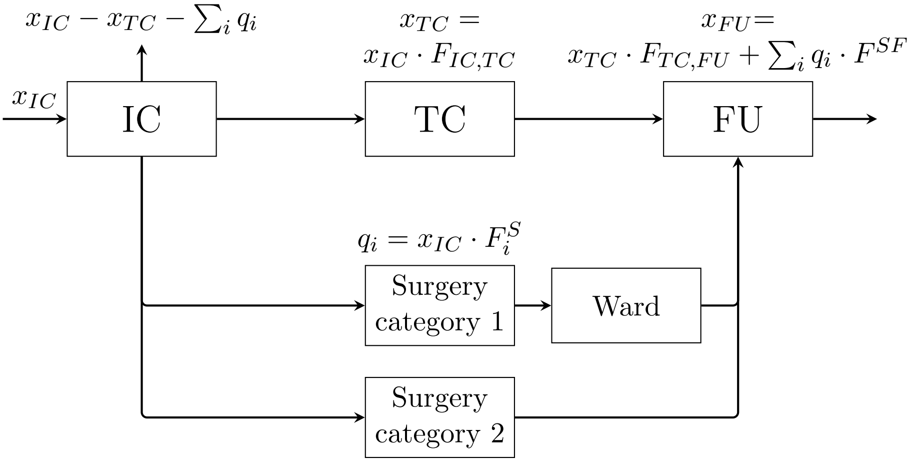

$x_{IC} - x_{TC} - \sum_i q_i$
$x_{TC} =$
$x_{IC} \cdot F_{IC,TC}$
$x_{FU}=$
$x_{TC} \cdot F_{TC,FU} + \sum_i q_i \cdot F^{SF}$
x_{IC}
IC
TC
FU
$q_i = x_{IC} \cdot F_i^S$
Surgery category 1
Ward
Surgery category 2

Figure 3.2: The activity types considered in the problem, and how they relate to each other. x_{IC}: Number of initial consultations scheduled. x_{TC}: Derived demand for treatment consultations that must be scheduled. q_i: Derived demand for surgery category i that must be scgeduled. x_{FU}: Derived demand for follow-up consultations that must be scheduled. $F_{IC,TC}$: Average fraction of initial consultations that yield a demand for treatment consultations. $F_{TC,FU}$: Average fraction of treatment consultations that yield a demand for follow-up consultations. F_i^S: Average fraction of initial consultations that yield a demand for surgery of category i. F^{SF}: Fraction of surgeries that yield a demand for follow-up consultations.

There is a given number of OC rooms and ORs. The opening hours of the rooms are divided into consecutive time slots. The slot duration can differ between the OC rooms and the ORs, but they are of constant length within each facility. The slots are synchronised, such that a surgeon can serve one of the facilities in the morning, and the other one in the afternoon. As a consequence of these requirements, there can be at most two slots available during a day in each of the facilities. Each slot can be scheduled for one specialty, and the number of patients that can be scheduled within a slot is limited by the slot duration. If a specialty covers two consecutive slots in a room, an activity may begin in the first slot and end in the other. There are several wards available to serve the inpatients that require a bed following surgery. In every ward, a given number of beds are available each day, and each ward can serve patients from a subset of the surgery categories.

The surgeons are categorised according to what specialties they master, and their level of experience. Surgeons can be trained to master several specialties, and surgeons that master a given specialty may provide all activity types to patients belonging to that specialty. Based on the level of experience, surgeons are either consultants or residents, and each surgery requires the presence of at least one consultant. Both consultants and residents can perform activity types in the OC. Each surgeon type has a fixed number of surgeons available each day of the cycle.

To schedule a surgery of a given category there must be enough time available in an OR and a slot scheduled for the corresponding specialty. Furthermore, there must be enough surgeons scheduled to the same room and slot to perform the surgery, and the bed capacity must be sufficient to cover the entire LOS of the patient. To schedule an OC activity there must be enough time available in an OC room and a slot scheduled for the corresponding specialty.

The expected arrival rate of initial consultations during a cycle is constant and known for each specialty. In addition, there is a queue of patients that have not yet received an initial consultation at the beginning of the planning horizon. To maintain a stable waiting list for each specialty, a minimum throughput of initial consultations should be set such that the service rate is at least incrementally higher than the expected arrival rate of new referrals. Furthermore, to decrease a potential queue of patients waiting for an initial consultation, a reward is given for scheduling more than the minimum throughput. However, the maximum number of initial consultations that can be scheduled for each specialty in a cycle cannot exceed the expected arrival rate of new referrals, plus the length of the waiting list divided by the number of cycles in the planning horizon. To avoid queues from building up within the system, we must ensure that the system can handle the downstream demand for services generated.

Our goal is to serve at least the expected number of new referrals, and more if possible, while making sure that there is sufficient capacities to handle the derived demand for downstream services.

3.4 The mathematical model

In this chapter, the mathematical model for solving the problem is presented. Tables 3.1, 3.2, and 3.3 include all the notation used in the mathematical model. To enhance readability, the constraints are introduced in thematic groups.

Table 3.1: Sets

Symbol	**Description**	
$\mathcal{A}$	Consultation types performed at the OC	$a \in \mathcal{A}$
$\mathcal{D}$	Days in a cycle	$d \in \mathcal{D}$
$\mathcal{I}$	Surgery categories	$i \in \mathcal{I}$
$\mathcal{J}$	Surgical specialties	$j \in \mathcal{J}$
$\mathcal{K}$	ORs	$k \in \mathcal{K}$
$\mathcal{L}$	OC rooms	$l \in \mathcal{L}$
$\mathcal{N}$	Number of surgeons that can be present during surgery	$n \in \mathcal{N}$
$\mathcal{P}$	Surgeon types	$p \in \mathcal{P}$
$\mathcal{S}$	Time slots	$s \in \mathcal{S}$
$\mathcal{W}$	Wards	$w \in \mathcal{W}$
$\mathcal{D}_{id}^{LOS}$	Days that a patient of surgery category i can have received surgery if the patient is still in a ward on day d	$d' \in \mathcal{D}_{id}^{LOS}$
$\mathcal{I}_j^J$	Surgery categories that can be handled by specialty j	$i \in \mathcal{I}_j^J \subseteq \mathcal{I}$
$\mathcal{I}_w^W$	Surgery categories that can rest in ward w	$i \in \mathcal{I}_w^W \subseteq \mathcal{I}$
$\mathcal{J}_k^K$	Specialties that can be scheduled to OR k	$j \in \mathcal{J}_k^K \subseteq \mathcal{J}$
$\mathcal{K}_j$	ORs that can be utilised by specialty j	$k \in \mathcal{K}_j \subseteq \mathcal{K}$
$\mathcal{P}_j^C$	Surgeon types that are consultants and can cover specialty j	$p \in \mathcal{P}_j^C \subseteq \mathcal{P}$
$\mathcal{P}_j^R$	Surgeon types that are residents and can cover specialty j	$p \in \mathcal{P}_j^R \subseteq \mathcal{P}$
$\mathcal{W}_i^I$	Wards that can serve patients from surgery category i	$w \in \mathcal{W}_i^I \subseteq \mathcal{W}$

Table 3.2: Parameters

Symbol	**Description**
A_{wd}	Number of beds available at ward w on day d
C_{pd}	Number of surgeons of type p available on day d
$\underline{D}_j$	Min. number of ICs of specialty j that must be scheduled during one cycle
$\overline{D}_j$	Max. number of ICs of specialty j that can be scheduled during one cycle
F_i^S	Fraction of ICs that yields a downstream demand for surgery category i
$F_{jaa'}$	Fraction of OC consultations of type a that yields a downstream demand for OC consultations of type *a’* for specialty j
F_j^{SF}	Number of FU consultations generated by a surgery of specialty j
H_{ja}^{OC}	Planned time needed for consultations of type a of specialty j
$\underline{N}_i$	Min. number of surgeons that must be present for a surgery of category i
$\overline{N}$	Max. number of surgeons that can be present during surgery
Q_i	Number of patients from surgery category i that are referred straight to surgery during a cycle
Q_j^I	Number of patients from specialty j waiting for an IC at the beginning of the planning horizon
$\overline{Q}_{nikd}$	Max. number of surgeries of surgery category i that can be scheduled for surgery with n surgeons present in OR k on day d
R_j	Reward obtained from scheduling an IC of specialty j
S_i	Planned surgery duration for patients of surgery category i
T	Number of cycles in the planning horizon
T_{ksd}^{OR}	Time available for surgeries in OR k, slot s on day d
T_{lsd}^{OC}	Time available for consultations in OC room l, slot s on day d
$\overline{X}_{jald}$	Max. number of consultations of type a of specialty j that can be scheduled to OC room l on day d

Table 3.3: Variables

Letter	**Description**
g_{pjsd}^{OC}	# of surgeons of type p assigned to specialty j in slot s in the OC on day d
g_{pksd}^{OR}	# of surgeons of type p allocated to OR k in slot s on day d
q_{nikd}	# of surgeries of category i scheduled in OR k with n surgeons on day d
u_{iwd}	# of beds occupied by patients of surgery category i in ward w on day d
x_{jald}	# of OC consultations of type a of specialty j scheduled in OC room l on day d
β_{jlsd}	Indicates if specialty j is assigned OC room l in slot s on day d
λ_{njksd}	Indicates if specialty j is assigned OR k in slot s, with n surgeons on day d

Demand constraints

$$\underline{D}_j \leq \sum_{l \in \mathcal{L}} \sum_{d \in \mathcal{D}} x_{j,IC,ld} \leq \overline{D}_j \qquad j \in \mathcal{J} \tag{3.1}$$

The demand constraints (3.1) ensure that we schedule between the planned demand and the upper limit of initial consultations for each specialty j. Here, $\overline{D}_j$ is the maximum number of initial consultations of specialty j that can be scheduled in a cycle, while $\underline{D}_j$ is the minimum throughput of new referrals for one cycle. $\overline{D}_j$ can be calculated by adding the number of patients from specialty j waiting for an initial consultation at the beginning of the planning horizon, Q^I_j, divided by the number of cycles in the planning horizon, T, to the minimum throughput:

$$\overline{D}_j = \underline{D}_j + \left\lceil \frac{Q^I_j}{T} \right\rceil \qquad j \in \mathcal{J} \tag{3.2}$$

Slot constraints

$$\sum_{j \in \mathcal{J}} \beta_{jlsd} \leq 1 \qquad l \in \mathcal{L}, s \in \mathcal{S}, d \in \mathcal{D} \tag{3.3}$$

$$\sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{J}^K_k} \lambda_{njksd} \leq 1 \qquad k \in \mathcal{K}, s \in \mathcal{S}, d \in \mathcal{D} \tag{3.4}$$

Constraints (3.3) make sure that at most one specialty j is assigned to each slot s in every OC room l on day d. Constraints (3.4) ensure that at most one specialty j with n surgeons is assigned to each slot s in OR k on day d.

Surgeon constraints

$$\sum_{l \in \mathcal{L}} \beta_{jlsd} \leq \sum_{p \in \mathcal{P}^C_j} g^{OC}_{pjsd} + \sum_{p \in \mathcal{P}^R_j} g^{OC}_{pjsd} \qquad j \in \mathcal{J}, s \in \mathcal{S}, d \in \mathcal{D} \tag{3.5}$$

$$\lambda_{njksd} \leq \sum_{p \in \mathcal{P}^C_j} g^{OR}_{pksd} \qquad n \in \mathcal{N}, j \in \mathcal{J}, k \in \mathcal{K}_j, s \in \mathcal{S}, d \in \mathcal{D} \tag{3.6}$$

$$n\lambda_{njksd} \leq \sum_{p \in \mathcal{P}^C_j} g^{OR}_{pksd} + \sum_{p \in \mathcal{P}^R_j} g^{OR}_{pksd} \quad n \geq 2, j \in \mathcal{J}, k \in \mathcal{K}_j, s \in \mathcal{S}, d \in \mathcal{D} \tag{3.7}$$

$$\sum_{j \in \mathcal{J}} g^{OC}_{pjsd} + \sum_{k \in \mathcal{K}} g^{OR}_{pksd} \leq C_{pd} \qquad p \in \mathcal{P}, s \in \mathcal{S}, d \in \mathcal{D} \tag{3.8}$$

Constraints (3.5) ensure that all OC rooms that are scheduled for specialty j in slot s on day d, must be covered by at least one surgeon each. Constraints (3.6) require that, if specialty j is assigned to OR k in slot s on day d with n surgeons, there should be at least one consultant of the same specialty assigned to that OR, at that point in time. Constraints (3.7) state that, if specialty j is assigned to OR k in slot s on day d with two or more surgeons, this OR must be covered by enough surgeons from that specialty, at that point in time. Constraints (3.8) make sure that the number of surgeons allocated to slot s from surgeon type p on a given day d does not exceed the number of surgeons available from that surgeon type on that day.

Time capacity constraints in the OC

$$\sum_{a\in\mathcal{A}} H^{OC}_{ja} x_{jald} \leq \sum_{s\in\mathcal{S}} T^{OC}_{lsd}\beta_{jlsd} \qquad j\in\mathcal{J}, l\in\mathcal{L}, d\in\mathcal{D} \tag{3.9}$$

The time capacity constraints (3.9) make sure that the total time scheduled for initial consultations, treatment consultations, and follow-up consultations for specialty j, in OC room l on day d, cannot exceed the time scheduled for that specialty in that room on that day.

Patient flow constraints

$$\sum_{l\in\mathcal{L}}\sum_{d\in\mathcal{D}} F^S_i x_{j,IC,ld} + Q_i \leq \sum_{n=\underline{N}_i}^{\overline{N}} \sum_{k\in\mathcal{K}_j}\sum_{d\in\mathcal{D}} q_{nikd} \qquad j\in\mathcal{J}, i\in\mathcal{I}^J_j \tag{3.10}$$

Constraints (3.10) state that the number of scheduled surgeries of surgery category i, is at least the same as the sum of the planned demand for surgery of category i, derived from the initial consultations scheduled for specialty j, and the expected number of surgeries of surgery category i referred from other instances.

$$\sum_{l\in\mathcal{L}}\sum_{d\in\mathcal{D}} F_{j,IC,TC} x_{j,IC,ld} \leq \sum_{l\in\mathcal{L}}\sum_{d\in\mathcal{D}} x_{j,TC,ld} \quad j\in\mathcal{J} \tag{3.11}$$

Constraints (3.11) ensure that we, for each specialty j, schedule enough treatment consultations in relation to initial consultations.

$$\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{I}^J_j}\sum_{k\in\mathcal{K}_j}\sum_{d\in\mathcal{D}} F^{SF}_j q_{nikd} + \sum_{l\in\mathcal{L}}\sum_{d\in\mathcal{D}} F_{j,TC,FU} x_{j,TC,ld} \leq \sum_{l\in\mathcal{L}}\sum_{d\in\mathcal{D}} x_{j,FU,ld} \quad j\in\mathcal{J} \tag{3.12}$$

Constraints (3.12) make sure that we schedule at least the required fractions of follow-up appointments from specialty j after surgery or after a treatment consultation.

Time capacity constraints for surgery

$$\sum_{i\in\mathcal{I}^J_j} S_i q_{nikd} \leq \sum_{s\in\mathcal{S}} T^{OR}_{ksd}\lambda_{njksd} \qquad n\in\mathcal{N}, j\in\mathcal{J}, k\in\mathcal{K}_j, d\in\mathcal{D} \tag{3.13}$$

Constraints (3.13) ensure that time capacity is respected in the ORs. The total time scheduled for surgery categories belonging to specialty j within OR k on day d cannot exceed the time scheduled for that specialty, in that room, on that day.

Ward constraints

$$\sum_{i\in\mathcal{I}_w^W} u_{iwd} \leq A_{wd} \qquad w\in\mathcal{W}, d\in\mathcal{D} \tag{3.14}$$

$$\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}}\sum_{d'\in\mathcal{D}_{id}^{LOS}} q_{nikd'} = \sum_{w\in\mathcal{W}_i^I} u_{iwd} \qquad i\in\mathcal{I}, d\in\mathcal{D} \tag{3.15}$$

Constraints (3.14) state that the number of beds occupied at ward w on day d does not exceed the available number of beds at the ward. In constraints (3.15), we count the number of patients of surgery category i still present at a ward on a given day d.

Objective function

$$\max \sum_{j\in\mathcal{J}}\sum_{l\in\mathcal{L}}\sum_{d\in\mathcal{D}} R_j x_{j,IC,ld} \tag{3.16}$$

In the objective function we maximise the reward generated from covering more than the expected demand of initial consultations at the OC. This is an attempt to decrease the queue of referrals for a specialty during the planning horizon.

Variable domains

$$x_{jald} \in \{0,1,...,\overline{X}_{jald}\} \qquad j\in\mathcal{J}, a\in\mathcal{A}, l\in\mathcal{L}, d\in\mathcal{D} \tag{3.17}$$

$$q_{nikd} \in \{0,1,...,\overline{Q}_{nikd}\} \qquad i\in\mathcal{I}, n\geq\underline{N}_i, k\in\mathcal{K}, d\in\mathcal{D} \tag{3.18}$$

$$u_{iwd} \in \{0,1,...,A_{wd}\} \qquad i\in\mathcal{I}, w\in\mathcal{W}, d\in\mathcal{D} \tag{3.19}$$

$$g_{pjsd}^{OC} \in \{0,1\} \qquad p\in\mathcal{P}, j\in\mathcal{J}, s\in\mathcal{S}, d\in\mathcal{D} \tag{3.20}$$

$$g_{pksd}^{OR} \in \{0,1,...,\overline{N}\} \qquad p\in\mathcal{P}, k\in\mathcal{K}, s\in\mathcal{S}, d\in\mathcal{D} \tag{3.21}$$

$$\beta_{jlsd} \in \{0,1\} \qquad j\in\mathcal{J}, l\in\mathcal{L}, s\in\mathcal{S}, d\in\mathcal{D} \tag{3.22}$$

$$\lambda_{njksd} \in \{0,1\} \qquad n\in\mathcal{N}, j\in\mathcal{J}, k\in\mathcal{K}, s\in\mathcal{S}, d\in\mathcal{D} \tag{3.23}$$

Constraints (3.17) to (3.23) give the domains for the variables.

Prioritising among the specialties

Between the minimum demand and the upper limit for new referral consultations, it is possible to prioritise among the specialties. The majority of patient generated income comes from surgical activity. The income is correlated with the expected surgery duration, such that surgeries with long expected duration generates a high income. We can incorporate the surgery duration when prioritising the specialties in the reward set for each specialty. A way to calculate the reward for each specialty is as the product of the fraction of surgeries and the expected surgery duration, as shown in equations (3.24).

$$R_j = \sum_{i\in\mathcal{I}_j} F_i^S S_i \qquad j\in\mathcal{J} \tag{3.24}$$

3.5 The simulation model

In this section, the discrete-event simulation (DES) model is introduced. To describe the simulation study and the DES model, the STRESS guidelines, introduced by Monks et al. (2019), are used. First, we describe the objectives of the simulation study, before presenting the logic of the model. The data is introduced in Section 3.6, and described in details in Appendix A.2.

3.5.1 Objectives

The purpose of the simulation study is to evaluate the performance of the tactical schedules provided by the optimisation model when including random arrivals of new referrals, and random paths of patients through the system. The main input for the simulation model is the schedules generated by the optimization model, while the main output is the development over time of the queues of patients waiting for both OC consultations and surgery, and the mean total service time of patients in the system. The length of the queues are recorded at the beginning of each simulated week. There are two main aims of experimentation: To evaluate the performance of the tactical schedules provided by the optimisation model, and to evaluate the effect of different operational scheduling policies.

We want to emphasise that the simulation model is not a replication of the complex system of the orthopaedic department. There are many events that are not considered, such as no-shows and absence of staff, and stochastic processes, such as surgery duration and patient length of stay, that are considered deterministic. To isolate the effects from implementing different schedules and scheduling policies, we have chosen to keep the system under study rather simple. For this reason, a model validation is not applicable in our case.

3.5.2 Logic

The entities of the model are the patients, and the attributes of the patients are presented in Table 3.4. An illustration of the system considered can be seen in Figure 3.3. The resources available are the OC rooms, and the combination of ORs and beds. The resource capacities available to the various specialties at different days are given by the schedules obtained from the solution of the optimisation model. The variables β_{jlsd} and λ_{njksd} indicate what specialty that has access to the different room slots during the week, while the variables x_{jald} and q_{nikd} indicate what activity types (type of OC consultation or surgery category respectively) that should be performed in the rooms. The daily bed capacity reserved for patients of surgery category i is given by the u_{iwd} variables. The activity types provided in the OC rooms are initial consultations, treatment consultations and follow-up consultations, while the activity types provided in the ORs are surgery categories and potentially a subsequent stay in a bed. There are four queues in the system, illustrated by the grey squares, one in front of each activity type. Each queue is split into subqueues, one for each specialty.

Table 3.4: The attributes of the patients

Attribute number	**Description**
1	Specialty
2	Surgery category
3	Treatment at OC, $\{0, 1\}$
4	Surgery, $\{0, 1\}$
5	Ward
6	Number of follow-ups, $\{0, 1, 2\}$

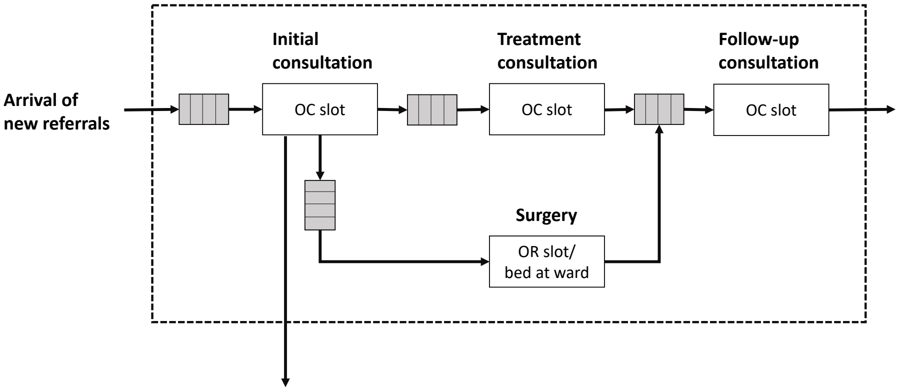
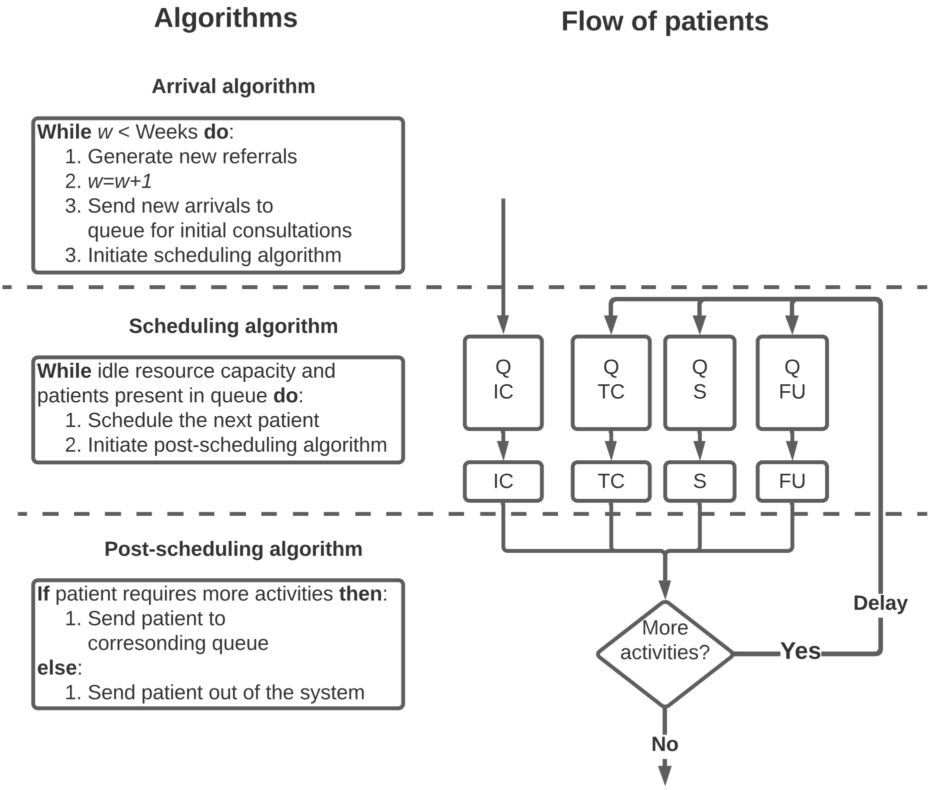


Figure 3.3: The flow of patients in the DES model.



Algorithms

Arrival algorithm

```
While w < Weeks do:
    1. Generate new referrals
    2. w=w+1
    3. Send new arrivals to
       queue for initial consultations
    3. Initiate scheduling algorithm
```

Scheduling algorithm

```
While idle resource capacity and
patients present in queue do:
    1. Schedule the next patient
    2. Initiate post-scheduling algorithm
```

Post-scheduling algorithm

```
If patient requires more activities then:
    1. Send patient to
       corresonding queue
else:
    1. Send patient out of the system
```

Figure 3.4: Overview of the DES model implementation.

In Figure 3.4, an overview of the DES model implementation can be seen. Fixed time increments of one week are applied, and we assume that no patients receive more than one consultation per week, and that patients are added to the queue for initial consultations the week following arrival. Although this may not always be the case in real life, it is a reasonable assumption when modelling elective patients whose waiting time limits are typically in the range of months. In the arrival algorithm, new referrals are generated each week and sent to the queue for initial consultations the week after. Then, the scheduling algorithm is initiated, which assigns patients to activities in the present week according to two scheduling policies, described below. When a patient is assigned for an activity, the post-scheduling algorithm is initiated. Here, patients are either sent to the queue for the subsequent activity, or, if no further activities are required, they leave the system. Patients do not join the queue for the subsequent activity before the following week. There can be a need for an additional delay between activities, and if so, a patient will not join the queue before the delay has passed.

We assume that the referrals arrive independently of each other, and model the arrivals as a Poisson process with expected arrival rates equal to the expected demand used in the optimisation model. The probabilities of generating patients of a given surgery category, and a given path, are set such that the flow of patients corresponds to the flow given by the F parameters in the optimisation model. In contrast to the arrival of patients and the paths required by patients, the planned service durations are deterministic. We acknowledge that there exist mechanisms that will impact the realised outcome of a schedule, and the importance of efficient rescheduling. However, these are not studied here.

In the scheduling algorithm, two different scheduling policies are used for assigning patients to activities. In both policies, all subqueues are sorted according to a FIFO principle, and when a patient of a given subqueue is to be scheduled, the first patient in the queue is chosen. Algorithms 1 and 2, presented in Appendix A.1, explain the two policies applied to the OC room activities. In the former, patients are scheduled in accordance with the optimisation model solution, that is the x_{jald} variables. If there are not enough patients present in the corresponding subqueue, the scheduled capacity is left idle. In the second scheduling policy, we schedule patients based on specialty and the remaining time available in an OC room. If a specialty is scheduled for a slot in an OC on a given day, and the remaining time in this room slot is sufficient to perform more OC consultations, the patient that has waited the longest for an OC consultation (either an initial, a treatment or a follow-up consultation) within the corresponding specialty is scheduled. If several patients have waited equally long, one is randomly chosen.

Corresponding policies are implemented for scheduling patients in the ORs. If, in the first scheduling policy, we cannot find a patient to schedule according to the solution of the optimisation model, the q_{nikd} variables, the corresponding OR capacity is left idle, and the bed capacity that was reserved for this patient is freed. In the second scheduling policy, we schedule patients based on specialty, the number of surgeons, and the remaining time available in an OR. For a patient to be scheduled, there must be enough time remaining in an OR scheduled for the right specialty, with enough surgeons available to perform the surgery, and there must be a bed available for the patient in a suitable ward for at least as many consecutive days as the LOS of the patient. The first patient (the one who has waited the longest) to fulfill these criteria is scheduled. Since the first scheduling policy applies information regarding the type of consultation (in the OC rooms) and the surgery category (in the ORs), we refer to this policy as the *Activity* (`Act`) policy. The second policy schedules patients based on specialty, and is therefore referred to as the *Specialty* (`Spec`) policy.

When applying the `Act` policy, we take advantage of the resource coordination provided by the optimisation model. If we compare the two scheduling policies, we can interpret the differences in outcome as the value of coordination. If the two scheduling policies are combined, and run successively, we may be able to utilise the capacity that is left idle after scheduling in accordance with the `Act` policy. In this case, the `Spec` policy may be thought of as having a list of patients that can be called and scheduled on short notice (the week before), if there is idle capacity. Performing activities on a short notice requires resource flexibility and responsiveness. Additional gains from combining the two policies can be interpreted as the value of flexibility.

When scheduling according to the `Act` policy, the solution from the optimisation model can be used to calculate the minimum possible queues that can be achieved. As a result, we may end up with queues that are shorter than the ones calculated based on the solution of the optimisation model.

3.6 Computational study

The aim of the computational study is to demonstrate how a department can apply the optimisation model to coordinate its OC and OR activities to decrease the queues of patients waiting to be served in either of the facilities. The study based on data from the Orthopaedic Department at St. Olav's Hospital. In addition to a base case, representing the present resource capacities at the Orthopaedic Department, we design multiple cases with slightly altered resource capacities to demonstrate how the department can temporarily alter its resources to enhance the system performance. Finally, we evaluate three operational scheduling policies to demonstrate the gains from scheduling activity types in addition to specialties when generating the master schedules. In accordance with the Orthopaedic Department, we use the term subspecialty instead of specialty throughout the computational study.

An Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz, 16 GB RAM computer is used when performing the computational study. The optimisation model is implemented in IVE Xpress 8.6, while the simulation model is written in Python 3.7, and the package SimPy. To perform the random sampling, the algorithms included in Python are used.

Table 3.5: The resource capacity cases

Case	Description
I_0	Base case
I_1	Homogeneous wards
I_2	Homogeneous ORs
I_3	All beds available during weekend
I_4	I_1 + I_3

3.6.1 Case descriptions

To establish the base case, data from the Orthopaedic Department at St. Olav's Hospital is used. All data necessary to define the base case are provided in Tables A.1 - A.8, in Appendix A.2. To sum up, there are 19 surgery categories divided among seven subspecialties. There are eight OC rooms and seven ORs, and one slot per day is used. The OC rooms can be used by all specialties, while the ORs are heterogeneous and can only be accessed by a subset of the specialties. The slot length at the OC is 240 min, while a slot lasts for 480 min at the operating theatre. Four wards are available, with a total capacity of 28 beds from Monday to Friday, and seven beds during the weekend. The wards are specialised, and an inpatient that has received surgery can only access a subset of the wards. The initial queues of patients waiting for an initial consultation at the start of the planning period are specified in Table A.9, in Appendix A.3.

In addition to the base case, labelled I_0, four resource capacity cases are investigated, and these are presented in Table 3.5. In I_1, the wards are treated as homogeneous, implying that each inpatient can be assigned to all wards. In practice, this would imply that the nurses, who serve the wards, must gain a wider competence such that they can handle patients outside their main field of competence. In I_2, homogeneous ORs are applied, meaning that each subspecialty can be assigned to all ORs. Room size and location can inhibit complete homogeneity between ORs, but by introducing similar equipment in all ORs, they can be more or less homogeneous. In I_3 and I_4, the bed capacity is not decreased during the weekend, and in the latter the wards are homogeneous.

3.6.2 The results from the optimisation model

One alteration has been made to the mathematical formulation when performing the optimisation study. To avoid unnecessary opening of the OC rooms, we introduce a small penalty of 0.1 for assigning specialties to the OC room slots (that is, we penalise the β_{jlsd} variables in the objective function).

The main results from running the five cases for three hours are presented in Table 3.6. The problem has not been solved to optimality in any of the cases, so we provide the best objective function values found, together with the upper bound and the dual gap. In addition, the aggregated number of activity types scheduled per week is presented. In the base case, 129 initial consultations are scheduled every week. By imposing homogeneous wards, we are able to increase the activity type by seven initial consultations, which is three more compared with leaving the bed capacity constant all week. Combining the two yields one more initial consultation compared with only having homogeneous wards. Also note that the number of surgeries increases in these instances, providing a higher reward in the objective function. By introducing homogeneous ORs, we can expect to provide two more initial consultations each week, compared with the base case, and one additional surgery. Based on these results, we can conclude that the beds are a scarce resource, and that the separation of wards imposes considerable restrictions for the patient throughput.

Table 3.6: Results from running the optimisation model for three hours. UB: Upper bound on objective value. IC: Number of ICs scheduled. TC: Number of TCs scheduled. FU: Number of FUs scheduled. S: Number of surgeries scehduled.

Case	**Obj func**	**UB**	**Gap**	**IC**	**TC**	**FU**	**S**
I_0	7058.13	7211.40	2.13%	129	31	138	65
I_1	7680.15	7918.00	3.00%	136	30	143	71
I_2	7231.16	7385.67	2.09%	131	32	141	66
I_3	7469.96	7859.53	4.96%	133	30	141	69
I_4	7846.27	8100.66	3.14%	137	30	146	74

In Tables 3.7 and 3.8, the number of initial consultations and surgeries scheduled are provided respectively. Compared with the base case, all other cases schedule more initial consultations, resulting in more surgeries being scheduled as well. When introducing homogeneous wards, less initial consultations from the hand and tumour subspecialties are scheduled, while the capacity increases for the remaining subspecialties. The reason for this is that these specialties impose less reward in the objective function. To take advantage of the increased bed capacity, specialties that provide a higher reward are prioritised. In the base case, the ward that houses the arthroplasty patients is closed during weekend. If the bed capacity is not reduced on Saturday and Sunday, the activity related to the arthroplasty patients can be increased. Also when applying homogeneous ORs, the capacity is increased for arthroplasty patients. In the base case, the arthroplasty subspecialty has access only to ORs six and seven, and due to the long LOS of these patients, all surgeries must be performed on Monday and Tuesday. When allowing for homogeneous ORs, the arthroplasty subspecialty gains access to all ORs, enabling the surgery of one additional patient.

In Table 3.9, the total resource consumption for one week in the different cases is given. Increasing the resource flexibility results in a higher resource consumption, which increases the surgeon workload. The total surgeon capacity is calculated as the number of surgeons available multiplied by five days. However,

Table 3.7: The number of initial consultations scheduled in the different cases

	Number of ICs							
Specialty	I_0	I_1	I_2	I_3	I_4	D_j	$\overline{D}_j$	R_j
Arthroscopy	17	19	17	17	17	17	20	42.4
Hand	23	20	23	23	20	19	23	35.1
Plastic	30	33	30	30	34	29	34	77.9
Arthroplasty	19	21	21	23	23	19	23	86.5
Reconstructive	18	20	18	20	20	18	22	46.1
Back	10	13	10	10	13	10	13	56.7
Tumour	12	10	12	10	10	10	12	13.2
Sum	**129**	**136**	**131**	**133**	**137**	**122**	**147**	

Table 3.8: The number of surgeries scheduled in the different cases

Subspecialty	**Surgery category**	I_0	I_1	I_2	I_3	I_4
Arthroscopy	Arthroscopy (aggregated)	2	3	2	2	2
Arthroscopy	ACL	1	2	1	1	1
Arthroscopy	Meniscus	1	1	1	1	1
Arthroscopy	Patellae	1	1	1	1	1
Hand	Hand (aggregated)	7	6	7	7	6
Hand	CTS	2	2	2	2	2
Plastic	Plastic (aggregated)	12	13	12	12	14
Plastic	Carsinoma	1	1	1	1	1
Plastic	BCC	3	3	3	3	4
Plastic	Malignant melanoma	5	5	5	5	6
Plastic	Cancer mammae	3	4	3	3	4
Plastic	SCC	2	2	2	2	2
Arthroplasty	Hip (primary)	7	8	8	9	9
Arthroplasty	Hip (revision)	2	2	2	3	3
Arthroplasty	Knee (primary)	5	5	5	5	5
Arthroplasty	Knee (revision)	1	1	1	1	1
Reconstructive	Reconstructive (aggregated)	6	7	6	7	7
Back	Back (aggregated)	2	3	2	2	3
Tumour	Tumour (aggregated)	2	2	2	2	2
Sum		**65**	**71**	**66**	**69**	**74**

Table 3.9: The scheduled use of resources. For the beds, some numbers are underlined to indicate that more beds are available in these cases.

	Resource usage					
Resource type	I_0	I_1	I_2	I_3	I_4	**Capacity per week**
OC room slots	40	40	40	40	40	40
OR slots	24	29	30	28	30	35
Bed days	134	154	138	149	161	154/196
Surgeon days	85	96	97	93	98	220

the surgeons have other duties to fulfill, such as serving the wards and the emergency department, and conducting research, so having 220 surgeon days available for OC and OR activities is not realistic. As can be seen from Table 3.10, there can be days where all the available surgeon hours are utilised for OC and OR activities. The table presents the maximum daily utilisation of surgeon hours available for each subspecialty. The surgeon types that can cover several specialties are added to the capacity of all the corresponding specialties when performing the calculation.

Table 3.10: The maximum daily utilisation of surgeon hours

Specialty	I_0	I_1	I_2	I_3	I_4
Arthroscopy	0.5	0.67	1.00	0.50	0.50
Hand	1.00	1.00	1.00	1.00	1.00
Plastic	1.00	1.00	1.00	1.00	1.00
Arthroplasty	0.64	0.55	0.64	0.55	0.55
Reconstructive	1.00	0.57	0.86	0.71	0.43
Back	1.00	0.60	1.00	0.60	0.60
Tumour	0.29	0.43	0.43	0.29	0.43

The expected OC room and OR utilisation for the different cases can be seen in Table 3.11. Here, the resource utilisation is given relative to the scheduled resource capacity. When regarding the OC rooms, we see that the utilisation increases when allowing for a more flexible use of resources. For these rooms, a utilisation of 100% is possible as we have set the duration of OC consultations to be 30 minutes, which is a multiple of the slot duration. For the ORs, the utilisation is low compared with in the OC rooms. The reason for this is the combination of surgery durations not adding up to full slots, and the bed capacity restricting the possible combinations of surgeries on a day.

Table 3.11: The planned utilisation of OC rooms and ORs. The values represent the utilisation of the scheduled time.

Resource type	I_0	I_1	I_2	I_3	I_4
OC room slots	93.13%	96.56%	95.00%	95.00%	97.81%
OR slots	66.82%	62.90%	54.22%	61.12%	61.97%

3.6.3 The simulation study

The aim of the simulation study is to evaluate the performance of the tactical master schedules provided by the optimisation model, and to test different scheduling policies that can be implemented for the operational scheduling of patients. Furthermore, the arrival of patients, and whether the patients require surgery, a treatment consultation in the OC, or if they leave the system following an initial consultation are modelled as stochastic processes. Because the number of follow-ups after an orthopaedic treatment is rather standardised, we assume that the number of follow-up consultations is deterministic. Furthermore, there is a delay of minimum three weeks before a follow-up consultation.

Experimental setup

For each of the cases analysed in the previous section, the solution obtained from the optimisation model is implemented in the simulation model. Then, 100 simulated weeks are run with the simulation model, and a total of 100 replications are performed for each case. We only report on the 25 first weeks, but a cool-down period is added to ensure that all patients arriving within the first 25 weeks have left the system by the end of the simulation.

As the system under study is not a steady state system, no warm-up is applied. Furthermore, the system is not empty when starting the simulation, and the queues are pre-filled with patients to mimic a realistic situation. For the two first weeks, the expected number of the different activities are pre-assigned. Then, for the two following weeks, half of the expected number of activities are pre-assigned. There is also an initial queue of patients waiting to be scheduled for an initial consultation at the beginning of week 1. The queues of pre-assigned activities and unscheduled initial consultations are specified in Tables A.9 and A.10 in Appendix A.3.

Figure 3.5, illustrates how the initial consultations performed in the first simulated week will cause a delayed downstream demand for the remaining services. From the first simulated week, the number of initial consultations performed is equal to the number of initial consultations performed in the solution of the optimisation model. The derived demand for treatment consultations and surgeries is evident from week two, while the derived demand for follow-up consultations is delayed to weeks five (first follow-up) and eight (second follow-up). As a consequence, the full impact of the altered scheduling regime will be evident from

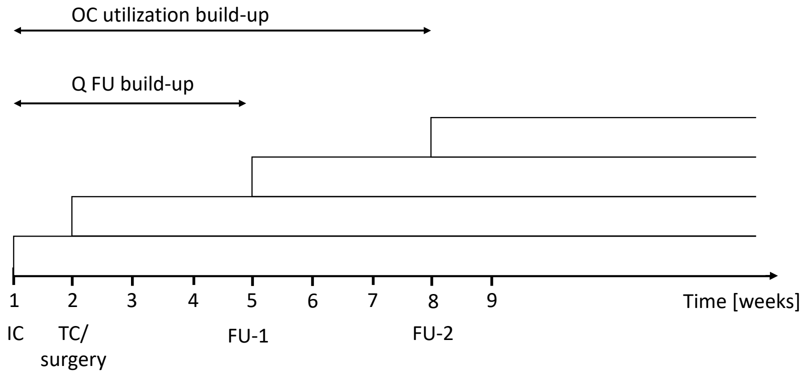


Figure 3.5: Illustration of the initial development of activity in the simulation model. The derived demand for treatment consultations and surgeries is evident from week two, while the derived demand for follow-ups consultations is delayed to weeks five (first follow-up) and eight (second follow-up).

week eight. Other consequences of the delay related to follow-up consultations is that the queue of follow-ups will increase the first five weeks, and that the OC utilisation will not peak until week eight.

We do not make assumptions about how long the patients that are already in queue have been waiting, so we only report on the waiting time of patients arriving after the start of the simulated time. Furthermore, for patients that belong to subspecialties that require two follow-up consultations, we sample with a 50% probability for each outcome, whether a prescheduled patient is scheduled for his first or second follow-up consultation.

Before comparing the different scheduling policies, we first present results where the `Act` policy is applied, as these can be compared directly to the results of the optimisation model.

The queues for OC consultations and surgery

In Figure 3.6, the mean queue length for OC consultations in total, initial consultations at the OC, and surgeries are provided for the five different cases. To calculate the total queue of OC consultations, the queue of initial, treatment and follow-up consultations are added together. Note that all follow-up consultations for a patient are added to the queue of follow-ups the week following either a treatment consultation or a surgery, even though there is a delay between these activities. Not surprisingly, the cases perform according to the rank of perfor-

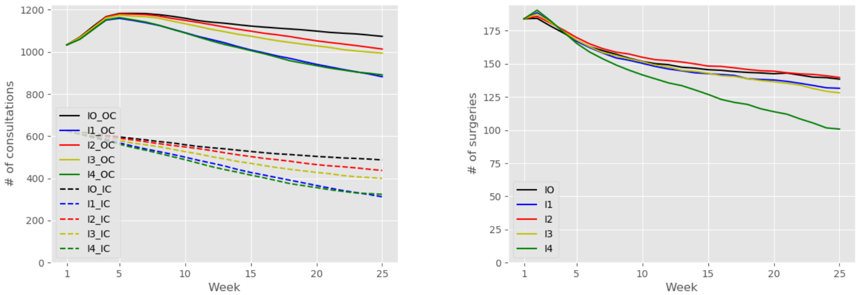


Figure 3.6: The mean queue length for OC consultations (left, solid lines), initial consultations (left, dashed lines) at the OC, and surgeries (right) in the five different cases

mance obtained from the solutions of the optimisation model. It is also evident that all queues decrease during the period. Initially, the queue of OC consultations increases, caused by the delay related to follow-up consultations. The number of initial consultations, and the relatively low throughput of surgeries in week one, leads to more patients being added to the queue for surgery than what is removed. Therefore, the queue for surgery increases from week one to two. However, from week two, the surgery rate increases and the queue decreases.

The resource utilisation and overall efficiency

The mean utilisation of the scheduled OC rooms, ORs, and the beds is displayed in Figure 3.7. Since we model the arrivals and paths of patients as stochastic processes, the demand for activities will deviate from the expected demand, causing some of the pre-scheduled activities to be unused. This tendency becomes more evident as the queues decrease, leading to a decrease in resource utilisation towards the end of the planning horizon.

When regarding the OC room utilization, the delayed demand for follow-up consultations yields a jump in utilization in weeks five and eight. As previously shown, the derived demand for surgeries will be present from week two, causing the jump in OR utilization from week one to two. The same effect causes similar behaviour for the bed utilization.

To measure the system efficiency achieved in the different cases, we register the mean time that patients, who have received all necessary activities, stayed in the system. Figure 3.8 illustrates the mean number of weeks that patients stay in the system as a function of when they arrive. In all cases, the patient waiting times decrease throughout the period, resulting in lower times spent in the system.

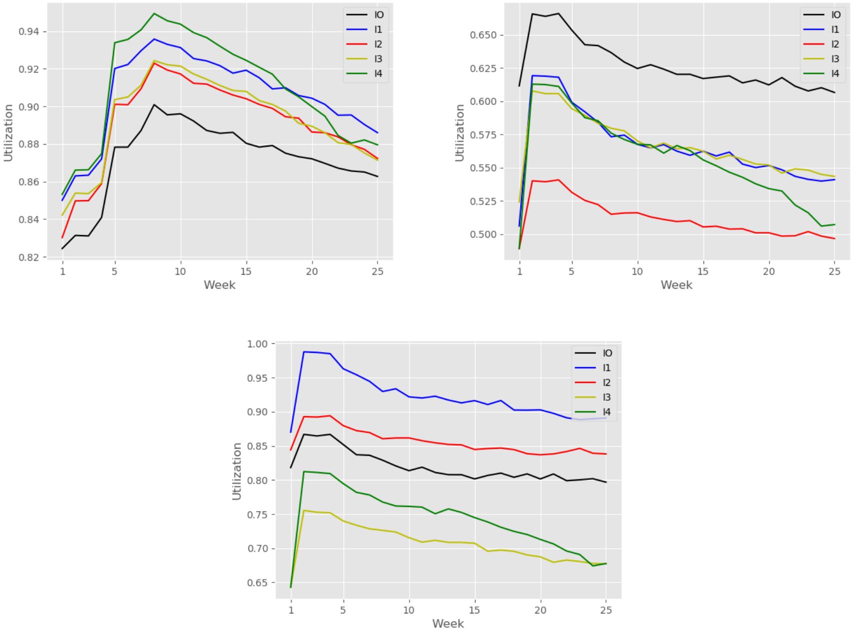


Figure 3.7: The mean utilisation of the scheduled OC rooms (top left), the scheduled ORs (top right), and the beds (bottom) in the five cases.

Evaluating different patient scheduling policies

In this section, three different scheduling policies are evaluated. The scheduling policies introduced in Section 3.5.2, the `Act` and the `Spec` policies, are evaluated individually. In the final scheduling policy, the two former policies are used successively, such that after having scheduled patients according to the optimisation model solution, we schedule additional activities if possible. This is equivalent to scheduling patients according to the `Act` policy, and then, if excess capacity is available, summon patients who can enter on a short notice. This can be achieved through establishing a calling list of patients who are willing and able to enter on a short notice. We refer to the final policy as the *Combined* (`Comb`) policy. When applying the `Act` policy, only the capacity that is pre-scheduled for activities can be utilised (see Table 3.11 to see how much of the scheduled capacity that can be utilised with this policy). However, when applying the two other policies, all the capacity scheduled for a specialty can be used, allowing for a higher resource utilisation in these policies. It is therefore not fair to compare the `Act` to the other policies, but we choose to include it to indicate the value of establishing a calling list.

The mean queue length for initial consultations, OC consultations, and surgery

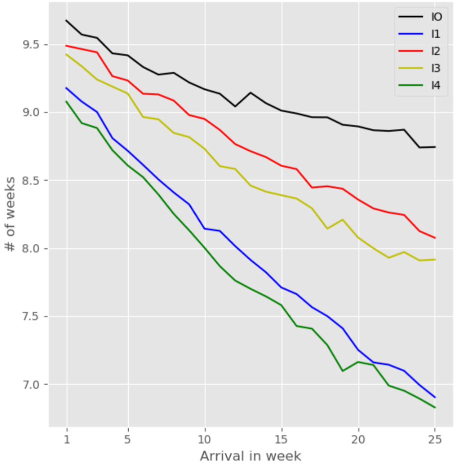


Figure 3.8: The mean number of weeks that patients stay in the system as a function of when they arrive

when applying the three different scheduling policies in I_1 can be seen in Figure 3.9. If patients are scheduled according to the Act policy, the queues obtained in a deterministic reality can be calculated from the solution of the optimisation model. This queue has also been added to the figure. Not surprisingly, the optimisation model solution outperforms the Act policy for all queues. Furthermore, the Comb policy outperforms the Act policy, indicating the value of flexibility in terms of a calling list.

When regarding the queues for initial and OC consultations in total, the Act performs worse than the two other scheduling policies because it has access to less resource capacity. However, when regarding the queue of surgeries, something interesting is observed. Due to the excessive scheduling of initial consultations during the first weeks, the queue of surgeries grows initially when applying either the the Spec or the Comb policy. Since the Comb policy ensures a coordination between initial consultations and surgeries, the queue for surgeries eventually decreases below the level seen with the Act policy. However, this is not the case with the Spec policy, where the queue of surgeries keeps growing. This clearly indicates the value of coordination, especially when downstream capacities are scarce.

To evaluate the efficiency obtained from the three scheduling policies, the mean number of weeks that patients stay in the system as a function of when they arrive, is displayed in Figure 3.10. The Comb policy clearly outperforms the Act policy, indicating the value of flexibility. As a consequence of poor coordination between initial consultations and surgeries, the Spec policy performs worse than the Act

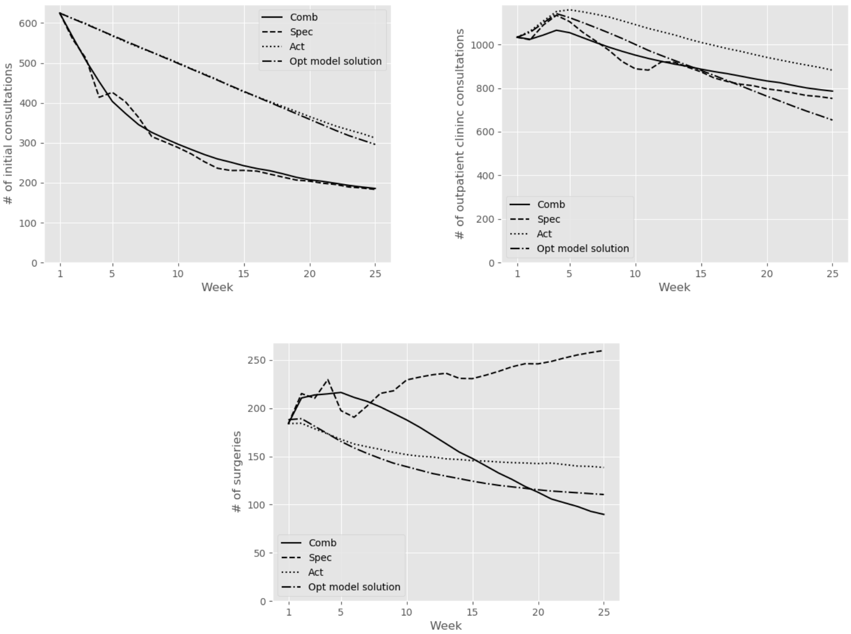


Figure 3.9: The mean queue length for initial consultations (top left), OC consultations (top right), and surgery (bottom) for the three scheduling policies in I_1.

policy when approaching the end of the period.

In Table 3.12, we present different measures from the end of the planning horizon [week 25] in the simulation model. In general, we observe that the `Comb` policy outperforms the other scheduling policies, stating the value of coordination in combination with resource flexibility. Furthermore, the `Spec` policy yields a longer queue of surgeries at the end of the planning horizon compared with the two other policies (except for I_4). The patients waiting for surgery in week 25, will eventually require one or two follow-up consultations in the OC. However, these are not counted in the total OC queue before the week following surgery. Therefore, there are relatively more OC consultations left to be performed in the `Spec` policy, compared with the other policies, than indicated by comparing the "Queue OCs". Note that the `Spec` policy performs well in I_4 where the bed capacity is much increased. This indicates that coordination is not crucial when downstream capacities are high.

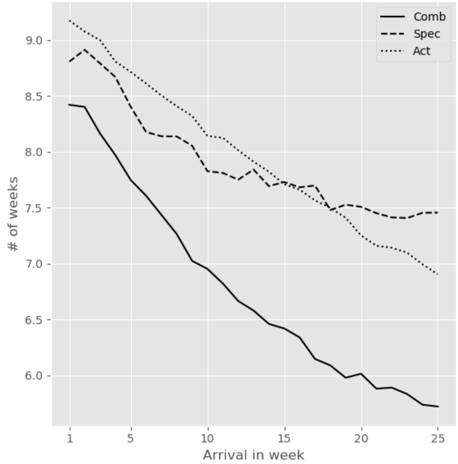


Figure 3.10: The mean number of weeks that patients stay in the system as a function of when they arrive for the three scheduling policies in I_1.

Table 3.12: Different measures in the last week of the planning horizon (week 25) for the `Act (A)`, `Spec (S)` and `Comb (C)` policies. The "Time in system" is the mean time [weeks] spent for patients arriving in week 25. "Throughput" is the mean number of patients that has left the system within week 25.

	Time in system			**Queue OR**			**Queue OC**			**Throughput**		
Case	A	S	C	A	S	C	A	S	C	A	S	C
I_0	8.74	10.58	6.33	138	288	201	1074	993	757	3175	3098	3421
I_1	6.90	7.46	5.72	132	260	90	882	751	778	3372	3382	3523
I_2	8.08	7.51	6.08	140	276	185	1013	781	738	3233	3335	3462
I_3	7.92	8.91	5.99	128	170	116	994	949	811	3269	3256	3470
I_4	6.83	5.91	5.82	101	64	60	891	825	812	3391	3505	3508

3.7 Managerial insights

In this section we list insights based on the findings of this paper, relevant for surgical departments where patients require both OC consultations and surgery.

- A hospital department is often measured by how fast it can provide initial consultations for its patients. Increasing the throughput of initial consultations without coordinating with downstream activities can harm the system efficiency and increase the total throughput time of patients. In our case, implementing the `Act` instead of the `Spec` policy decreases the average time in the system by 0.38 weeks (averaging over all resource capacity cases) even though the `Spec` has access to more capacity.

- Coordination between resources is particularly important when downstream resources are scarce. In our case, I_0 is most limited in terms of bed- and OR capacity, and this is the case where the throughput time of patients differs the most (1.84 weeks).

- Including information about what activity types to perform within each time slot is valuable when coordination is an issue. In our example, scheduling the surgery categories in addition to the specialty and the number of surgeons made it easier to achieve a high bed utilisation and to provide a high throughput of patients.

- Scheduling patients according to a pre-defined pattern settled at a tactical level can suffer from inflexibility causing unused capacities. Establishing a patient calling list or other flexible mechanisms can increase efficiency. In the case studied here, implementing the `Comb` instead of the `Act` policy decreases the average throughput time of patients by 1.71 weeks (averaging over all resource capacity settings).

3.8 Conclusion

In this paper we argue that surgical departments should consider the OC and the operating theatre simultaneously to ensure efficient handling of patients. Furthermore, we demonstrate how optimisation can be used as a tool to develop efficient master schedules for both facilities. Through simulation, we conclude that scheduling activity types, in addition to specialties, in the master schedules allow for a simple and efficient scheduling of individual patients. The value of scheduling activity types in the master schedules is to a large extent caused by the scheduling of patients for surgery, which is a complex task due to the need for coordination between the surgical activities and the bed capacity. Furthermore, a successful implementation of a patient calling list, which enables patients to be summoned for a consultation on short notice (the coming week), is beneficial for increasing patient throughput.

In our problem formulation, the resource capacities are leveled based on a demand for initial consultations, and a derived demand for surgery and downstream OR consultations. However, there can be situations where there are considerable queues also for surgery and downstream OC consultations at the beginning of a planning horizon. If the queue for initial consultations is long, and we schedule the capacities to decrease this queue, the corresponding capacities scheduled for surgery and downstream OC consultations will be sufficient to handle more than expected demand for treatment, and these queues will decrease as well. However, if the queues for initial consultations are short, while the others are long, these queues should be handled as separate demands, independent of the derived downstream demand.

There are some limitations in our study. First, we apply a deterministic optimisation model to solve a problem that has several stochastic parameters. Applying a stochastic model could allow us to capture more of the inherent uncertainties when generating the master schedules, and introduce flexible mechanisms to handle the uncertainties. Furthermore, due to our relatively simple simulation model, we cannot give the complete picture of how the schedules and the scheduling policies will perform in a real-life setting. A recommendation for future research is to consider both uncertain patient arrivals and activity demands in the optimisation model. Capturing these uncertainties and proposing mechanisms for handling them will be of great value to hospitals that face variations in demand.

Bibliography

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.

A. Ahmadi-Javid, Z. Jalali, and K. J. Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3 – 34, 2017.

I. A. Bikker, N. Kortbeek, R. M. van Os, and R. J. Boucherie. Reducing access times for radiation treatment by aligning the doctor's schemes. *Operations Research for Health Care*, 7:111 – 121, 2015.

A. Braaksma, N. Kortbeek, G.F. Post, and F. Nollet. Integral multidisciplinary rehabilitation treatment planning. *Operations Research for Health Care*, 3(3): 145 – 159, 2014. ISSN 2211-6923.

B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201 (3):921 – 932, 2010.

N. Freeman, M. Zhao, and S. Melouk. An iterative approach for case mix planning under uncertainty. *Omega*, 76:160 – 173, 2018.

A. Fügener, E. W. Hans, R. Kolisch, N. Kortbeek, and P. T. Vanberkel. Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, 239(1):227 – 236, 2014.

S. Hahn-Goldberg, M. W. Carter, J. C. Beck, M. Trudeau, P. Sousa, and K. Beattie. Dynamic optimization of chemotherapy outpatient scheduling with uncertainty. *Health Care Management Science*, 17(4):379–392, 2014. ISSN 1572-9389.

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

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

A. G. Leeftink, I. A. Bikker, I. M. H. Vliegen, and R. J. Boucherie. Multidisciplinary planning in health care: a review. *Health Systems*, 9(2):95–118, 2020.

X. Li, N. Rafaliya, M. F. Baki, and B. A. Chaouch. Scheduling elective surgeries: the tradeoff among bed capacity, waiting patients and operating room utilization using goal programming. *Health Care Management Science*, 20(1):33–54, 2017.

B. Liang, A. Turkcan, M. E. Ceyhan, and K. Stuart. Improvement of chemotherapy patient flow and scheduling in an outpatient oncology clinic. *International Journal of Production Research*, 53(24):7177–7190, 2015.

J. Marynissen and E. Demeulemeester. Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, 272(2):407 – 419, 2019. ISSN 0377-2217.

T. Monks, C. S. M. Currie, B. S. Onggo, S. Robinson, M. Kunc, and S. J. E. Taylor. Strengthening the reporting of empirical simulation studies: Introducing the stress guidelines. *Journal of Simulation*, 13(1):55–67, 2019.

A. Moosavi and S. Ebrahimnejad. Robust operating room planning considering upstream and downstream units: A new two-stage heuristic algorithm. *Computers & Industrial Engineering*, 143:106387, 2020.

T. B. T. Nguyen, A. I. Sivakumar, and S. C. Graves. A network flow approach for tactical resource planning in outpatient clinics. *Health Care Management Science*, 18(2):124–136, 2015.

A.J. T. Schneider, J. T. van Essen, M. Carlier, and E. W. Hans. Scheduling surgery groups considering multiple downstream resources. *European Journal of Operational Research*, 282(2):741–752, 2020. ISSN 0377-2217.

J.T. Van Essen, E.W. Hans, J.L. Hurink, and A. Oversberg. Minimizing the waiting time for emergency surgery. *Operations Research for Health Care*, 1(2): 34 – 44, 2012.

M. B. Vik, H. Finnestrand, and R. L. Flood. Systemic problem structuring in a complex hospital environment using viable system diagnosis – keeping the blood flowing. *Systemic Practice and Action Research*, 35:203–226, 2022.

Chapter A

Appendices

A.1 Algorithms for describing the DES model

Algorithm 1: The algorithm for scheduling patients to the OC rooms according to the solutions of the optimisation model. Referred to as the `Act` scheduling policy.

```
input: Queue_referred, Queue_treatment, Queue_follow_up,
       OC_availability
for $d \leftarrow 1$ to $5$ do
    for $l \leftarrow 1$ to $Number\_of\_OC$ do
        for $j \leftarrow 1$ to $Specialties$ do
            $N_1 = min\{x_{j1ld}, length(Queue\_referred)\}$;
            $N_2 = min\{x_{j2ld}, length(Queue\_treatment)\}$;
            $N_3 = min\{x_{j3ld}, length(Queue\_follow\_up)\}$;
            for $i \leftarrow 1$ to $N_1$ do
                Subtract the consultation duration from OC_availability;
                Remove the first patient from Queue_referred;
            end
            for $i \leftarrow 1$ to $N_2$ do
                Subtract the consultation duration from OC_availability;
                Remove the first patient from Queue_treatment;
            end
            for $i \leftarrow 1$ to $N_3$ do
                Subtract the consultation duration from OC_availability;
                Remove the first patient from Queue_follow_up;
            end
        end
    end
end
```

Algorithm 2: The algorithm for utilising idle OC room capacity after having scheduled patients according to the solutions of the optimisation model. Referred to as the ***Spec*** scheduling policy.

```
input  : Queue_referred, Queue_treatment, Queue_follow_up,
         OC_availability
for $d \leftarrow 1$ to 5 do
    for $l \leftarrow 1$ to Number_of_OC do
        for $j \leftarrow 1$ to Specialties do
            check=1;
            while length(List_of_candidates) $> 0$ or check $= 1$ do
                check = 0;
                List_of_candidates=[];
                if consultation duration of Queue_referred[0] fits into
                  remaining OC_availability then
                    Append patient to List_of_candidates;
                end
                if consultation duration of Queue_treatment[0] fits into
                  remaining OC_availability then
                    Append patient to List_of_candidates;
                end
                if consultation duration of Queue_follow_up[0] fits into
                  remaining OC_availability then
                    Append patient to List_of_candidates;
                end
                Choose the patient from List_of_candidates that has
                  waited the longest;
                Subtract the consultation duration of the chosen patient
                  from OC_availability;
                Remove the chosen patient from List_of_candidates;
                Remove the chosen patient from the corresponding queue;
            end
        end
    end
end
```

A.2 Data for the base case

Table A.1: Base case: the sets

Set	**# of elements**
Days	7
Subspecialties	7
Surgery categories	19
OC rooms	8
ORs	7
Wards	4
Activity types performed at the OC	3
Surgeons present	2
Surgeon types	17
Slots	1

Table A.2: Base case: the subspecialties

	Availability		**Duration [min]**			
Specialty	**OCRs**	**ORs**	**IC**	**TC**	**FC**	**Reward**
Arthroscopy	All	3 and 4	30	30	30	42.4
Hand	All	3	30	30	30	35.1
Plastic	All	1, 2 and 3	30	30	30	77.9
Arthroplasty	All	6 and 7	30	30	30	86.5
Reconstructive	All	1 and 5	30	30	30	46.1
Back	All	5	30	30	30	56.7
Tumour	All	5	30	30	30	13.2

Table A.3: Base case: the surgeons

Surgeon type	Subspecialty	# of surgeons	Experience
Arthroscopy 1	Arthroscopy	4	Consultants
Arthroscopy 2	Arthroscopy	2	Residents
Hand 1	Hand	2	Consultants
Hand 2	Hand	1	Residents
Plastic 1	Plastic	4	Consultants
Plastic 2	Plastic	4	Residents
Arthroplasty 1	Arthroplasty	6	Consultants
Arthroplasty 2	Arthroplasty	3	Residents
Reconstructive 1	Reconstructive	4	Consultants
Reconstructive 2	Reconstructive	2	Residents
Back 1	Back	4	Consultants
Back 2	Back	1	Residents
Tumour 1	Tumour	3	Consultants
Tumour 2	Tumour	1	Residents
Cons 1	Arthroplasty and tumour	1	Consultant
Cons 2	Arthroplasty and tumour	1	Consultant
Cons 3	Reconstructive and tumour	1	Consultant

Table A.4: Base case: the number of beds available

Ward	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Trauma	4	4	4	4	4	2	2
Reconstructive	5	5	5	5	5	3	3
Elective	3	3	3	3	3	2	2
FT	16	16	16	16	16	0	0

Table A.5: Base case: the surgery categories

Surgery cat.	**Subspecialty**	**Duration [min]**	**# of Surgeons**	**Ward**	**LOS [days]**
Arthro. (agg.)	Arthroscopy	174	2	El.	2
ACL	Arthroscopy	173	1	El.	2
Meniscus	Arthroscopy	103	2	-	0
Patellae	Arthroscopy	176	2	El.	1
Hand (agg.)	Hand	107	2	-	0
CTS	Hand	54	2	Tr.	1
Plastic (agg.)	Plastic	108	2	Tr., Rec.	2
Carsinoma	Plastic	52	1	Rec.	1
BCC	Plastic	59	2	Tr., Rec., FT.	1
Mal. melanoma	Plastic	85	1	-	0
Cancer mammae	Plastic	146	1	Rec.	1
SCC	Plastic	65	2	Rec.	1
Hip (primary)	Arthroplasty	110	2	FT.	4
Hip (revision)	Arthroplasty	152	2	FT.	4
Knee (primary)	Arthroplasty	122	2	FT.	4
Knee (revision)	Arthroplasty	165	2	FT.	4
Recon. (agg.)	Reconstructive	145	2	Rec.	2
Back (agg.)	Back	309	2	El.	6
Tumour (agg.)	Tumour	93	1	Rec.	1

Table A.6: Base case: the flow of patients at the OC

Subspecialty	**Exp. demand**	**Max demand**	**Share to TC after I**	**Share to FU after T**
Arthroscopy	17	20	0.01	1
Hand	19	23	0.55	2
Plastic	29	34	0.20	1
Arthroplasty	19	23	0.01	1
Reconstructive	18	22	0.06	2
Back	10	13	0.01	2
Tumour	10	12	0.53	2

Table A.7: Base case: the flow of patients at the operating theatre

Surgery category	**Share to surgery after IC**	**Share to FU after surgery**	**Add surg demand**
Arthroscopy (aggregated)	0.11	1	0
ACL	0.06	1	0
Meniscus	0.04	1	0
Patellae	0.05	1	0
Hand (aggregated)	0.29	2	0
CTS	0.08	2	0
Plastic (aggregated)	0.39	1	0
Carsinoma	0.02	1	0
BCC	0.09	1	0
Malignant melanoma	0.15	1	0
Cancer mammae	0.10	1	0
SCC	0.03	1	0
Hip (primary)	0.37	1	0
Hip (revision)	0.09	1	0
Knee (primary)	0.21	1	0
Knee (revision)	0.04	1	0
Reconstructive (aggregated)	0.32	2	0
Back (aggregated)	0.18	2	0
Tumour (aggregated)	0.14	2	0

Table A.8: Base case: slots

Location	**# of slots**	**Time available per slot [min]**
Outpatient clinic	1	240
Operating theatre	1	480

A.3 Initial conditions for the DES model

Table A.9: The queue of patients for the OC when starting the simulation. Except from the queue of initial consultations that are not yet scheduled, the remaining consultations are scheduled within the 4 first weeks.

Subspecialty	**Scheduled for IC**	**Scheduled for TC**	**Scheduled for FU**	**IC not yet scheduled**
Arthroscopy	52	0	12	23
Hand	58	30	102	42
Plastic	88	18	84	37
Arthroplasty	58	0	42	42
Reconstructive	56	6	42	44
Back	32	0	12	43
Tumour	30	18	42	20

Table A.10: The queue of patients for surgery when starting the simulation. The surgeries are scheduled within the 4 first weeks.

Surgery category	**# of patients in queue**
Arthroscopy (aggregated)	6
ACL	4
Meniscus	4
Patellae	4
Hand (aggregated)	16
CTS	6
Plastic (aggregated)	34
Carsinoma	4
BCC	6
Malignant melanoma	12
Cancer mammae	12
SCC	4
Hip (primary)	22
Hip (revision)	6
Knee (primary)	12
Knee (revision)	4
Reconstructive (aggregated)	18
Back (aggregated)	6
Tumour (aggregated)	6

Paper III

T. R. Bovim, A. N. Gullhav, H. Andersson, A. Riise:

A framework for integrated resource planning in surgical clinics

This paper is submitted for publication and is therefore not included.

Submitted to an international journal

Chapter A

Appendices

A.1 The notation related to the LMSP

The notation used in the LMSP is presented in Tables A.1 to A.4.

Table A.1: Sets

Symbol	**Description**	
$\mathcal{I}$	Planning periods in the planning horizon	$i \in \mathcal{I}$
$\mathcal{D}$	Days covering the maximum activity delay, the planning delay and the planning horizon	$d \in \mathcal{D}$
$\mathcal{D}^{PD}$	Days covering the planning delay and the planning horizon	$d \in \mathcal{D}^{PD} \subseteq \mathcal{D}$
$\mathcal{D}^T$	Days in the planning horizon	$d \in \mathcal{D}^T \subseteq \mathcal{D}$
$\mathcal{D}^I_i$	Days in planning period i	$d \in \mathcal{D}^I_i \subseteq \mathcal{D}^T$
$\mathcal{D}^C$	Days in a planning cycle	$d \in \mathcal{D}^C$
$\mathcal{D}^T_{d'}$	Days in the planning horizon that correspond to cycle day d'	$d \in \mathcal{D}^T \mid d \mod \lvert D^C \rvert = d'$
$\mathcal{D}^Q$	Days when we measure the waiting lists	$d \in \mathcal{D}^Q \subseteq \mathcal{D}^T$
$\mathcal{D}^{LOS}_{ad}$	Dys that a patient who stays in a ward on day d can have had a surgery of type a	$d' \in \mathcal{D}^{LOS}_{ad}$
$\mathcal{U}$	Units	$u \in \mathcal{U}$
$\mathcal{J}$	Surgical specialties	$j \in \mathcal{J}$
$\mathcal{P}$	Surgeon types	$p \in \mathcal{P}$
$\mathcal{W}$	Wards	$w \in \mathcal{W}$
$\mathcal{B}$	OR activity blocks	$b \in \mathcal{B}$
$\mathcal{A}$	Activity types	$a \in \mathcal{A}$
$\mathcal{A}^{OC}$	OC activity types	$a \in \mathcal{A}^{OC} \subseteq \mathcal{A}$
$\mathcal{A}^{OT}$	Surgery activity types	$a \in \mathcal{A}^{OT} \subseteq \mathcal{A}$
$\mathcal{K}$	Waiting list intervals	$k \in \mathcal{K}$
$\mathcal{P}^C_j$	Consultant types that can cover specialty j	$p \in \mathcal{P}^C_j \subseteq \mathcal{P}$
$\mathcal{P}^R_j$	Resident types that can cover specialty j	$p \in \mathcal{P}^R_j \subseteq \mathcal{P}$
$\mathcal{W}^A_a$	Wards that can house patients who received activity type a	$w \in \mathcal{W}^A_a \subseteq \mathcal{W}$
$\mathcal{B}^J_j$	OR activity blocks available for specialty j	$b \in \mathcal{B}^J_j \subseteq \mathcal{B}$
$\mathcal{A}^J_j$	Activity types that can be handled by specialty j	$a \in \mathcal{A}^J_j \subseteq \mathcal{A}$
$\mathcal{A}^{OT}_j$	Surgery activity types that can be handled by specialty j	$a \in \mathcal{A}^{OT}_j \subseteq \mathcal{A}^J_j$
$\mathcal{A}^W_w$	Surgery activity types that can rest in ward w following surgery	$a \in \mathcal{A}^W_w \subseteq \mathcal{A}^{OT}$

Table A.2: Parameters

Symbol	Description
R_u	Number of rooms available in unit u
V_{ujd}	Number of rooms that can be accessed in unit u by specialty j on cycle day d
C_u^N	Number of room-days that can be accessed in unit u during the planning cycle
B_u^F	Number of room-days that must be assigned as flexible in unit u through the planning cycle
T^{OC}	Time available in an OC room-day
C_{pd}	Number of surgeons available of surgeon type p on cycle day d
C_{pi}^{MAX}	Maximum number of days available for surgeon type p during planning period i
N_b^B	Number of surgeons that must be present to assign OR activity block b
A_{wd}	Number of staffed beds available in ward w on day d in the planning cycle
X_{jad}	Number of activities of type a and specialty j (expected to be) performed on day d, before the planning horizon
L_{ja}	Expected external arrival rate of activity type a and specialty j
$F_{jaa'}$	Fraction of activity of type a that yields a downstream demand for activity of type a' for specialty j
D_{ja}^{OC}	Duration of OC activity type a, specialty j
A_{bja}^B	Number of patients from specialty j and activity type a that are assigned to OR activity block b
D^P	Number of days in the planning delay
D_{ja}^A	Number of days in the activity delay after activity type a, specialty j
Q_{ja}^0	Number of patients on the waiting lists for specialty j and activity type a on at the day of planning
Q_{jak}	Maximum number of patients that can be assigned to the waiting list of specialty j, activity type a and interval k
$\overline{C}_{jak}$	Penalty coefficient associated with the waiting list of specialty j, activity type a and interval k

Table A.3: The high-level variables.

Symbol	Description
β_{ujd}	Number of rooms assigned to unit u and specialty j on cycle day d
y_{ud}	Number of rooms in unit u assigned as flexible on cycle day d
μ_{jd}^{OR}	Maximum number of ORs that can be assigned as flexible for specialty j on cycle day d

Table A.4: The low-level variables.

Symbol	Description
λ_{ujd}	Number of rooms in unit u used by specialty j on day d
y_{ujd}	Number of flexible rooms in unit u assigned for specialty j on day d
g_{pjd}	Number of surgeons from surgeon type p that cover specialty j on day d
x_{bd}^{OR}	Number of OR blocks of type b assigned to day d
x_{jad}	Number of activities of type a assigned to specialty j on day d
u_{awd}	Number of beds occupied in ward w on day d, by patients who received activity type a
q_{jad}	Number of patients on the waiting list of specialty j and activity type a on day d
$\overline{q}_{jadk}$	Number of patients on the waiting list of specialty j and activity type a on day d, within interval k

A.2 Input data for the LMSP

In Tables A.5 to A.12, we provide the data applied to define the Small, Medium and Large cases in the LMSP.

Table A.5: The main sets

Set	**Symbol**	**# of elements**		
		Small	Medium	Large
Planning periods	$\mathcal{I}$	3	3	3
Days in planning horizon	$\mathcal{D}^T$	84	84	84
Days in planning cycle	$\mathcal{D}^C$	7	7	7
Days when we measure the waiting lists	$\mathcal{D}^Q$	12	12	12
Surgical specialties	$\mathcal{J}$	3	3	7
OC activity types	$\mathcal{A}^{OC}$	3	3	3
Surgery activity types	$\mathcal{A}^{OT}$	7	12	19
Wards	$\mathcal{W}$	3	3	4
Surgeon types	$\mathcal{P}$	6	6	14
Waiting list intervals	$\mathcal{K}$	3	3	3

Table A.6: The availability of surgeons on weekdays, and the maximum number of days available for clinical work in a planning period.

Surgeon type	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**	**Days in planning period**		
						Small	Medium	Large
Arthroscopy C	2	2	1	3	2	30	30	20
Arthroscopy R	2	1	2	1	2	15	15	15
Hand C	2	1	3	1	1	30	30	24
Hand R	1	2	1	2	1	16	16	16
Plastic C	2	2	2	1	3		35	30
Plastic R	3	1	2	2	2		34	34
Arthroplasty C	2	2	2	1	0			24
Arthroplasty R	2	2	1	1	0			24
Reconstructive C	2	2	0	1	1	20		20
Reconstructive R	0	1	1	1	0	16		26
Back C	1	2	1	2	1			24
Back R	2	1	2	1	1			16
Tumour C	1	0	1	0	1			15
Tumour R	0	1	0	1	1			12

Table A.7: The number of beds available

Ward	**Small**							**Medium**							**Large**						
	M	**T**	**W**	**T**	**F**	**S**	**S**	**M**	**T**	**W**	**T**	**F**	**S**	**S**	**M**	**T**	**W**	**T**	**F**	**S**	**S**
Trauma	4	4	4	4	4	2	2	4	4	4	4	4	2	2	4	4	4	4	4	2	2
Reconstructive	5	5	5	5	5	3	3	4	4	4	4	4	3	3	5	5	5	5	5	3	3
Elective	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
FT															16	16	16	16	16	0	0

Table A.8: The rooms

Location	**Availability [min]**	**Number of rooms**			**Number of room-days**		
		Small	Medium	Large	Small	Medium	Large
Outpatient clinic	240	4	4	8	14-17	17-20	34-37
Operating theatre	480	3	4	7	4-8	7-11	16-20

Table A.9: OC activity types

Activity type	**Specialty**	**Duration [min]**	Q_{ja1}	Q_{ja2}	Q_{ja3}	$\overline{C}_{ja1}$	$\overline{C}_{ja2}$	$\overline{C}_{ja3}$
IC	Arthroscopy	30	34.0	17.0	∞	5	15	100
TC	Arthroscopy	30	1.7	0.9	∞	2	5	50
FU	Arthroscopy	30	26.6	13.3	∞	0	1	30
IC	Hand	30	38.0	19.0	∞	5	15	100
TC	Hand	30	20.1	10.1	∞	2	5	50
FU	Hand	30	78.8	39.4	∞	0	1	30
IC	Plastic	30	58.0	29.0	∞	5	15	100
TC	Plastic	30	11.6	5.8	∞	2	5	50
FU	Plastic	30	90.1	45.0	∞	0	1	30
IC	Arthroplasty	30	38.0	19.0	∞	5	15	100
TC	Arthroplasty	30	1.9	1.0	∞	2	5	50
FU	Arthroplasty	30	29.3	14.6	∞	0	1	30
IC	Reconstructive	30	36.0	18.0	∞	5	15	100
TC	Reconstructive	30	2.2	1.0	∞	2	5	50
FU	Reconstructive	30	41.0	20.6	∞	0	1	30
IC	Back	30	10.0	5.0	∞	5	15	100
TC	Back	30	0.5	0.3	∞	2	5	50
FU	Back	30	13.8	6.9	∞	0	1	30
IC	Tumour	30	10.0	5.0	∞	5	15	100
TC	Tumour	30	5.3	2.7	∞	2	5	50
FU	Tumour	30	40.2	20.1	∞	0	1	30

Table A.10: Surgery types

Surg. type	**Specialty**	**Dur. [min]**	**# surgeons**	**Ward**	**LOS [days]**	Q_{ja1}	Q_{ja2}	Q_{ja3}	$\overline{C}_{ja1}$	$\overline{C}_{ja2}$	$\overline{C}_{ja3}$
Arthro. (agg.)	Arthroscopy	174	2	El.	2	5.1	7.7	∞	2	5	50
ACL	Arthroscopy	173	1	El.	2	4.1	6.1	∞	2	5	50
Meniscus	Arthroscopy	103	2	-	0	3.7	5.6	∞	2	5	50
Patellae	Arthroscopy	176	2	El.	1	3.4	5.1	∞	2	5	50
Hand (agg.)	Hand	107	2	-	0	11.4	17.1	∞	2	5	50
CTS	Hand	54	2	Tr.	1	3.8	5.7	∞	2	5	50
Plastic (agg.)	Plastic	108	2	Tr., Recon.	2	17.4	26.1	∞	2	5	50
Carsinoma	Plastic	52	1	Recon.	1	5.8	8.7	∞	2	5	50
BCC	Plastic	59	2	Tr., Recon.	1	2.9	4.4	∞	2	5	50
Mal. mel.	Plastic	85	1	-	0	8.7	13.1	∞	2	5	50
Cancer m.	Plastic	146	1	Recon.	1	5.8	8.7	∞	2	5	50
SCC	Plastic	65	2	Recon., El.	1	2.9	4.4	∞	2	5	50
Hip (primary)	Arthroplasty	110	2	FT	4	14.1	21.1	∞	2	5	50
Hip (revision)	Arthroplasty	152	2	FT	4	3.4	5.3	∞	2	5	50
Knee (primary)	Arthroplasty	122	2	FT	4	8.0	12.0	∞	2	5	50
Knee (revision)	Arthroplasty	165	2	FT	4	1.9	2.9	∞	2	5	50
Recon. (agg.)	Reconstructive	145	2	Recon.	2	5.8	8.6	∞	2	5	50
Back (agg.)	Back	309	2	El.	6	1.8	2.7	∞	2	5	50
Tumour (agg.)	Tumour	93	1	Recon.	1	1.4	2.1	∞	2	5	50

Table A.11: The flow of patients at the OC

Specialty	**Expected # of new IC per day**	**Share to TC after IC**	**Share to FU after TC**
Arthroscopy	2.43	0.05	1
Hand	2.71	0.53	2
Plastic	4.14	0.20	1
Arthroplasty	2.71	0.05	1
Reconstructive	2.57	0.06	2
Back	1.43	0.05	2
Tumour	1.43	0.53	2

Table A.12: The flow of patients at the operating theatre

Surgery category	**Share to surgery after IC**	**Share to FU after surgery**
Arthroscopy (aggregated)	0.15	1
ACL	0.12	1
Meniscus	0.11	1
Patellae	0.10	1
Hand (aggregated)	0.30	1
CTS	0.10	1
Plastic (aggregated)	0.30	1
Carsinoma	0.10	2
BCC	0.05	1
Malignant melanoma	0.15	1
Cancer mammae	0.10	2
SCC	0.05	2
Hip (primary)	0.37	1
Hip (revision)	0.09	1
Knee (primary)	0.21	1
Knee (revision)	0.05	1
Reconstructive (aggregated)	0.32	2
Back (aggregated)	0.18	2
Tumour (aggregated)	0.14	2

Paper IV

T. R. Bovim, A. N. Gullhav, H. Andersson, J. Dale, K. Karlsen:

Simulating emergency patient flow during the COVID-19 pandemic

Published in Journal of Simulation

Chapter 5

Simulating emergency patient flow during the COVID-19 pandemic

Abstract

The work presented in this paper is based on two projects that were conducted at St. Olavs Hospital (Norway) when preparing for the COVID-19 pandemic. During this period of time, there was a growing concern in the hospital management that the emergency department (ED) and the ambulance services would collapse under the increased demand for service related to the testing and transportation of COVID-19 patients.

Three discrete event simulation models are provided to evaluate the resource requirements during the peak of the pandemic. In the first model, we estimate the number of beds needed in the ED. In the second model, we estimate the number of ambulances required to maintain prepandemic response times for urgent emergency patients. The third model is an implicit coupling of the two former models, and it is used to study the effects of ED boarding time for patients that are transferred from the ED to COVID-19 ward. To reduce the modeling complexity, the capacity available in the COVID-19 ward is managed through counting rules. This method can be used under the assumption that the boarding time does not affect the total length of stay of patients.

The models are used to analyze the resource needs under different COVID-19 testing policies. A strict testing policy increases the need for beds in the ED, while it has the opposite effect in terms of the number of ambulances required. Using the third model, two distinct mechanisms causing boarding time are analyzed: *excessive-flow-induced* boarding and *ambulance-induced* boarding. The relative effect of these depends on the testing policy implemented by the hospital management. We also find that the effects from extended boarding time are most prominent during night and weekends.

5.1 Introduction

The COVID-19 pandemic has put the health care sector in many countries under pressure. In Norway, societal restrictions, such as closing down public institutions and instructing social distancing, were imposed on the 12th of March 2020. Moreover, the hospitals reduced the elective patient activity to free resource capacity, resulting in a decrease in the number of inpatient stays in March and April by 39% compared with the same period in 2019. Furthermore, presumably due to less accidents and the fact that people are reluctant towards seeking medical assistance in danger of becoming infected, the activity related to emergency patients decreased by 19% in the same period compared with 2019 (The Norwegian Directorate of Health, 2020).

The main contribution of this paper is to demonstrate how discrete event simulation (DES) can be used to provide decision support for the hospital management when preparing for the pandemic. The second contribution is a novel approach to model boarding time in the emergency department (ED). Boarding occurs when downstream units are not able to serve patients at the rate at which the patients are ready to leave the ED, causing additional demands for beds in the ED. Boarding time is defined as the time between the decision made by a physician to admit a patient and the time the patient leaves the ED to an inpatient unit (Tang et al., 2015).

St. Olavs Hospital is a university hospital located in Trondheim, Norway, treating about 60 000 inpatients each year. The work presented in this paper is based on two projects that were conducted at St. Olavs Hospital between the 17th of March and the 29th of March 2020. The first project was conducted for the ED, and the second for the ambulance services. In each project, one DES model was developed, and eventually these were implicitly combined into a third model. During this period of time, the hospital management proposed that all COVID-19 suspected patients that enter the hospital should be tested for COVID-19 in the emergency department (ED). Furthermore, these patients must be transported by ambulance when going to the hospital, and this also applies to patients that are not confirmed to be COVID-19 negative upon departure from the hospital. To evaluate these proposals, the hospital management required to estimate the need for both additional beds in the ED, and additional ambulances during the peak of the pandemic.

On the 12th of March 2020, the Norwegian Institute of Public Health (NIPH) released a ”recommended planning scenario” for the evolvement of the COVID-19 pandemic in Norway, which aimed to provide the Norwegian hospitals with support when preparing for the pandemic. On the 24th of March 2020, the recommended planning scenario was updated with a higher number of COVID-19 patients hospitalized at peak of the pandemic. Both scenarios were used as input for our three models.

The rest of the paper is outlined as follows. In Section 5.2, relevant literature is presented to provide a context for our contribution. Then, in Section 5.3 we present the objectives of the study, the basic assumptions, the logic of the models and the data used to perform the studies. The scenarios considered for analysis are presented in Section 5.4, while the simulation results are provided in Section 5.5. Finally, in Section 5.6 we discuss the main implications of our findings and conclude the paper.

5.2 Literature

ED crowding, a consequence of a simultaneous increase in the demand for health care and a deficit in available hospital and ED beds, has become a significant public health problem (Bair et al., 2010). A growing body of evidence suggests that ED crowding is linked to adverse quality of care, such as medication errors, patient dissatisfaction and staff burnout (Valipoor et al., 2021). One cause of ED crowding is boarding patients that experience a delay in transfer to hospital wards (Tang et al., 2015). Only 7% of the papers reviewed by Vanbrabant et al. (2019) include boarding time as a key performance indicator, but they were all published in the last 10 years. This confirms the growing interest in ED boarding as a research topic within operations research.

To model ED boarding time, downstream units should be regarded. However, this adds modeling complexity, and some authors sample boarding times to omit this complexity (De Boeck et al., 2019; Carmen et al., 2014; Bair et al., 2010). Other contributions, like Kolb et al. (2007, 2008) explicitly include the inpatient unit to obtain realistic boarding patterns. Kolb et al. (2007) investigate the effect of the inpatient unit utilization on ED crowding, while Kolb et al. (2008) evaluate the effect of different buffer concepts. Wood and Murch (2020) develop a continuous Markov chain to model a stroke pathway with different units. Unit capacities are part of the model formulation, and capacity shortage induces delays in patient transfer.

In this paper, we model ED boarding through an implicit coupling of two models, where output data from one model is used as input for the other. This data is used to model the downstream ward capacity through simple counting rules, allowing us to obtain realistic boarding patterns in the ED, and maintain a low model complexity.

DES has also been used to evaluate ambulance systems. Aboueljinane et al. (2013) review the literature on simulation models applied to such systems and find that most simulation studies focus on medium-term decisions such as the deployment problem and long-term decisions such as dimensioning of resources. Lam et al. (2014) use DES to evaluate different strategies for reducing ambulance response times, defined as the time it takes for a dispatched ambulance to arrive on scene. Lutter et al. (2016) use DES to compare different strategies for ambulance location planning. They compare five optimization models that are used to facilitate ambulance location, and use simulation to compare the solutions in terms of the proportion of calls that are served within the time threshold.

Currie et al. (2020) address how simulation modeling can help reduce the impact of COVID-19. The authors present different problems where simulation can be used as decision support. One of the problems they highlight is related to capacity of inpatient hospital beds and critical care.

Several authors apply DES to provide decision support in relation to the COVID-19 pandemic. Wood (2020), and Melman and Cameron (2021) both consider the trade-offs related to decreasing the activity for nonCOVID-19 patients during the pandemic. Mallor et al. (2020) aim to predict the number of beds needed by COVID-19 patients both in the Intensive Care Unit and in the rest of the hospital for the coming weeks. In addition to estimating the bed requirements imposed by the COVID-19 patients, Le Lay et al. (2020) also evaluate different policies for managing the increased demand for beds. Wood et al. (2020) aim to predict the number of deaths, which they divide into capacity-dependent and capacity-independent deaths, caused by the COVID-19 pandemic. They analyze different scenarios with regards to both the loading of COVID-19 positive patients and the number of intensive care beds available. Finally, Asgary et al. (2020) apply DES to evaluate different settings related to a drive-through facility for mass vaccination.

This paper adds to the literature on how DES can be a viable tool for decision support when preparing for a state of pandemic. In addition, we extend on the literature on ED boarding by proposing a new method for modelling a downstream ward and the ambulance waiting time experienced by inpatients leaving this ward. In this specific problem, two sources of ED boarding are identified and quantified, but we believe that similar methods can be applied to identify and quantify mechanisms that cause boarding in other systems.

5.3 Materials and methods

Three cases are considered in this paper; the ED, the ambulance and the combined case, and one DES model is developed for each case. These are referred to as the *ED model*, the *ambulance model* and the *combined model*. Before describing the models, the objectives of the study and a set of basic assumptions are presented. To describe the three DES models, the STRESS guidelines proposed by Monks et al. (2019) are used.

5.3.1 The objectives of the study

The purpose of the study is to provide decision support for the hospital management when preparing for a state of pandemic. The first objective is to estimate the number of beds that must be present in the ED to host COVID-19 suspected patients that wait for a COVID-19 test result (the ED model). The second objective is to estimate the number of ambulances required to obtain similar response times for the most urgent patients as in a prepandemic state (the ambulance model). We here define response time as the time it takes from the transport request emerges to an ambulance is assigned the mission. The third objective is to estimate the additional number of (boarding) beds required in the ED when considering the delayed transfer of patients from the ED to the COVID-19 ward, due to the lack of available beds in the COVID-19 ward (the combined model). All estimates should reflect the demand during the peak of the pandemic, and different COVID-19 testing policies.

5.3.2 Basic assumptions

In this section, we specify the assumptions that were made at the time when the two projects were performed.

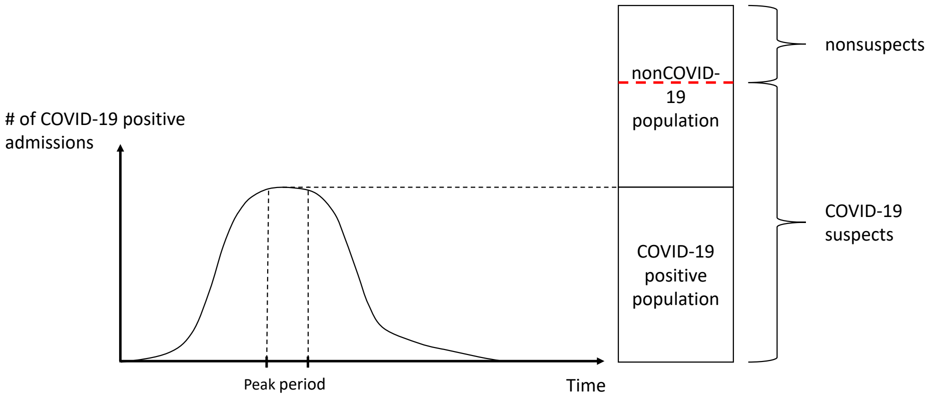


Figure 5.1: The two patient populations considered, and how they are divided into COVID-19 suspects and nonsuspects. All COVID-19 positive patients are COVID-19 suspects when arriving at the hospital, so is a share of the patients from the nonCOVID-19 patient population.

Patient groups

In all three cases, the emergency patients are considered. We define that the patients are divided into two groups: those that require a stay at the hospital due to their COVID-19 disease, and the rest. In Figure 5.1, the groups are labelled as the COVID-19 positive and the nonCOVID-19 population, respectively.

As we cannot know to what group a patient belongs before receiving the test results, the patients arriving at the hospital are divided into two categories: those with a COVID-19 suspicion and those without. All patients that are labelled as COVID-19 suspects must be treated as if they belong to the COVID-19 positive population until they are potentially clarified as belonging to the nonCOVID-19 population. We assume that all patients in the COVID-19 positive population have symptoms that place them in the COVID-19 suspicion category. In addition, we assume that a share of the patients that belong to the nonCOVID-19 patient population have symptoms that qualify for placing them in the COVID-19 suspicion category. The testing policy is decided on by the hospital management, and a strict testing policy implies that the threshold for testing is low and that a large share of patients are labelled as COVID-19 suspects. The red dashed line in Figure 5.1 illustrates how the nonCOVID-19 patient population is separated into either COVID-19 suspects or nonsuspects.

The development of the pandemic

When regarding the development of the pandemic over time, initially, the number of COVID-19 positive admissions increases. At a point in time, a peak period of activity is reached, followed by a period of decreasing incidence. At the time when the projects were conducted, we did not know for how long the peak period would last. To obtain a conservative estimate of the resource requirements, we assumed a peak period lasting longer than the average patient LOS. This implies that there is a stationary period when the number of COVID-19 admissions is equal to the number of COVID-19 patients leaving the hospital, and this period represents the peak of COVID-19 positive patients present in the hospital simultaneously. In comparison to the arrival peak, this peak is delayed by the time equal to the average patient LOS, and we refer to it as the *delayed peak period*.

In addition to the two ”recommended planning scenarios”, NIPH provided estimates of the average LOS of the COVID-19 positive patients. Based on this information, and since we assume a stationary system (with days as the time resolution) during the delayed peak period, Little’s formula (Little, 1961) is used to derive the daily arrival rate of COVID-19 positive patients entering the hospital.

The flow of patients through the hospital

All COVID-19 suspects are admitted to the *COVID-19 area* upon arrival at the ED, where testing is performed. If the test results indicate a COVID-19 disease, the patient is transferred to a hospital ward for treatment. If no beds are available in the downstream ward, patients remain in the ED until a bed becomes vacant. This additional waiting time is referred to as the boarding time, and patients require a bed while waiting to be admitted in the downstream ward. As a simplification, we aggregate the total bed capacity devoted for the COVID-19 positive patients to a common resource, referred to as the COVID-19 ward. The COVID-19 positive patients stay in the COVID-19 ward until they leave the hospital by ambulance. Furthermore, each patient that is not confirmed to be COVID-19 negative in the ED requires an ambulance upon departure, also those that were labeled as nonsuspects upon arrival at the ED.

At the time when the projects were conducted, it was decided by the hospital management to assume that the bed capacity for treating COVID-19 patients is sufficient during the peak period. The total bed capacity at St. Olavs Hospital is approximately 1000 beds, and elective patient activity will be adjusted to provide beds for the COVID-19 positive patients. We therefore assume that the bed capacity in the COVID-19 ward is sufficient and constant during the delayed peak period.

The arrival process of COVID-19 positive patients

Even though the number of COVID-19 patients resting in the COVID-19 ward is assumed to be stationary during the delayed peak period (on a daily basis), the number of COVID-19 patients present in the ED is non-stationary (on a hourly basis). We assume that patients arrive independently of each other and with varying intensity, and we therefore model the patient arrival processes as nonhomogeneous Poisson processes. We assume that the arrival process of COVID-19 positive patients to the ED is similar to the arrival process of semi-urgent patients, who mainly arrive at the ED during daytime. This is based on the assumption that the progression of symptoms is gradually increasing, which makes it possible to avoid traveling during night.

The arrival processes of requests in the ambulance model are also modelled as nonhomogeneous Poisson processes. We assume that all patients that will prove to be COVID-19 positive are transported with an ambulance to the ED. Together with the assumption that the time between a request for ambulance and the arrival at the ED is generally small, this justifies the choice to model the arrival process of patients belonging to the COVID-19 positive population with the same process as we used to generate the arrival of these in the ED model. The patients that are not confirmed to be COVID-19 negative upon departure are assumed to be discharged mainly during daytime, and the the same process is used again to model the discharge process. Even though the processes are the same, the intensities are adjusted to fit the associated expected arrival/ discharge rates.

5.3.3 The logic of the models

In this section, the logic of the three models are presented.

The ED model

In the ED model, we consider the flow of emergency patients with a COVID-19 suspicion entering the ED. These patients must be isolated from the nonsuspects,

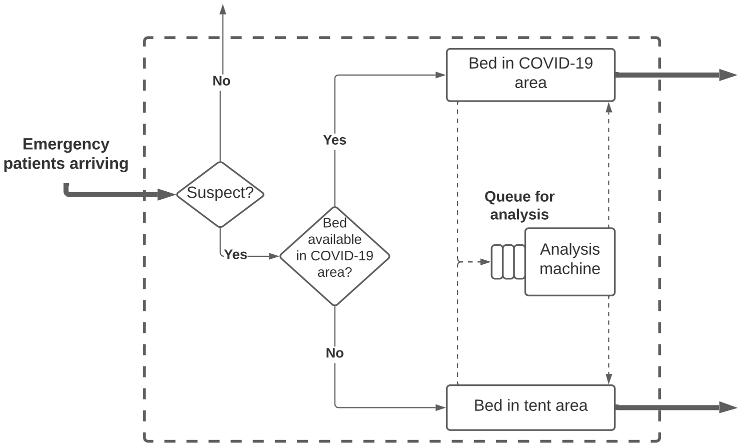


Figure 5.2: The system modelled in the ED case. The dashed arrows illustrate the flow of tests that are taken immediately after the patient is assigned a bed. The tests queue up in front of the analysis machine, and patients cannot leave the ED before receiving the outcome of the test analysis.

and enter an area referred to as the COVID-19 area.

In Figure 5.2, the system considered in the ED model is illustrated. There is a number of beds available in the COVID-19 area, and each patient is assigned a room and a bed upon arrival. A COVID-19 test is performed just after the arrival, and the patients must remain in the COVID-19 area until their test results are ready. If a patient enters the ED, and no beds are available in the COVID-19 area, the patient is escorted to a buffer area, referred to as the *tent area*, with additional beds. Tests are also performed in the tent area, and the process is not delayed for patients that stay in these beds. We assume that patients who are placed in the tent area are not transferred to the COVID-19 area if beds become vacant there.

The COVID-19 test samples are batched together, and analyzed in a machine. Only one machine is available, and only one batch can be analyzed at a time. This means that patients that enter just after a batch is initiated must wait to have their tests analyzed until this batch is done.

We assume that all ED activities required by the patients are undertaken while the patients wait for the test results. When the test results are ready, the patients leave the ED. After a patient leaves, the room must be sterilized independently of the test result.

The entities of the simulation model are the COVID-19 suspected patients entering the ED, while the resources are the beds in the COVID-19 and the tent area, and the machine for analyzing the COVID-19 tests. The state of the system is given by the number of patients in the COVID-19 and the tent area. The events in the simulation model are patients arriving at the COVID-19 area, patients being assigned to a bed either in the COVID-19 or the tent area, starting the analysis of a COVID-19 test batch, ending the analysis of a COVID-19 test batch, patients leaving the COVID-19 or the tent area, starting the cleaning of a room after a patient has left and finishing the cleaning of a room.

The ambulance model

The system modelled in the ambulance case is presented in Figure 5.3. There are two categories of patient transports considered: the normal and the *COVID-19 transports*. All patients that are either COVID-19 suspects when going to the hospital or that are not confirmed to be COVID-19 negative upon departure, require a COVID-19 transport. Patients that are not confirmed to be COVID-19 negative constitute of those that were confirmed to be COVID-19 positive, and those that were not tested for COVID-19 in the ED (the nonsuspects). The remaining transports are normal transports.

A COVID-19 transport requires additional transportation time, because the ambulance workers must wear an anti-infection coat, and the ambulance must be cleaned after the delivery of the patient. All transports are characterized by an urgency level and a required service time. If two patients request an ambulance at the same time, and only one ambulance is vacant, the most urgent patient is served first. We do not consider the position of the ambulances or the pick-up destinations in the model, but the service times are stochastic to reflect a variation in driving distances.

There is a number of ambulance cars available for patient transportation. Each car can only transport one patient at a time, and a car is unavailable for new missions during the entire service time of the patient that it is carrying. The ambulance personnel are not explicitly considered in the model, but the number of ambulances available through the day depends on the number of ambulance personnel on duty at different times during the week.

The entities of the simulation model are the transport requests, and the resources are the ambulances. The state of the system is given by the number of patients in transportation and the number of patients waiting to be assigned an ambulance (number of patients in queue). The events are a new transport request, an ambulance being assigned a patient, a patient leaving the ambulance, starting the cleaning of an ambulance after a COVID-19 transport and finishing the cleaning of an ambulance.

The combined model

This model extends the ED model, and the system under consideration is illustrated in Figure 5.4. After initial testing in the ED, all COVID-19 positive patients are transferred to the COVID-19 ward to receive treatment, while the COVID-19 negative patients leave the system. If no beds are available in the COVID-19 ward,

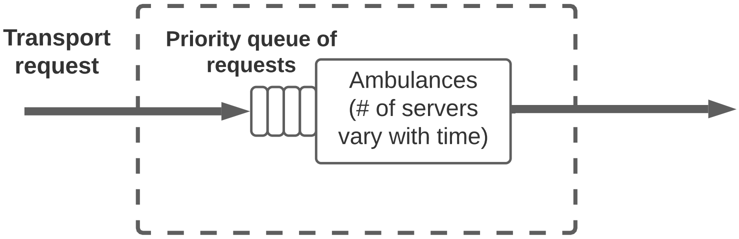

Transport request
Priority queue of requests
Ambulances (# of servers vary with time)

Figure 5.3: The system modelled in the ambulance case

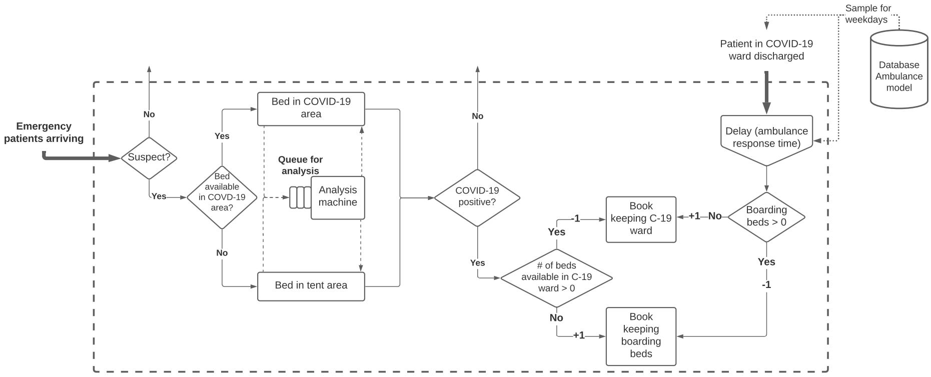


Figure 5.4: The system modelled in the combined case. The discharge time from the COVID-19 ward and the the ambulance response time are both sampled from the ambulance model output database.

the COVID-19 positive patients must wait in a boarding bed in the ED until a bed becomes vacant.

Under the assumption that a stay in a boarding bed does not extend the LOS of a patient, we do not have to model the individual patients stay in the COVID-19 ward. However, it is sufficient to know the number of beds available at a given point in time, and keep track of the relative difference from this point as patients enter and leave the COVID-19 ward. This saves computational effort, as the length of stay of COVID-19 patients is typically in the range of days and weeks, while the stay in the ED is in terms of hours.

Both the number of patients present in the boarding beds and the COVID-19 ward beds are handled via counting. At the beginning of the simulated time, a given number of beds are available in the COVID-19 ward. This is represented with a counter that is set to equal the number of available beds. If a patient is transferred from the ED to the COVID-19 ward, the counter is decreased by 1 as one less bed becomes vacant. There is also a counter for the boarding beds, representing the number of patients resting in a boarding bed. If the counter representing the COVID-19 ward is 0 (no beds available), and yet another patient should be transferred to the COVID-19 ward, the boarding bed counter is increased by 1.

When a COVID-19 positive patient is no longer in need of hospital services, an ambulance is requested to transport the patient out. Then, following a delay, an ambulance arrives to transport the patient home. When a patient leaves the COVID-19 ward, the corresponding counter is increased by 1, as one more bed becomes vacant. If there are patients resting in the boarding beds when a patient leaves the COVID-19 ward, one patient is transferred from a boarding bed to the vacant bed in the COVID-19 ward. The net change of patients in this process is -1 in the boarding beds, and 0 in the COVID-19 ward. At the same time, the boarding time of the patient who has stayed in a boarding bed the longest is recorded.

Two mechanisms that cause boarding are identified. First, having many patients ready to leave the ED at the same time (following a test batch) may cause prolonged boarding time if not enough beds are vacant in the COVID-19 ward. We refer to this as *excessive-flow-induced boarding.* Second, additional boarding time might occur when patients that are ready to leave the COVID-19 ward cannot leave because no ambulance is available. This is referred to as *ambulance-induced boarding.*

5.3.4 Data and experimentation

In the following we present the input data and specify the number of replications used to conduct the studies. The outcome variables that we want to study are referred to as the dependent variables, while the input variables that affect the dependant variables and that we alter through the sensitivity analysis are called the independent variables. Before performing the simulation study, preliminary testing is performed to decide on the length of warm-up necessary to avoid transient effects, and the number of replications needed to ensure accurate results (Law, 2015).

The ED model

The expected arrival rate of the nonCOVID-19 patients at different hours of the week is calculated based on historical data from St. Olav's Hospital, 2019. The weeks 37-47 were chosen by the ED management to represent normal weeks. As stated in Section 5.3.2, the arrival process of semi-urgent emergencies is used to model the arrival process of COVID-19 positive patients. The intensity is however altered to make sure that the weekly number of arrivals equals the estimates provided by the NIPH scenarios.

There are 27 beds available in the COVID-19 area. Since we want to estimate the need for additional beds required in the ED, the tent area is treated as having infinite capacity. The analysis machine is used for evaluating tests taken both in the ED and in other locations in the region. Each batch has a capacity of approximately 100 test samples, and the tests performed in the ED are prioritized. Even during the peak period, the test intensity in the ED will not require the entire batch capacity. We therefore assume that a test batch has infinite capacity with regards to the tests performed in the ED. Each batch is analyzed for 4 hours before receiving the results. The cleaning of a room after a patient has left the ED takes 30 minutes.

For each scenario presented in Section 5.4, 200 replications of one simulated week are performed, and one week warm-up is applied. In each replication, the output data is aggregated to an hourly resolution, implying that we calculate the average number of beds used during each hour of the simulated week. Based on the 200 samples, we calculate the hourly mean and hourly 90th percentile bed loading during a week. The independent variable is the arrival intensity of COVID-19 suspects, while the dependent variable is the number of beds used in the tent area.

The ambulance model

Six subgroups of transport requests are considered in the model, and each subgroup is associated with an urgency level. Sorted by decreasing urgency, the levels are red, yellow, green and planned transports. For the nonCOVID-19 patient population, we consider red (37%), yellow (36%), green (9%) and planned transports (18%) going to the hospital. The fifth subgroup are patients that will prove to be COVID-19 positive when tested in the ED. These patients request a transport to the ED due to experiencing COVID-19 related symptoms, and they are categorized as yellow transports. The last subgroup are patients that are not confirmed to be COVID-19 negative when leaving the hospital, and these are categorized as planned transports. In Figure 5.5, the subgroups are displayed, and we include whether they require a normal or a COVID-19 transport.

The expected arrival rate of requests generated by the nonCOVID-19 population at different hours of the week is calculated based on historical data from St.

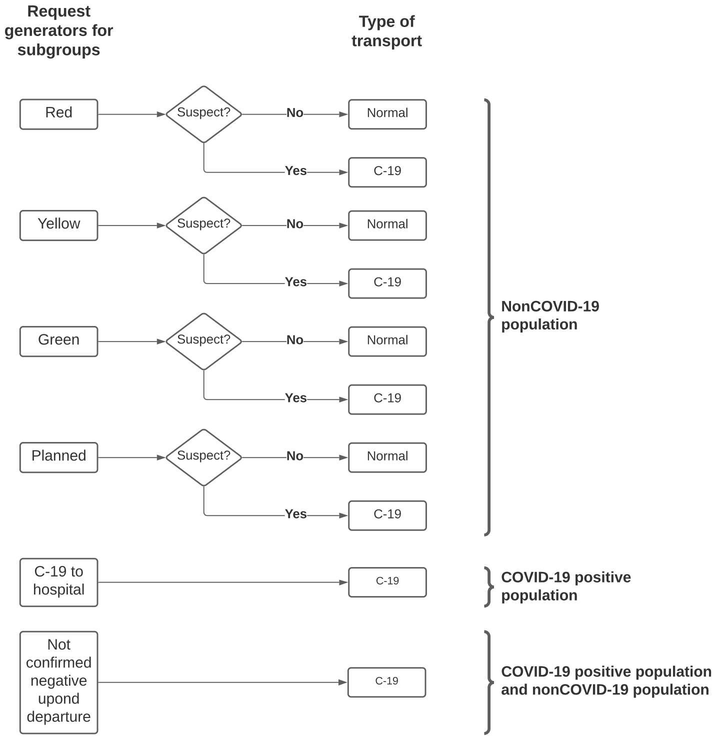


Figure 5.5: The six patient subgroups and what transport they require.

Olav's Hospital, 2019. The weeks covering January to March were chosen by the management at the ambulance services to represent a normal period. As stated in Section 5.3.2, to generate requests from subgroups five and six, the arrival process of semi-urgent (green) emergencies to the ED is used. The intensity is however altered to fit the scenarios of the sensitivity analysis.

To obtain realistic transport durations, the time spent for each transportation is sampled from the set of historical transport durations from 2019. For COVID-19 transports, 45 minutes are added to the sampled duration to include the cleaning of the ambulance. The number of ambulances available during the week is identical to the ambulance schedule that was present when the project was conducted.

For each scenario presented in Section 5.4, 300 replications of one simulated week are performed, and one week warm-up is applied. For each patient request, the response time is recorded. The records are used to calculate the mean response times for patients within each urgency category. The independent variables are the arrival intensity of transport requests and the number of ambulances available, while the dependent variable is the ambulance response time.

The combined model

All input data used for the ED model is also applied in the combined model. In addition, two input parameters are used to model the discharge process of patients leaving the COVID-19 ward: the desired discharge time of patients and the corresponding ambulance response time. This data is collected from running the ambulance model for one week (following one week warm-up) with the number of ambulances necessary to obtain prepandemic waiting times for red and yellow emergency patients. The ambulance model is run 500 times, producing a data set containing 500 replications of both the desired discharge times and the corresponding ambulance response times through the week. The data is stored according to the simulated replication (1 to 500) and weekday (1 to 7). For example, on Tuesday in replication 30 there may be 18 COVID-19 positive patients leaving the COVID-19 ward. Patient 13 is discharged at 15:00 and gets an ambulance at 15:20, yielding an ambulance response time of 20 minutes for this patient.

For each simulated day in the combined model, one replication is sampled, and the data from the corresponding weekday in the sampled replication is used to generate the discharge time of patients in the COVID-19 ward, and the corresponding ambulance waiting time.

One adjustment is made to the input data of the ambulance model when producing the data base. Recall that the last subgroup of patients introduced in Section 5.3.4 contains both the COVID-19 positive patients resting in the COVID-19 ward, and patients that were not tested for COVID-19 in the ED. In the combined model, we are not interested in the latter group of patients. To exclusively model the requests coming from the COVID-19 ward, the last subgroup of patients introduced in Section 5.3.4 is therefore split in two when collecting and storing data from the ambulance model. Furthermore, since we model the delayed peak period, identical distributions are used to generate COVID-19 positive patients entering the ED and COVID-19 positive patients that are discharged from the COVID-19 ward. As we want to estimate the need for boarding beds during the peak of the pandemic, we model the boarding bed capacity as unlimited and evaluate the usage of these.

For each scenario presented in Section 5.4, 200 replications of one simulated week are performed with the combined model. The independent variables are the arrival intensity of COVID-19 suspects, the departure intensity from the COVID-19 ward and the ambulance response time, while the dependent variable is the boarding bed requirement. In contrast to previous simulations, we are here interested in the transient period starting with the delayed peak, so warm-up is not applied. Furthermore, each simulated replication is initiated at 00:00 in the night with no patients in the COVID-19 area and 3 vacant beds in the COVID-19 ward. The experiment is run for two modes. In the first mode, the ambulance response time is set to zero, implying that we only observe excessive-flow-induced boarding time. In the second mode, ambulance response time is added.

5.4 Implementation and the setup of the sensitivity analysis

An Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz, 16 GB RAM computer is used when performing the simulations. The simulation models are written in Python 3.7 and the package SimPy. To perform the random sampling, the algorithms included in Python is used. To reduce output variance, common random numbers are applied when performing sensitivity analysis. In the first replication, a seed is set, and it is then increased by one for each subsequent replication.

A scenario tree is constructed to guide the sensitivity analysis. The tree contains three parameters that represent aspects of uncertainty that are common for the cases:

- The number of COVID-19 positive patients arriving for the ED each day
- The size of the nonCOVID-19 population
- The testing policy, defining the share of nonCOVID-19 patients that will be labelled as suspects

In the first branching, the daily arrival rate of COVID-19 positive patients to the hospital during the peak period is represented. In the second branching, the loading intensity of patients that belong to the nonCOVID-19 patient population, in relation to the reference loading, is represented. The reference loading is the expected number of emergency patients that entered the ED or required an ambulance each day in a normal prepandemic week. The third branching represents the testing policy, describing the threshold of categorizing patients as COVID-19 suspects. In reality, the threshold can be related to what symptoms that should trigger a test. The policy levels are given as the percentage of individuals from the

Table 5.1: The 16 scenarios applied in the models.

Scenario	# of COVID-19 positive (μ^{C19})	nonCOVID-19 relative to normal (α)	Share of suspects in nonCOVID-19 (β)	$\boldsymbol{E[suspects]}$	$\boldsymbol{E[transports]}$
1	12	80%	33%	31	96
2	12	80%	50%	41	88
3	12	80%	67%	51	81
4	12	80%	100%	70	67
5	12	100%	33%	36	114
6	12	100%	50%	48	105
7	12	100%	67%	60	96
8	12	100%	100%	84	78
9	21	80%	33%	40	114
10	21	80%	50%	50	106
11	21	80%	67%	59	99
12	21	80%	100%	79	85
13	21	100%	33%	45	132
14	21	100%	50%	57	123
15	21	100%	67%	69	114
16	21	100%	100%	93	96

nonCOVID-19 patient population that are labelled as COVID-19 suspects when entering the ED or requesting an ambulance to the hospital.

One split is applied in the first and the second branch, while we have four levels of testing policies in the third branch. The split in the first branch reflects the two scenarios provided by NIPH, with 12 and 21 COVID-19 positive patients entering each day respectively. The split in the second branch was discussed with the hospital management, and set to be 80% and 100%. Also the last split was discussed with the hospital management, and the values 33%, 50%, 67% and 100% were applied to cover a wide range of testing policies. In total this yields 16 scenarios. The scenarios are listed in Table 5.1. For each scenario we obtain the expected number of both COVID-19 suspects arriving for the ED, and transport requests each day. Note that there are intra-day variations in the expectations, but these are not shown in the table. For more information on how the the expected number of both COVID-19 suspects arriving for the ED and transport requests are calculated, see Appendix A.2.

5.5 Results

In this section, the results from all three models are presented. The hospital management was mostly interested in the scenarios with the high COVID-19 positive patient loading, which are represented by scenarios 9 to 16. These are therefore emphasized in the following. The main results for all scenarios are presented in Table A.1 in Appendix A.1.

5.5.1 Results for the ED case

Figure 5.6 provides results for the ED bed loading in scenarios 9 to 12, which differ in the testing policy. The light shaded area represents the beds in the COVID-19 area, while the dark shaded area is the beds in the tent area. The borders of the shaded areas indicate the mean and the 90th percentile measures for each hour of the week. The mean represents the mean bed requirement over the 200 replications, while the 90th percentile indicates a threshold where only 20 out of 200 measured bed requirements for a given hour of the week equal or exceed the threshold.

Note how the testing policy impacts the number of beds that must be established in the tent area. In scenario 12, all patients are tested upon arrival to the ED. In this case, the tent capacity should be similar to the capacity of the COVID-19 area. Furthermore, the need for additional beds is much less during the weekends. In all scenarios, the use of a tent area emerges at around 12:00 (noon) and the peak number of patients in the tent area is observed between 16:00 - 19:00. The number of patients in the COVID-19 area falls towards the evening, implying that the patients resting in the tent area can be moved inside (although this is not done in the simulation model). The total number of additional beds needed, if we allow for patients to transfer from the tent area to the COVID-19 area, can be derived from the simulated results. This is done by adding the beds used in the COVID-19 area and the tent area, and subtract the capacity of 27 beds (if this becomes negative, the value is set to zero). The resulting number of additional beds in the 90th percentile level can be seen as the solid red line in Figure 5.6. Depending on the testing policy and the size of the nonCOVID-19 population at the peak of the pandemic, there is a need for between 0 to 41 additional beds during the weekdays, and 0 to 13 additional beds in the weekend.

5.5.2 Results for the ambulance case

For each scenario, we estimate the minimum number of additional ambulances required to ensure mean response times for red and yellow emergency patients that are equal to or shorter than those of the prepandemic state. To represent the prepandemic situation, the model is first run for a base case. That is, we only include requests from the nonCOVID-19 patient population and apply the ambulance capacity available in a prepandemic situation.

Figure 5.7 illustrates the mean utilization of ambulances and the mean response time for different patient categories during the week from simulating the base case. During the weekdays, except from Friday, the ambulance capacity is satisfying yielding short response times. On Friday, the combination of more requests and less capacity available during the evening causes significant waiting times. The

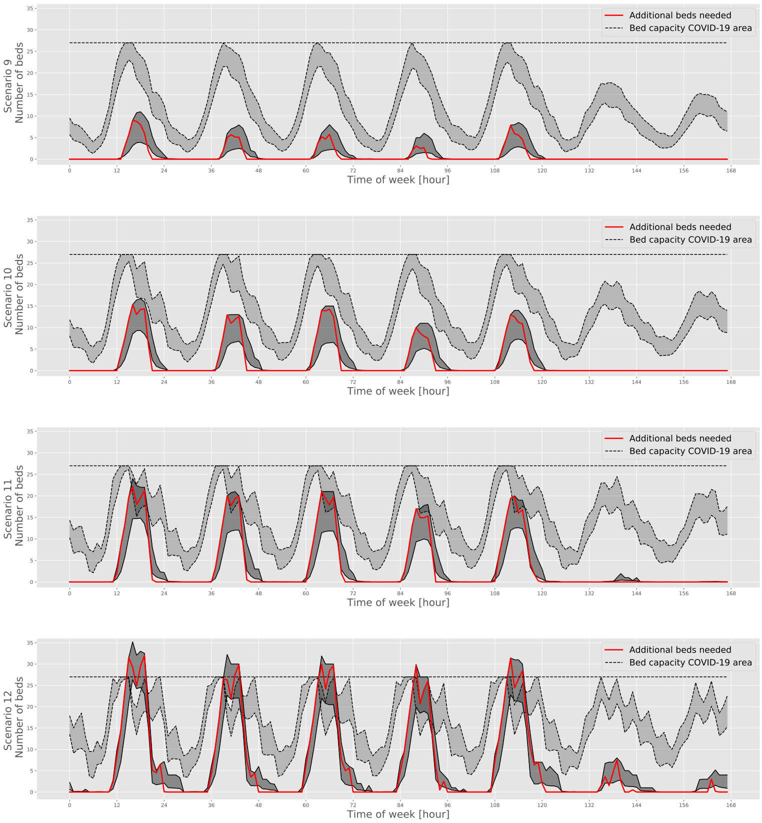


Figure 5.6: Results from the ED model: The bed loading in the COVID-19 area (dashed lines) and the tent area (solid black lines) through the week for scenarios 9 to 12. The bands cover the area between the mean and the 90th percentile. The solid red line indicate the 90th percentile bed requirement in the tent area if patients can be transferred from the tent area to the COVID-19 area. The horizontal dashed line indicates the planned bed capacity in the COVID-19 area.

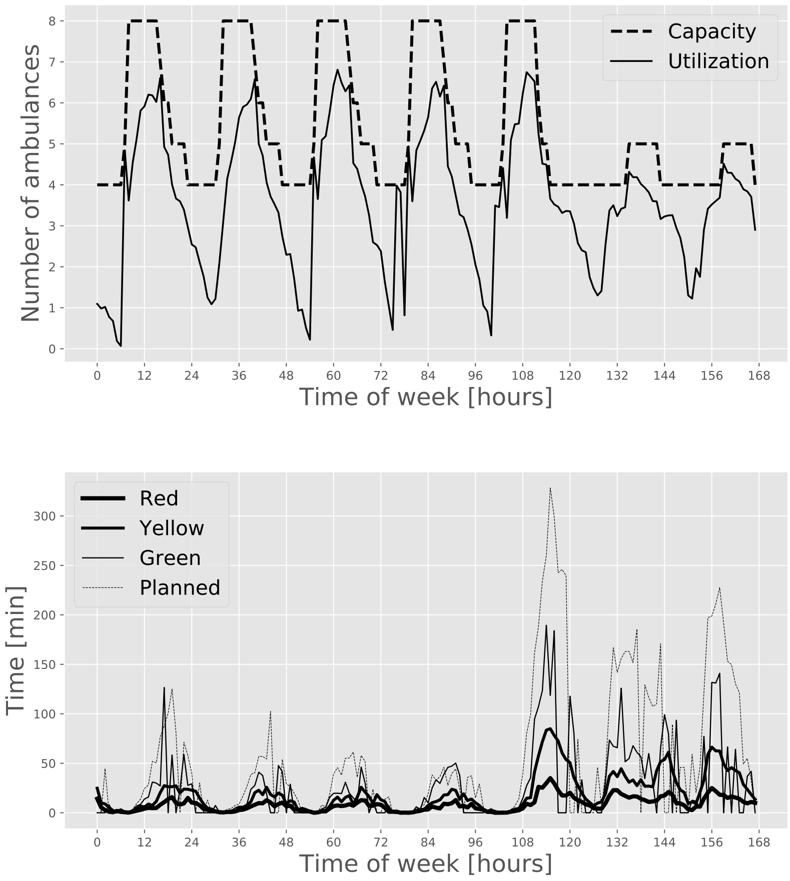


Figure 5.7: Results from the ambulance model: The base case. Top: Mean utilization of the ambulance capacity. Bottom: Mean response time during the week

Table 5.2: Results from the ambulance model: The number of ambulances added in scenarios 9 to 16 to obtain similar mean response times as in the base case

Scenario	Red	Yellow	Green	Planned
9	5	5	6	10
10	5	5	6	10
11	5	5	6	9
12	5	5	6	7
13	6	6	7	12
14	6	6	7	12
15	6	6	7	11
16	6	6	7	9

waiting times are also prolonged during the weekend because less ambulances are available.

To obtain the preferred response times in the 16 scenarios, the ambulance resources are added flat. That is, for each additional ambulance, the resource is available through the entire week. When the number of ambulances is increased, the response time decreases towards the base level. The resulting number of additional ambulances needed in scenarios 9 to 16 is presented in Table 5.2. In general, because of the queue prioritization rules, the base level response times for the most urgent patient groups are easier obtained compared with the less urgent patients. Adding more ambulances than what is suggested from just regarding the response times for red and yellow requests should be considered, as it dramatically decreases the expected response time for the planned requests. If we consider Scenario 13, going from 6 to 12 additional ambulances yields a decrease in mean response time for planned patients from 585 to 63 minutes. The corresponding values for red and yellow patients are 11 to 3, and 21 to 4 minutes respectively.

As for the ED case, the results in the ambulance case are sensitive to the testing policy. The planned patient category is most sensitive to the policy level. A strict testing policy yields fewer COVID-19 transports leaving the hospital, causing relatively short response times for planned patients in these scenarios since the demand for planned transports is reduced. Conversely, in the ED case, a strict testing policy yields a high demand for additional beds in the ED, making those scenarios more demanding.

5.5.3 Results for the combined case

When generating the ambulance waiting time data, 5 and 6 additional ambulances were added to scenarios 9 to 12 and 13 to 16 respectively. Furthermore, we assume that the boarding beds are located in the ED.

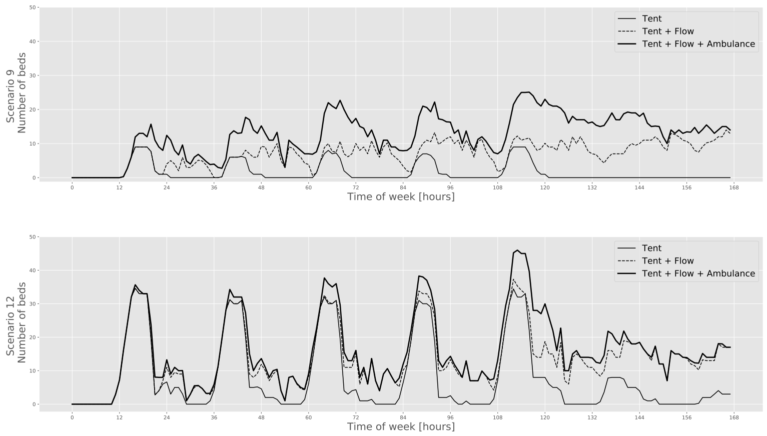


Figure 5.8: Results from the combined model: 90th percentile number of additional beds when considering boarding for scenarios 9 and 12. The label Tent represents the number of beds needed in the tent area. Tent + Flow is the number of additional beds when adding the excessive-flow-induced boarding. Tent + Flow + Ambulance represents the number of additional beds when also adding the ambulance-induced boarding.

Figure 5.8 illustrates, for scenarios 9 and 12, the 90th percentile number of additional beds needed in the ED when considering ED boarding and compares it with the result when boarding is disregarded. The results clearly indicate the need for excessive bed capacity in the ED when entering the peak period. The excessive-flow-induced boarding results in an increased bed loading primarily during night, as the patients must wait until the next morning for beds to become vacant in the COVID-19 ward. When adding ambulance waiting time, the problems related to boarding starts earlier in the day because patients leaving the COVID-19 ward during the day are delayed. During night, the ambulance waiting time is short and the effect of ambulance-induced boarding is less prominent. Note that because Friday is a busy day for the ambulance service (see Figure 5.7), the ambulance-induced boarding is most prominent on this day. Finally, the effect of ambulance-induced boarding is less in scenario 12, caused by shorter ambulance waiting times due to the strict testing policy.

Figure 5.9 illustrates the requirement for additional beds if we allow for a transfer of patients from the tent area and the buffer beds to the COVID-19 area in scenarios 9-12, and compares it with the results when boarding is disregarded. As the simulations with boarding are run without a warm-up period and starting from an empty system, the results cannot be directly compared. However, they illustrate some important aspects, like the fact that excessive ED boarding will cause an additional need for beds both during the late evening and in the week-

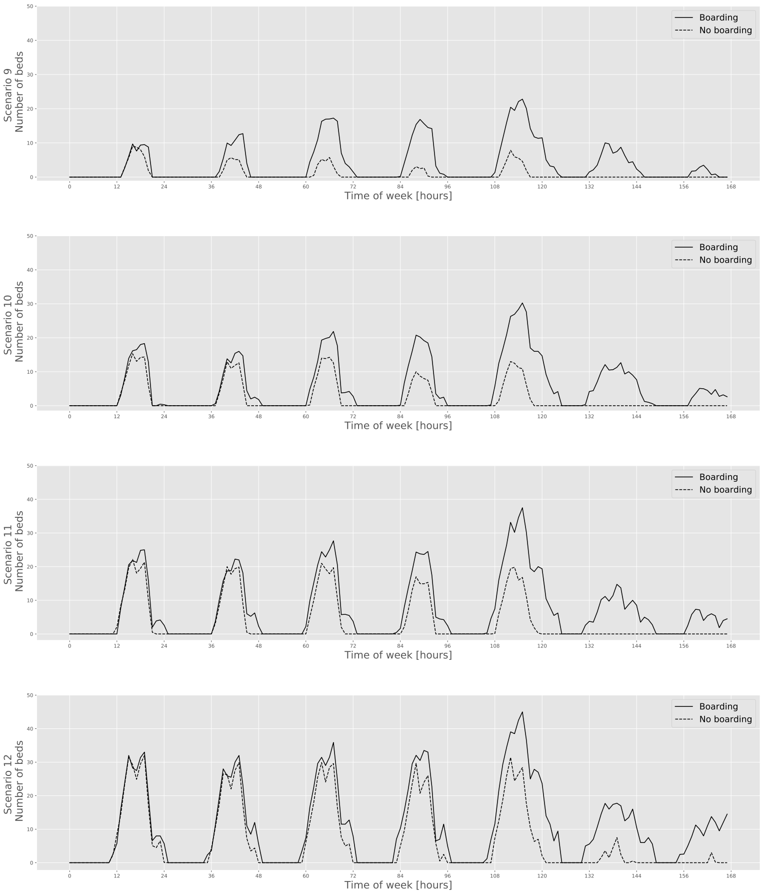


Figure 5.9: Results from the ED model and the combined model: Comparison of bed usage when boarding is considered and not for scenarios 9-12. The results illustrate the 90th percentile number of additional beds required in the ED when allowing for patients to transfer from the tent area and the buffer beds to the COVID-19 area

ends. Furthermore, we see that the difference between the results when regarding boarding and not is larger for scenario 9 compared with scenario 12, reflecting the shorter ambulance response times in scenario 12.

5.5.4 Managerial implications

The results from the ED and the ambulance model were used to inform the hospital management, partly through presentations for the hospital pandemic committee and partly as input for a managerial report on how to perform the ambulance planning through the pandemic. Based on these results, the following decisions were made when preparing for a state of pandemic:

- Outpatient clinic examination rooms close to the ED were used to provide additional bed capacity for COVID-19 suspects that required testing in the ED.
- Additional resources for transporting patients to and from the hospital were established, including Red Cross ambulances, and military ambulances operated by the Home Guard.

In August 2020, some months after the first peak in Norway, the management requested updated analysis on the bed requirements in the the ED. At this point in time, new testing regimes had become available, including the option to buy tests that could provide answers within 90 minutes instead of 4 hours. The management wanted to know how the bed requirements would change given different levels of available 90-minutes-tests. To provide decision support, the ED model were extended and new results were presented for the hospital management.

5.6 Discussion

In this paper we have shown how a set of DES models can be applied to provide decision support for the hospital management when time is limited. Even if the models presented are rather simple, the analyses performed proved to be of great value to the hospital management. The results are highly sensitive to the NIPH planning scenarios, and the relative loading of emergency patients compared with the prepandemic situation. In contrast to the testing policy, these cannot be controlled by the hospital management.

When regarding the number of beds needed in the ED, the results are very sensitive to the testing policy. A strict testing policy increases the need for additional beds in the ED considerably, and consequently the number of nurses required. As a consequence, resources must be reallocated from elective activity, or the capacity must be increased. When regarding the ambulance response times of red and yellow transports, these decrease with a strict testing policy. However, the differences are small and the number of ambulances required to obtain prepandemic response times are not affected by the testing policy. Based on these observations, a less strict testing policy seems reasonable. However, the consequences of admitting a COVID-19 positive patient into a non COVID-19 ward can be fatal, and the costs related to increased resource capacity must be weighted against the potential of ignoring a COVID-19 positive patient in the ED.

When boarding is considered, the bed requirement increases, especially during night and in the weekends. If a less strict testing policy is implemented, the boarding time is to a large extent affected by the ambulance response times of patients discharged from the COVID-19 ward, that are categorized as planned transports. Based on this observation, increasing the ambulance capacity further to decrease the waiting times for planned transports seems reasonable. This will have less effect if a strict testing policy is implemented.

We have demonstrated how boarding can be modelled with simple counting rules. This saves computational effort, as we can omit the explicit modelling of patient stay in the downstream ward. Furthermore, initiating the model is very simple, as the vacant bed capacity is set by a single number. We assumed that the rate of patients leaving the COVID-19 ward was equal to the rate of patients entering the ward. This assumption implies that a stay in a boarding bed does not affect the LOS, meaning that we regard the boarding beds as a server and not as a queue, and consider an infinite server system. In the opposite case, where a stay in the boarding bed delays the healing process, the boarding beds should be considered as a queue for service at the ward. Then, the rate of patients leaving the COVID-19 ward depends on the bed capacity and we may have rates that are unequal.

The counting approach is not appropriate if extended boarding affects the LOS of patients. Extended boarding can sometimes cause misplacement of patients and delay the treatment process. However, boarding time is often measured in the range of minutes and hours, while the LOS is typically several days. In many cases it should therefore a fair assumption that the LOS is not affected by extended boarding time.

Bibliography

L. Aboueljinane, E. Sahin, and Z. Jemai. A review on simulation models applied to emergency medical service operations. *Computers Industrial Engineering*, 66(4):734 – 750, 2013.

A. Asgary, M. M. Najafabadi, R. Karsseboom, and J. Wu. A drive-through simulation tool for mass vaccination during covid-19 pandemic. *Healthcare*, 8(4), 2020.

A. E. Bair, W. T. Song, Y. Chen, and B. A. Morris. The impact of inpatient boarding on ed efficiency: A discrete-event simulation study. *Journal of Medical Systems*, 34:919–929, 2010.

R. Carmen, M. Defraeye, B. C. Aydin, and I. Van Nieuwenhuyse. Modeling emergency departments using discrete-event simulation: A real-life case study including patient boarding. Technical report, KU Leuven - Faculty of Economics and Business, Leuven, Belgium, 2014.

C. S.M. Currie, J. W. Fowler, K. Kotiadis, T. Monks, B. S. Onggo, D. A. Robertson, and A. A. Tako. How simulation modelling can help reduce the impact of covid-19. *Journal of Simulation*, 14(2):83–97, 2020.

K. De Boeck, R. Carmen, and N. Vandaele. Needy boarding patients in emergency departments: An exploratory case study using discrete-event simulation. *Operations Research for Health Care*, 21:19 – 31, 2019.

E. M. W. Kolb, T. Lee, and J. Peck. Effect of coupling between emergency department and inpatient unit on the overcrowding in emergency department. In *2007 Winter Simulation Conference*, pages 1586–1593, 2007.

E. M. W. Kolb, J. Peck, S. Schoening, and T. Lee. Reducing emergency department overcrowding - five patient buffer concepts in comparison. In *2008 Winter Simulation Conference*, pages 1516–1525, 2008.

S. S. W. Lam, Z. C. Zhang, H. C. Oh, Y. Y. Ng, W. Wah, M. E. H. Ong, and on behalf of the Cardiac Arrest Resuscitation Epidemiology (CARE) Study Group. Reducing ambulance response times using discrete event simulation. *Prehospital Emergency Care*, 18(2):207–216, 2014.

A. M. Law. *Simulation Modeling and Analysis.* McGraw-Hill Education, 2 Penn Plaza New York, NY 10121, 2015.

J. Le Lay, V. Augusto, X. Xie, E. Alfonso-Lizarazo, B. Bongue, T. Celarier, R. Gonthier, and M. Masmoudi. Impact of covid-19 epidemics on bed requirements in a healthcare center using data-driven discrete-event simulation. In *2020 Winter Simulation Conference (WSC)*, pages 771–781, 2020.

J. D. C. Little. A proof for the queuing formula: L = w. *Operations Research*, 9 (3):383–387, 1961.

P. Lutter, D. Degel, L. Wiesche, and B. Werners. Analysis of ambulance location models using discrete event simulation. In *Operations Research Proceedings 2014*, pages 377–383, Cham, 2016. Springer International Publishing.

F. Mallor, D. García-Vicuña, and L. Esparza. Planning ward and intensive care unit beds for COVID-19 patients using a discrete event simulation model. Technical report, Smart Cities Institute, Public University of Navarre, Pamplona, Spain, 2020.

A.K. Melman, G.J.and Parlikad and E.A.B. Cameron. Balancing scarce hospital resources during the covid-19 pandemic using discrete-event simulation. *Health Care Management Science*, 2021.

T. Monks, C. S. M. Currie, B. S. Onggo, S. Robinson, M. Kunc, and S. J. E. Taylor. Strengthening the reporting of empirical simulation studies: Introducing the stress guidelines. *Journal of Simulation*, 13(1):55–67, 2019.

C. Tang, Y. Chen, and S. Lee. Non-clinical work counts: facilitating patient outflow in an emergency department. *Behaviour & Information Technology*, 34 (6):585–597, 2015.

The Norwegian Directorate of Health. Changes in activity of healthcare services for March and April 2020. 2020. Published in Norwegian.

S. Valipoor, M. Hatami, H. Hakimjavadi, E. Akçalı, W. A. Swan, and G. D. Portu. Data-driven design strategies to address crowding and boarding in an emergency department: A discrete-event simulation study. *HERD: Health Environments Research & Design Journal*, 14(2):161–177, 2021.

L. Vanbrabant, K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse. Simulation of emergency department operations: A comprehensive review of kpis and operational improvements. *Computers Industrial Engineering*, 131:356 – 381, 2019.

R. M. Wood. Modelling the impact of covid-19 on elective waiting times. *Journal of Simulation*, pages 1–9, 2020.

R. M. Wood and B. J. Murch. Modelling capacity along a patient pathway with delays to transfer and discharge. *Journal of the Operational Research Society*, 71(10):1530–1544, 2020.

R. M. Wood, C. J. McWilliams, M. J. Thomas, C. P. Bordeaux, and C. Vasilakis. Covid-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. *Health Care Management Science*, 23(3):315 – 324, 2020.

Chapter A

Appendices

A.1 Main results

Table A.1: Main results for the three cases, scenarios 1 to 16. For the ED and the combined case, the maximum 90th percentile number of additional beds both during the week and the weekend is presented. For the ambulance case, the number of additional ambulances required to maintain base case response times are included. In the combined case, the analysis is performed with the number of additional ambulances as given in the table.

Scen.	Max beds week	Max beds weekend	Additional ambulances	Max beds week (boarding)	Max beds weekend (boarding)
1	0	0	4	5	0
2	7	0	4	16	3
3	15	0	4	26	6
4	28	2	3	37	13
5	5	0	5	13	2
6	15	0	5	26	8
7	23	0	5	33	11
8	38	9	4	46	18
9	9	0	5	23	10
10	16	0	5	31	13
11	23	0	5	38	15
12	32	8	5	45	18
13	12	0	6	31	18
14	22	0	6	39	19
15	29	3	6	41	21
16	41	13	6	51	23

A.2 Calculating the expected number of suspects and transports

Here, we describe how we calculate the expected number of both COVID-19 suspects arriving for the ED and transport requests each day.

The ED and the combined cases share the same scenario tree, and the expected number of COVID-19 suspects that enter the ED each day in each scenario is calculated by the following formula:

$$E[suspects] = \mu^{C19} + \mu^{Non} \cdot \alpha \cdot \beta \tag{A.1}$$

Here, μ^{C19} is the expected number of COVID-19 positive patients entering the ED each day at the peak of the pandemic, and μ^{Non} is the expected number of emergency patients belonging to the nonCOVID-19 patient population that enter the ED each day. This number depends on the weekday. The parameter α is used to adjust the expected patient activity (the second branching), while β represents the share of patients belonging to the nonCOVID-19 patient population that are categorized as COVID-19 suspects (the third branching). Note that since μ^{Non} depends on the weekday, the expected number of suspects given here represents the average day, but the number varies between weekdays.

To calculate the daily total number of ambulance transports in each scenario, the following equation is used:

$$\begin{aligned} E[transports] &= (\mu^{Non,A} \cdot \alpha \cdot \beta) + (\mu^{Non,A} \cdot \alpha \cdot (1-\beta)) \\ &\quad +2\mu^{C19} + (\mu^{Non,A} \cdot \alpha \cdot (1-\beta)) \\ &= 2\mu^{Non,A} \cdot \alpha \left(1 - \frac{\beta}{2}\right) + 2\mu^{C19,A} \end{aligned} \tag{A.2}$$

Here, $\mu^{Non,A}$ is the expected number of patient transports to the hospital generated by the nonCOVID-19 patient population, and its value depends on the weekday. The parameters α, β and μ^{C19} have the same interpretation as in the ED case. The first term represents the expected number of COVID-19 transports to the hospital generated by the nonCOVID-19 patient population, while the second term is the number of normal transports generated by the same population. The COVID-19 positive patients require a Covid transport both to and from the hospital, which is ensured by the third term. The final term represents the transportation of COVID-19 suspects from the nonCOVID-19 patient population that require an ambulance when leaving the hospital. Note that this equals the second term and represents the fact that all patients that are not tested for COVID-19 in the ED require an ambulance when leaving the hospital.

ISBN 978-82-326-7500-5 (printed ver.)
ISBN 978-82-326-7499-2 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology