NTNU

Norwegian University of
Science and Technology

Stochastic Master Surgery Scheduling for the orthopaedic department at St. Olav's Hospital

Thomas Reiten Bovim

Industrial Economics and Technology Management

Submission date:	June 2018
Supervisor:	Marielle Christiansen, IØT
Co-supervisor:	Anders Nordby Gullhav, IØT Troels Martin Range, IØT Lars Hellemo, SINTEF

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

Preface

The submission of this thesis completes my Master of Science degree in Industrial Economics and Technology Management at The Norwegian University of Science and Technology (NTNU). The work presented is meant to serve as tactical decision support for the management at the orthopaedic department at St. Olav's Hospital when developing the semiannual Master Surgery Schedule, and the thesis builds upon a report delivered last semester in the course TIØ4500 - Managerial Economics and Operations Research.

Several individuals deserve special acknowledgements for their contribution to this thesis. First and foremost, I would like to thank my supervisors, Anders Nordby Gullhav, Troels Martin Range, Lars Hellemo and Marielle Christiansen, for sharing their knowledge and dedicating much time and effort into the thesis.

I also would like to thank the management at the orthopaedic department, and in particular Vigleik Jessen, for providing insight in the department and for being easily accessible throughout the project. Furthermore, I would like to thank Trude Mittet, Mette Røsbjørgen and Steinar Havik for providing useful information, crucial to understand the way things are done at the department.

Finally, I would like to thank Ingeborg for her support and patience throughout my academic studies, and I look forward to having more time to spend with her from now on.

Trondheim, 2018

Thomas Reiten Bovim

Thomas Reiten Bovim

Sammendrag

Det regionale helseforetaket, Helse Midt-Norge (HMN) tilbyr sykehustjenester til fylkene Trøndelag og Møre og Romsdal. Prognoser estimerer at antallet innbyggere i de to fylkene som er eldre enn 67 år vil øke med 42 % innen 2030, noe som vil medføre en betraktelig økt etterspørsel etter sykehustjenester de neste årene (Helse-midt.no, 2016). St. Olavs Hospital er det største sykehuset i HMN, og i denne oppgaven vil vi ta utgangspunkt i operasjonsstueplanlegging ved klinikk for ortopedi, revmatologi og hudsykdommer (videre kalt klinikk for ortopedi) ved St. Olavs Hospital. Klinikk for ortopedi er ansvarlig for å behandle alle ortopediske pasienter, både akutte og elektive. Å tilby behandling til ortopediske akuttpasienter innen sykehusets interne tidsfrister har vært en utfordring for klinikken i flere år. Lang ventetid er uheldig for akuttpasientene, men det påvirker i tillegg flyten av elektive pasienter. I perioder når køen av akuttpasienter bygger seg opp så vil flere av akuttpasientene måtte tilbys operasjonsstuekapasitet i elektive operasjonsstuer, hvilket betyr at elektive operasjoner må strykes, og den elektive pasienten må tildeles et nytt operasjonstidspunkt. De fleste akuttpasienter opptar senger på sengepostene mens de venter på operasjon. I perioder hvor mange akuttpasienter venter på operasjon vil derfor mange senger være opptatt. Dette kan medføre at man må stryke elektive pasienter som behøver en seng etter operasjon. Den gruppen av akuttpasienter som det haster minst å operere, de grønne akuttpasientene, er de første som overflyttes fra akuttstuene til elektive operasjonsstuer i perioder med stort press av akuttpasienter. I tillegg er det slik at måloppnåelsen, altså andelen av pasienter som opereres innen de interne tidsfristene, ligger på omlag 70 % for de grønne pasientene. Vi ønsker derfor å undersøke effektene av å dedikere mer operasjonsstuekapasitet for de grønne akuttpasientene.

Hovedproblemet som vi ønsker å løse i denne oppgaven er det overordnede operasjonsplanleggingsproblemet (OPP) som går ut på å lage en syklisk masterplan for bruken av operasjonsstuene, hvor de ulike kirurgiske undergruppene er tidsplanlagt til de ulike operasjonsstuene gjennom uken. De fleste akademiske bidragene som tar for seg dette problemet har valgt å se bort fra akuttpasientene, og argumenterer for at disse tas hånd om av dedikerte ressurser. Selv om noen akademiske bidrag inkluderer usikkerhet relatert til for eksempel usikker liggetid etter operasjon eller varierende etterspørsel etter operasjoner, så er de aller fleste formuleringene deterministiske. Som et bidrag til den eksisterende litteraturen på området så foreslår vi en stokastisk tostegsmodell for å modellere den usikre tilstrømningen av akuttpasienter.

I optimeringsmodellens første steg bestemmer vi hvilke operasjonsstuetidsluker som skal være fleksible og hvilke som skal være dedikerte for elektive pasienter, og vi planlegger elektive operasjoner til de elektive tidslukene. I tillegg bestemmer vi antallet senger som skal være tilgjengelige på de aktuelle sengepostene hver dag gjennom uken. De usikre parameterne som blir tilgjengelige mellom de to stegene er antallet grønne pasienter som må opereres i løpet av uken, samt antallet akuttpasienter som opptar en seng på de ulike sengepostene hver dag gjennom uken. Den fleksible operasjonsstuekapasiteten som vi fastsatte i første steg brukes til å operere alle de grønne pasientene som tilkommer. Dersom vi har planlagt alle de grønne pasientene til en fleksibel operasjonsstuetidsluke og det fremdeles er

mer fleksibel stuetid igjen, vil den ekstra operasjonsstuekapasiteten bli brukt til å operere gule akuttpasienter (akuttpasienter som det haster mer med enn de grønne). Dersom det i første steg er satt av for lite fleksibel kapasitet vil dette medføre at grønne pasienter må planlegges til elektive operasjonsstuetidsluker, som igjen kan medføre strykninger av elektive operasjoner. I tillegg vil vi få elektive strykninger dersom det er fult på sengepostene og vi har planlagt flere elektive pasienter som trenger et sengeopphold etter operasjon.

Ved å kjøre optimeringsmodellen med virkelighetsnære instanser, finner vi at denne modellen gir ekstra verdi kontra den deterministiske formuleringen for samme problem. Videre utarbeider vi noen generelle retningslinjer for planlegging av fleksibel operasjonsstuekapasitet til elektive operasjonsstuer. Dersom operasjonsstuekapasiteten er ansett som bedre enn sengepostkapasiteten anbefales det å anvende en relativt stor andel fleksibel kapasitet. Dersom operasjonsstuekapasiteten derimot anses som knappere enn sengepostkapasiteten er rådet å anvende relativt lite fleksibel operasjonsstuekapasitet.

Da vi anvender en tostegs formulering for å modellere usikkerheten i problemet gjøres dette basert på en forenklet antagelse om at vi mottar all nødvendig informasjon vedrørende usikkerheten på et gitt tidspunkt (like før hver syklus). Dette, samt at vi genererer en syklisk plan for å håndtere en varierende etterspørsel, gjør at vi bør teste planen i et virkelighetsnært miljø for å undersøke hvor godt den fungerer. For å gjøre dette så utvikler vi en diskret hendelsesbasert simuleringsmodell for å representere driften ved en virkelig operasjonsklinikk. Entitetene som inkluderes i modellen er de elektive og de akutte pasientene, sengepostene samt både de akutte og de elektive operasjonsstuene. Det er enda en grunn for å anvende en simuleringsmodell i denne oppgaven: Vi behøver en måte å generere scenarier for å representere usikkerheten i optimeringsmodellen. Ved å trekke scenarier fra den simulerte virkeligheten vil vi kunne generere scenarier med utgangspunkt i ulike planleggingsregimer som implementeres i simuleringsmodellen.

Hovedhensikten med oppgaven er å gi taktisk beslutningsstøtte, vedrørende operasjonsstueplanlegging, til ledelsen ved klinikk for ortopedi ved St. Olavs Hospital. For å gjøre dette, anvender vi modellene for å generere alternative masterplaner for klinikken. Det viser seg at ved å dedikere seks fleksible operasjonsstuetidsluker til de elektive operasjonsstuene, vil man klare å operere både de grønne og de gule akuttpasientene raskere enn i dag. Videre vil tilføring av fleksibel operasjonsstuekapasitet gjøre klinikken i stand til å bedre håndtere perioder med stor etterspørsel etter akutte operasjoner. Som et resultat av dette vil det være mye mindre behov for replanlegging av elektive pasienter. En innføring av fleksible tidsvinduer til en viss grad redusere antallet elektive operasjoner som planlegges hver uke. Det vil derfor være en balansegang mellom antallet elektive pasienter man planlegger å operere hver uke og antallet elektive operasjoner som må replanlegges.

Tema for videre forskning vil være å videreutvikle modellen med fokus på mer effektive formuleringer. Dersom vi finner måter å løse problemet raskere på vil vi kunne anvende flere steg i formuleringen og slik kunne få en mer realistisk modell.

Summary

The regional health authority, Helse Midt-Norge (HMN), provides hospital services to the two counties Trøndelag and Møre og Romsdal. Forecasting estimates that the number of people living in these two counties that exceed 67 years old will increase by 42 % within 2030, yielding a massive increase of the demand for health care in the years to come (Helse-midt.no, 2016). St. Olav's Hospital is the largest hospital in HMN, and in this thesis we will consider the surgery scheduling at the orthopaedic department at St. Olav's Hospital. The orthopaedic department is responsible for treating all orthopaedic patients entering the hospital, both electives and emergencies. Treating the orthopaedic emergency patients within the dead lines proposed by the hospital has been an issue for many years. In addition to being unfortunate for the emergencies, delaying the emergencies also affects the elective patients. As the queue of emergencies grows, more emergency patients are scheduled for the elective ORs implying elective rescheduling. In addition, most of the emergency patients cover beds at the wards while waiting for surgery. In periods of many emergencies waiting for surgery, the bed capacity may become scarse, yielding rescheduling of elective inpatients. The least urgent of the emergency patients, the green emergencies, are the first ones to be displaced from the emergency ORs in periods of excessive emergency demand. In addition, only about 70 % of these patients receive surgery within the dead line, and dedicating more OR capacity for the green emergency patients should be considered.

The main problem faced in this report is the Master Surgery Scheduling Problem (MSSP). This problem consists of developing a cyclic Master Surgery Schedule (MSS), linking the surgical subspecialties to the different ORs through the cycle, which is typically set to one week. The majority of authors on the field focus exclusively on the elective patients, arguing that the emergencies are treated by dedicated resources. Although some authors address uncertain aspects relevant to surgery scheduling, the majority of optimization models provided on the field are deterministic. As a contribution to the existing literature, we propose a two-stage stochastic recourse formulation to address the stochastic arrival of emergency patients when solving the MSSP.

In the first stage of the optimization model, we schedule the OR capacity available as either flexible or dedicated for electives, and we schedule elective patients for surgery in the elective OR slots. In addition, we schedule the amount of beds to be available at the wards on each day of the cycle. The stochastic parameters in the problem are the number of green emergencies that need to be scheduled in the cycle, and the daily number of emergency patients covering beds at the wards. The flexible OR capacity is dedicated to handle the weekly demand of green patients, and we require that all green patients should be scheduled for surgery in the second stage. If all the green emergency patients are scheduled and we still have excess capacity of flexible slots, more urgent emergencies are scheduled for these slots. If there are too few flexible slots available to treat all green emergencies present in the cycle, these patients need to be scheduled for the elective ORs, which may imply elective cancellations. In addition, if the bed capacity is reached and there are still elective inpatients left to be treated, these surgeries are cancelled.

From running the optimization model on realistic size instances we find that applying a stochastic formulation provides additional value compared to the deterministic counterpart. Furthermore, we develop some general advises regarding the scheduling of flexible slots to elective ORs. If the OR capacity is regarded as better than the ward capacity, a relatively high share of the ORs should be made flexible. If however the bed capacity is good, and the OR capacity is scarse, less OR capacity should be scheduled as flexible.

Applying a two-stage formulation to model the uncertainty means that we assume that all necessary information regarding the uncertainty will be made available to us at one specific point in time (before each cycle). This assumption, together with the fact that we generate a cyclical schedule to handle real life fluctuations over time calls for some way to test the robustness of the schedules proposed. To do this, we develop a discrete event simulation model that represents a real life hospital department. The entities considered in the system are the elective patients, the emergency patients, the wards, the elective ORs and the emergency ORs. There is yet another reason for developing the simulation model: We require a way of generating scenarios representing the stochastic parameters applied in the optimization model. The scenarios are generated from the simulation model, allowing us to generate scenarios that are dependent on the scheduling regime implemented in the simulation model.

The main purpose of this thesis is to provide tactical decision support for the management at the orthopaedic department at St. Olav's Hospital, and we perform a case study of the department, applying both the models in a loop. By scheduling six flexible slots of four hours to the elective ORs we show that both the green and yellow emergencies receive surgery faster compared to the historical data. In addition, the flexible slot capacity makes the department better prepared for handling fluctuations in the demand for emergency surgeries. Because this, far less elective rescheduling is needed. However, scheduling flexible slots yields a decrease in the number of electives scheduled, and there exist a trade-off between the number of electives scheduled versus the amount of elective rescheduling that need to be made when the system is exposed to emergency patients.

Topics for further research include issues regarding symmetry in the optimization model formulation. If we can provide more efficient formulations, we may have the opportunity to impose more stages in the stochastic formulation.

Contents

List of Figures

List of Tables

1 Introduction

The regional health authority, Helse Midt-Norge (HMN), provides hospital services to the two counties Trøndelag and Møre og Romsdal. By the end of 2015, there were a total of 710,000 inhabitants living in these two counties, and forecasting estimates that the number will increase by 93,000 people by the year of 2030. In the same period, the amount of people living in the area who are older than 67 years old will increase by 42 % (Helse-midt.no, 2016).

Numbers from the two counties covered by HMN reveal that 10 % of the population consume 70 % of the health care services. The main part of these 10 % consist of people spanning the age from 65-79 years (Helse-midt.no, 2016). As the population grows and the inhabitants get older, the demand for health care services will increase in the years to come. Numbers from HMN suggest that by the year of 2030, there will be an increase of hospital health care demand corresponding to a 25 % increase of full time equivalents in the hospital sector. Having such a large part of the society working within health care will not be sustainable, so to cope with the demographic changes the hospitals need to treat more patients per full time equivalent.

One of the major activities at a hospital, is providing surgery to patients. Freeman et al. (2017) state that 60-70 % of all patients admitted to a hospital require some surgical intervention, while van 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 of efficient surgery scheduling will therefore be positive in terms of economy, but also necessary to treat more patients in the years to come.

In this thesis we will consider the surgery scheduling at the orthopaedic department at the biggest hospital in HMN, St. Olav's Hospital. The orthopaedic department is responsible for handling all orthopaedic patients entering the hospital, both electives and emergencies. In order to treat both electives and emergencies, the flow of patients is separated. The emergencies are mainly treated in three of the operating rooms (ORs), while eight ORs are dedicated for electives. In contrast to the elective patients that may wait for months to have their surgery, the emergency patients require surgery within some hours or a few days. Treating the emergencies within time has been an issue at the orthopaedic department ever since the hospital was built in the early 2000s. Failing to provide surgery for emergency patients within a short time will also affect the scheduling of elective patients. In periods of high emergency demand, the emergency OR capacity is not sufficient to handle the work load, implying that emergencies need to be scheduled for the elective ORs. As a consequence of this, elective surgeries need to be rescheduled in order to provide OR capacity for the emergency patients. When waiting for surgery, most of the emergency patients cover beds at the wards. Having many emergencies waiting for surgery may therefore lead to elective rescheduling as no beds are available to host the electives following surgery. Finding ways of handling the emergency patients more efficiently will therefore yield less elective rescheduling in addition to avoid excessive waiting time for the emergencies.

There is much literature available on the topic of surgery scheduling within operations research. The main problem faced in this report is the so called *Master Surgery Scheduling Problem* (MSSP) which is about scheduling the different surgical specialties (or subspecialties) to the different operating rooms in order to provide surgery to as many patients as possible. The schedule is normally made for one cyclic week, and this weekly schedule is typically applied for a period of six to twelve months. In addition to including the surgical specialties and the operating rooms, some authors like for example Adan et al. (2011) and Testi and Tànfani (2008) include other resources like the wards and the intensive care unit as they may impose bottlenecks for an efficient flow of patients. Most of the literature provided on the MSSP consider only the elective patients, arguing that the emergencies are taken care of by dedicated resources. There are however some contributions, like Adan et al. (2011) that consider the emergencies by estimating their arrivals based on a Poisson process and reserve capacity for the patients based on this.

The main purpose of this report is to provide tactical decision support for the management at the orthopaedic department when developing the Master Surgery Schedule (MSS). To do this, we develop a planing tool consisting of both an optimization model and a simulation model. These two models are set up in a loop, where the optimization model generates a MSS, while the simulation model runs this MSS in a real life environment in order to investigate how the schedule performs at an operational level. The simulation is also providing information back to the optimization model in order to alter the MSS based on the operational outcomes. Our main contribution to the department is the presentation and analysis of three MSSs that are proposed by the model.

We also provide contributions of academic interest. In order to deal with the uncertain arrival of emergency patients, a stochastic two-stage formulation is proposed. Bruni et al. (2014) propose a stochastic two-stage model for an operational-level surgery scheduling problem, but to our knowledge, such a model framework has not been provided by any authors on the MSSP before. The stochastic parameters applied in our stochastic two-stage formulation is the amount of emergencies covering beds at the wards each day of the cycle, and the amount of sub-urgent emergencies (referred to as the green emergencies) that need to be scheduled for surgery within the cycle. These parameters are generated by the simulation model, and by applying different scheduling regimes in the simulation model we may alter the scenarios that are fed to the optimization model.

The rest of the thesis is structured as follows. In Section 2, background on St. Olav's Hospital and the orthopaedic department is provided and we introduce relevant methods for modelling uncertainty. Then, in Section 3 the MSSP is introduced, and in Section 4 we provide related literature. The mathematical two-stage formulation is presented in Section 5, while the simulation model is introduced in Section 6. Then, in Section 7 we describe how the simulation model and the optimization model are connected, and the scenario generation procedure is presented. Following this, in Section 8 a computational study is performed, including both technical studies of the optimization model, and managerial studies relevant to hospital departments that perform surgeries and in particular the orthopaedic department at St. Olav's Hospital. Finally, in Section 9, we wrap up the thesis with concluding remarks and proposals for further research.

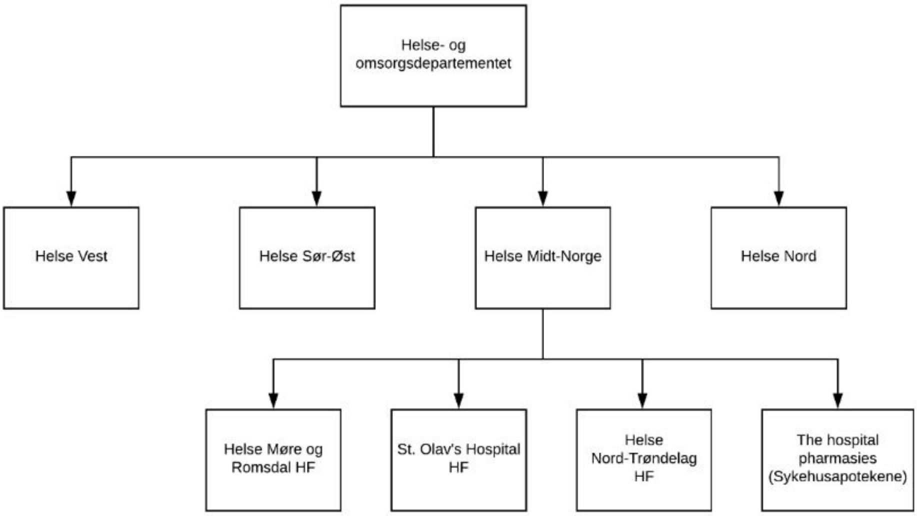


Figure 1: The organization of governmental-owned health care in Norway

2 Background

In this section we introduce St. Olav's Hospital, and the orthopaedic department which is the case department in this thesis. We also provide relevant hospital terminology, and finally we introduce relevant methods for modelling uncertainty.

2.1 Introducing St. Olav's Hospital

Figure 1 illustrates the organization of the governmental-owned hospitals in Norway. The highest level of health care management belongs to the ministry of health care, *Helse- og omsorgsdepartementet.* The ministry of health care manages the four regional health care authorities, that again are responsible for the governmental-owned hospitals throughout the country. *Helse Midt-Norge* is one of these regional health care authorities, and they own St.Olav's Hospital which is a university hospital and the largest hospital in its region.

St. Olav's Hospital is a health authority situated at several locations in the county of Trøndelag, including Trondheim, Røros, and Orkdal. The largest hospital unit is situated at Øya in Trondheim. St. Olav's Hospital consists of 20 departments, that are listed in Table 1, and it is considered a large hospital by Norwegian standards. In 2017 there were 10,483 employees at the hospital performing a full-time equivalent of 8,063. The same year there were roughly 670,000 outpatient consultations undertaken, yielding a decrease of 3 % compared to 2016 (Stolav.no, 2017). Table 2 includes additional key numbers for St. Olav's Hospital.

Table 1: The different departments at St. Olav's Hospital

Departments
Department of mental health care
Department of operations service St. Olav's
Department of children and youth
Surgical department
Emergency department
Department of anesthesia and intensive medicine
Department of imaging diagnostics
Department of physical medicine and rehabilitation
Department of heart medicine
Department of clinical service operations
Department of lunge and occupational medicine
Department of orthopedics, rheumatology and skin diseases
Department of intoxication and addiction medicine
Department of thoracic surgery
Department of ear nose and throat, jaw, and eye diseases
Cancer department
Women department
Department of laboratory medicine
Department of medicine
Neurological department

Table 2: Key numbers for St. Olav's Hospital, 2017

Outpatient consultations	669,427
Consultations (somatic)	453,059
Consultations (psychiatry (adults))	138,550
Consultations (psychiatry (children))	55,770
Intoxication	22,048
Mean waiting time	56 days
Mean length of stay (somatic patients)	4.25 days
Maximum beds available	983
Beds available (somatic patients)	983
Employees	10,483
Full-time equivalent	8,063

2.2 Relevant hospital terminology

In this subsection all hospital terminology relevant for the thesis is provided. Where it comes naturally, the terminology is related to the orthopaedic department and the orthopaedic patients.

2.2.1 The diagnosis related groups system

In order to measure the activities at a hospital, and to compare hospitals, a common system based on diagnosis related groups (DRG) has been developed. The DRG is a patient classification system where all patient-related interventions at a somatic hospital is categorized. The classification is based on diagnostic similarity, and the amount of resources required in the treatment of different patients, and all patients and interventions give rise to an amount of DRG points. The main variables that influence the DRG points are: The diagnosis, the procedure performed, the sex of the patient, the age of the patient, and the state of the patient when leaving the hospital. By using DRG points as a measure of activity, one is able to compare hospitals independent of which patients are treated at the hospital (Helsedirektoratet.no, 2017). Since the DRG system is a quantitative measure reflecting both quality and quantity of hospital services it is commonly used when developing annual budgets both at the hospital as a whole, and also for the individual departments.

2.2.2 The dividing of patients

The most fundamental dividing of patients, from a planning point of view, is the separation of patients into elective and emergency patients. The elective patients suffer from some slow-developing or chronic disease, and will typically require surgery in some months. As there is no urgency to provide surgery for these patients, the scheduling of the elective patients is quite predictable. The emergency patients typically suffer from some trauma or fast-developing condition and require more or less immediate intervention. It is common to further divide these patients into subgroups based on the degree of urgency for treatment. Due to the acute nature of the conditions harming these patients, they arrive at the hospital randomly. This, together with the fact that they require immediate treatment makes the scheduling of the emergencies hard.

At a hospital it is crucial to know for how long a patient will stay to recover after surgery. This gives birth to another classification of patients into either inpatients or outpatients. The inpatients require some time, typically some days, to recover in a bed following the surgery. This recovery is done at a ward within the department. The outpatients enter and leave the hospital at the day of surgery, making these patients less demanding in terms of resources.

2.2.3 The staff required for surgeries

In order to perform surgeries, a surgical team is needed. A surgical team typically consists of two surgeons, three to four nurses and an anesthesiologist.

The surgeons are the surgical staff actually performing the surgery, and these are often divided into subspecialties. For instance there may be some orthopaedic surgeons specialized on performing knee surgeries, while others are experts at performing shoulder surgeries. The surgeons are not only performing surgeries, they also need to serve the wards in order to discharge those patients that have recovered from surgery, and to follow up on patients who are still recovering. In addition, these doctors need to serve the outpatient clinic in order to examine patients entering the hospital and decide on who should receive surgery.

The anesthesiologists are the doctors responsible for providing anesthesia to the patients that are to receive surgery. The number of operations starting at the same time may therefore be limited by the amount of anesthesiologists available. Some patients require only local anesthesia, while others need full narcosis, making the demand for anesthesiologists fluctuating. At St. Olav's Hospital the anesthesiologists belong to the department of anesthesiology and intensive medicine, but they are serving the other departments and should therefore be regarded as a shared capacity.

The nurses are central in all processes regarding the flow of patients through the hospital system. They are managing the wards and they are the only personnel usually present through the entire surgery.

2.2.4 The operating theatre

At some departments, like the orthopaedic and the surgical department, surgeries are among the main activities. These departments usually have an operating theatre which is the place where all the operating rooms (ORs) are located. An operating theatre may consist of several ORs, preparation rooms and storage rooms for surgical equipment. The preparation room is a room used to prepare patients for surgery by cleaning the patient and providing anesthesia. If utilized in a proper manner, this room can be of great importance in order to attain an efficient flow of patients: By doing all preparations in the preparation room, rather than inside the OR, one is able to start surgery of the next patient as soon at the OR is cleaned after the former patient.

The ORs within an operating theatre are often used for different surgeries based on properties related to the room or the location of the room. If an OR is to be used for infected patients there should for example be a decontaminator for sterilizing equipment close by. In addition, there should be a small room in between of these ORs and the rest of the operating theatre to provide isolation. Some ORs are carefully equipped in order to perform certain kinds of surgery. On example is a back-table used for certain types of back surgeries.

2.2.5 The surgery and the postoperative recovery

The surgery consists of three phases: pre-surgery time, knife time, and post-surgery time. In Figure 2 the three phases are illustrated along a time line. In the first phase there is typically one anesthesiologist providing the anesthesia, and three to four nurses preparing the equipment needed and positioning the patient for the surgery to come. These nurses are present throughout the surgery. When the anesthesia has started working and the patient is ready for surgery, the surgeon enters the OR, and the anesthesiologist leaves. When the surgeon is done she also leaves the room, and the nurses are responsible for monitoring the patient (if the patient has received a lot of anesthesia, an anesthesiologist is responsible for the monitoring) before he is transported out. Following the surgery, the OR should be cleaned before the next patient enters.

Following surgery the patient is transported to the post anesthesia care unit (PACU) in order to be monitored for some time. If the patient is medically unstable and needs to be closely monitored for some prolonged period, he is sent to the intensive care unit (ICU). After the initial monitoring, the inpatients are sent to the wards for further recovering. Different patients go to different wards, and different wards have different capacities in terms of beds available. The length of stay of inpatients at the wards has been a hot topic at hospitals around the world for some years now, and the mean length of stay has significantly decreased in order to treat more patients. In 2017, the mean length of stay of patients at St. Olav's Hospital was 4.25 days (Stolav.no, 2017).

2.2.6 Uncertainty related to surgery scheduling

There are several aspects regarding surgery scheduling that are prone to uncertainty. The arrival of emergency patients may vary much from one period to the next, affecting both the activity at the ORs, but also at down-stream activities at the PACU, the ICU and the wards. Furthermore, the surgery duration of patients will vary depending on the diagnosis, the age of the patient, the level of experience obtained by the surgeon and unforeseen complications during the surgery. The length of stay of patients following surgery may also vary depending on the diagnosis and the age of the patient. In addition, aspects of a more social character may influence on the length of stay: If the patient has no one at home that may look after him, he may need some more time at the hospital, also if the patient is not supposed to go home, but rather to an municipal care-institution, the patient has to wait at the hospital until there is room for him at the care-institution.

Dealing with parts of the uncertainty is very important when solving real life problems. Including all aspects of uncertainty is seldom expedient or necessary, but failing to include uncertainty that is relevant to the problem will lead to solutions that tend to be irrelevant and overly optimistic.

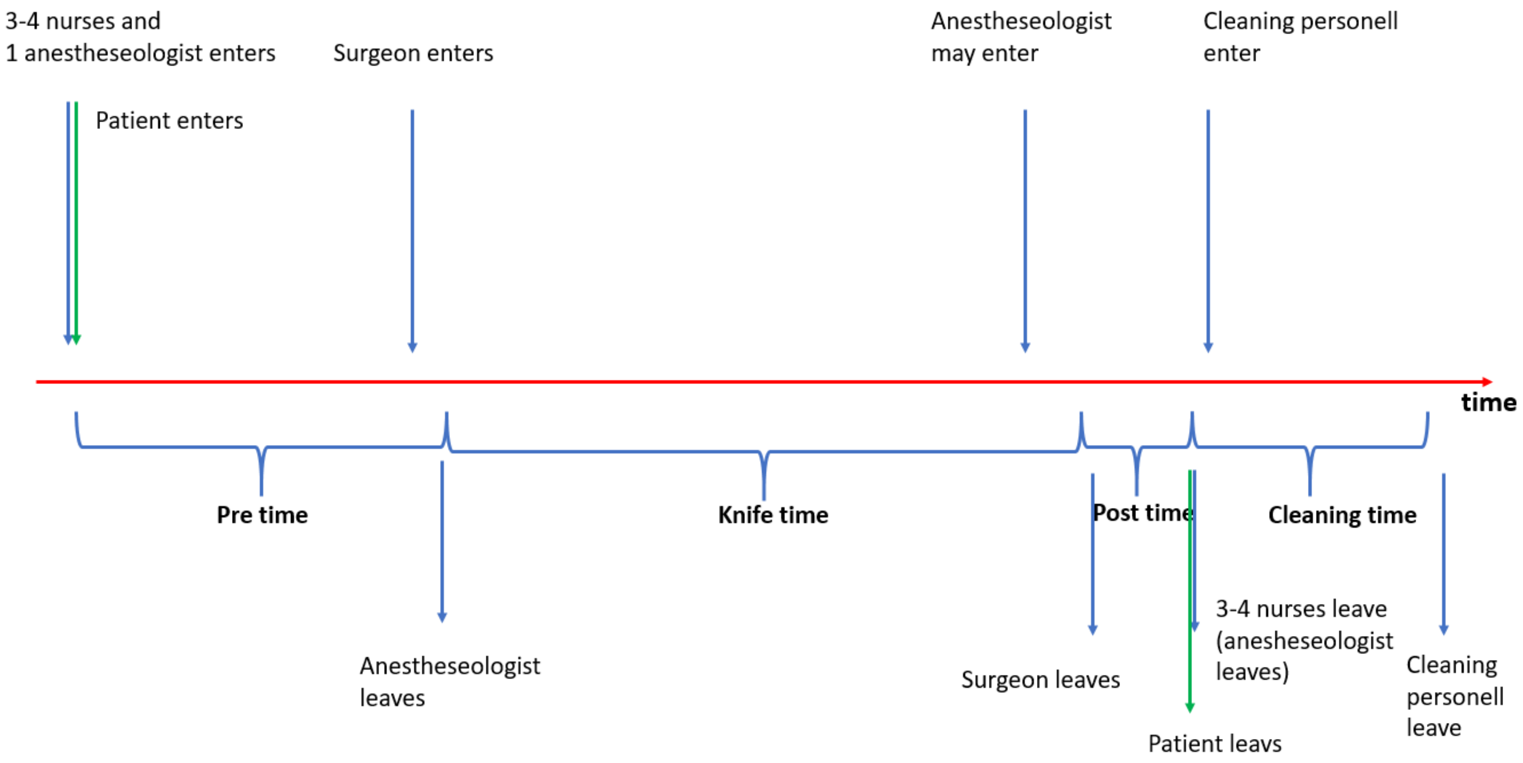


Figure 2: Time line for a surgery

2.3 The orthopaedic department

The case department considered in this thesis is the orthopaedic department at St. Olav's Hospital, Øya. The orthopaedic department is responsible for treating the orthopaedic patients entering St. Olav's Hospital, and it is located in a building called *Bevegelsessenteret* (BVS). In addition, they have access to ORs in two other buildings, named *Akutten- og hjerte-lunge-senteret* (AHL) and *Kvinne- og barnsenteret* (KBS). Figure 3 illustrates the hospital buildings at Øya, and the three buildings where the orthopaedic department disposes ORs can be seen.

The surgeons working at the orthopaedic department are divided into subspecialties making them able to perform a wide variety of surgeries. The subspecialties are as follows: *Elective foot, plastics, reconstructive, elective trauma, hand, arthroscopic, back, prosthesis and children.* The surgeons representing the plastic subspecialty are actually not part of the orthopaedic department, but the surgical department. However, some of the ORs at BVS are partly dedicated to the plastic patients, and so are some of the beds at the wards belonging to the orthopaedic department.

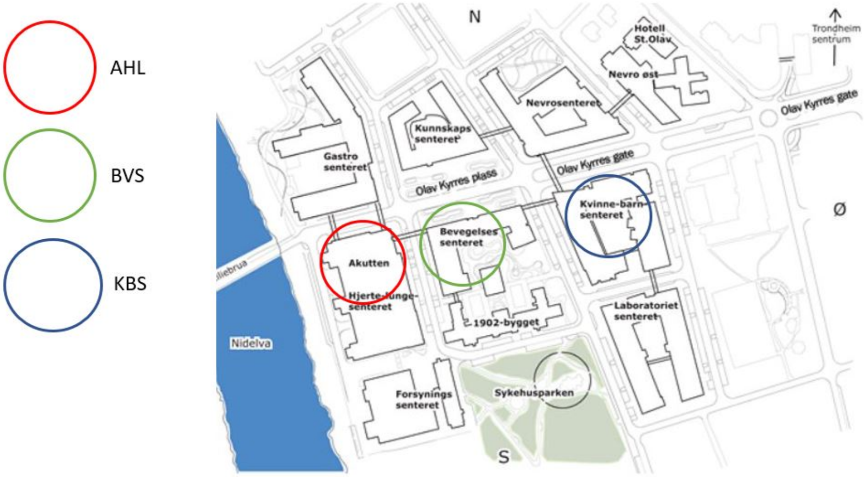


Figure 3: Map of St. Olav's Hospital, Øya.

2.3.1 Operating rooms available

There are a total of eleven ORs available to the orthopaedic department, and eight of these are located at BVS. In addition, there are two ORs located at AHL, and yet another one located at KBS. Seven of the ORs located in BVS (OR-2 - OR-8), and the one located in KBS, are dedicated for elective surgeries, while the rest are dedicated for emergencies. In Figure 4 the regular opening hours for the ORs at disposal for the orthopaedic department are illustrated. All the ORs are open between 07.45-15.30 during the weekdays, except of the two rooms at AHL. At AHL, one of the rooms is dedicated for the orthopaedic department between

07.45-22.00 from Monday to Thursday, while the other one is dedicated for the orthopaedic department between 07.45-15.30. From 15.30-22.00 this room serves as a shared capacity accessible to all departments with emergency patients. From 22.00-07.45 there is one room available at AHL to perform surgeries to the most urgent patients who can't wait till the next day. In the weekends the ORs at BVS and KBS are closed, but there is one OR available at AHL in order to perform surgeries on emergencies entering in the weekend. This room is a shared capacity with the rest of the hospital.

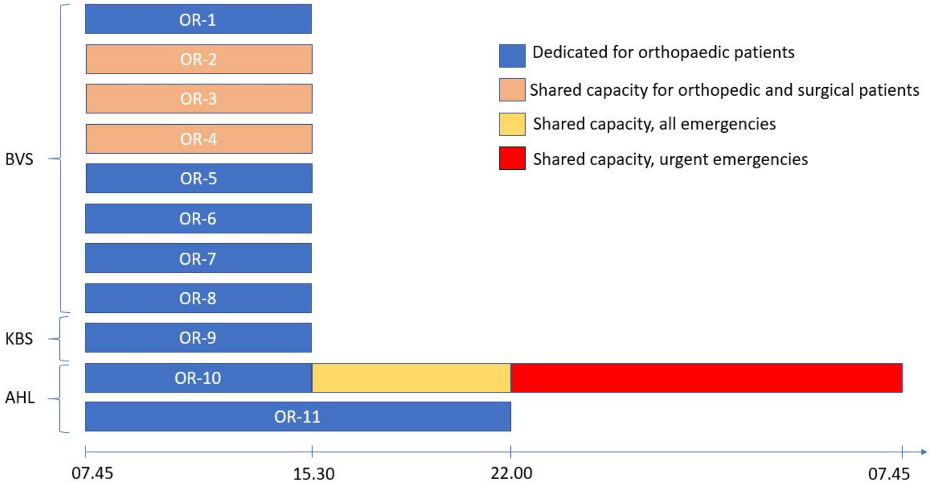


Figure 4: The opening hours of the ORs as BVS, KBS and AHL from Monday to Friday.

2.3.2 The wards available

There are a total of six wards wards available to the inpatients at the orthopaedic department. The wards host different patient categories, and part of the capacity is shut down during weekends and holidays (see Table 3). Note that the capacities provided in Table 3 are not the total number of beds available, but the staffed beds. The total bed capacity available to the orthopaedic department is about 90 beds, but not all of these are staffed, leaving about 20 beds unused during the weekdays. Since there is a decrease of staffed beds during the weekend, there is a policy at the department that no elective inpatients should receive surgery on Friday as these patients consume beds following surgery. There may be times when some of the patients resting at the wards that close down during the weekends need one or two additional nights at the hospital. These patients are then moved to some other ward for the weekend. If the surgeons covering the weekend shift do not know the patients that are left at the wards, they may not feel confident in sending the patient home during the weekend. As a consequence of this, there may be patients covering beds on Monday morning that were meant to leave during the weekend.

Table 3: The wards at the orthopaedic department

Name	Floor	Capacity weekday	Capacity weekend	Patient categories
Fast-track hip and knee	4th (west)	16	0	Hip and knee prosthesis
Hotel-day	5th (north)	5	0	Prepare electives before surgery, buffer capacity
Elective	5th (north)	10	12	Infected hip/knee, back, arthroscopics
Plastic	5th (west)	3	3	Large plastic surgeries
Reconstructive	5th (west)	13	13	Amputations, fire, skin- and muscle
Trauma	6th (north and west)	20	16	Fractures and trauma

The dynamic at the wards is crucial to understand when developing tools for efficient flow of patients. There are basically three elements that are challenging when dealing with the wards at the orthopaedic department: Emergencies occupying beds while waiting for surgery, emergencies waiting to be dismissed following surgery, and the fluctuating inflow of emergencies. If the OR capacity at AHL is insufficient (which may be the case in periods of many emergencies entering the hospital), the emergency inpatients will occupy beds while waiting for surgery. If the municipal health-care institutions are struggling to provide beds to host the emergencies following the hospital stay, these patients have to wait at the wards in the hospital. Finally, the inflow of emergencies to the orthopaedic department fluctuates from one day to the next. Because of these three factors, the number of beds covered at a ward each day is both fluctuating and unpredictable.

At the orthopaedic department they take different actions as the loading of beds approaches the limiting capacity, which may be seen in Figure 5. First, they reorganize the patients at the different wards to better utilize the capacity at hand. This may result in patients belonging to one ward staying a night or two at some other ward. They also increase the bed capacity during weekends by not lowering the capacity as much as usual. Then, elective inpatients are rescheduled or cancelled, and elective outpatients are prioritized as they do not require any stay at the wards following surgery. The last patient categories to be cancelled are the fast-track patients, as these generate the most DRG-points. Finally, if there is still a chance of capacity limitations, extra staff is called in to increase the number of beds available.

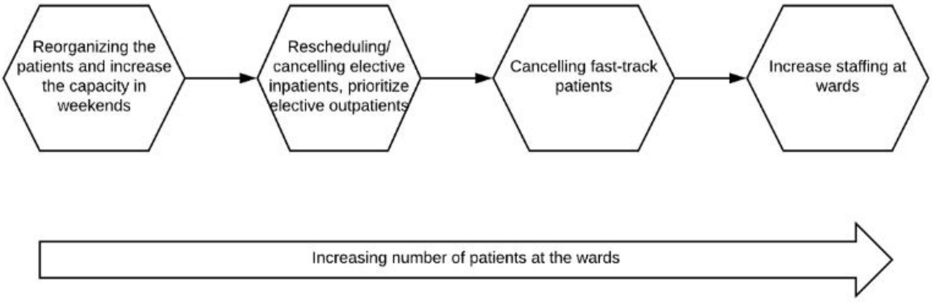

Reorganizing the patients and increase the capacity in weekends
Rescheduling/ cancelling elective inpatients, prioritize elective outpatients
Cancelling fast-track patients
Increase staffing at wards
Increasing number of patients at the wards

Figure 5: Strategies to deal with increased bed loading.

2.3.3 The orthopaedic patients

The orthopaedic patients are divided into five different groups based on the medical urgency of their condition, and the different groups may be seen in Figure 6. The emergency patients are divided into three groups, where the most urgent ones should receive treatment within six hours, the intermediate ones within 24 hours, and the least urgent group should receive surgery within five days. The three groups are divided from each other by referring to a traffic light system: *Red* for the most urgent ones, then *yellow* and *green*. This system, and the limits proposed, are decided on by St. Olav's Hospital and are serving as guidelines rather than absolute limits.

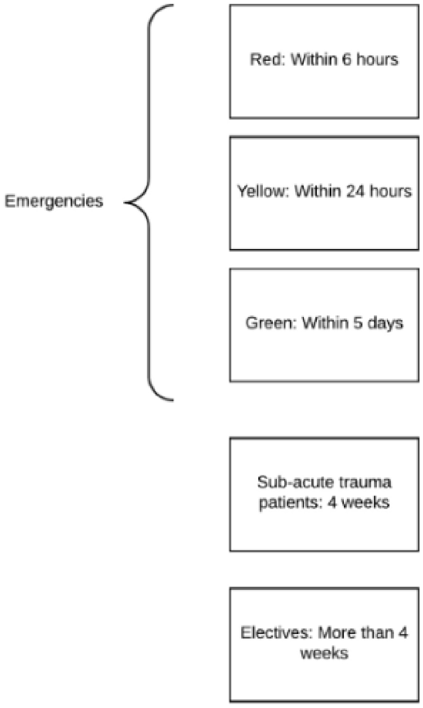


Figure 6: Medical urgency of the orthopaedic patients

The orthopedic elective patients typically suffer from slow-developing diseases like osteoarthritis or other chronic diseases harming the mobility and potentially the quality of life of these patients. These patients are in need of surgery to regain their daily function, but the need of surgery is not as urgent as for the emergency patients. As a rule of thumb these patients should be scheduled for surgery about three months ahead.

The last group of patients, the sub-urgent trauma patients, should receive surgery within four weeks. These patients typically suffer from pain and dysfunction due to some complication following their last surgery. The limit of four weeks is not an absolute medical limit, but it indicates that these patients, because of the pain condition they are in, should not have to wait several months for surgery. However, the four week limit has another very important meaning: It represents the minimum amount of time required to schedule elective patients to the ORs at BVS. If the limit is set to four weeks or below, these patients are, from a scheduling perspective, categorized as emergency patients, implying that they are scheduled to OR-1 at BVS or to the ORs at AHL instead of to the elective ORs at BVS.

In Table 4 the distribution of the orthopaedic patients receiving surgery in the period 01.01.15-27.04.17 may be seen. We were not able to identify the sub-urgent trauma patients from the data provided. Note that almost half of the patients are emergency patients.

Table 4: The distribution of orthopaedic patients that received surgery in the period from 01.01.15-27.04.17

Urgency	Inpatient	Outpatient
Elective	3424	4885
Green	1172	1380
Yellow	3284	204
Red	1351	35

2.3.4 The flow of orthopaedic patients

The flow of patients at the orthopaedic department may be split into six: the elective inpatients, the elective outpatients, the sub-urgent trauma patients, the emergency inpatients, the emergency outpatients and the emergency outpatients that require a bed following surgery. However, for the purpose of this thesis we may aggregate the flows into the four following: the inpatient and outpatient flows of both elective and emergency patients. Next, these four flows are presented.

The flow of the elective patients

The flow of elective patients is illustrated in Figure 7. All elective orthopaedic patients enter the hospital through the orthopaedic outpatient clinic. Here, they are assessed by an orthopaedic surgeon who decides whether surgery is necessary, and if so, by when surgery is needed. The surgeon also decides whether the patient should be an inpatient or an outpatient. Regardless of this, the patient is sent back home. The surgeon puts the patient on a list for surgery, and delivers this to the scheduling coordinators sitting at the patient intake office, who schedule the patient for surgery. When a date for surgery is set, the patient receives a message about this.

If the patient is an inpatient, he is summoned to a preoperative assessment some days prior to the surgery. Here, the patient meets the surgeon who will be performing the surgery, he is assessed by an anesthesiologist and he is provided with information from both a physiotherapist and a nurse about the process following the surgery. After the preoperative assessment the patient is sent back home before reentering the hospital at the day of surgery. On the day of surgery the patient enters through the hotel-day ward where a general preoperative preparation is performed. Afterwards, the patient is transported to the operating theatre where the final preparations, such as providing anesthesia, are made. Then, the surgery is performed, and afterwards the patient is sent to the recovery area to recover form the anesthesia. When the patient is medically stable he is sent to one of the wards to rest for some days.

The elective outpatients are not summoned for a preoperative assessment before surgery as these patients are generally in a better condition than the inpatients, and the surgeries are often less comprehensive. On the day of surgery these patients enter through the outpatient recovery area, where they are prepared for surgery.

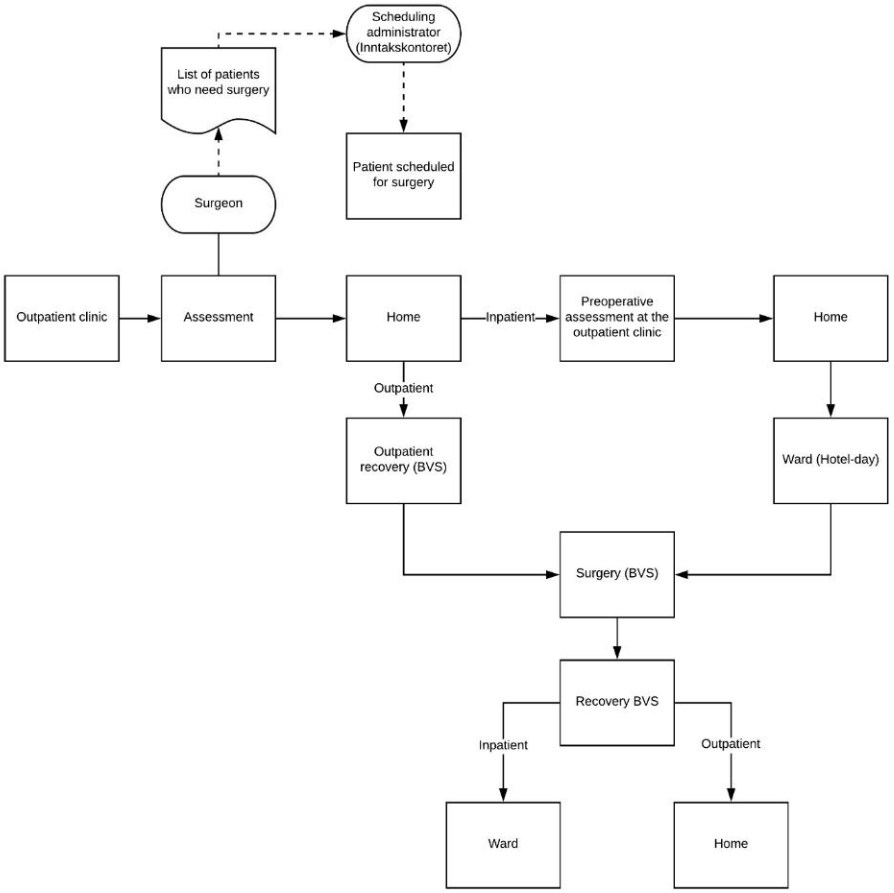

Scheduling administrator (Inntakskontoret)
List of patients who need surgery
Patient scheduled for surgery
Surgeon
Outpatient clinic
Assessment
Home
Inpatient
Preoperative assessment at the outpatient clinic
Home
Outpatient
Outpatient recovery (BVS)
Ward (Hotel-day)
Surgery (BVS)
Recovery BVS
Inpatient
Outpatient
Ward
Home

Figure 7: The flow of elective inpatients and outpatients

Following surgery they are recovering from the anesthesia in the recovery area, before they are sent back home.

The flow of emergency patients

The flow of emergency patients is illustrated in Figure 8. Almost all yellow and red patients, and a little less than half of the green patients, are inpatients, and they enter the hospital mainly through the emergency department. After arrival, the patient is first assessed by a nurse, and later by an orthopaedic surgeon deciding the degree of urgency according to the traffic-light system. When the degree of urgency is decided on, the patient is sent to a ward to wait for surgery, and an electronic note is sent to a scheduling administrator at OTS who schedules the patient for surgery, either at AHL or to OR-1 at BVS. Almost all red and yellow patients are scheduled for AHL, while a fair part of the green patients are scheduled for OR-1 at BVS. Following surgery the patient stays for some time in the recovery area before being sent to the wards at BVS for further recovery. Because the recovery area at AHL is a shared capacity for all emergency patients receiving surgery at AHL, this recovery tend to be crowded. Therefor, a fair part of the orthopaedic patients who have received surgery at AHL are sent to the recovery area at BVS.

Almost all emergency outpatients are green patients. The emergency outpatients mainly enter through the trauma outpatient clinic, where they are assessed by an orthopaedic surgeon. If surgery is needed, an electronic note is sent to OTS, and here the patient is scheduled for surgery at either AHL or BVS. On the day of surgery the patient enters the hospital through OTS, before being transported to the operating theatre at either BVS or AHL. If the patient is scheduled for AHL he may be displaced by more urgent patients, and if no OR is idle through the day, he is sent back home and need to be rescheduled for some later day. Following surgery, the patient recovers at the recovery area before leaving the hospital.

2.3.5 The surgery scheduling

An illustration of the scheduling process at the orthopaedic department may be seen in figure 9. The hospital has an economical budget to fulfill every year, and the management of the hospital decide on an annually target DRG that each department should fulfill in order to deliver on budget. Besides fulfilling the DRG target, the orthopaedic department also need to cope with the fluctuating patient demand for surgery. In periods this means treating a fair part of patients that are not very attractive in terms of DRG, but that have to have their surgery. However, these short-term fluctuations in patient demand tend to equal out over the horizon of one year, meaning that the department is usually delivering on budget at the end of the year.

In order to run the ORs effectively, a *Master surgery schedule* (MSS) (see Figure 10) is developed. The MSS is a schedule that links the different subspecialties to the different ORs through the week. For each OR there is at most one subspecialty scheduled on a given day. The MSS developed is a cyclic schedule, where the weekly schedule is repeated throughout the planning horizon. The planning horizon is set

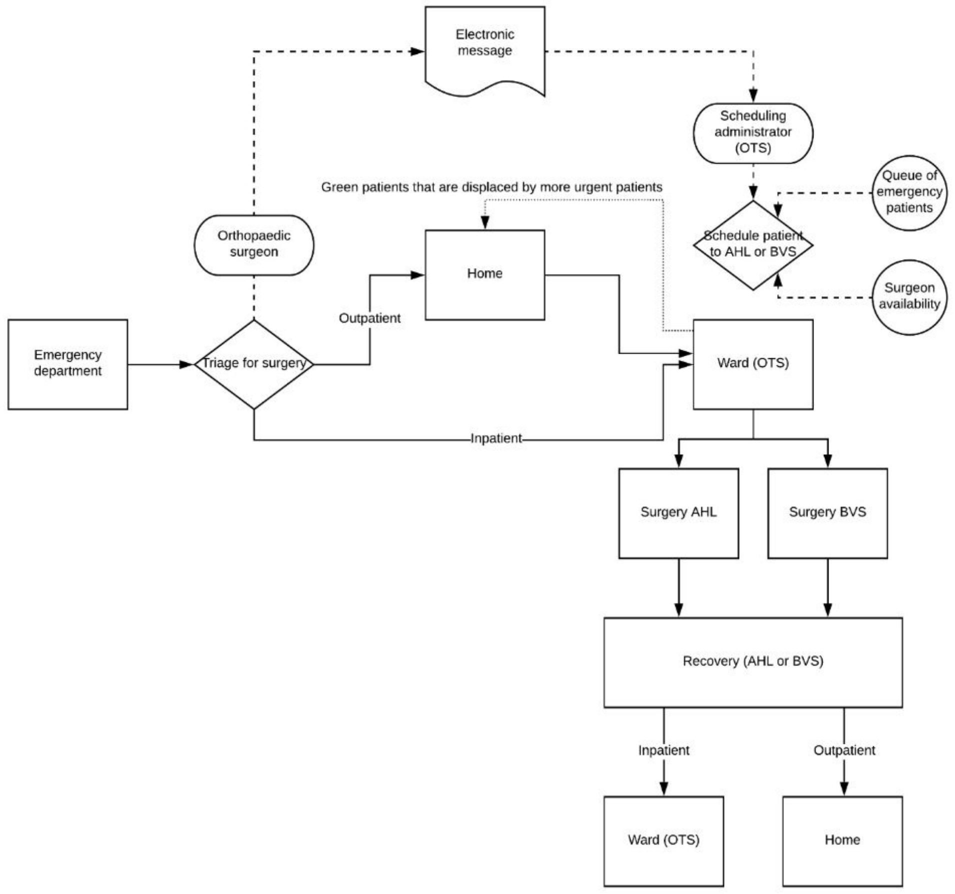


Figure 8: The flow of emergency patients

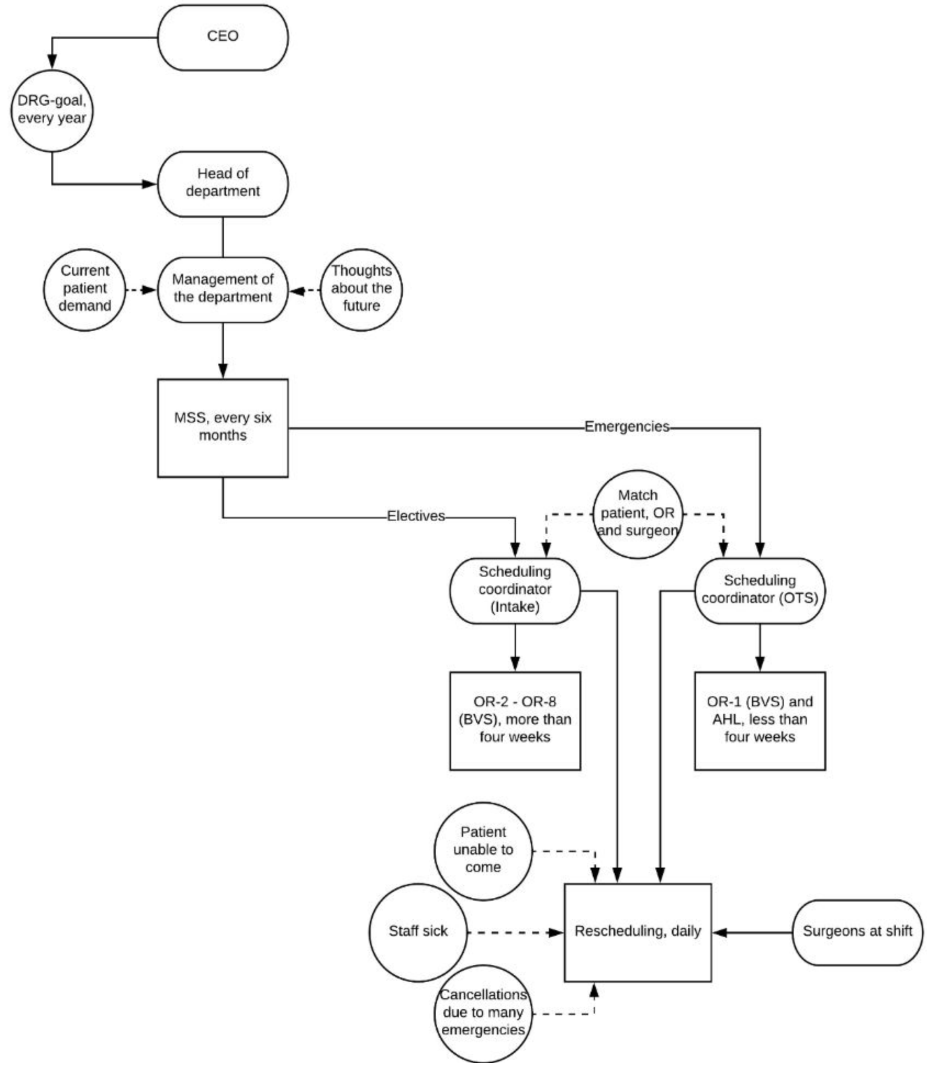
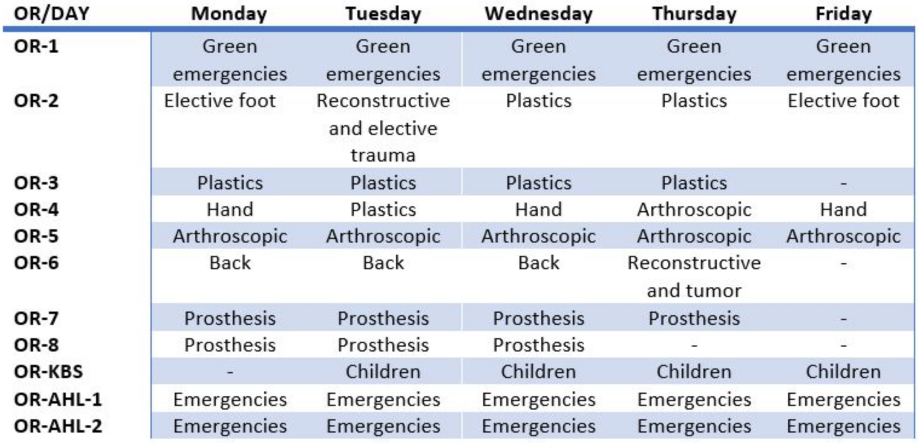

CEO
DRG-goal, every year
Head of department
Current patient demand
Management of the department
Thoughts about the future
MSS, every six months
Emergencies
Electives
Match patient, OR and surgeon
Scheduling coordinator (Intake)
Scheduling coordinator (OTS)
OR-2 - OR-8 (BVS), more than four weeks
OR-1 (BVS) and AHL, less than four weeks
Patient unable to come
Staff sick
Rescheduling, daily
Surgeons at shift
Cancellations due to many emergencies

Figure 9: The scheduling process at the orthopaedic department

to be half a year in order to, if necessary, reallocate resources in order to cover the patient demand for surgery.



OR/DAY	Monday	Tuesday	Wednesday	Thursday	Friday
OR-1	Green emergencies	Green emergencies	Green emergencies	Green emergencies	Green emergencies
OR-2	Elective foot	Reconstructive and elective trauma	Plastics	Plastics	Elective foot
OR-3	Plastics	Plastics	Plastics	Plastics	-
OR-4	Hand	Plastics	Hand	Arthroscopic	Hand
OR-5	Arthroscopic	Arthroscopic	Arthroscopic	Arthroscopic	Arthroscopic
OR-6	Back	Back	Back	Reconstructive and tumor	-
OR-7	Prosthesis	Prosthesis	Prosthesis	Prosthesis	-
OR-8	Prosthesis	Prosthesis	Prosthesis	-	-
OR-KBS	-	Children	Children	Children	Children
OR-AHL-1	Emergencies	Emergencies	Emergencies	Emergencies	Emergencies
OR-AHL-2	Emergencies	Emergencies	Emergencies	Emergencies	Emergencies

Figure 10: The master surgery schedule at the orthopaedic department

In order to decide on the MSS, the management of the department sit down for a whole day, trying to allocate OR capacity to the different subspecialties based on the current situation regarding patient demand, and thoughts regarding the months to come. Another important aspect of this meeting is to go through every week of the following period to detect days where the capacity will be less due to staff being at conferences or other arrangements. The planning of holiday periods is also an issue at this meeting as this will affect the production of surgeries in that period.

The operational planning of surgeries is done by allocating the patients waiting for surgery to the different ORs, and surgeons. For elective patients, this is typically done two to three months before the actual surgery is scheduled, and the scheduling is performed by scheduling coordinators at the intake office. The scheduling of emergency patients is performed by a scheduling coordinator at OTS, and this scheduling is performed soon after the patients have entered the emergency department or the trauma outpatient clinic.

Occasionally, rescheduling the surgeries on the day of surgery is necessary. Reasons may be that the patient was unable to show up on the day intended, or the fact that many emergencies are occupying beds at the wards, leaving no beds available for the elective inpatients. The daily rescheduling of patients is performed in a collaboration between the scheduling coordinators and the surgeons on shift.

2.4 Modeling of uncertainty: An introduction

In this thesis we aim to develop tactical surgery schedules for a hospital department. As described above, there are several sources of uncertainty that may be relevant to include when developing such schedules. As we aim to deal with parts of the uncertain aspects, we should apply methods that are well suited for modelling of uncertainty. In the following we introduce the methods we have chosen to apply.

2.4.1 Introduction to stochastic recourse models

Stochastic programming is the part of mathematical programming and operations research that studies how to incorporate uncertainty into decision problems (King and Wallace, 2010), and stochastic recourse models are one of the major frameworks within stochastic programming (Sahinidis, 2004). The term "recourse" refers to the opportunity to adapt a solution to a specific outcome observed, and these problems are always presented as problems in which there are two or more decision stages (Higle, 2005). According to King and Wallace (2010), stochastic models apply well in problems where decisions must be made at some point in time, but important information will not be available until after the decisions are made.

Components of a recourse problem

Each recourse problem can be characterized by its scenario tree, its scenario problems, and its nonanticipativity constraints. A scenario is one specific, complete realization of the stochastic parameters that might appear during the course of the problem, while a scenario tree is a structured representation of the stochastic parameters and the manner in which they may evolve over time. Depending on the manner in which the problem is formulated, it may be necessary to include specific conditions to ensure that the decision sequence honors the information structure associated with the scenario tree. These conditions are known as the nonanticipativity constraints, and impose the condition that scenarios that share the same history until a particular decision should also make the same decisions (Higle, 2005).

Two-stage recourse models

A two-stage recourse model consists of a first stage problem and a second stage, or a recourse, problem. The first stage decisions are determined before any information regarding the stochastic parameters has been obtained, while the second stage decisions are decided on after observing the stochastic parameters. The goal of a two-stage model is to identify a first stage solution that is well positioned against all possible outcomes of the stochastic parameters (Higle, 2005). The second-stage problem may often be an operational-level decision problem following a first-stage plan and the uncertainty realization (Sahinidis, 2004).

As an example of a two stage recourse problem, we may regard a simplified nurse scheduling problem. Let us assume that a department want to make a schedule for the nurses covering a ward for the coming period, but, due to the uncertain arrival of emergency patients, they do not know exactly how many nurses they will need on the different days. Here, a typical first stage decision will be to decide on a schedule that distributes the nurses to different shifts through the period, and that is robust against fluctuations in emergency demand for beds. The second stage decision may be to call in extra staff in periods of many emergencies arriving, or to send the patients to other wards.

A recourse model is said to have fixed recourse if the constraint matrix in the recourse problem is not subject to uncertainty (Higle, 2005). For the nurse scheduling problem, this imply that the cost of calling in an extra nurse, or sending a patient to another ward is the same no matter what happens. Complete recourse is another property that may be obtained by a recourse problem. If there exist a second-stage solution to all, possible and impossible, outcomes of the first-stage variables, the problem is said to have complete recourse. Another, more restrictive property is relative complete recourse. A problem is said to have relative complete recourse if there exist a second-stage solution to all possible outcomes of the first-stage variables.

For the nurse scheduling example, having relative complete recourse means that there exist ways in the problem formulation to handle all possible amounts of emergencies arriving. If there is a maximum of beds available that may be staffed by additional staff and the number of beds at other wards are restricted by an upper limit, and no other mechanisms are present in the problem formulation to deal with extreme arrival cases, the problem do not have relative complete recourse (and certainly not complete recourse).

The value of the stochastic solution

Kall and Wallace (1994) state that stochastic programs are much more complicated, technically, than the corresponding deterministic programs. Hence, from a practical point of view, there must be good reasons for turning to the stochastic models.

Birge (1982) introduces the value of the stochastic solution (VSS) in linear programs with fixed recourse. This value may be found by calculating the difference between the objective function value obtained by solving the stochastic model and the one obtained by solving the mean value (deterministic) problem (MVP). In the MVP we fix all the uncertain parameters to their expected value, and solve the deterministic problem. This yields a set of first stage variables corresponding to the optimal solution to the deterministic problem. Now, we fix the first stage variables in the stochastic model to the values obtained in the deterministic model and solve this (only the second stage variables are free to change now) for each scenario, obtaining one objective value for each scenario. Finally we weight the different solutions with the probability of the corresponding scenario taking place, yielding the mean value solution (MVS). In order to find the VSS, we calculate the difference between the solution obtained from solving the two-stage stochastic program (the stochastic solution (SS)) and the MVS. Note that the SS is always equal to, or better than the MVS.

Evaluating the scenario generation procedure

In order to apply a two-stage stochastic model, the distribution of the stochastic parameters are assumed to be known. Based on the distribution, a scenario-tree should be generated in order to represent the distribution. Note that if there are 10 stochastic parameters, and if each of these may have three different outcomes, we end up with a total of $3^{10} = 59,049$ separate data scenarios (Higle, 2005).

According to King and Wallace (2010), the goal of the scenario generation procedure should be to produce a subset of scenarios that well represent the true distribution of the stochastic parameters. Two problems regarding scenario generation are mentioned by the authors: Numerical problems and representation problems. Numerical problems arise if too many scenarios are needed in order to well represent the original distribution. This may lead to the optimization model being very time inefficient, or even unsolvable. A representation problem arises if the scenarios generated are not able to mimic the original distribution. This may lead to the conclusion that we need more scenarios, which may again leave us with a numerical problem.

When solving our two-stage stochastic model we need to make sure that we are not just testing the scenario generation procedure, but that the solutions are in fact a result of the algebraic model. To do this, we should perform in-sample and out-of-sample stability tests. The first represents a test of the internal consistency of the model (including the scenario tree generation procedure), while the second is related also to model quality (King and Wallace, 2010).

To test for in-sample stability, King and Wallace (2010) state that we have to generate several scenario trees, where each of these trees are named Ω_i. Afterwards, we run the optimization model equally many times, one for each scenario tree, and obtain one optimal solution $\hat{x}_i$ for each run. If the optimal objective function, $f(\hat{x}_i; \Omega_i)$ is about the same for all runs, that is if

$$f(\hat{x}_i; \Omega_i) \approx f(\hat{x}_j; \Omega_j) \tag{1}$$

we have in-sample stability.

Having in-sample stability, we are confident that if we run our model with the same data several times, we will produce more or less the same objective function value each time, implying that our model is internally consistent.

Stochastic programs tend to have flat objective functions, meaning that several different solutions may provide approximately the same objective function value. To achieve stability in terms of solutions is very hard for problems with very flat objective functions. As ending up with different solutions that all serve us equally well is not a problem, the objective function value is used to measure similarity (King and Wallace, 2010).

Out-of-sample stability means that if we calculate the true objective function value corresponding to the solutions coming from the different scenario trees, we get about the same value (King and Wallace, 2010). Let Ψ denote the original probability distribution. To have out-of-sample stability we require that:

$$f(\hat{x}_i;\Psi) \approx f(\hat{x}_j;\Psi) \tag{2}$$

This test aims to detect whether the scenario generation procedure has generated a stability that is not really there. This may happen if the scenario generation procedure consequently avoids difficult tails of the distribution, so that the in-sample stability observed is just over a part of the support of the random variables. This may be revealed through out-of-sample stability testing, as the true objective function value will be significantly different for the different solutions generated by the model.

Even though we are fixing the fist stage variables when testing for out-of-sample stability, there may be far too many scenarios to solve all the second-stage problems available. If this is the case, we should approximate the distribution by generating a large scenario tree and let this serve as the true distribution(King and Wallace, 2010).

According to King and Wallace (2010), the correct way to calculate the out-of sample value may often be to construct a simulation model of the problem. This will not only take care of the fact that the scenario tree generated is an approximation of the real distribution, but also the fact that the objective function most likely only approximates the true problem as well.

2.4.2 Introduction to simulation

According to J. Banks (1996), a simulation is the imitation of the operation of a real-world process or system over time. The behaviour of a system as it evolves over time is studied by developing a simulation model. This model usually takes the form of a set of assumptions concerning the operation of the system, and these assumptions are expressed in mathematical, logical, and symbolic relationships between the objects of interest in the system. Once developed and validated, a model can be used to investigate a wide variety of ”what if” questions about the real-world system. Potential changes to the system can first be simulated in order to predict their impact on the system performance.

Components of a system

In order to understand and analyze a system, a number of terms should be defined (J. Banks, 1996). An entity is an object of interest in the system, and an attribute is a property of an entity. An activity represents a time period of specified length. At a hospital department, the patients may be an entity, the diagnose may be an attribute, and performing a surgery may be an activity.

The state of a system is defined as the collection of variables necessary to describe the system at any time. An event is defined as an instantaneous occurrence that may change the state of the system. Events are divided into endogenous and exogenous events, where the former are events occurring within the system, while the latter are events in the environment that affect the system. At a hospital department, a state may be the number of emergency patients waiting to receive surgery, an endogenous event may be the completion of a surgery, and an exogenous event may be an emergency patient arriving to the system (J. Banks, 1996).

Discrete and continuous systems

Systems can be categorized as either discrete or continuous. A discrete system is one in which the state variables change only at a discrete set of points in time. The orthopaedic department is an example of a discrete system, as all events such as patients arriving, patients entering the OR or patients leaving the OR, are happening at discrete points in time. A continuous system is one in which the state variables change continuously over time. An example of such a system may be the water-level behind a dam. The level may raise due to rain and the melting of snow, and it may decrease due to evaporation and draining of water to power production (J. Banks, 1996).

Verification and validation

Verification deals with whether the computer program performs properly. For complex models this involves debugging and carefully testing of the applied algorithms. If the input parameters and logical structure of the model are correctly represented in the computer, verification has been completed. Validation is the determination that a model is an accurate representation of the real system. Validation is usually achieved through the calibration of the model, an iterative process of comparing the model to actual system behaviour and using the discrepancies between the two, and the insights gained, to improve the model. This process is repeated until model accuracy is judged acceptable (J. Banks, 1996).

The use of discrete-event simulation in health care planning

Günal and Pidd (2010) review the use of discrete-event simulation for performance modelling in health care. The use of simulation within health care has a long history, and the first models were developed in the 1960s. However, few studies from the 1960s and 1970s report any successful use of models, due to the lack of economic incentives, no vested authority, non-quantifiable data and no commitment to follow up. The authors of the review state that we today have access to large electronic data sets, but that the other issues remain.

Accident and emergency departments units are the most popular area for simulation modelling in health care. The authors state that this is probably due to the fact that these departments are relatively self-contained and have easily observable processes that cover relatively short periods of a few hours. The modelling of inpatient care has also been an active research area for many years, and discrete-event simulation models are primarily used for testing mathematical models developed. Patient flow to hospital beds, bed occupancy and length of stay are commonly investigated within inpatient care.

Outpatient clinics are also commonly modelled using discrete-event simulation as they have some common characteristics with the accident and emergency departments. Most of the studies on outpatient clinics include scheduling and capacity planning, focusing on micro waiting; that is the waiting time from the patient arrives till he is summoned for treatment. Other units are also frequently modelled by discrete-even simulation, and especially popular are ORs and critical care units. When applying discrete-event simulation to ORs, macro waiting is commonly investigated. Here, macro waiting refers to the time interval between the point when it is decided that a patient should receive surgery, and the time when the surgery takes place.

When it comes to implementation of simulation models, the authors state that choosing the right level of detail and client involvement are crucial for success. Choosing the right level of detail is especially important for saving time in the model development phase, and also for convincing stakeholders on the use of the model. Two barriers to successful implementation is mentioned. First, a simulation project is often initiated by decision-makers who seek urgent solutions to their problems. As a result, the simulation analysts are expected to generate quick solutions, which they fail to do due to the time spent in collecting and analyzing data. Second, when quick solutions are expected, modellers tend to oversimplify the models, which may cause decision makers lack of confidence.

The authors also mention another issue related to health care simulation. Modelling a single unit, may imply missing the big picture. At a hospital there are often many factors that have impact on the operations of different units, and excluding too much of these factors in the modelling may cause unrealistic results. However, there have been very few attempts on modelling a hospital as a whole system.

2.4.3 Introducing queueing models

Simulation is often used in the analysis of queueing models (J. Banks, 1996). In a typical queueing model customers arrive from time to time and join a queue, are eventually served, and finally leave the system. The term ”customer” refers to any type of entity that can be viewed as requesting ”service” from a system. When regarding health care systems, the patients are typical the customers that arrive at the hospital to receive some kind of treatment or service, like for example surgery.

Typical measures of system performance include server utilization, length of waiting line, and delays of customers. Often, when attempting to improve the queueing system, the analyst is involved in trade-offs between server utilization and customer satisfaction in terms of line length and delays. Queueing theory and simulation analysis are used to predict measures of system performance as a function of the input parameters. The input parameters include the arrival rate of customers, the service demands of customers, the rate at which a server works, and the number and arrangements of servers. For relatively simple systems, the performance measures can be computed mathematically at great savings in time and expense compared to the use of a simulation model. But for realistic models of complex systems, simulation is usually required(J. Banks, 1996).

Characteristics of queueing systems

The key elements of a queueing system are the customers and the servers. The population of potential customers may be assumed to be finite or infinite. In systems with a large population of potential customers, such as hospitals, the population is usually assumed to be infinite. The main difference between finite and infinite population model is how the arrival rate is defined. In an infinite population model, the arrival rate is not affected by the number of customers who have left the population and joined the queueing system. In many queueing systems there is a limit to the number of customers that may be in the waiting line or system, while other systems are considered as having unlimited capacity(J. Banks, 1996).

The arrival process for infinite-population models is usually characterized in terms of interval times of successive customers. The most important model for random arrivals is the Poisson arrival process. If A_n represents the inter arrival time between customer n-1 and customer n, then for a Poisson arrival process, A_n is exponentially distributed with mean $1/\lambda$ time units. The arrival rate is λ customers per time unit(J. Banks, 1996).

The queue behaviour refers to the customer actions while in a queue waiting for service to begin. In some situations, there is a possibility that incoming customers may see the line and decide not to enter if the queue is too long. Also, there may be situations when customers may leave the queue when they see that the line is moving too slow, or the customers may move from one queue to another. Queue discipline refers to the logical ordering of customers in a queue and determines which customer will be chosen for service when a server becomes free. Common queue disciplines include first-in-first-out (FIFO), last-in-first-out (LIFO), service in random order (SIRO), and service according to priority (PR)(J. Banks, 1996). Regarding patient queues, emergency patients typically have prioritization over electives, so that these patients move to the head of the line upon arrival.

The service times my be either constant or random. For random service times, the exponential, Weibull, gamma, lognormal and truncated normal distributions have all been successfully as models of service times in different situations. Sometimes the service duration may be identically distributed for all customers of a given type or class, while customers of different types may have completely different service time distributions(J. Banks, 1996). Patients may be characterized by the diagnosis, and service times (such as surgery duration) are typically identically distributed for patients of a given diagnosis.

A queueing system consists of a number of service centers and interconnecting queues. Each service center consists of some number of servers working in parallel, that is, upon getting to the head of the line, a customer takes the first available server. Parallel service mechanisms are either single server, multiple server or unlimited servers(J. Banks, 1996). At a hospital department, the service center may be the operating theatre, consisting of several ORs working as parallel servers.

3 The Master Surgery Scheduling Problem

The overall goal when taking on the Master Surgery Scheduling Problem (MSSP) is to produce a cyclic schedule where each orthopaedic subspecialty is scheduled to surgery slots in one or several of the ORs through the cycle in order to perform surgeries efficiently. By efficiently we mean scheduling many elective patients for surgery, and at the same time be able to handle a fluctuating demand of emergency patients without having to cancel the elective surgeries. The cycle is usually set to one week, and this weekly schedule repeats itself for a prolonged period of typically six to twelve months.

To avoid cancellations of elective patients in periods of high emergency demand for surgery, we want to devote surgery slots for the green emergency patients in the elective ORs. The green patients included in the problem are the ones that are scheduled for the elective ORs due to the limiting capacity at the emergency ORs. We call these patients the excess demand of green patients. The OR slots scheduled for the green patients are referred to as the flexible slots, as these may be used for yellow emergency patients in periods of low demand for green surgeries. The resources considered in the MSSP are the ORs, the wards, the surgeons and the anesthesia staff.

3.1 The patients

The patients included in the problem are the electives, and the green and yellow emergencies. The red emergencies are implicitly included as they may cover beds at the wards, but these patients are not scheduled for surgery as they are normally treated at the emergency ORs which are not considered a part of this problem. Both the electives and the green emergencies are sub divided into patient categories based on diagnostic similarities, and each of these categories are either inpatients or outpatients. The yellow patients are treated as one aggregated patient category.

We assume that the target throughput of elective patients to be scheduled within each patient category for the cycle is known, and we aim to schedule as many of these patients as possible. In addition, we demand that all the excess demand of green patients that enter in a cycle should be scheduled for surgery within the cycle. As the emergencies enter the hospital in a random manner, the excess demand of green patients may change from one cycle to the next, providing uncertainty into the problem. Another uncertain aspect caused by the random entrance of emergencies is the bed loading at the wards, providing uncertainty regarding the number of beds available for elective patients during each cycle.

There are two properties related to each patient category: The expected surgery duration, and the expected length of stay at a ward following surgery. The total number of patients that might be scheduled to the same OR on a given day is restricted by the expected surgery duration of the different patient categories. Furthermore, we may not schedule an inpatient for surgery if there is not capacity at the wards to host the patient following surgery. The expected length of stay may vary depending on the day of surgery. A patient receiving surgery on a Friday may be offered a bed to Monday morning because of poor guidelines for dismissing the patient makes the surgeon covering the weekend shift insecure on whether to let the patient go or not. If the same patient receives surgery on another weekday, say Tuesday, there will be a trained surgeon available to dismiss the patient as soon as he is regarded medically prepared.

In order to facilitate a scheduling regime where some patients from each elective patient category are scheduled, rather than all patients from some categories and none from others, we demand that a given share of each elective patient category should be scheduled for surgery. As some patient categories may provide more DRG-points to the department than others, we may give different values of both scheduling and cancelling patients from the different patient categories.

3.2 The operating rooms

There is a set of available ORs, and each of these rooms may be used for different surgeries depending on their characteristics. The available time at an OR is divided into time slots, and the surgery teams from the different subspecialties are scheduled to the slots available. There is usually one or two slots available at each OR during a day. The number of patients belonging to a specific subspecialty that may be scheduled for surgery on a given day is restricted by the slot-time scheduled for that surgical subspecialty on that day. Scheduling green patients to the elective surgery slots may lead to elective cancellations, but note that cancellation may be avoided. This is if there is excess capacity exceeding the expected surgery duration of a green patient at an elective surgery slot, after having treated all electives scheduled to the slot. If no excess capacity is available at the elective slots, and we need to schedule more green patients, there will be cancellations of elective surgeries.

3.3 The wards

There is a set of wards available, and at each ward there is a given number of staffed beds available each day. In addition, there may be additional beds at the wards that are not staffed. The wards are heterogeneous, and may host different patient categories. There is a bed assigned to each inpatient entering the hospital at the day of arrival, and this bed will be occupied by that patient throughout the stay. In periods with shortage of beds at some of the wards, there is a possibility to let patients rest at some of the other wards available. However, this is not optimal as the nursing staff covering the different wards have some patient categories that they are more trained to handle than others. If the total bed capacity is utilized one day, and more elective inpatients that demand a bed on that day are scheduled, these surgeries will be cancelled.

3.4 The surgeons and the anesthesia staff

The surgeons represent different subspecialties who are trained to perform surgeries to some of the patient categories. In reality, the surgeons may be able to perform surgeries to patients outside of their subspecialty, but this will not be the case in this problem. A surgeon of a particular subspecialty may have the opportunity to perform surgery to different patient categories, but each category may only receive surgery from surgeons of one subspecialty. The amount of surgeon resources available are restricted by the total work load (number of slots) allowed to schedule for a subspecialty on a given day, and also the total work load (number of slots) allowed to schedule for a subspecialty within a cycle. Furthermore, a surgeon can only be assigned to a time slot in an OR meant for the subspecialty which the surgeon represents. Note that we do not include the individual surgeons, but rather the surgeon capacity available to each subspecialty. The anesthesia staff may cover a number of ORs each day. The number of surgeries undertaken within each OR is irrelevant as long as the room is covered by the anesthesia staff. An OR that is not covered by the anesthesia staff may not be used for surgery.

3.5 The objectives in the problem

There are multiple objectives relevant to the MSSP. The first is to maximize the number of elective patients scheduled for surgery. Furthermore, we aim to minimize both the number of elective cancellations, the number of green patients to receive surgery in the elective slots, and the number of patients not resting at the ward meant for them. Finally, we want to maximize the number of yellow patients to receive surgery in the slots scheduled for green emergency patients.

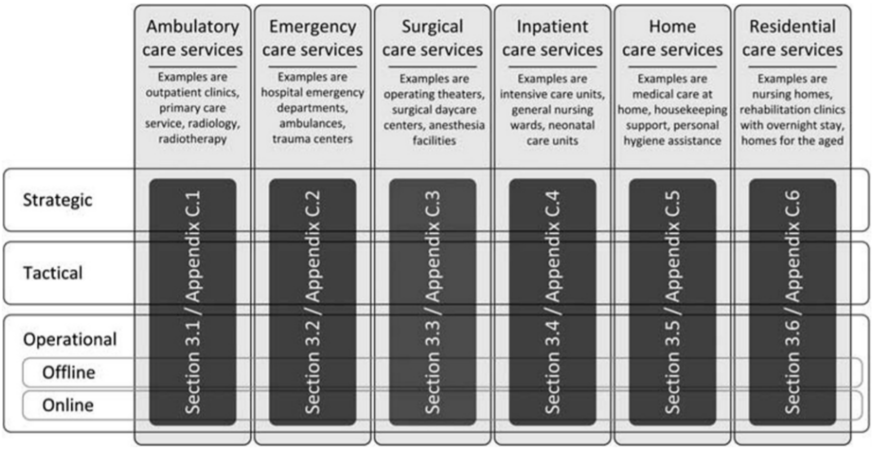

Ambulatory care services
Examples are outpatient clinics, primary care service, radiology, radiotherapy
Emergency care services
Examples are hospital emergency departments, ambulances, trauma centers
Surgical care services
Examples are operating theaters, surgical daycare centers, anesthesia facilities
Inpatient care services
Examples are intensive care units, general nursing wards, neonatal care units
Home care services
Examples are medical care at home, housekeeping support, personal hygiene assistance
Residential care services
Examples are nursing homes, rehabilitation clinics with overnight stay, homes for the aged
Strategic
Tactical
Operational
Offline
Online
Section 3.1 / Appendix C.1
Section 3.2 / Appendix C.2
Section 3.3 / Appendix C.3
Section 3.4 / Appendix C.4
Section 3.5 / Appendix C.5
Section 3.6 / Appendix C.6

Figure 11: Illustration of the taxonomy provided by Hulshof et al. (2012).

4 Related literature

The literature presented in this section is related to surgery scheduling in operations research. First, a theoretical framework is provided, and then we introduce some literature covering trade-offs related to OR scheduling regimes. Finally, we investigate some of the literature available on the Master Surgery Scheduling Problem (MSSP).

4.1 A theoretical framework

To classify the literature within the field, we refer to the work of Hulshof et al. (2012). The authors present a bi-axial, taxonomic classification of planning decisions in health care. The vertical axis reflects the hierarchical nature of decision making in resource capacity planning and control, and builds upon the work of Hans et al. (2012). On the horizontal axis, the different services within health care are positioned. See Figure 11 for an illustration of the taxonomy provided by Hulshof et al. (2012).

There are four decision levels to consider: Strategic, tactical, offline operational, and online operational. Strategic planning involves defining the organization's mission, and the decision making to translate this into design, dimension and development of the health care delivery process. Inherently, strategic planning has a long planning horizon and is based on highly aggregated information and forecasts. An example of strategic planning can be deciding on the number of ORs to build in order to serve a population with surgeries in the years to come.

Tactical planning translates strategic planning decisions to guidelines that facilitate operational planning decisions. As a first step in tactical planning, patient categories are characterized based on diagnosis, urgency and resource requirements. As a second step, the available resource capacities, settled at the strategic level, are divided among these categories. The MSSP is a classic problem within tactical surgery planning.

Operational planning involves the short-term decision making related to the execution of the health care delivery process. At this level, elective demand is entirely known but the demand for emergency treatment still must be forecasted. Examples of decisions made at this level can be the explicit planning of an elective patient to an OR, and the rescheduling of elective patients due to incoming emergencies. The operational decision level is further divided into offline and online decisions. The offline decisions are all operational decisions made up until the day of surgery. Online decisions, on the other hand, are all decisions made on the day of surgery. An example of an offline decision is the rescheduling of a surgery due to a patient calling in and saying that he is prohibited from coming on the day he was scheduled for. If this patient did not call in some days prior to the surgery, but did not show up at the day of surgery without any notice, the rescheduling performed on this day would have been an online rescheduling process.

The services presented in the taxonomy are as follows: Ambulatory care services, emergency care services, surgical care services, inpatient care services, home care services and residential care services. The service category most relevant to this report is the surgical care services which cover all literature relevant to surgery scheduling, from the strategical case mix planning problem to the online surgical case rescheduling problem.

Samudra et al. (2016) argue that the decision levels described by Hulshof et al. (2012) vary considerably for different settings and hence are often perceived as vague and interrelated. They propose a structure that is based on descriptive fields in order to categorize the literature. These fields include the patient types included, the different performance measures used, the decisions that must be made, the integration of up-and downstream units of the OR, the incorporation of uncertainty, the operations research methodology and the testing phase and application. Next, a brief introduction to some of these topics are provided.

Two major patient classes are considered in the literature: elective patients and emergency patients (Samudra et al., 2016). Another classification used by some authors is the dividing of patients into inpatients and outpatients. At the hospitals, there is an ongoing shift of services going from inpatient towards outpatient care. However, Samudra et al. (2016) show that the literature considering outpatients is not increasing, but so is the literature on inpatients.

The performance measures used will favor the interests of some stakeholders over others. The management at the hospital could be interested in high profits, while medical staff care less about cost factors and rather aim to achieve low overtime. The patients may care little of the aforementioned aspects, and rather care about short waiting times and little chances of cancellation. Samudra et al. (2016) distinguish between the following major performance measures: waiting time, utilization, leveling, idle time, throughput, preferences, financial measures, make-span and patient deferral. The authors find that the most used performance measure is overtime followed by waiting time. Overtime in the OR may result in dissatisfaction of the surgical staff, in excessive costs for the hospital and in disruption of the schedule in downstream departments. Waiting time mostly refers to the waiting time for patients (on the day of operation and the size of the waiting list), but some authors have also developed models to minimize the waiting time for the surgeons.

As OR scheduling is not made in isolation of other departments at the hospital, it may be wise to include both upstream and downstream facilities to produce realistic schedules and to improve their combined performance. Facilities often included is the post anesthesia care unit (PACU), the intensive care unit (ICU) and the wards. The upstream facilities, such as the preoperative assessment, are not usually included in the literature.

One of the major problems associated with OR scheduling is the uncertainty inherent to surgical services. The two primary areas where uncertainty is considered in the literature is the arrival of emergency patients and the surgery duration. Out of all papers investigated by Samudra et al. (2016), 44 % take surgery duration uncertainty into account, while 28 % consider arrival uncertainty. Other areas where uncertainty may be considered is the length of stay of an inpatient following surgery. Although some papers include aspect of uncertainty to their model, there are only a limited number of authors who apply stochastic programming to solve real-life problems.

Many researchers obtain real life data when solving their problems. However, Samudra et al. (2016) report that less than 7 % of the methods are applied in practice. Unfortunately, simply testing of procedures or tools on real data does not imply that the methods get implemented in practice. Lagergren (1998) indicates that the lack of implementation in the health care service seems to have improved considerably, but this cannot be stated within the field of OR planning.

4.2 Scheduling policies in surgery scheduling

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 higher elective waiting time and costly resource overtime and rescheduling. Meanwhile, 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., 2014b). These trade-offs are very relevant to our thesis, as we aim to provide a mix of elective OR capacity and flexible slot capacity. In the following we present some literature considering OR scheduling policies and relevant trade-offs faced when deciding on the scheduling policy.

Ferrand et al. (2014b) provide a literature survey, investigating literature related to scheduling policies in surgery scheduling. The authors find that there is mixed opinions on whether the ORs should be dedicated with a subset focusing on elective surgeries and a subset focusing on emergency surgeries, or whether the ORs should be made flexible and shared between emergency and elective surgeries. Furthermore, they state that almost all literature focus on either the dedicated or the flexible policy, and only a few papers investigate a mixed policy.

The authors find that three broad categories of approaches have been proposed to handle the mix of elective and emergency surgeries:

- Decide how many elective surgeries to schedule so that a fraction of OR time is reserved for emergencies.
- Schedule elective surgeries and insert a short fraction of OR time (slack) between them.
- Do not allocate time for emergency surgeries but schedule elective surgeries with the goal of limiting the waiting time of emergency surgeries or limiting the perturbations to schedule when inserting emergency surgeries.

The first scheduling approach is well suited for a dedicated OR policy, while the two latter are typically adopted for the flexible OR policy. If applying a mixed OR policy, all three approaches may be used.

Ferrand et al. (2014b) propose the following topics for future research:

- Study what kinds of changes that can be made to the OR schedule in order to respond to elective cancellations, and highlight the interactions between elective cancellations and access to OR emergency.
- Study the distribution, the magnitude and frequency of emergency waiting time as a result of the scheduling regime applied.

Ferrand et al. (2014a) develop a discrete event simulation tool to investigate the use of the mixed policy. They simulate the flow of patients through 20 ORs for one 8-hour shift, allowing for overtime work. 75 elective patients arrive through the day in batches of 15, and the batches arrive with 90 minutes in between. The goal is to spread the workload of the 75 elective cases evenly throughout the day. In addition, they include emergencies that arrive according to a Poisson process with an arrival rate of 1.5 patients per hour.

To investigate the mixed policy, the authors vary the amount of ORs that are flexible, how many that are dedicated for electives, and how many that are dedicated for emergencies. Upon an emergency arrival, if an emergency OR is available, the patient is sent to the dedicated OR with the lowest number. If no emergency ORs are available, the patient is sent to the flexible OR with the lowest number. If no emergency OR and no flexible OR are available, the patient has to wait in a single queue for the first available OR.

To schedule the elective patients a simple heuristic is applied:

- If the number of elective ORs are less than the batch size, all elective ORs are scheduled, and then the remaining elective patients are assigned cyclically to the flexible ORs.
- If the number of elective ORs are larger than the batch size for elective arrivals, then the electives are assigned cyclically to both the flexible and the

elective ORs.

The authors conclude that the mixed policy outperforms both the completely flexible and completely dedicated policy in terms of patient waiting time and staff overtime. Furthermore, they find that when dedicating a few ORs, the rooms should be dedicated for emergencies, and if a few ORs are to be made flexible, these ORs should be taken from the set of elective ORs.

Zonderland et al. (2010) propose a model for planning and scheduling of semi-urgent surgeries, aiming to investigate the trade-offs between cancellations of elective surgeries due to semi-urgent surgeries, and unused OR time due to excessive reservation of OR time for semi-urgent surgeries. By semi-urgent the authors refer to surgeries that should be performed either within one or two weeks. The patients are further categorized according to the surgery duration, given as the a number of surgery slots required.

The authors propose a three-stage model for the scheduling of semi-urgent emergencies. Firstly, using a queuing theory framework, they evaluate the OR capacity needed to accommodate every incoming semi-urgent surgery. Secondly, they introduce another queuing model that enables a trade-off between the cancellation rate of elective surgeries and unused OR time. Finally, based on Markov decision theory, they develop a decision support tool that assists the scheduling process of elective and semi-urgent surgeries. In the first two models the semi-urgent surgeries are treated as one group according to the degree of urgency.

Based on the the first model, the authors argue that the minimum number of slots needed to handle the emergencies should equal the expected number of emergency slots arriving each week. In the next model they assign costs, both for leaving one semi-urgent OR slot empty, and for the cancellation of one elective slot. Furthermore, they derive expressions for the expected amount of empty semi-urgent OR slots and for the expected amount of elective slots cancelled, and then minimize the total cost by varying the amount of scheduled semi-urgent slots above the minimum level decided on in the first model. To develop the third model the authors develop scheduling rules for the semi-urgent surgeries. At the beginning of the week, the one-week-emergencies are scheduled for the emergency slots decided on in the second model. If the emergency OR capacity is insufficient, electives are cancelled to provide more capacity. If all the ORs are utilized, and there are still one-week-emergencies left to schedule, these are performed in overtime. If there is emergency OR capacity left after treating all one-week-emergencies, the two-week-emergencies are scheduled to these slots. If all emergency OR capacity is used, and there are still two-week-emergencies left to schedule, we may wither cancel electives, or we may postpone the two-week-emergencies to next week (they will then become one-week-emergencies the next week). All electives that are cancelled become two-week-emergencies the next week. In the third model, the authors aim to minimize the total expected costs of cancelling electives, leaving emergency OR-capacity unused and adding overtime work.

The authors present the results from applying the model on a real life hospital department, and they highlight the trade-offs by altering the values of the objective function coefficients. However, they do not compare the results to the real world scheduling regime obtained at the case department.

4.3 The Master Surgery Scheduling Problem

Next, literature on the MSSP is presented. Inspired by Samudra et al. (2016) we report on the following aspects for all the papers: The patient types considered, the objective function, the clinical restrictions including the ORs, the handling of patients, and the inclusion of up- and downstream facilities, the inclusion of uncertainty, and the real-life applications. Note that three of the papers included (Freeman et al. (2017), Ma and Demeulemeester (2013) and Testi et al. (2007)) provide models that perform surgery scheduling at both a strategic, a tactical and an operational level. Here, developing a MSS is done at the tactical level. In Table 5 we provide a summary of the papers investigated on the MSSP.

4.3.1 The patient classes included

Most of the literature on the MSSP include only elective patients, and the authors give different reasons for excluding the emergencies. van Oostrum et al. (2008) include only elective patients in their model, but argue that capacity should be reserved for emergencies. However, they are not stating whether it should be done according to a flexible, a dedicated or a mixed policy. Testi and Tànfani (2008) also disregard emergencies, assuming that these patients go to dedicated ORs, according to the dedicated policy. Santibáñez et al. (2007) exclude emergencies, assuming that these patients are treated after the time designated for elective surgeries. The authors argue that because of the highly unpredictable arrival of emergencies, it would make little sense to fit probability distributions for these. Mannino et al. (2012) schedule for electives only, and do not mention the emergency patients. Adan et al. (2011), on the other hand, include both elective and emergency patients, and they make estimates of emergency arrivals based on a Poisson process. Freeman et al. (2017) estimate, using discrete event simulation, the number of emergencies entering, and based on this they reserve capacity for these patient in the ORs.

4.3.2 The objective function

There are numerous objective functions presented in the literature on the MSSP. Common examples may be objective functions focusing on the bed capacity at the ward, objective functions maximizing the stakeholders preferences, or multiple criteria objective functions.

Both Testi and Tànfani (2008), Testi et al. (2007) and Penn et al. (2017) include aspects of welfare into their objective functions. In the two latter, the surgeons preferences are maximized. In Testi et al. (2007) these preferences are assumed to depend on two parts: Exclusion of particular days due to other engagements (like teaching activities), and the expected length of stay of patients on the waiting list. If patients with an expected length of stay less than or equal to five days are admitted on the first days of the week, they can be discharged before the weekend, allowing for the short stay area to be closed during the weekends. In order to achieve this, the surgeon preference for a given ward and a given day is increasing in the number of short-stay patients on the waiting list, and decreasing in the number of days counting from Monday.

Testi and Tànfani (2008) propose an objective function that minimizes loss of welfare among the patients. To accomplish this, the authors introduce urgency coefficients for each diagnostic category, and then a priority is given to each patient as the product of time waited computed at a day and the urgency coefficient. According to this system, the priority is increasing in both days from referral and in the urgency coefficient. In the objective function the authors minimize the priority among both the patients who will have their surgery and those still waiting at the end of the period.

A variety of objective functions regarding the bed capacity is proposed in the literature. The objective function developed by van Oostrum et al. (2008) aims to minimize both the number of ORs used and the maximum demand for hospital beds during the planning cycle. The beds are divided into categories, representing for example beds at the wards and at the intensive care department, and a priority parameter is introduced for each category. They also state that different patients require different lengths of stay in each of the bed types, according to the procedure performed. Ma and Demeulemeester (2013) also consider the beds at the wards in their objective function. The authors aim to minimize the total bed deficit, the maximum daily spare bed volume and the maximum variance of the bed occupancy. They use weights in order to prioritize between the different wards and the different measures.

In order to capture several aspects, the inclusion of multiple criteria objective functions seem to be very popular. By including several measures into the objective function the authors aim to satisfy several stakeholders. The challenge by including several criteria is how the different parts are to be inter-prioritized. Most of the authors impose weighting parameters obtained for example by interviewing different stakeholders, while one of the papers (Li et al., 2017) propose a goal programming approach.

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. In order to give prioritization to the three parts, weighting parameters are provided. The authors make no attempt to explain how these should be weighted. Li et al. (2017) include four parts in the objective function: minimizing the number of patients not being scheduled, minimizing the under utilization 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. Then, two approaches are suggested, one applying lexicographic prioritization, and one including weights.

Some authors propose several individual objective functions instead of one single objective function consisting of many parts. Santibáñez et al. (2007) propose five individual objective functions, within two categories: minimizing the sum of maximum usage of post-surgical beds at the hospitals included (1) and maximizing total throughput of patients (4). In order to minimize the usage of post-surgical beds, the authors include the expected days of stay in different beds following surgery for the different surgical categories. For each day the consumption of different beds is measured, and the aim is to minimize the consumption of beds on the day where most beds are used. Mannino et al. (2012) propose two quite similar model formulations with two different objective functions. In the first model they aim to minimize the patient queues, and they develop a piece wise linear objective function to penalize harder if the queue is above some predefined thresholds. In the other model they demand the queue of the different patient categories to be below a predefined threshold, while minimizing the overtime work for the staff.

The objective function provided by Adan et al. (2011) aims to minimize the absolute deviation of resource consumption compared to a predefined target. The model is a further development of the models provided by Adan and Vissers (2002) and Adan et al. (2008), but unlike the objective functions provided there, the authors now allow for the consumption of resources to go beyond the maximum capacity available. However, this is penalized in the objective function. Also here, weighting parameters are used in order to give prioritization to the resources considered.

4.3.3 The operating rooms

The OR capacity is central in every MSSP. The rooms usually have a given amount of available opening hours, often provided as a number of slots available, which can be booked by the different subspecialties. Some authors, like Mannino et al. (2012) include surgery blocks of different lengths, and allow for several subspecialties to book slots on the same OR on a given day. However, the most common approach is to consider surgery slots of a given time interval, and only allow for one subspecialty to book a room on a given day. When a subspecialty has booked an OR, the surgeons belonging to this category are allowed to perform surgeries to their patients within the time interval that they are scheduled to. Some authors, like Testi and Tànfani (2008), Adan et al. (2011) and Mannino et al. (2012) allow for some overtime work at the ORs. While most of the papers include the ORs by means of restrictions, some contributions, like van Oostrum et al. (2008) include the ORs in the decision variables, aiming to minimize the rooms that have to be opened.

It is quite surprising to see that almost all the papers investigated on the MSSP include homogeneous ORs. Of the contributions presented here, only Testi et al. (2007) and Penn et al. (2017) explicitly regard the ORs as heterogeneous. Penn et al. (2017) are also the only contribution to include constraints on surgery equipment.

4.3.4 Handling of the patients

Many authors, like Adan et al. (2011), van Oostrum et al. (2008), and Ma and Demeulemeester (2013) demand that the total amount of patients waiting are scheduled for surgery (hard constraints). Others, like Santibáñez et al. (2007) use soft restrictions to provide an interval for the number of operations required for each subspecialty. Beliën et al. (2008) and Penn et al. (2017) does not explicitly include the number of patients to receive surgery, but require that all surgeons receive a given number of slots during the time horizon in order to fulfill the number of operations required.

In their second model formulation, Mannino et al. (2012) introduce a robust formulation in order to deal with the uncertain patient demand. They propose a two-stage model where they first minimize the amount of overtime needed in order to keep the queue of patients under a certain predefined threshold when the patient demand is taking the mean value. In the second stage they let the patient queue increase to some maximum level, and then they aim to cope with this the best way as possible by imposing a restriction that requires the overtime cost in the second stage not to deviate from the solution obtained in the first stage by more than a predefined level.

4.3.5 The intensive care unit, the wards and the staff

Including the ICU is not very common in the MSSP, but some authors like Li et al. (2017) and Adan et al. (2011) consider this unit in their model. The latter are also the only one to explicitly include the nurses, by imposing restrictions on the amount of nursing hours available at the ICU.

The wards, on the other side, are very commonly included. Of the papers investigated on the MSSP, only Testi et al. (2007), Li et al. (2017) and Mannino et al. (2012) completely exclude handling of the wards in their model under the assumption that this resource is not imposing any bottleneck to the efficient flow of patients. Testi and Tànfani (2008) include the wards and the length of stay of patients, in a simplified way. The authors assume that no beds are available through the weekends, and to deal with this they prohibit scheduling of patients on days that will imply that the patients have to stay at the hospital in the weekends (based on the expected length of stay). They also include restrictions stating that there cannot be performed more surgeries on a given day than there are beds available on that day. These restrictions are sufficient to handle the ward restrictions without having to count days.

All the rest of the papers explicitly handle the wards (either as a common pool, or as heterogeneous wards) in more sophisticated ways. Most of the contributions include the expected length of stay for the different patient categories, and use counting mechanisms in order to count the total bed occupancy at the wards on given days. Beliën et al. (2008) impose an alternative way of doing this. They state that scheduling different surgeons to an OR will impose a given contribution to the mean and variance of the bed occupancy (derived from historical data), and aim to minimize this in the objective function.

Few of the papers include explicitly the surgeon capacity as restrictions. Santibáñez et al. (2007) restrict the number of ORs scheduled to a subspecialty to the number of surgeons available for that subspecialty. Similarly, Testi et al. (2007) restrict the number of ORs dedicated to one ward to the number of surgeon teams available to serve patients from the given ward.

4.3.6 The uncertainty included

None of the papers investigated include stochastic optimization. However, some of the authors include probabilities when calculating the lengths of stay at the wards and the surgery duration. Including probabilities makes the model more realistic as it provides a more accurate estimates of these values, but it do not add any stochastic value.

van Oostrum et al. (2008) deals with the uncertainty of surgery duration by providing probability distributions for running into over time at an OR as a function of the number of surgeries scheduled to that room. In Adan et al. (2011) the authors provide probability distributions for the lengths of stay in both the intensive care unit and in the wards. They also aim to decide on the number of emergencies entering each day by multiplying the arrival rate of emergencies times the probability that an emergency patients enters during day time. Ma and Demeulemeester (2013) and Li et al. (2017) calculate the lengths of stay for the different patients by using the probability that a patient is still at the ward a certain number of days after the surgery.

Both Freeman et al. (2017), Ma and Demeulemeester (2013) and Testi et al. (2007) develop models containing several stages of decision making, going from the strategic Case-Mix Planning Problem, via the MSSP, and ending up at the operational level represented by a discrete event simulation model. Although the optimization models included are deterministic, the simulation model allows for testing the model in a stochastic framework. The model developed by Testi et al. (2007) is not set up in a loop, so the information gained by the simulation model is not used to alter the optimization models. Both Ma and Demeulemeester (2013) and Freeman et al. (2017) apply a loop setting to their models. Ma and Demeulemeester (2013) state that after performing simulations at the operational level, the obtained valuable information is fed back to the upper stages to further affect the case mix and capacity decision and to further influence the operational performance. However, they do not explicitly state what information that is fed back to the upper stages. Freeman et al. (2017) are not altering the input to the optimization models based on the solutions obtained in the simulation model. However, they utilize the loop setting to run their model several times to obtain different solutions which they perform statistical measures on. Freeman et al. (2017) are also the only ones to use simulation when generating the waiting list of elective patients and use this as input to their optimization models.

4.3.7 The real-life applications

Although all of the papers included in this section report on receiving data from real life hospitals none of them report on their model being implemented at the collaborating hospitals.van Oostrum et al. (2008) seem to be closest to actually implement their model at a hospital. The authors test their model on data from the Erasmus Medical Center in Rotterdam, and they state that the hospital management is pleased with the outcomes of the model, and that they want to initiate further research into practical implementation of the MSS-approach proposed.

4.4 Our contribution related to the literature

The simulation model proposed by Ferrand et al. (2014a) is fairly simple, which makes it appropriate for comparing different scheduling policies. However, the elective scheduling heuristic seem quite far from a real life scheduling procedure. In our problem formulation we include a more realistic scheduling of elective patients, as we develop a MSS relating the scheduling of electives to different resources available. Furthermore, as Ferrand et al. (2014a) only simulate one day, the long term effect of the scheduling regimes are not investigated. The work presented by Zonderland et al. (2010) is very theoretical, and the strict scheduling rules essential for formulating the third model may be to simplistic. However, used as a tactical decision tool, the model may be appropriate to provide the number of ORs to dedicate for the semi-urgent emergencies. Also this contribution includes only the ORs, neglecting the ward capacity and the proper scheduling of elective patients, making it less attractive to real-life applications.

Our main contribution to the present literature is the inclusion of emergency patients and the development of a two-stage linear stochastic model in order to handle both the uncertain demand of semi-urgent emergencies (the green emergencies) and the uncertain bed loading imposed by the emergency patients when developing a MSS. The only papers detected to include emergencies for the MSSP is Adan et al. (2011) and Freeman et al. (2017). However, their optimization models are deterministic implying that there is no flexibility embedded in the model to handle deviations from the expected surgery demand for the emergencies. We also aim to close the gap between the literature focusing on OR scheduling regimes and the literature concerning the MSSP.

Table 5: Categorizing the investigated papers, MSSP

Paper	Objective function	Patients included	Operating rooms	Handling of patients	ICU included	Wards included	Staff included	Uncertainty included
Ma and Demeulemeester (2013)	Minimizing fluctuations in bed utilization	Electives	Homogeneous	Hard constraints	No	Yes	The surgeon categories, but not the individual surgeons	None
Testi et al. (2007)	Maximizing surgeon preferences	Electives	Heterogeneous	Hard constraints	No	No	The surgery teams	None
Penn et al. (2017)	Maximizing excess capacity of beds and surgeon preferences	Electives	Heterogeneous	The patients are not explicitly included	No	Yes	The surgeons	None
Santibáñez et al. (2007)	Minimizing the maximum usage of beds or maximizing patient throughput	Electives	Homogeneous	Between an upper and a lower bound	No	Yes	The surgeons	None
Beliën et al. (2008)	Minimizing the mean and variance of the bed occupancy	Electives	Homogeneous	The patients are not explicitly included	No	Yes	The surgeons	None
Li et al. (2017)	Minimizing number of patients not scheduled, the under utilization of ORs ans the load of patients to the recovery unit	Electives	Homogeneous	The patients are not explicitly included	Yes	No	The surgery teams	Apply probabilities when calculating LOS
Adan et al. (2011)	Minimizing deviation from target resource utilization	Electives and emergencies	ORs only represented by available OR-time, homogeneous	Hard constraints	Yes	Yes	Nursing hours	Probabilities when calculating LOS and emergency arrivals
van Oostrum et al. (2008)	Minimizing use of ORs and minimizing the maximum demand for beds	Electives	Homogeneous	Hard constraints	No	Yes	None	Probabilities when calculating the chance of overtime
Testi and Tànfani (2008)	Minimizing patient welfare loss	Electives	Homogeneous	Not more than the demand	Yes, but only at the day of surgery	Yes, but only at the day of surgery	Surgical subspecialties, but not the individual surgeons	None
Mannino et al. (2012)	Minimizing patient queue and overtime work	Electives	Homogeneous	Predefined queue thresholds	No	No	Surgical specialties, but not the individual surgeons	Yes, uncertain patient demand
Freeman et al. (2017)	Maximizing reimbursement	Electives and emergencies	Homogeneous	Not more than the demand	No	Yes	The surgical specialties, but not the individual surgeons	Apply simulation to generate patient arrival and probabilities when calculating LOS
Testi et al. (2007)	Maximizing benefit	Electives	ORs only represented by blocks available, homogeneous	Not less than a lower bound	No	No	None	None
Ma and Demeulemeester (2013)	Maximizing reimbursement	Electives	ORs only represented by blocks available, homogeneous	Between an upper and a lower bound	No	Yes	The surgeon categories, but not the individual surgeons	None

5 Modelling approach and model description

This section is twofold: First we present the two-stage modeling approach used to handle the uncertainty related to emergency patients, and afterwards the mathematical model for solving the problem is provided.

5.1 A two-stage modeling approach

As stated in the problem description, we have to decide on the surgery slot scheduling, and the scheduling of elective patients, before knowing the excess demand of green emergency patients or the amount of emergencies resting at the wards in each cycle. When we receive this information, we need to schedule all the green patients for surgery, and cancel elective surgeries if necessary. Including the emergencies in the MSSP yields a problem that is suitable for a stochastic two-stage formulation.

The first stage decision in the model are as follows:

- On each day of the cycle we need to decide which of the ORs that should be available, by scheduling anesthesia resources for those ORs
- Decide which of the OR slots that should be scheduled as flexible, and schedule the subspecialties for these slots
- Schedule the subspecialties for elective slots
- Schedule the elective patients for surgery in the elective slots
- Decide on the number of beds to staff at each ward on every day through the cycle

The second stage decisions are the following:

- Schedule the excess demand of green patients to the flexible slots
- Schedule the excess demand of green patients to elective slots (if no more flexible slots are available)
- Cancel elective surgeries if necessary
- Send inpatients to the wards and let patients rest at wards not meant for them if necessary due to short bed capacity
- Perform surgeries to yellow patients in elective slots if excess flxible capacity

Choosing a two-stage model to represent the reality has some simplifying aspects in the way we treat the uncertainty in the problem. The first stage decisions are made without knowing the actual excess demand of green patients or the number of emergencies that will be resting at the different wards, and applies well to the real world problem. However, we assume that we receive all the information related to the amount of emergencies present during the cycle at one specific point in time (just before the cycle starts). This is a simplification of the reality, where we will receive new information every day. For an illustration of this simplified world, see Figure 12 where we provide an example where the cycle length is set to one week and the information regarding the excess demand of green patients is aggregated to the Sunday prior to the planning cycle. We want to stress the fact that this is a model for generating a MSS, and not a tool for the day to day scheduling of patients, so allowing for some more aggregated view of the uncertainty involved may be acceptable.

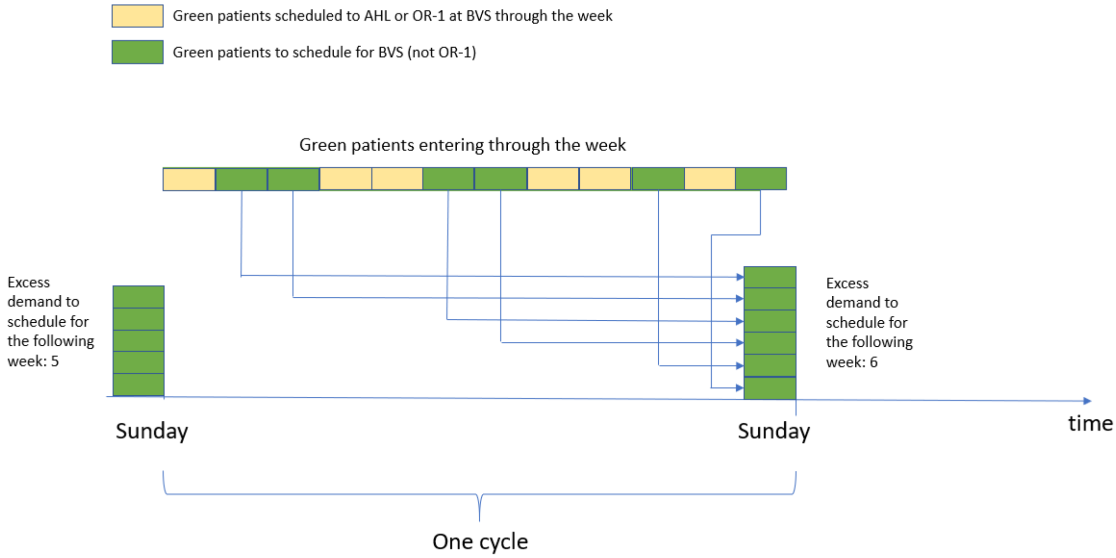

Green patients scheduled to AHL or OR-1 at BVS through the week
Green patients to schedule for BVS (not OR-1)
Green patients entering through the week
Excess demand to schedule for the following week: 5
Excess demand to schedule for the following week: 6
Sunday
Sunday
time
One cycle

Figure 12: Illustration of the aggregated two-stage approach.

Note that in the optimization model we send yellow patients to the flexible slots if all green patients are scheduled and there still is flexible capacity left. Scheduling yellow patients to the flexible slots represent an arbitrary alternative use of the flexible slots. However, this is just one alternative way of utilizing the flexible slots, and different departments may have different ways of filling the spare capacity. One alternative use of the flexible slots may be to reschedule electives that were previously cancelled to these slots.

5.2 The mathematical formulation

Next, the mathematical formulation of the problem is provided. See Tables 6, 7, 8, and 9 for all the notation used in the formulation.

Table 6: Indices used in the mathematical formulation

Letter	Description
d	Days
i	Patient categories
j	Surgical subspecialties
k	Operating rooms
s	Scenarios
w	Wards
$w^{'}$	Wards

Table 7: Sets used in the mathematical formulation

Letter	Description	
$\mathcal{D}$	Set of days in a cycle	$d \in \mathcal{D}$
$\mathcal{I}$	Set of patient categories	$i \in \mathcal{I}$
$\mathcal{J}$	Set of surgical subspecialties	$j \in \mathcal{J}$
$\mathcal{K}$	Set of operating rooms	$k \in \mathcal{K}$
$\mathcal{W}$	Set of wards	$w \in \mathcal{W}$
$\mathcal{S}$	Set of scenarios	$s \in \mathcal{S}$
$\mathcal{I}^{EL}$	Set of elective patient categories	$i \in \mathcal{I}^{EL} \subseteq \mathcal{I}$
$\mathcal{I}^{IN}$	Set of elective inpatients	$i \in \mathcal{I}^{IN} \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^{J}_{j}$	Set of elective patient categories that can be treated by subspecialty j	$i \in \mathcal{I}^{J}_{j} \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^{K}_{k}$	Set of elective patient categories that can be scheduled to operating room k	$i \in \mathcal{I}^{K}_{k} \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^{W}_{w}$	Set of elective patient categories meant for ward w	$i \in \mathcal{I}^{W}_{w} \subseteq \mathcal{I}^{EL}$
$\mathcal{I}^{GR}$	Set of green emergency patient categories	$i \in \mathcal{I}^{GR} \subseteq \mathcal{I}$
$\mathcal{I}^{GRJ}_{j}$	Set of green emergency patient categories that can be treated by subspecialty j	$i \in \mathcal{I}^{GRJ}_{j} \subseteq \mathcal{I}^{GR}$
$\mathcal{I}^{GRK}_{k}$	Set of green emergency patient categories that can be scheduled to operating room k	$i \in \mathcal{I}^{GRK}_{k} \subseteq \mathcal{I}^{GR}$
$\mathcal{I}^{GRW}_{w}$	Set of green emergency patient categories meant for ward w	$i \in \mathcal{I}^{GRW}_{w} \subseteq \mathcal{I}^{GR}$
$\mathcal{K}_{j}$	Set of operating rooms that can be managed by surgeons with subspecialty j	$k \in \mathcal{K}_{j} \subseteq \mathcal{K}$

Table 8: Parameters used in the mathematical formulation

Letter	Description
A_w^{MAX}	Maximum amount of beds available on ward w
A_{wd}	Amount of staffed beds available at ward w on day d
B_{kd}	Time (hours) available for surgery in one slot in operation room k at day d
C_i^C	Penalty for cancelling an elective patient of category i
C^{GR}	Penalty for scheduling a green patient to an elective surgery slot
$C_{ww'}^{W}$	Penalty for putting a patient belonging to ward w in ward $w^{'}$
E_{id}	Expected length of stay (days) of patient category i scheduled for surgery on day d
E_{id}^{GR}	Expected length of stay for green emergency patient of category i scheduled on day d
H_d	Number of elective inpatients allowed to schedule for surgery on day d
M_{kd}^{OR}	Maximum number of slots that can be reserved at an operating room k on day d within the opening hours of the operating room
M_d^A	Number of operating rooms covered by anesthesia staff on day d
M^{CYCLE}	Total amount of slots available through one cycle
N_j	Maximum surgeon capacity (slots) of subspecialty j in one cycle
N_{jd}^D	Maximum surgeon capacity (slots) of subspecialty j at day d
P_i	Reward for scheduling more patients from patient category i than the lower limit
P_s^S	The probability of ending up in scenario s
P^Y	Reward for scheduling a yellow emergency patient to a flexible slot
R_i	Parameter used to define the number of elective patient category i that have to be scheduled for surgery
S_i	Expected surgery duration of elective patient category i
S_i^{GR}	Expected surgery duration of green emergency patient category i
S^Y	Expected surgery duration of yellow patients
T_i	Target throughput of elective patients belonging to patient category i
T_{is}^{GR}	Excess demand of green emergency patient category i
U_{wds}^{EM}	Amount of emergencies resting at ward w on day d in scenario s
$\bar{X}_{ikd}$	Maximum number of patients of category i that can be scheduled to operating room k on day d
$\bar{X}_{ikd}^{GR}$	Maximum number of green emergency patients of category i that can be scheduled to operating room k on day d
$\bar{X}_{kd}^{Y}$	Maximum number of yellow patients that can be scheduled to operating room k on day d

Table 9: Variables used in the mathematical formulation

Letter	Description
a_{wd}	Number of staffed beds available at ward w on day d
n_{jkd}	Number of slots scheduled as flexible for subspecialty j in operating room k on day d
v_i	Number of elective patients from patient category i scheduled beyond the lower limit
x_{ikd}	Number of elective patients of patient category i scheduled to an elective slot in operating room k on day d
y_{jkd}	Number of elective slots scheduled for subspecialty j in operating room k on day d
α^A_{kd}	Indicates whether operating room k is covered by anesthesia staff on day d or not
q_{kd}	Total sum of elective patients, or surgery duration, scheduled to operating room k on day d (related to symmetry)
$b_{ww'ds}$	Number of beds occupied at ward $w^{'}$ by patients belonging to ward w on day d and scenario s
e_{ijkds}	Number of green emergency patients of category i scheduled to subspecialty j in a flexible surgery slot in operating room k on day d in scenario s
e^{EL}_{ijkds}	Number of green emergency patients of category i scheduled to subspecialty j in an elective surgery slot in operating room k on day d in scenario s
e^{Y}_{jkds}	Number of yellow emergency patients scheduled to subspecialty j in a flexible surgery slot within operating room k on day d in scenario s
u_{iwds}	Number of elective patients of patient category i resting at ward w on day d in scenario s
u^{GR}_{iwds}	Number of green emergency patients from category i resting at ward w on day d in scenario s
x^{C}_{ikds}	Number of elective patients of patient category i scheduled to an elective slot within operating room k on day d that are cancelled

$$max \sum_{i \in \mathcal{I}^{EL}} P_i v_i - \sum_{s \in \mathcal{S}} P^S_s \Bigg[\sum_{i \in \mathcal{I}^{EL}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} C^C_i x^C_{ikds} + \sum_{w \in \mathcal{W}} \sum_{w' \in \mathcal{W}} \sum_{d \in \mathcal{D}} C^W_{ww'} b_{ww'ds} + \\ \sum_{i \in \mathcal{I}^{GR}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} C^{GR} e^{EL}_{ijkds} - \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} P^Y e^Y_{jkds} \Bigg] \tag{3}$$

subject to:

$$\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}_j} \sum_{d \in \mathcal{D}} (n_{jkd} + y_{jkd}) \leq M^{CYCLE} \tag{4}$$

$$\sum_{k \in \mathcal{K}} \alpha^A_{kd} \leq M^A_d \qquad d \in \mathcal{D} \tag{5}$$

$$\sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} \alpha^A_{kd} \leq \sum_{d \in \mathcal{D}} M^A_d \tag{6}$$

$$\sum_{j \in \mathcal{J}} (n_{jkd} + y_{jkd}) \leq M^{OR}_{kd} \alpha^A_{kd} \qquad k \in \mathcal{K}, \quad d \in \mathcal{D} \tag{7}$$

$$\sum_{k \in \mathcal{K}_j} B_{kd}(n_{jkd} + y_{jkd}) \leq N^D_{jd} \qquad j \in \mathcal{J}, \quad d \in \mathcal{D} \tag{8}$$

$$\sum_{k \in \mathcal{K}_j} \sum_{d \in \mathcal{D}} B_{kd}(n_{jkd} + y_{jkd}) \leq N_j \qquad j \in \mathcal{J} \tag{9}$$

$$\sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} x_{ikd} \geq \left\lceil \frac{T_i}{R_i} \right\rceil \qquad i \in \mathcal{I}^{EL} \tag{10}$$

$$\sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} x_{ikd} \leq T_i \qquad i \in \mathcal{I}^{EL} \tag{11}$$

$$\sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} x_{ikd} - v_i = \left\lceil \frac{T_i}{R_i} \right\rceil \qquad i \in \mathcal{I}^{EL} \tag{12}$$

$$\sum_{i \in \mathcal{I}^J_j} S_i x_{ikd} \leq B_{kd} y_{jkd} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D} \tag{13}$$

$$\sum_{i \in \mathcal{I}^{\mathcal{IN}}} \sum_{k \in \mathcal{K}} x_{ikd} \leq H_d \qquad d = Friday \tag{14}$$

$$a_{wd} \leq A_w^{MAX} \qquad d \in \mathcal{D}, \quad w \in \mathcal{W} \tag{15}$$

$$\sum_{w \in \mathcal{W}} a_{wd} \leq \sum_{w \in \mathcal{W}} A_{wd} \qquad d \in \mathcal{D} \tag{16}$$

$$\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} (e_{ijkds} + e_{ijkds}^{EL}) = T_{is}^{GR} \qquad i \in \mathcal{I}^{GR}, s \in \mathcal{S} \tag{17}$$

$$\sum_{i \in \mathcal{I}_j^J} S_i (x_{ikd} - x_{ikds}^C) + \sum_{i \in \mathcal{I}_j^{GRJ}} S_i^{GR} e_{ijkds}^{EL} \leq B_{kd} y_{jkd}$$
$$j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{18}$$

$$\sum_{i \in \mathcal{I}_j^{GRJ}} S_i e_{ijkds} + S^Y e_{jkds}^Y \leq B_{kd} n_{jkd}$$
$$j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{19}$$

$$x_{ikds}^C \leq x_{ikd} \qquad i \in \mathcal{I}^{EL}, \quad k \in \mathcal{K}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{20}$$

$$\sum_{k \in \mathcal{K}} \sum_{d'=1}^{E_{id}} (x_{ik(d-d'+1)} - x_{ik(d-d'+1)s}^C) \leq u_{iwds}$$
$$w \in \mathcal{W}, \quad i \in \mathcal{I}_w^W, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{21}$$

$$\sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d'=1}^{E_{id}^{GR}} (e_{ijk(d-d'+1)s} + e_{ijk(d-d'+1)s}^{EL}) \leq u_{iwds}^{GR}$$
$$w \in \mathcal{W}, \quad i \in \mathcal{I}_w^{GRW}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{22}$$

$$\sum_{i \in \mathcal{I}_w^W} u_{iwds} + \sum_{i \in \mathcal{I}_w^{GRW}} u_{iwds}^{GR} + \sum_{w' \in \mathcal{W} | w' \neq w} b_{w'wds} - \sum_{w' \in \mathcal{W} | w' \neq w} b_{ww'ds} \leq a_{wd} - U_{wds}^{EM}$$
$$w \in \mathcal{W}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{23}$$

$$y_{jkd} \in \{0, 1, ..., M_{kd}^{OR}\} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D} \tag{24}$$

$$n_{jkd} \in \{0, 1, ..., M_{kd}^{OR}\} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D} \tag{25}$$

$$x_{ikd} \in \{0,1,...,\bar{X}_{ikd}\} \qquad k \in \mathcal{K}, \quad i \in \mathcal{I}_k^K, \quad d \in \mathcal{D} \tag{26}$$

$$a_{wd} \in \{0,1,...A_w^{MAX}\} \qquad w \in \mathcal{W}, \quad d \in \mathcal{D} \tag{27}$$

$$\alpha_{kd}^A \in \{0,1\} \qquad k \in \mathcal{K}, \quad d \in \mathcal{D} \tag{28}$$

$$x_{ikds}^C \in \{0,1,...,\bar{X}_{ikd}\} \qquad k \in \mathcal{K}, \quad i \in \mathcal{I}_k^K, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{29}$$

$$e_{ijkds} \in \{0,1,...,\bar{X}_{ikd}^{GR}\} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_|, \quad i \in \mathcal{I}_w^{GRW}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{30}$$

$$e_{ijkds}^{EL} \in \{0,1,...,\bar{X}_{ikd}^{GR}\} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_|, \quad i \in \mathcal{I}_w^{GRW}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{31}$$

$$e_{jkds}^Y \in \{0,1,...,\bar{X}_{kd}^Y\} \qquad j \in \mathcal{J}, \quad k \in \mathcal{K}_j, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{32}$$

$$v_i \in \{0,1,...,T_i - \left\lceil \frac{T_i}{R_i} \right\rceil\} \qquad i \in \mathcal{I}^{EL} \tag{33}$$

$$u_{iwds} \in \{0,1,...,A_{wd}\} \qquad i \in \mathcal{I}^{EL}, \quad w \in \mathcal{W}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{34}$$

$$u_{iwds}^{GR} \in \{0,1,...,A_{wd}\} \qquad i \in \mathcal{I}^{GR}, \quad w \in \mathcal{W}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{35}$$

$$b_{ww'ds} \in \{0,1,...,A_{w'd}\} \qquad w \in \mathcal{W}, \quad w^{'} \in \mathcal{W}|w^{'} \neq w, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{36}$$

The objective function

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 at wards not originally intended for them, and scheduling green patients to elective slots. In addition, we maximize the amount of yellow patients scheduled for surgery in the second stage.

The first stage decisions

Constraint (4) requires that the sum of all slots scheduled through the cycle, both elective slots and flexible slots, may not exceed the slots available in the cycle.

The anesthesia constraints are given by constraints (5) to (7), where the former require that we can not schedule surgeries to more operating rooms during a day than there are rooms covered by anesthesia staff on that day. The second constraint demands that the total number of operating rooms used for scheduling patients through the week have to be less than or equal to the total number of operating rooms covered by the anesthesia staff throughout the week. Note that constraints (5) are more restricting than constraint (6), and that constraint (6) will not be interesting as long as constraints (5) are present. The reason for providing both is that it may be interesting to investigate whether there may be other ways of distributing the total anesthesia resource throughout the cycle. The third set of anesthesia constraints state that there may be no more slots scheduled for each operating room on a given day than there are slots available in that operating room on that day, and that an operating room has to be covered by anesthesia staff in order to schedule surgeries there.

Constraints (8) state that the surgeons of a given subspecialty may not be scheduled to more slots during a day than the amount of slots available to that subspecialty on that day, while (9) require that for the whole cycle, the surgeons of one subspecialty may not be scheduled to more slots than than there are slots available to that subspecialty during a cycle.

Constraints (10) define the minimum number of elective patients from each category that have to be scheduled for surgery. Constraints (11) state that for each elective patient category, there can not be scheduled more patients for surgery than the target throughput of that category. Constraints (12) state that the sum of all elective patients scheduled through the cycle for one patient category, minus the amount of patients scheduled above the lower limit for the same patient category, has to equal the lower limit of patients scheduled for that patient category.

Constraints (13) limit the total expected surgery duration scheduled to elective slots within an operating room to not exceed the elective slot time available in that operating room. They also require that a trained subspecialty has to be scheduled to the room in order to schedule patients there.

The constraints given by (14), demand the number of elective inpatients scheduled for Friday to be less than or equal to a given limit in order to provide less loading on the wards during the weekend.

Constraints (15) and (16) are the first stage bed restrictions. Through these, we aim to distribute the existing bed capacity in an optimal manner. Constraints (15) restrict the number of staffed beds at a ward on a given day to be less than or equal to the maximum of beds available on that ward, while constraints (16) restrict the total number of staffed beds on a given day to be less than or equal to the amount of staffed beds available on all wards on that day.

The second stage decisions

The second stage constraints have to hold for all scenarios, and has the first-stage decision variables as input parameters.

Constraints (17) state that for a given scenario, and a given green emergency patient category, the sum of patients scheduled to both the flexible slots and the elective slots have to equal the weekly excess demand of that category. Constraints (19) require that the expected surgery duration of elective patients scheduled for all elective slots in an operating room, minus the surgery duration of electives cancelled, plus the surgery duration of green emergencies scheduled to the same slots have to be less than or equal to the elective slot time scheduled in that operating room. Constraints (19) state that the expected surgery duration of both green and yellow patients scheduled for flexible slots in an operating room, have to be less than or equal to the flexible slot time scheduled in that operating room. We may not cancel surgeries that are not scheduled, which is formulated in constraints (20).

Constraints (21) to (23) are the second stage bed constraints. The former sum all elective patients resting at a ward each day, while the second does the same for the green emergencies. The last set of constraints sum all patients resting at a ward on a day, plus the beds provided from the ward to host patients belonging to other wards, minus the beds obtained at other wards to host patients belonging to the ward, and demand that the sum should be less than the beds available for the ward on that day minus the beds covered by emergencies on the ward on that day.

Constraints (24) to (36) give the domains for the different variables.

5.3 Symmetry in the model formulations

An integer linear problem is symmetric if its variables can be permuted without changing the structure of the problem. When symmetry is present, even relatively modestly sized problems may become difficult to solve using traditional branch-and-bound or branch-and-cut algorithms (Jünger et al., 2010). If we apply homogeneous beds in our problem formulation, symmetry will be present.

Jünger et al. (2010) present several methods to avoid symmetry in mathematical formulations, such as perturbation of coefficients, fixing variables and symmetry breaking inequalities. To handle the symmetry related to homogeneous beds, we propose the following symmetry breaking inequalities:

$$\sum_{i \in \mathcal{I}^{EL}} x_{ikd} - q_{kd} = 0 \qquad k \in \mathcal{K}, \quad d \in \mathcal{D} \tag{37}$$

$$\sum_{i \in \mathcal{I}^{EL}} S_i x_{ikd} - q_{kd} = 0 \qquad k \in \mathcal{K}, \quad d \in \mathcal{D} \tag{38}$$

$$q_{kd} - q_{k+1,d} \geq 0 \qquad k \in \mathcal{K} | k \leq |K|, \quad d \in \mathcal{D} \tag{39}$$

In constraints 37 we require that q_{kd} have to equal the number of elective patients scheduled for operating room k on day d, while in constraints 38 we require that q_{kd} have to equal the total surgery duration scheduled for operating room k on day d. In constraints 39 we state that the total amount of workload scheduled for surgery in operating room k on day d has to be at least as big as for operating rooms $k+1$ on the same day. Note that constraints 37 and 38 should generally not be used together, as this may cut away feasible solutions. Furthermore, constraints 39 should be applied when either equations 37 or 38 are used.

6 Simulation

In the MSSP we aim to deal with both the uncertain bed loading at the wards and the fluctuating demand of green emergency surgeries. However, the optimization model neglects other important issues regarding uncertainty, such as the uncertain surgery duration and length of stay within a patient category. Furthermore, as we aim to develop a cyclic MSS we should verify how well this cyclic schedule handles real life fluctuations over time. In order to include more uncertainty into the modelling, and to analyze the performance of the MSS in a real life environment, we develop a simulation model.

There is yet another reason for developing a simulation model. The historical data we have access to reflect the scheduling policy that was present at the time when the data was generated and stored. As we want to investigate a scheduling policy with flexible slots, the historical data available may not provide suitable scenarios for the optimization model. We therefore want to use the simulation model to generate the data used in the scenarios fed to the optimization model.

6.1 System description

The system under consideration is a hospital department, and the flow of both elective and emergency patients from arrival at the wards to being sent home following surgery. The flow of both elective and emergency patients are based on the descriptions provided in Section 2.3.4. For the elective patients we exclude all activities prior to entering the hospital at the day of surgery, and for the emergency patients we exclude all activities before, and including the triage-process. The system may be described as a queueing network where the patients are the customers, and the ORs and the wards are the servers.

6.1.1 The components of the system

The entities considered in the system are the following:

- The elective patients
- The emergency patients
- The wards
- The elective ORs
- The emergency ORs

For the elective patients, the category is the attribute which is governing both the surgery duration and the length of stay following surgery. The attribute relevant for the emergency patients is the level of urgency, as this governs the prioritization in the queue, the surgery duration and the length of stay of these patients. The attributes linked to the wards are whether or not the specific ward may host patients of a certain category, and the number of beds available at each ward on each day of the cycle. For the ORs, the attributes are the opening hours, and the slots scheduled for each elective OR.

There are five activities to consider in the system:

- The preoperative stay at the wards for the emergency inpatients
- The preoperative stay at home for the emergency outpatients
- The transportation of emergency patients from the ward to the OR
- The surgery (including cleaning)
- The postoperative stay at the wards for the inpatients

The preoperative stay at the wards, or at home, for emergencies represent the queue, while the three other activities are performed by the servers. Due to shortage of capacity at the post anesthesia care unit (PACU) some patients may have to wait in the OR before being transported out and receive a bed in the wake up area. Because of this, there may also be a queue of patients waiting for a bed at the PACU.

There states of the system are as follows:

- The number of emergency inpatients resting at each ward, preoperative
- The number of emergency outpatients waiting at home, preoperative
- The number of patients resting at each ward, postoperative
- Whether an OR is busy or idle

The events of the model are the following:

- The arrival of emergency patients
- The completion of a surgery (including cleaning)
- An inpatient leaving the ward after receiving surgery

6.2 Modelling the system

To model the system, we develop a discrete-event simulation program in MATLAB. In Figure 13, a simplified illustration of the flow of patients in the model is provided. The fundamental flow of all emergencies is starting with the arrival to the hospital and the preoperative ward. When a suitable OR is ready, the most prioritized patient is admitted for surgery, and afterwards the patient is either sent to a ward for rest, or he is sent home. Note that the PACU is not included in the model.

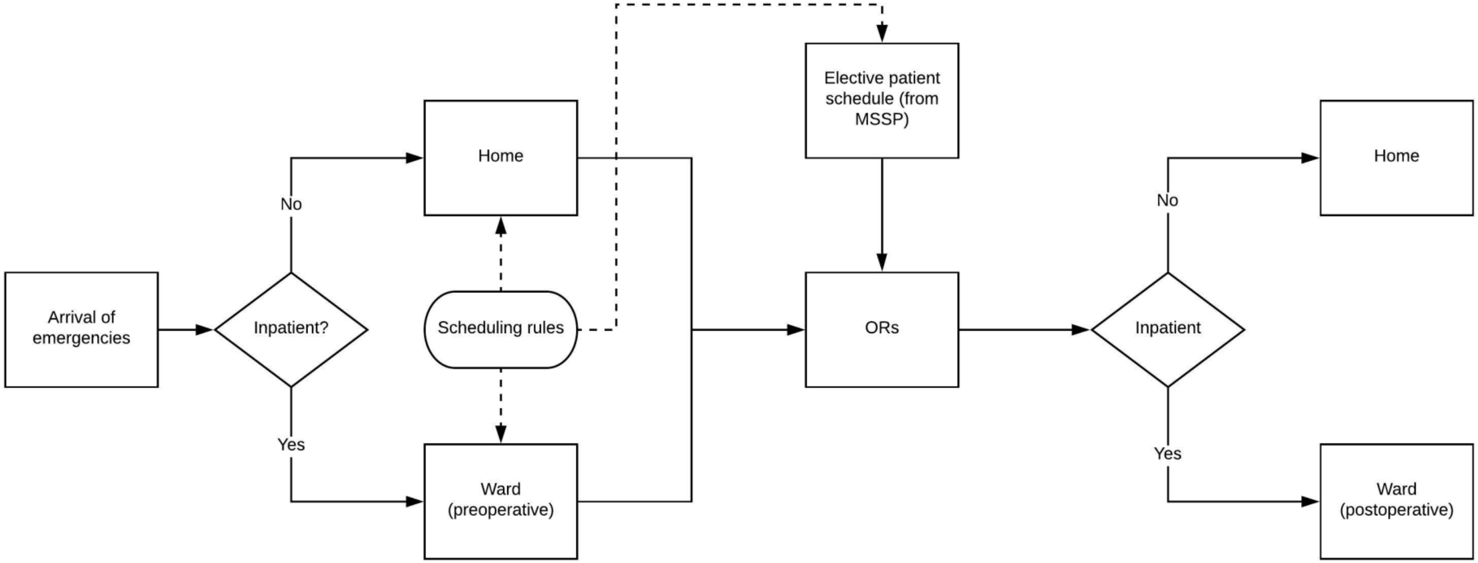


Figure 13: A flow chart describing the simulation model

The flow of elective patients starts at the OR, as these patients are not going to the ward prior to surgery. Following surgery, also these patients are either going to the ward or home. To govern the flow of patients in the system, some scheduling rules (queue disciplines) are applied.

6.2.1 Assumptions made in the model

The wards are assumed to have infinite capacity, implying that no rescheduling are done as a result of the wards being overloaded. This assumption also means that all patients may leave the OR immediately following surgery, implying that no queue is formed between the OR and the ward. Furthermore, all emergency inpatients return to the same ward that they were resting at prior to surgery. For the elective patients, all patients go to the ward they were meant to according to the schedule provided by the optimization model.

We schedule the patients based on expected surgery duration. We may not schedule for overtime, but overtime may occur as a result of the realized surgery duration. Following each surgery, the OR should be cleaned, implying that the room is unavailable for some time following surgery. For the urgent emergency patients that are not scheduled ahead, we assume that there will be some delay between two surgeries (in addition to the cleaning of the OR), due to transportation and logistical issues. The surgery of elective patients are performed according to the schedule provided by the first stage decisions in the optimization model. Furthermore, no electives are delayed or prevented from showing up on the day of surgery. We assume that all weeks have the same input of elective patients, implying that all vacations and similar activities are excluded.

When scheduling green emergencies to the flexible slots, we schedule the ORs in increasing order. This means that the flexible slots available in OR-1 need to be fully booked before scheduling the green patients for the flexible slots in OR-2 and so on. If elective OR capacity is necessary to have all green emergencies scheduled for surgery, we need to choose both a day and an OR where the green patient may be scheduled. To choose the day of rescheduling, we pick the first day available where there are elective patients scheduled. Then, to find a patient to reschedule, we iterate over the elective patient categories, choosing the first patient available for the chosen day.

The red patients are the only group of patients that may receive surgery after 22:00. Outside the opening hours of the ORs, the red patients have to compete for the OR with other urgent emergencies from other departments. Furthermore, no green emergencies are scheduled for surgery during the weekend.

6.2.2 Scheduling rules and the flow of patients

The process of prioritizing patients for surgery is a very sophisticated process depending on many factors such as the staff present, the number and types of patients waiting, the situation at the different wards, and the situation at the hospital as a whole, which make the process impossible to model exact. Here, we provide the scheduling rules applied in the model in order to mimic these processes at orthopaedic department. We also provide a more detailed description of the flow related to the orthopaedic department.

Within all emergency categories, the patients are scheduled according to a first-come-first-serve rule (FCFS). Furthermore, the red patients will always have prioritization over the yellow and green patients if all are present at the same time. If no red patients are present, there are some basic rules for prioritization between the yellow and green patients. If none of the candidates have waited beyond the limits proposed by the traffic light system, the yellow patient will go first. If only one of the patients has waited beyond the limit, this patient will go first. If both patients have waited beyond the limits, the one that has exceeded the limit the most will go first. There is however cases where the yellow patient should go first independent of the rules.

For a detailed illustration of the flow of patients in the simulation model, see Figure 39 in Appendix C. All yellow and red patients may be summoned for surgery immediately after arrival to the ward, while the scheduling of green patients is done on the morning after arrival. The green inpatients are primarily put in line for surgery at the emergency ORs, but to keep the waiting time down for these patients, there is a limit on the amount of green inpatients that may wait in this line. If the limit is reached, the green inpatients are scheduled for the flexible slots. If no flexible slots are available within a given number of days, elective patients need to be rescheduled in order to provide OR capacity for the green patients. All elective patients that are displaced will be rescheduled to a flexible slot some days ahead. There is a lower limit on the number of days that we are inhibited from rescheduling elective patients, as rescheduling a patient just before surgery is not preferable. There is also a lower limit on the number of days ahead that we may not reschedule the elective patients after having been displaced.

The scheduling of the green outpatients is quite similar to the one for the green inpatients. However, we would not like to have these patients waiting in line for surgery in the emergency ORs, as this will imply uncertainty around the exact time for surgery for the green outpatients. To achieve this, we first try to find an idle flexible slot available within a limited number of days. If no flexible slot is available, we have to send the green outpatients in line for surgery at the emergency ORs. If the maximum limit of green outpatients in queue for the emergency ORs is reached, we have to reschedule elective patients to provide OR capacity for the green outpatients.

For each scheduling day, we limit the number of elective patients that may be rescheduled. If this limit is reached, we have to put the green patients in queue for the emergency ORs, even though the maximum limit of emergencies in line for the emergency ORs is reached.

In Appendix C we report on the validation of the simulation model, comparing the simulated outcomes to the historic data obtained from the orthopaedic department.

6.3 Output parameters of interest

There are several output parameters that are of interest when analyzing the outcomes of the simulation model, and next, a short description of these parameters are provided.

Gaining information about the number of emergencies waiting for surgery at a given time of the day through the week is of great interest. Having a large amount of emergency patients resting at the wards prior to surgery may result in over crowded wards and the cancellation of elective surgeries as there may not be room for them following surgery. To analyze the queue of emergencies at a given point of time, we sum all emergencies waiting for surgery at all wards, plus the amount of green outpatients waiting at home at that time.

The total number of patients resting at the wards is interesting to analyze because the wards are assumed to have infinite capacities in the model, and no rescheduling is done in order to keep the number of inpatients below a certain limit. Having periods in the simulated reality where the total bed capacity is exceeded will represent a point of time where some rescheduling/cancellation of electives would have been made in order to keep the bed loading under the maximum level. If a MSS yields repeatedly periods where the amount of inpatients exceed the total bed capacity available at the orthopaedic department, the schedule is probably to optimistic regarding the number of elective patients scheduled, and a lot of rescheduling will be necessary in order to cope with the limited bed capacity.

We are not only interested in whether the emergency patients receive surgery within the limits proposed by the traffic-light system. It is even more interesting to know the distribution of waiting times the different emergency patients. The waiting time for surgeries are counted in days for the green patients, and in minutes for the red and yellow patients. The waiting time is measured as the time from arrival to the system, and to entering the OR.

The OR utilization at a given day is measured as the time used for all surgery related activities at the OR, including the postoperative cleaning of the OR, divided by the opening hours available at the OR on that day. A high OR utilization indicates that the OR is being used most of the time, and that the resource is well utilized. However, increasing the OR utilization will often imply increasing the chances of overtime work at the OR, and reduce the flexibility to treat more patients in busy periods.

7 Implementation

To generate good MSSs we arrange the optimization model and the simulation model in a loop, as may be seen in Figure 14. Each iteration in the loop begins with running a MSS in the simulation model to generate scenarios that are fed into the optimization model. Then, the optimization model produces a new MSS that may again be run in the simulation model. For each iteration a new MSS is generated. The optimization model is implemented in the Mosel language and it is solved in Xpress IVE 8.3. The computer used to solve the model is an HP Intel (R) Core (TM) i7-7700 CPU, 3.6 GHz, 32 GB RAM.

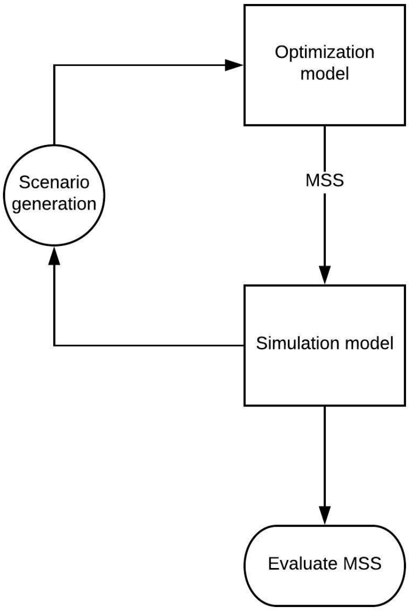


Figure 14: The loop set-up including both the optimization model and the simulation model.

As the simulation model starts out as an empty system, a warm-up period should be implemented in order to reach steady state before recording data. The simulation model includes many stochastic variables, and the outcomes of each run may differ a lot. To generate scenarios that are representative of the simulated reality, and such provide stable scenario trees, we should run the simulation model for several times in each iteration of the loop.

The scenarios are gathered from the data produced by the simulation model. One scenario represent one week, and to keep the scenarios independent of each other we draw the weeks with one month in between. When drawing a scenario, we exclude the postoperative length of stay for the green inpatients, as these will be added by the simulation model. Note that we exclude the postoperative length of stay of the green inpatients based on the day when they received surgery in the simulation model. This may not be the same day as they are scheduled for surgery in the optimization model. We may for example remove the patient from covering beds on Wednesday and Thursday, while the patient may be scheduled for surgery on Monday in the optimization model, meaning that we add the postoperative days to Monday and Tuesday instead of to Wednesday and Thursday. An alternative way of doing this may be to remove all the length of stay of the green patients, both before and after surgery, and then add both in the simulation model when scheduling the patients for surgery. The positive about this approach is that it provides a continuous length of stay for the green patients, where we know that the postoperative days are joint to the preoperative days for the same patient. However, when removing only the postoperative days we introduce some randomness regarding the length of stay of the green patients as we do not know whether the postoperative days are joint with the preoperative days from the simulation model. This may represent a reality where the length of stay is not the same for all patients. In this approach we get the same amount of days added in total, but we add some randomness that may provide some value.

For the yellow patients we do not remove either the preoperative or the postoperative length of stay as these patients are only scheduled for surgery in the flexible slots in periods of low excess demand of green patients. Therefore, when these patients are scheduled for surgery in the optimization model, we do not add any postoperative length of stay as these days are already accounted for in the scenarios. Note that this approach may also be applied for the green patients, and it might provide a good alternative. As most of the green inpatients cover a bed prior to surgery, performing surgery to a green inpatient do not mean that more beds are covered after surgery than before.

Because the simulation model treats the wards as having unlimited capacity, we may end up drawing scenarios that exceed the total bed capacity defined in the optimization model. To overcome this challenge, we introduce some new variables in the optimization model, β_{wds}. These variables represent additional beds at ward w, on day d, in scenario s, so that we are able to solve the model without altering the scenarios. Since these new variables represent beds that are not available, we need to punish these beds hard in the objective function. The objective function applied to handle all scenarios may be seen in equation (40), where the parameter C^{β} is the cost of adding new beds to the wards. An altered version of equations (23) are provided in equations (41).

$$\begin{aligned} max \sum_{i \in \mathcal{I}^{EL}} P_i v_i - \sum_{s \in \mathcal{S}} P^S_s \Bigg[& \sum_{i \in \mathcal{I}^{EL}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} C^C_i x^C_{ikds} + \sum_{w \in \mathcal{W}} \sum_{w' \in \mathcal{W}} \sum_{d \in \mathcal{D}} C^W_{ww'} b_{ww'ds} + \\ & \sum_{i \in \mathcal{I}^{GR}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} C^{GR} e^{EL}_{ijkds} - \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} \sum_{d \in \mathcal{D}} P^Y e^Y_{jkds} + \sum_{w \in \mathcal{W}} \sum_{d \in \mathcal{D}} C^{\beta} \beta_{wds} \Bigg] \end{aligned} \tag{40}$$

$$\sum_{i\in\mathcal{I}_w^W} u_{iwds} + \sum_{i\in\mathcal{I}_w^{GRW}} u_{iwds}^{GR} + \sum_{w'\in\mathcal{W}|w'\neq w} b_{w'wds} - \sum_{w'\in\mathcal{W}|w'\neq w} b_{ww'ds} \\ \leq a_{wd} + \beta_{wds} - U_{wds}^{EM} w \in \mathcal{W}, \quad d \in \mathcal{D}, \quad s \in \mathcal{S} \tag{41}$$

When applying the models to provide a MSS for a real life department, we should first alter the simulation model to mimic the department of interest. Running the simulation model should then provide us with scenarios that are representative for the department. Based on these scenarios we may now obtain a new MSS from running the optimization model. From iteration two we may want to alter the simulation model again. Imposing flexible slots at the elective ORs may require that the scheduling rules are changed, and these changes should be implemented in the simulation model. When the new scheduling regime is implemented we may run the loop for some iterations, aiming to generate a good MSS.

8 Computational study

In this section we firstly provide the input data for the models. Then, we perform a technical study of the optimization model, including stability testing and the value of the stochastic solution. Afterwards, we aim to provide some general managerial insight from running the optimization model on small instances, and finally we perform a case study on the orthopaedic department at St. Olav's Hospital.

8.1 Input data for the optimization model

In this subsection we present the input data applied when solving the optimization model for the orthopaedic department at St. Olav's Hospital. The values obtained for the different parameters are not provided here, but these may be found in Appendix A.2. The values applied for the orthopaedic department will form the basis for all the instances presented in this Section, and when changes are made to the input data, this will be noted. The data are obtained from two databases, OPPLAN and Nimes, and a short introduction of these are given in Appendix A.1.

8.1.1 The patients

The properties related to the patients are the different patient categories, the target throughput of electives, the expected surgery duration and the expected length of stay of inpatients at the wards. The data gathered for each patient category can be seen in Table 26 in Appendix A.2.

In order for the model to yield realistic results, we need to find a reasonable level of aggregation when sorting the patients. The sorting is based on the subspecialties that are present at the orthopaedic department today, and we aim to fit the patients into these. To divide the patients into groups that fit to the subspecialties, both the procedure codes and a one-line description of the recorded surgeries are used. Some of the subspecialties are hard to sort out from the data, and are therefore left out. In order to provide some more granularity to the data, we further divide many of the patient groups belonging to a subspecialty into smaller groups according to diagnostic similarities (referred to as the patient categories in the models). The emergency patients are treated more aggregated. For the green patients we have two categories, the green inpatients and the green outpatients, while for the yellow patients we only use one category for each.

Figure 15 illustrates the amount of orthopaedic surgeries that are performed each week. The mean number of elective surgeries performed each week is 68, while the emergency counterpart is 61. In some periods there are less elective surgeries performed. These periods are for example the summer vacations, where the department plan for less elective production. Excluding the weeks where production is low, the mean number of elective surgeries performed each week increases to 78, representing a normal, high-production week. This number is interesting, as it provides us with a reasonable number for the target throughput of elective patients for the model.

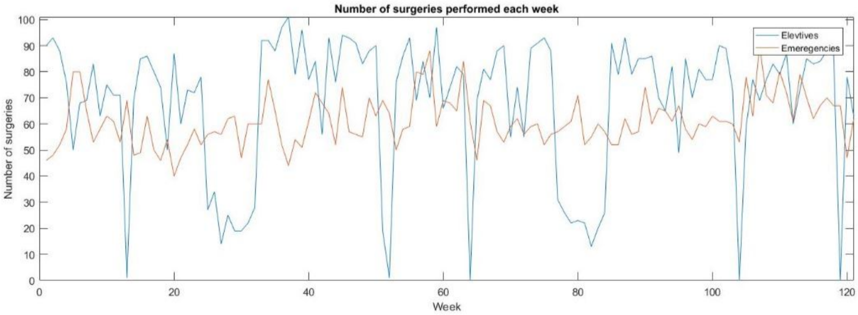


Figure 15: The number of orthopaedic surgeries performed every week at the orthopaedic department

The surgery duration used for input in the optimization model is calculated as the mean surgery duration for each patient category. As not all patients are inpatients, only a fraction of the patients belonging to each patient category need a bed following surgery. To obtain the average length of stay for each patient category, we therefore multiply the expected length of stay of all inpatients belonging to a patient category with the fraction of the patients from the given patient category that are inpatients. For the yellow patients present in the optimization model we set the length of stay equal to zero. This is because the length of stay of these patients are covered by the scenarios that provide the number of emergencies resting at the wards each day of the week. For the green inpatients we exclude the preoperative length of stay as these are covered by the scenarios. We estimate the postoperative length of stay of the green inpatients to be two days. Note that in the optimization model we allow for the expected length of stay of a patient to vary depending on the day of surgery, however, in the implementation we do not include any variation depending on day of surgery.

8.1.2 The operating rooms and the wards

Parameters relevant for the ORs in the optimization model are the opening hours, the number of slots available in each OR through the day, and what subspecialty that may be scheduled for the individual rooms. In Table 27, in Appendix A.2, the information regarding the seven elective ORs is provided. The ORs are not homogeneous, and we assume that the rooms can only be scheduled for the same subspecialties that they are today. We assume that there are two slots available at all ORs every weekday, and that the rooms are closed during the weekends.

Both the number of ORs that may be covered by the anesthesiologists each day, and the number of OR slots available to each subspecialty each day and through the week are set equal to the present situation at the orthopaedic department. Tables 28 and 29 available in Appendix A.2 provide the values chosen for these parameters.

There are five wards available in the optimization model, and each ward may host different patient categories and have different capacities. Three of the wards are available through all days of the week, while the others are closed during the weekends. Table 30 in Appendix A.2, presents an overview of the five wards and the beds available.

8.1.3 The objective function parameters and other parameters

The objective function parameters related to the elective patients may be seen in Table 31, while the rest of the objective function parameters are provided in Table 32, both available in Appendix A.2. When it comes to the scheduling of elective patients we do not value some diagnostic groups over others, but we value the scheduling of inpatients some more then the scheduling of outpatients as the inpatients consume more resources and provide more DRG points. The cost of cancelling elective patients are the same for all patient categories, except for the prosthesis patients that are more expensive to cancel as these generate much DRG points.

The cost of sending patients to other wards than intended should be positive, to prevent it from happening if it is unnecessary. However, this is something they often do, and should not be penalized hard. The gains from providing surgery to yellow patients in flexible slots should be positive such that the yellow patients are sent to flexible slots if no green patients are present. However, the gains should be low so that we do not provide excessive slot capacity for emergencies at the elective ORs. The parameter C^{β} is the cost added in the objective function if we exceed the total bed capacity available, which may happen as a result of the scenario generation procedure.

The parameters not yet covered in the text above may be found in Table 33 in Appendix A.2. Note that we have set the parameter H_5 to zero, indicating that no inpatients may be scheduled for Friday. We do not include H_d for any other days than Friday.

8.2 Input data for the simulation model

As the MSS generated by the optimization model is fed into the simulation model, the parameters presented above are also relevant for the simulation model. However, in the simulation model we treat both the arrival of emergencies, the surgery duration and the length of stay of the patients as stochastic variables.

To model the arrival of emergency patients, we sort all emergencies according to the degree of urgency based on the traffic-light system. We divide each day into three time intervals, 00.00-08.00, 08.00-16.00 and 16.00-24.00, and for each day of the week, and for each time interval we find the expected inter-arrival time of each urgency category. Note that many of the emergencies have been registered to arrive at 00.00 as a default value. For these patients we draw a random time for arrival between 08.00 and 22.00.

For the surgery duration we have much data for each patient category, so we let these serve as an empirical distribution. For each patient we draw random, independent realizations from the distribution belonging to the respective patient category.

As we apply relatively aggregated patient categories we have a large variation in the surgery duration for each patient category. When scheduling electives in the real life, the scheduling personnel work with more fine-grained patient categories, and they have a good feeling for the surgery duration of different patients. This enables them to schedule patients more precise than based the mean surgery duration of aggregated patient categories. To avoid unrealistic overtime peaks at the ORs we exclude the upper 1/6 of the distributions for surgery duration of elective patients. Also, to avoid drawing realizations that on average are much lower than the expected surgery duration, we exclude the lower 1/10 of the distributions for the electives. This is not done for the emergency patients because these are harder to schedule efficiently for the scheduling personnel.

For the simulation model we use a Kernel density estimator available in MATLAB to create an approximation of the empirical distribution of the length of stay for each elective patient category, and then we draw realizations from this distribution to decide on the length of stay of each inpatient following surgery. For the emergency inpatients we create one common Kernel-distribution for all the urgency categories. When drawing a realization from this distribution we subtract the preoperative length of stay, making sure that we do not end up with a postoperative length of stay of less than one day.

In addition to the seven elective ORs included in the optimization model, we also include OR-1 at BVS and the two emergency ORs at AHL. OR-1 is assumed to be a flexible OR. We apply the opening hours given in 2.3.1, but the ORs open at 08.00 (instead of 07.45), and the elective ORs close at 16.00 (instead of 15.30). Note that we do not include the OR at KBS, and that one emergency OR is available to the red and yellow emergencies through the weekend. This OR is a shared capacity with all emergencies at the hospital, and may not always be available to the orthopaedic emergencies.

To govern the scheduling rules and the flow of patients through the ORs, some additional parameters are needed. These parameters, and the values applied both when trying to mimic the real world, and when developing the three MSSs in Section 8.5, may be found in Table 36 in Appendix A.2. Note that the values of the parameters are the same for all the four cases, except for the parameters deciding on the maximum number of green inpatients and green outpatients that may wait in the queue for the emergency ORs before we start sending these patients to the elective ORs. When we aim to mimic the real life, we allow for three green inpatients and one green outpatient to wait in this queue. For the three other cases, we allow for no green outpatients and less green inpatients to wait in the queue.

8.3 Technical study of the optimization model

To perform the technical study of the optimization model we apply much of the input data described above. However, all subspecialties are given a maximum of 16 OR slots available through the week, and the target throughput of electives is increased to 105 in total. In addition, we do not imply restrictions regarding the amount of inpatients to be scheduled for Friday. As for the case study in Section 8.5, we do not allow for flexible slots to be scheduled for OR-7 and OR-8.

An illustration of the set-up used in this subsection may be found in Figure 16. Firstly, run the simulation model once to provide input scenarios for the optimization model. Then, the optimization model is run once, and a MSS is produced. For this MSS, we run the simulation model four times, to get four different scenario trees based on the same MSS. Then, for each of these four scenario trees, the optimization model is run twice. In the second run (these runs are indicated with a B), we add more beds to avoid the big penalty when the total bed capacity is exceeded because of the scenario generation procedure.

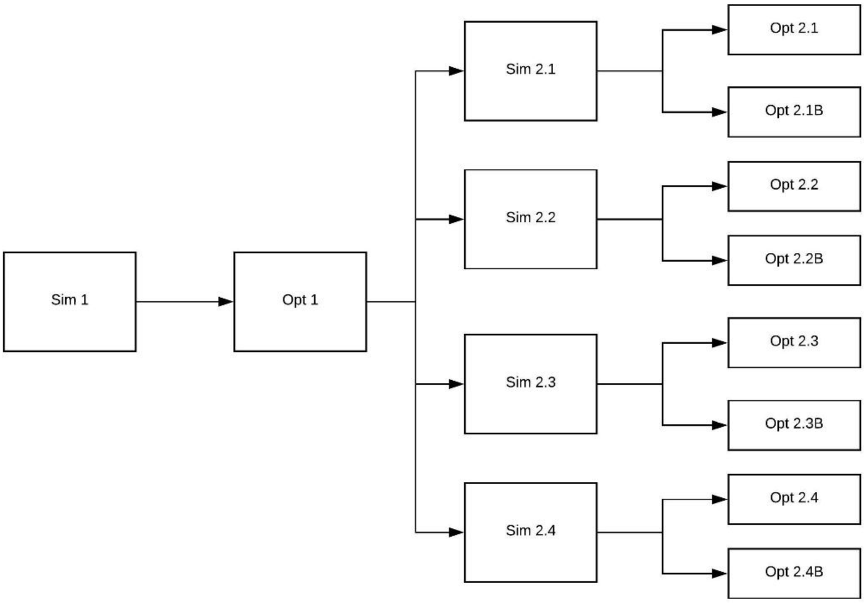


Figure 16: The set-up for testing the optimization model. A "B" at the end represents runs where the bed capacity has been increased to avoid the big penalty associated with the number of patients resting at the wards exceeding the total number of beds available.

A technical summary from running the optimization model is presented in Table 10. Each instance is run for 3 hours, and the *-sign indicates that the time limit is reached before the optimal solution is found. Xpress uses the Branch-and-Bound (B & B) algorithm for solving Mixed Integer Programs (MIP). The B & B algorithm is an exact method that guarantees optimality. However, if the model to be solved has a large number of integer variables, the running time with this method may be lengthy. We may not always have time to wait for the algorithm to provide the optimal solution, but we are still left with useful information. As the algorithm evolves it provides an upper and a lower bound yielding an interval where the optimal solution is found. This bound is referred to as the LP-gap as it provides us with the difference between the best active solution to the linear relaxation of the problem (LP-solution), and the best integer solution (IP-solution) obtained so far. For a maximization problem, as we are facing in our model, the lower bound is provided by the best IP-solution found so far, and the upper bound is represented by the best, active, LP-solution. To calculate the LP-gap, we subtract the objective value of the IP-solution from the objective value of the LP-solution, and divide this on the objective value of the IP-solution. The rows equal the number of constraints for the given instance, while the columns equal the number of variables present. We see that the number of B & B- nodes visited differ a lot for the same running time. This may be due to the fact that we run two-and-two instances in parallel on the computer.

Table 10: Technical summary of the optimization model

Run	Obj. val.	Time to solution	LP-Gap [%]	Rows	Columns	B&B-nodes	Int. sol. found
2.1	-68.97	10,800*	0.46	117,777	72,935	50,186	40
2.2	52.68	10,800*	1.03	117,777	72,935	34,617	65
2.3	71.46	10,800*	0.51	117,777	72,935	23,318	69
2.4	-243.91	10,800*	0.12	117,777	72,935	45,196	45
2.1B	89.84	10,800*	0.37	117,777	72,935	73,545	77
2.2B	90.68	10,800*	0.43	117,777	72,935	76,564	64
2.3B	89.45	10,800*	0.60	117,777	72,935	107,778	59
2.4B	85.58	10,800*	0.30	117,777	72,935	77,703	99

The model formulation is prone to symmetry, especially when the ORs are treated as homogeneous. As only the prosthesis subspecialty has access to OR-7 and OR-8, these ORs are homogeneous. To avoid symmetric solutions for these ORs, we add restrictions demanding that the number of slots scheduled for OR-8 has to be at least as high as the number of slots scheduled for OR-7. We also demand that the number of patients scheduled for the slots at OR-8 is at least as high as the number of patients scheduled for the sots at OR-7. These restrictions may be added simultaneously since both the hip and knee prosthesis patients require the same amount of bed capacity following surgery, and the surgery duration of the two categories allow for a maximum of one patient scheduled per slot, and a maximum of two patients scheduled for two slots.

8.3.1 Effects of symmetry breaking inequalities for homogeneous operating rooms

In Table 11 we present the results when having homogeneous ORs for the run 2.1. Note that when we have homogeneous ORs, the number of rows and columns increase a lot. Since we have the opportunity to send all subspecialties and all patients to each OR, we now generate more variables and more constraints for each OR. In 2.1C we apply no symmetry breaking inequalities. In 2.1D we add constraints 37 and 39 presented in Section 5. Finally, in run 2.1E we add constraints 38 and 39 provided in Section 5. The fact that the problem increases in size, and that more symmetry is introduced in the formulation, makes the problem harder to solve. However, we see that introducing symmetry breaking inequalities do not provide better IP- or LP-solutions after 24 hours of running. We do not make any further attempts to make more efficient formulations.

Table 11: Technical summary when applying homogeneous ORs

Run	Obj. val.	Time to solution	Upper bound	Rows	Columns	B&B-nodes	Int. sol. found
2.1C	-81.43	86,400*	-63.68	258,557	206,105	27,358	39
2.1D	-91.12	86,400*	-63.89	258,622	206,200	13,010	72
2.1E	-112.83	86,400*	-63.95	258,622	206,200	23,305	72

8.3.2 Stability testing

When performing the stability testing of the optimization model we, need to generate scenarios from one specific scheduling regime and one MSS in the simulation model. As each scenario contains 37 different stochastic parameters that may take on values in a relatively wide range, the total support of the random variables is very large. However, only a relatively narrow part of the parameter values are relevant for each scheduling regime and each MSS. As an example there will typically be more patients resting at the wards for a MSS with no flexible slots compared to a MSS with many flexible slots, given that everything else is equal. The reason for this is that many flexible slots provide the opportunity to perform more emergency surgeries in periods of high emergency demand, yielding less emergencies covering beds while waiting for surgery compared to if no flexible capacity is added. As a consequence of this we, we will not have stability when the scenarios are generated based on different MSSs and different scheduling regimes. However, we would like to have a piece-wise stability, where we have stable solutions if the scenarios are generated based on the the same MSS and the same scheduling regime. If the objective function is steep, or have breaking points, similar solutions may give very different objective function values making it hard to test for in-sample stability based on the objective function value as described in Section 2.4.1. Remember that we penalize hard if we exceed the total bed capacity in a scenario. This gives rise to breaking points in the objective function. To avoid these breaking points, we use the runs 2.1B-2.4B from Table 10 when testing for stability.

We see from Table 10 that the objective values of the best integer solutions found in the three first runs (2.1B-2.3B) are very similar, while the fourth optimal objective value is less than the others. However, the difference between the largest and the smallest objective value is only 6.0 % (when divided by the smallest value), so judging from these four runs we have in-sample stability. Note that for the runs 2.1-2.4 the optimal objective values differ a lot more, illustrating in-sample instability due to the big penalty added when exceeding the bed capacity in some scenarios.

To test for out-of-sample stability we would like to have a scenario tree representing the total support of the random parameters. For this problem, such a tree is way too big to generate and test, so we need to generate an approximation. We use the simulation model to generate a tree consisting of 500 scenarios to approximate the total scenario tree. This is still a relatively small tree, but it should provide insight to whether the model is out-of sample stable. The objective function value from solving the four second-stage problems with the big scenario tree may be found in Table 12. We see that the objective values are very similar, providing confidence that the model is out-of-sample stable.

Table 12: Out-of-sample stability results. The objective values are obtained from solving the second-stage problem with 500 scenarios using the first-stage variables obtained from run 2.1B-2.4B as input parameters

Run	Obj val.
2.1B	90.19
2.2B	90.04
2.3B	89.95
2.4B	90.18

8.3.3 The value of the stochastic solution

To explore the value of the stochastic solution (VSS), we use all the eight runs provided in Table 10. In the mean value problem (MVP), the stochastic parameters take their expected values. The expected values of these parameters are typically not integer values, so we round the expected values to the nearest integer. None of the MVPs are solved to optimality within 3 hours, but the gaps are small. As none of the stochastic problems are solved to optimality either, we may not calculate the true VSS. However, we may provide intervals for the VSS. The upper limit is calculated as the difference between the objective value of the upper bound obtained from the stochastic solution (SS) and the objective value of the mean value solution (MVS), while the lower limit is calculated as the difference between the objective value of the IP-solution from the SS and the objective value of the MVS. Since the MVS is based on the solution obtained from the MVP that was not proven to be optimal, the lower bound of the VSS is only the lower bound available after three hours of running the model, and it may be less (but never below zero).

In Table 13 we list the SS, the MVS and the VSS. For the instances 2.1-2.4 we see that the VSS is large. Remember that in the MVP the problem is deterministic, so it is meaningless to schedule bed capacity that we know will be left unused. As a consequence of this, there are some days where we do not schedule the maximum amount of beds available. When exposed to the scenarios, the solution obtained from solving the MVP has to add additional beds at a high cost in order to deal with the excessive loading of beds in some scenarios. Adding restrictions to the MVP stating that we should schedule the total capacity of beds each day would provide a smaller VSS. In the instances 2.1B-2.4B we have access to more beds during the weekend, yielding the chance to schedule more electives for surgery in the first stage. Scheduling more electives for surgery provide incentives for scheduling more beds every day. As a result, the total bed capacity is scheduled almost every day, making the solution more robust to tolerate fluctuations in the bed loading. As a result, the VSS is much lower for these instances.

Despite the relatively small VSSs in instances 2.1B-2.4B we see that the SSs provide less cancellations and more elective surgeries than the MVSs. The VSS is highly dependent on the values of the objective function parameters. To show this, we have solved instance 2.1 over again. This time we have excluded the opportunity of scheduling the beds in the first stage, but we have set the cost of cancelling electives to ten times the values just used. The SS and the MVS from solving this instance may be seen in Table 13 as instance 2.1F. When setting the cost of cancelling electives high, we obtain solutions that are more risk averse towards the scheduling of electives. In this setting, avoiding elective cancellations is important, and we see that the SS has far less cancellations compared to the MVS, and we obtain a large VSS.

Table 13: 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, the number of cancellations made in the second stage for all 100 scenarios, and the number of elective patients (out of 10500) to receive surgery in the second stage for all 100 scenarios.

Run	Flex slots (SS)	Flex sots (MVS)	Canc. (SS)	Canc. (MVS)	El. sched. (SS)	El. sched. (MVS)	SS	MVS	VSS
2.1	6	5	461	577	9239	9223	[-68.97, -68.66]	-613.27	[544.3, 544.61]
2.2	5	5	466	686	9334	9114	[52.68, 53.22]	-364.83	[417.51, 418.05]
2.3	6	5	446	652	9254	9148	[71.46, 71.82]	36.86	[34.6, 34.96]
2.4	5	5	629	696	9171	9104	[-243.91, -243.61]	-346.35	[102.44, 102.74]
2.1B	5	5	260	453	9540	9347	[89.84, 90.17]	84.28	[5.56, 5.89]
2.2B	5	5	244	455	9556	9345	[90.68, 91.08]	84.28	[6.40, 6.80]
2.3B	5	5	273	498	9527	9302	[89.45, 89.99]	82.65	[6.80, 7.34]
2.4B	5	5	403	619	9397	9181	[85.58, 85.84]	79.20	[6.38, 6.64]
2.1F	10	5	85	525	8315	8675	[-115.67, -115.30]	-237.30	[121.63, 122]

8.4 General insight

In this subsection we want to provide some general insight regarding the scheduling of flexible slots for the elective ORs. To do this, we run the model for some smaller instances, covering several topics of interest. We regard a small hospital department with a total of 8 ORs, two wards, three surgery subspecialties and five elective subspecialties. In addition, we include the emergency patients, and these are again divided into three urgency categories. Three of the ORs are dedicated for emergency patients and are only included in the simulation model. The rest are elective ORs and they are included in both the simulation model and the optimization model. We treat the five elective ORs as homogeneous, and on Friday we only have access to four of these ORs due to less anesthesia resources. The five elective patient categories have the same properties as the five first patient categories presented in Section 8.1, and the emergency patients are treated as before. Note that we allow for inpatients to be scheduled on Friday.

In Table 34 in Appendix A.2, we list the five elective patient categories and include the two target-levels that we want to apply. Three levels of emergency patient loading are applied. The normal emergency loading level level is generated based on the same arrival rate as we have at the orthopaedic department today. For the high loading case, the expected inter-arrival time is multiplied with 0.94, and for the low loading case, the expected inter-arrival time is multiplied with 1.1. For each of the emergency loading cases we want to apply two sets of target throughput of electives, and for each of these targets we want to test three bed capacities, resulting in 18 different instances. The 18 instances are listed in Table 14. We apply four different bed capacities to provide three levels of bed capacity for each of the three emergency loading cases. The bed capacities may be seen in Table 35 in Appendix A.2. For the low emergency loading we apply the bed configurations W1, W2 and W3, and for the medium- and high emergency loading we apply W2, W3 and W4.

Table 14: The small instances applied to provide general insight

Instance	Description
LE-LT-W1	Low emergency, low target, few beds
LE-LT-W2	Low emergency, low target, normal beds
LE-LT-W3	Low emergency, low target, many beds
LE-HT-W1	Low emergency, high target, few beds
LE-HT-W2	Low emergency, high target, normal beds
LE-HT-W3	Low emergency, high target, many beds
NE-LT-W2	Normal emergency, low target, few beds
NE-LT-W3	Normal emergency, low target, normal beds
NE-LT-W4	Normal emergency, low target, many beds
NE-HT-W2	Normal emergency, high target, few beds
NE-HT-W3	Normal emergency, high target, normal beds
NE-HT-W4	Normal emergency, high target, many beds
HE-LT-W2	High emergency, low target, few beds
HE-LT-W3	High emergency, low target, normal beds
HE-LT-W4	High emergency, low target, many beds
HE-HT-W2	High emergency, high target, few beds
HE-HT-W3	High emergency, high target, normal beds
HE-HT-W4	High emergency, high target, many beds

In Table 15 we provide the outcomes from running the 18 instances, and in the the Figures 18 - 20 we provide graphical views of some central outcomes. In these Figures, the three first pairs of bars indicate the three low emergency loading cases, the three following pairs of bars represent the three normal emergency loading cases, while the last three pairs of bars indicate the three high emergency loading cases.

Table 15: Output from solving the stochastic problem for the 18 instances. We include the number of flexible slots and the number of elective patients scheduled in the first stage. For the second stage we include the number of green patients scheduled for flexible slots, the number of elective cancellations, the number of electives treated and the number of beds moved for all 100 scenarios.

Instance	Flex. slots	El. scheduled	Green pat. in flex. slots	El. cancelled	Tot. el. treated	Beds moved	Obj. val. of best int. sol.	Best bound
LE-LT-W1	9	91/91	236/240	508	8592/9100	576	20.35	22.37
LE-LT-W2	9	91/91	240/240	27	9073/9100	141	114.62	116.38
LE-LT-W3	9	91/91	240/240	0	9100/9100	0	115.51	117.11
LE-HT-W1	2	111/115	120/240	989	10111/11500	631	17.78	22.09
LE-HT-W2	1	112/115	60/240	318	10882/11500	273	117.26	125.52
LE-HT-W3	0	114/115	0/240	262	11138/11500	5	123.89	127.25
NE-LT-W2	9	91/91	480/524	452	8648/9100	605	-250.47	-247.12
NE-LT-W3	9	91/91	483/524	36	9064/9100	49	-18.26	-15.34
NE-LT-W4	9	91/91	482/524	47	9053/9100	21	111.49	114.76
NE-HT-W2	2	111/115	183/524	932	10168/11500	1015	-255.84	-249.66
NE-HT-W3	3	109/115	240/524	392	10508/11500	43	-22.41	-16.06
NE-HT-W4	2	111/115	182/524	468	10632/11500	20	108.28	114.34
HE-LT-W2	10	90/91	886/988	1165	7835/9100	937	-6676.65	-6672.26
HE-LT-W3	10	87/91	903/988	106	8594/9100	62	-704.54	-689.19
HE-LT-W4	10	89/91	902/988	101	8799/9100	57	-697.67	-689.21
HE-HT-W2	5	104/115	515/988	1605	8795/11500	929	-6697.12	-6687.24
HE-HT-W3	7	98/115	705/988	447	9353/11500	251	-1778.81	-1764.44
HE-HT-W4	7	101/115	712/988	398	9702/11500	70	-710.72	-702.62

As a general tendency we may see that the number of flexible slots increases, and the share of electives scheduled for surgery decreases as the emergency loading increases. Despite a shift towards scheduling more emergency patients when the loading of these increases, the total amount of patients to receive surgery, which may be seen in Figure 17, is quite stable for the two first emergency loading levels, with a slight increase when the bed capacity increases. For the high emergency loading level, we see that the total number of patients scheduled for surgery decreases for the high target instances and increases a bit for the low target instances compared to the two other emergency loading levels. For the low target instances the OR capacity is relatively good allowing us to schedule many flexible slots and at the same time schedule almost all electives for surgery. In instance HE-LT-W2 this yields excessive cancellations due to the scarse bed capacity, but for the two following instances we see that far less patients are cancelled yielding many patients treated. For the high emergency loading and high target instances, the OR capacity is scarse, but we schedule relatively many flexible slots compared to the high target instances for the two other emergency loading levels, to avoid excessive elective cancellations. Scheduling many flexible slots will allow us to treat a large share of the green patients in the flexible slots, but having much flexible slot capacity, increases the chances of having idle slot capacity in weeks of less emergencies arriving. This is not a big problem for the low target instances, but for the high target instances the idle OR capacity is expensive in terms of the lost opportunity to perform elective surgeries.

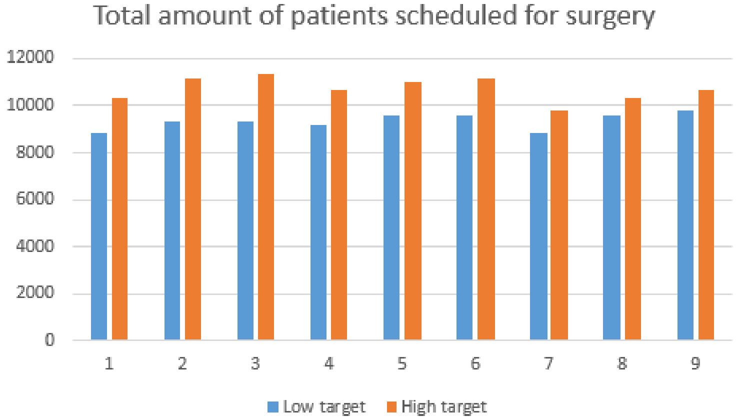


Figure 17: The total number of patients scheduled for surgery. The first three pairs of bars represent the low emergency loading level, the three following pairs of bars represent the normal emergency loading level, while the final three pairs of bars represent the high emergency loading level. Within each emergency loading level, the lowest number represents the lowest ward capacity and the highest number represents the highest ward capacity.

We see from Figure 18 that the number of green patients scheduled per flexible slot increases as the emergency loading increases, implying that each flexible slot is more valuable as the number of emergencies increases. We may also see that less green emergencies are scheduled for each flexible slot in the low target instances compared to the high target instances. This indicates that when the OR capacity is relatively good we may schedule excessive flexible slot capacity as this will not harm the scheduling of electives. The shares of green emergencies scheduled for flexible slots are seen in Figure 19. Since we may schedule excessive flexible slot capacity in the low target instances, we see that the shares of green patients treated in flexible slots are both stable and steady over 90 % for these instances. For the high target instances we see that the shares of green patients treated in flexible slots are far less as more OR capacity is dedicated for the electives. Note that when going from the instances NE-HT-W2 and HE-HT-W2 to the instances NE-HT-W3 and HE-HT-W3 we are able to treat a greater share of the green patients in the flexible slots as the bed capacity increases. This seems strange as increasing the bed capacity should yield more electives scheduled, as we may see examples of for all other high elective targets where we move towards more beds. The logic supporting that increasing the bed capacity should yield more electives scheduled and less flexible slots may be argued for in the following three ways. Firstly, the flexible slots may be utilized for either green outpatients or yellow patients if the bed capacity is scarse, while the green inpatients may be moved to elective slots where they may be treated without exceeding the bed capacity (this is done in instance LE-LT-W1, and will be discussed later in this subsection). Secondly, in periods of low emergency loading, the flexible slots will have idle capacity, which is positive if we isolate the effect on a ward that is fully loaded: We avoid excessive cancellation of elective patients that would have been scheduled for surgery if the flexible slot was scheduled as an elective slot. Finally, scheduling flexible slots will reduce the amount of electives scheduled, implying less demand for beds.

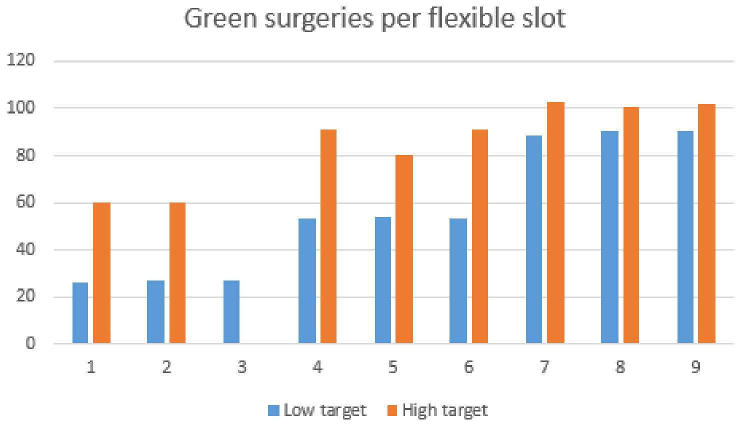


Figure 18: Total number of green patients scheduled per flexible slot. The first three pairs of bars represent the low emergency loading level, the three following pairs of bars represent the normal emergency loading level, while the final three pairs of bars represent the high emergency loading level. Within each emergency loading level, the lowest number represents the lowest ward capacity and the highest number represents the highest ward capacity.

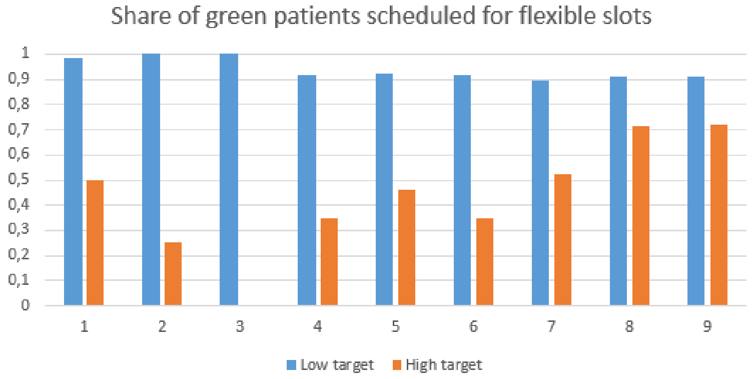


Figure 19: The share of green emergencies scheduled for flexible slots. The first three pairs of bars represent the low emergency loading level, the three following pairs of bars represent the normal emergency loading level, while the final three pairs of bars represent the high emergency loading level. Within each emergency loading level, the lowest number represents the lowest ward capacity and the highest number represents the highest ward capacity.

We now want to explain why less elective slot capacity is scheduled in the instances

NE-HT-W2 and HE-HT-W2 compared to the instances NE-HT-W3 and HE-HT-W3. In the instances NE-HT-W2 and HE-HT-W2, the bed capacity is very scarse resulting in many elective cancellations. Cancelling elective inpatients because of the scarse bed capacity will provide much spare elective OR capacity. This idle OR capacity may be used to schedule green outpatients or green inpatients (if they are less demanding in terms of beds than the electives that were cancelled) without having to make additional cancellations, decreasing the need for flexible slots and such allowing for scheduling many electives in the first stage. This mechanism represents the real life process where inpatients are cancelled, and outpatients and emergencies are prioritized, in periods of short bed capacity. However, choosing not to schedule flexible slots because we know there will be elective cancellations due to short bed capacity is not a good way of scheduling as this will result in excessive cancellations of electives.

The reason why the model allows for this mechanism to take place is that we have set the cost of cancelling elective patients very low. Next, we want to show that if we increase the penalty of cancelling electives from three to 10 for the instances NE-LT-W2 - NE-HT-W4, the solutions will behave according to the logic presented above. In addition we increase the maximum of beds available at the wards to 40 and 35, 50 and 50, and 65 and 65 for the wards W2, W3 and W4 respectively. The results from running the model for three hours may be found in Table 16. Here, we may see that the amount of flexible slots scheduled is steady or falling when the bed capacity increases.

Comparing the outcomes when increasing the penalty of cancelling electives to the outcomes presented in Table 15, we see that there are more flexible slots scheduled and less cancellations made when increasing the cost of cancelling patients. However, less elective patients receive surgery because the schedule made in the first stage is more risk averse when the cost of cancellations is high. The differences in the outcomes highlight one of the main issues faced by the management at a surgery department: Scheduling more capacity for electives and less for emergencies will result in more elective cancellations. However, having many flexible slots may result in excessive idle time at the ORs and lost opportunities to perform elective surgeries.

Table 16: Output from solving the stochastic problem when increasing the cost of cancelling electives. We include the number of flexible slots and the number of elective patients scheduled in the first stage. For the second stage we include the number of green patients scheduled for flexible slots, the number of elective cancellations, the number of electives treated and the number of beds moved for all 100 scenarios.

Instance	Flex. slots	El. scheduled	Green pat. in flex. slots	Cancel. of electives	Tot. el. treated	Beds moved	Obj. val. of best int. sol.	Best bound
NE-LT-W2	11	89/91	510/524	378	8522/9100	487	-278.29	-275.87
NE-LT-W3	9	90/91	481/524	35	8965/9100	36	-22.53	-15.87
NE-LT-W4	9	91/91	482/524	30	9070/9100	21	109.87	114.14
NE-HT-W2	7	99/115	424/524	418	9482/11500	426	-298.18	-289.21
NE-HT-W3	6	103/115	411/524	172	10128/11500	43	-38.82	-27.70
NE-HT-W4	3	109/115	254/524	404	10496/11500	27	107.45	114.32

For instance LE-LT-W1 we may note that four of the green patients are treated in elective ORs even though there are excess capacity in the flexible slots. Here, the green patients have been scheduled to elective ORs on other days to avoid exceeding the total bed capacity, as this yields a big penalty. In the real life, this situation will seldom happen. Most of the emergency inpatients are covering a bed prior to surgery, so performing surgery to these patients do not yield a higher bed consumption. As a matter of fact we would like to perform surgeries to emergency inpatients in such situations, as this means that we may send these patients home faster, releasing bed capacity for elective patients. We may avoid this situation in the optimization model if we do not exclude the postoperative bed loading of green patients in the scenarios, but rather give the green inpatients a length of stay equal to zero like we do for the yellow patients.

We see from Figure 20 that the number of patients resting at wards not meant for them (beds moved) increases much when the bed capacity decreases. This indicates that applying a shared bed capacity where the staff and equipment at all wards are more or less homogeneous should be considered if the bed capacity is scarse as this allows for a better utilization of the beds. If differentiated wards are applied, having some beds at each ward that are reserved as a shared capacity among the different wards will be valuable if the total bed capacity is scarse.

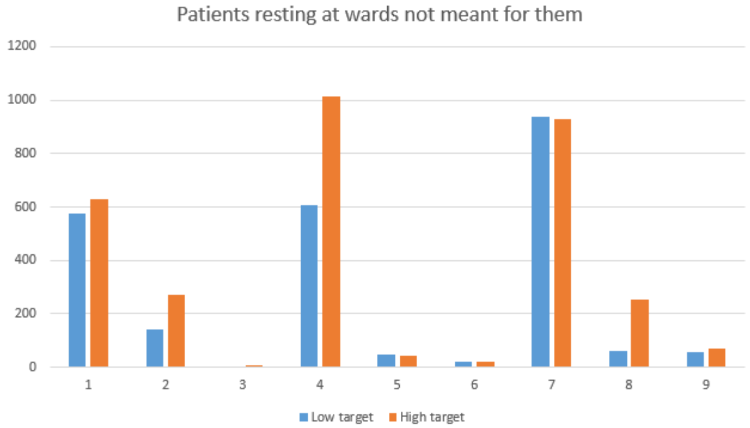


Figure 20: Total number of inpatients resting at wards not meant for them. The first three pairs of bars represent the low emergency loading level, the three following pairs of bars represent the normal emergency loading level, while the final three pairs of bars represent the high emergency loading level. Within each emergency loading level, the lowest number represents the lowest ward capacity and the highest number represents the highest ward capacity.

Based on the discussions above we propose the following statements regarding the scheduling of flexible slots to the elective ORs:

- The system has a given capacity of surgeries it may provide, and to reach this capacity we aim to have the right balance of electives scheduled and electives cancelled through the cycle

- Having excess capacity of flexible slots will yield less elective cancellations, but too few electives scheduled for surgery
- Having scarse capacity of flexible slots will yield many electives scheduled for surgery, but many elective cancellations
- If the OR capacity is regarded as better than the ward capacity, a relatively high share of the ORs should be made flexible
- If the ward capacity is regarded as better than the OR capacity, a relatively low share of the ORs should be made flexible
- Having a scarse bed capacity calls for the use of more homogeneous wards

8.5 Case study: The orthopaedic department at St. Olav's Hospital

In this subsection we aim to provide decision support for the management at the orthopaedic department. Firstly, we develop two MSSs (MSS 1 and MSS 2), representing two different scheduling regimes. In the first scheduling regime, we allow for two green inpatients to wait in line for the emergency ORs at AHL before we start buffering these patients to the flexible slots at BVS. All green outpatients are sent straight to the flexible slots. In the second scheduling regime, we aim to send all green emergencies to the flexible slots, leaving the emergency ORs exclusively for the red and yellow emergencies. An illustration of the generation procedure to develop the two MSSs may be seen in Figure 21. We start out with one simulation, representing the system as it is today. Then, we run the optimization model to generate an optimal MSS based on the scenarios provided by the first simulation. From the second simulation the generation procedures are separated, in order to implement the two scheduling regimes.

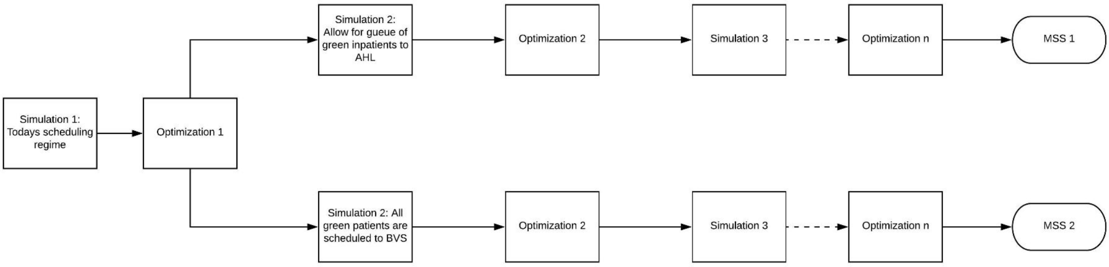

Simulation 1: Todays scheduling regime
Optimization 1
Simulation 2: Allow for gueue of green inpatients to AHL
Optimization 2
Simulation 3
Optimization n
MSS 1
Simulation 2: All green patients are scheduled to BVS
Optimization 2
Simulation 3
Optimization n
MSS 2

Figure 21: Procedure for generating MSS 1 and MSS 2

After developing and analyzing the two MSSs, we aim to produce one final MSS that may be more realistic to implement at the orthopaedic department.

8.5.1 Analyzing the outcomes of the optimization model for MSS 1 and MSS 2

When running the loop, the optimization model is run for three hours in each iteration. Two alterations are made to the optimization model exclusively for the case study. First, we have excluded the opportunity to distribute the bed capacity in the first stage, implying that we stick to the distribution of beds that are present at the department today. Furthermore, on request from the management at the department, we demand that the prosthesis subspecialty has access to as many surgery slots as it has today (14 slots). In the simulation model we do not allow for cancellations of the prosthesis patients, because these cancellations result in much DRG points lost. In order to generate the two MSSs, we run the loop for six iterations. In Tables 17 and 18, we report on the outcomes of the optimization model for each iteration.

Table 17: Outcomes form the optimization model when generating MSS 1

	It. 1	It. 2	It. 3	It. 4	It. 5	It. 6
Obj. func. val. of best int. sol.	-3157.83	-45.90	23.95	25.37	64.33	23.59
Upper bound	-3154.76	-43.78	25.67	27.23	66.07	25.31
Gap	0.11 %	4.62 %	6.68 %	6.84 %	2.62 %	6.80 %
Number of flexible slots	13	12	12	12	12	12
Electives scheduled each week	71/80	74/80	74/80	74/80	74/80	74/80
Green patients to flexible slots	1109/1171	673/687	615/625	584/590	660/667	702/704
Electives cancelled	679	103	114	100	113	117
Yellow pat. to receive surgery	526	871	946	993	852	856

Table 18: Outcomes form the optimization model when generating MSS 2

	It. 1	It. 2	It. 3	It. 4	It. 5	It. 6
Obj. func. val. of best int. sol.	-3157.83	38.89	-2.50	59.44	9.57	60.27
Upper bound	-3154.76	43.03	1.86	63.92	13.12	63.72
Gap	0.11 %	9.62 %	234.4 %	7.91 %	27.06 %	5.42 %
Number of flexible slots	13	12	12	12	12	12
Electives scheduled each week	71/80	74/80	74/80	74/80	74/80	74/80
Green patients to flexible slots	1109/1171	1168/1250	1260/1335	1204/1300	1209/1265	1219/1284
Electives cancelled	679	119	138	106	112	86
Yellow pat. to receive surgery	526	337	219	332	300	298

In the first simulation we obtain the scheduling regime present at the orthopaedic department today, resulting in a high number of emergencies resting at the wards. This is reflected in the scenarios fed to the optimization model, yielding many elective cancellations due to the scarse bed capacity. To handle the situation, the MSS provided in the first iteration has 13 flexible slots, one more than the rest of the MSSs produced, yielding less electives scheduled for surgery.

In the second scheduling regime we send all green patients to the ORs at BVS, resulting in twice the load of green emergencies to be scheduled in this scheduling regime compared to for the first. As 12 flexible slots are scheduled for both cases indicates that the OR-capacity is relatively good, and that the bed capacity is scarse, limiting the scheduling of elective inpatients. The number of yellow patients to receive surgery at the flexible slots are almost three times as high for the first scheduling regime compared to the second, indicating that the flexible slot capacity is excessive for most scenarios in the first scheduling regime.

The number of electives scheduled and cancelled from running the optimization model in each iteration may be seen in Tables 37 and 38 in Appendix B.1. In all iterations, only 2/4 of the back patients are scheduled for surgery, as these are the most bed demanding patient category. Furthermore, there are never scheduled more than 14/18 prosthesis patients, because the OR-capacity of 14 slots do not allow for more than 14 prosthesis patients to be scheduled.

Table 19: MSS 1

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

Table 20: MSS 2

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

The MSSs generated in the last iteration may be seen in Tables 19 and 20. Here, the green letters indicate the flexible slots. Note that the two schedules are very similar, and that the flexible slots are scheduled to the same subspecialties in both schedules. We also see that Friday is a popular day to schedule flexible slots. Since we only allow for elective outpatients to be scheduled for Friday, there is less competition for the OR capacity on this day. Furthermore, the flexible slots serve as an option in terms of the bed capacity: If the bed capacity is fully utilized on the day that the flexible slot is scheduled, or the next, we may schedule yellow patients (or green outpatients if there are any) for this slot as these patients have a length of stay of zero days. If however beds are available, we may utilize this by scheduling green inpatients for the flexible slot.

In Table 21 we provide the sum of elective patients that are scheduled for surgery each day of the week in the two MSSs. A more detailed overview may be found in Tables 40 and 41 provided in Appendix B.1. We see that Monday is the day with most elective patients scheduled for surgery, while only a few electives are scheduled for Friday. However, as may be seen from Figure 22, the number of beds covered by elective patients peaks on Thursday for both the schedules.

Table 21: Total amount of electives scheduled each day for MSS1 and MSS2

	Mon	Tue	Wed	Thu	Fri
MSS 1	24	14	14	17	5
MSS 2	24	17	14	16	3

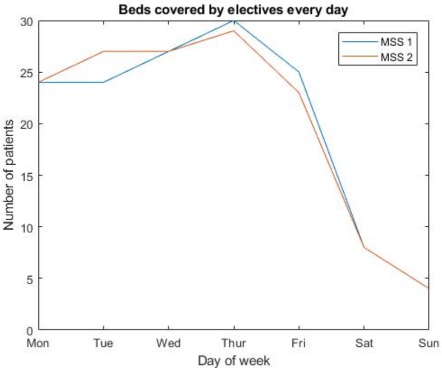


Figure 22: The expected number of beds covered by elective patients every day for MSS 1 and MSS 2

8.5.2 Evaluating MSS 1 and MSS 2

Outcomes form running the two MSSs for six months (following the warm-up period), 20 times in the simulation model may be found in Tables 43 and 44 in Appendix B.2. For MSS 1, no elective patients are cancelled in any of the 20 runs, and all green patients receive surgery within time in each run. Furthermore, 71 % of the yellow and 80 % of the red patients receive surgery within time. For MSS 2, there are two runs where a few elective cancellations are made. On average, 96 % of the green patients receive surgery within the dead line, and for the yellow and red patients there are 84 % and 81 % that receive surgery within time respectively. For the green patients, the expected waiting time to receive surgery is almost one day longer for the second MSS compared to the first one. However, for the yellow patients, the expected waiting time is more than five hours less for MSS 2. In the following, we compare the two MSSs on several areas of interest based on the outcomes from one run in the simulation model.

The impact on the emergency patients

In Figure 23, we illustrate the cumulative waiting time to receive surgery for the yellow and green emergencies for MSS 1 and MSS 2, and the outcomes are compared to the historical data. We do not include the red patients as the scheduling of these are independent of the MSS because they are always prioritized over the other emergencies. For both the MSSs we see that the outcomes regarding waiting time to surgery clearly outperform the historical data. For MSS 1, almost 60% of the green patients receive surgery the day following referral, and no green patients are delayed beyond five days. For MSS 2, almost 25 % of the green patients receive surgery the day after referral, and all green patients receive surgery within five days. For the yellow patients, we see that the waiting time is slightly less for the second MSS.

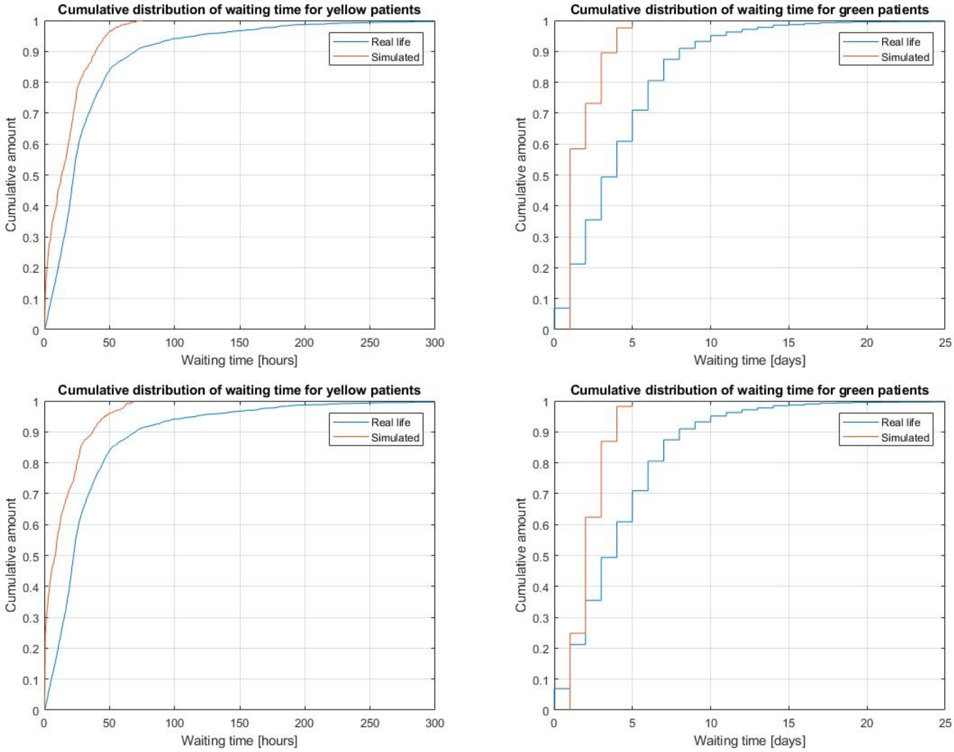


Figure 23: The cumulative distribution of waiting time to receive surgery for the yellow and green emergency patients, MSS 1 (top) and MSS 2 (bottom)

Figure 24 illustrates the mean number of emergencies waiting for surgery through the week for the two MSSs. As a general tendency we see that the queues of yellow and green emergencies reach a minimum on Friday afternoon, before increasing through the weekend when the OR capacity is low. For MSS 1, we see that the queue of green emergencies decreases steadily from Tuesday to Friday, and on average there are four green patients waiting for surgery when entering the weekend. For MSS 2, we see that the queue of green patients is more stable through the week as no green emergencies are allowed to go the the emergency ORs. Recall that there are only one more flexible slot scheduled for Friday in MSS 2 compared to for MSS 1. However, the average number of patients waiting for surgery falls much more on Friday in MSS 2, indicating that the flexible slots are better utilized on this day in

MSS 2. Because so many green emergencies receive surgery on Friday in MSS 2, there are four green patients waiting for surgery when entering the weekend also here. As the emergency ORs are reserved for the red and yellow patients in MSS 2, we see that the queue of yellow patients is less when entering the weekend in this MSS compared to the first.

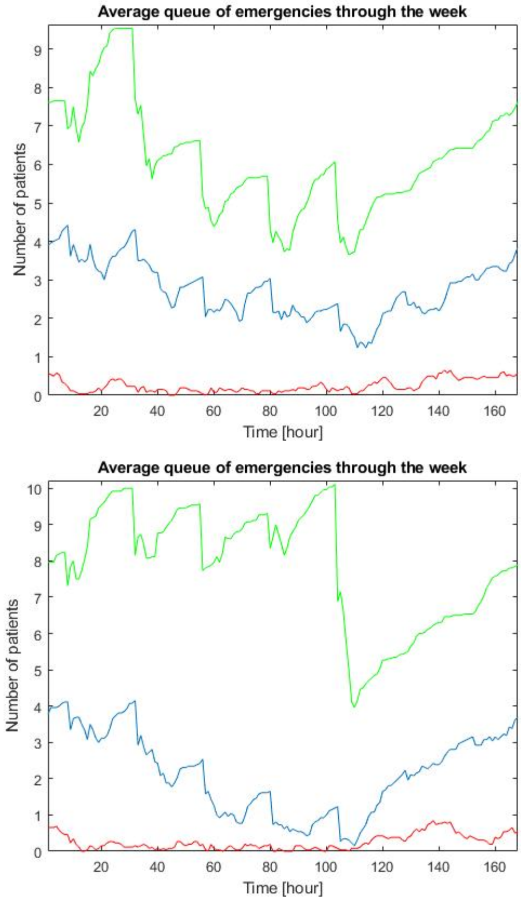


Figure 24: The mean number of emergency patients waiting for surgery through the week, MSS 1 (top) and MSS 2 (bottom)

In Figure 25 we illustrate the cumulative number of green and yellow patients waiting for surgery at 08.00 for the two MSSs, and compare the results to those from the historical data. Again, we see that the queue is less for both emergency categories in both MSS 1 and MSS 2 compared to the historical data. Note also that there is far less variation in the number of emergencies waiting for surgery, indicating that when scheduling flexible slots we are well prepared to handle a fluctuating emergency demand for surgery.

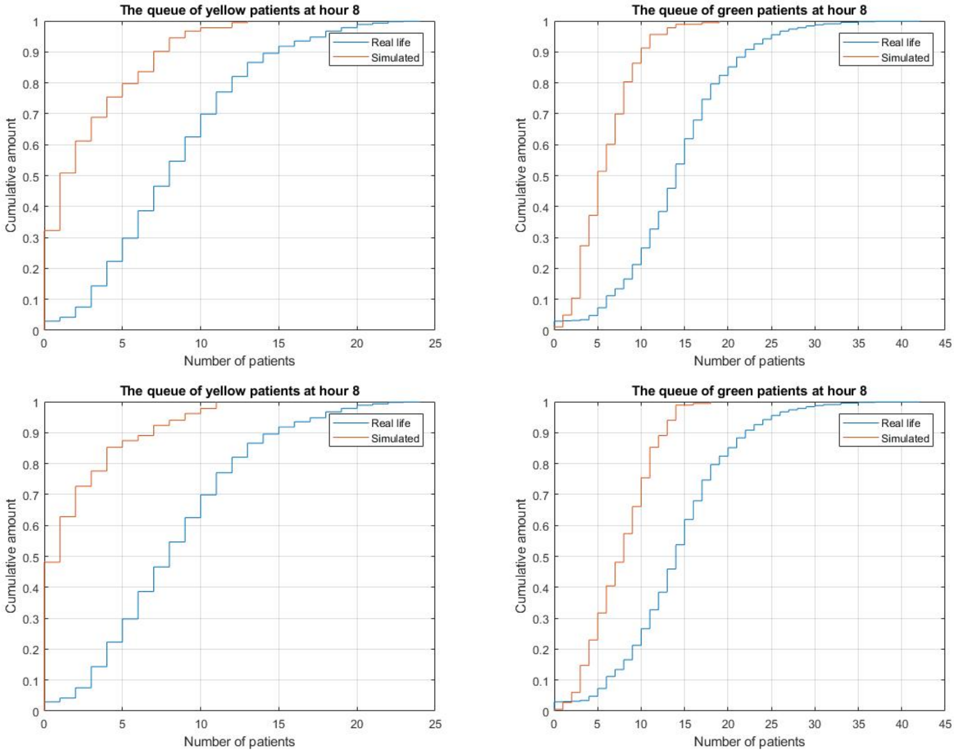


Figure 25: The cumulative distribution of the number of yellow and green emergencies waiting for surgery at 08.00, MSS 1 (top) and MSS 2 (bottom)

In Figure 26, we illustrate the number of green emergencies that receive surgery in flexible slots at BVS each week for both the MSSs. We see that on average, more green emergencies receive surgery in the flexible slots in MSS 2, which is not surprising given that all green emergencies are sent to BVS in this scheduling regime. Since the number of flexible slots are the same for both MSSs, the flexible slot utilization is less for MSS 1. To avoid idle OR capacity in times of low demand for green surgeries, strategies for utilizing idle flexible slots should be implemented. Having a low OR utilization is undesirable from an economical perspective. However, having a low utilization in some of the ORs is positive if we want to handle the queue of emergencies efficiently, as it provides buffer capacity to handle periods when many emergencies are entering.

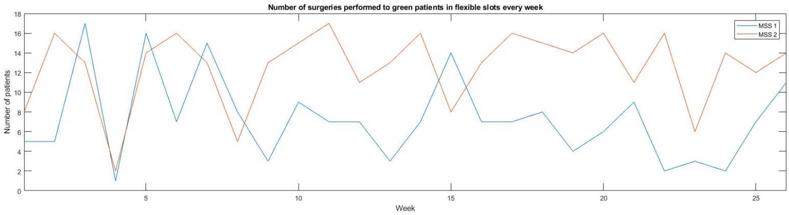


Figure 26: The weekly number of green patients to receive surgery in the ORs at BVS, MSS 1 and MSS 2

In Figure 27, the daily average number of green patients to receive surgery in each OR at BVS is illustrated for both MSSs. Recall that we load the ORs in increasing order, explaining why OR-1 is more loaded than the other ORs on the same day. We see that the average number of green patients to receive surgery in the flexible slots are quite similar for all days except from Friday, where MSS 2 have a much higher utilization of the available slot capacity. This was also noted in Figure 24, and the weak OR utilization on Friday for MSS 1 indicates that there is too much slot capacity dedicated for the green emergencies here. If we did allow for scheduling of green emergencies on the day of arrival, these slots would be better utilized in MSS 1, and the number of green emergencies waiting for surgery would be less towards the weekend.

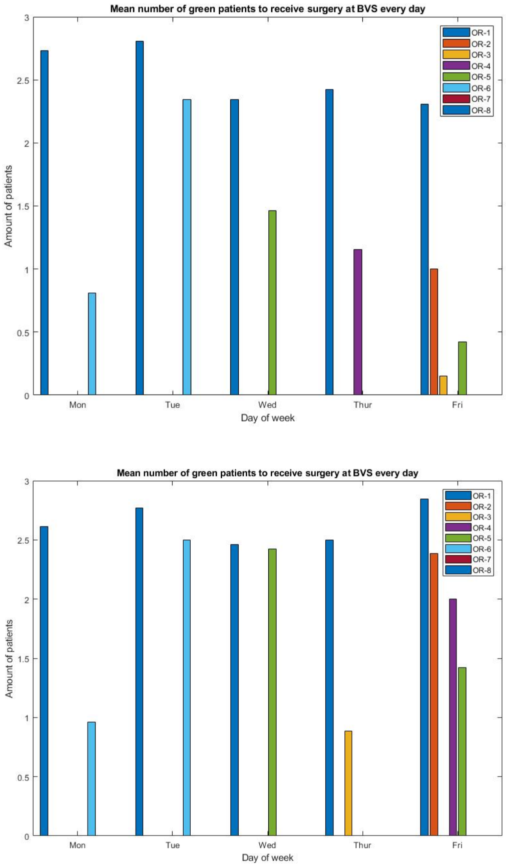


Figure 27: The average number of green surgeries performed in the ORs at BVS every week, MSS 1 (top) and MSS 2 (bottom).

OR utilization and ward loading

In Figures 28 and 29 the cumulative distribution of the working hours at the elective ORs are illustrated for the two MSSs, and we compare the results to the historical data. We see that scheduling based on the expected surgery duration yields much overtime at some of the ORs, even after removing the tails of the distributions for surgery duration. For OR-1 - OR-6 we see that the graphs are steeper for the real life distributions, meaning that the ORs close at approximately the same time on most of the days. As we draw independent realizations of the surgery duration, and use aggregated patient categories, we are not able to provide the same stability in the simulation model, resulting in long tails towards both overtime and under time. The vertical tails that may be seen for some of the simulated graphs are examples of OR capacity being unused, as a result of excessive flexible slot capacity. Applying another scheduling algorithm when scheduling green emergencies in the simulation model, for example scheduling the patients in a cyclic manner or scheduling patients to the flexible slot with most unused capacity, would yield less overtime work in

OR-1 and a better utilization of the other flexible slots.

The utilization of the emergency ORs at AHL for both MSS 1 and MSS 2 may be seen in Table 22. Since no green emergencies are sent to the emergency ORs in MSS 2, we have a lower utilization of these ORs in the second scheduling regime. The first scheduling regime results in a high emergency OR utilization, and a low utilization of the flexible slots, while the opposite is observed for the second scheduling regime. This gives rise to two different buffering mechanisms in periods of many green emergencies entering. In the first scheduling regime we will use the flexible ORs as a buffer capacity, while in the second scheduling regime we may use the emergency ORs to handle excessive demand for green surgeries. However, the buffering capacity at the emergency ORs are less predictable than the capacity offered in the flexible slots because the emergency slots are also used by the red and yellow emergencies.

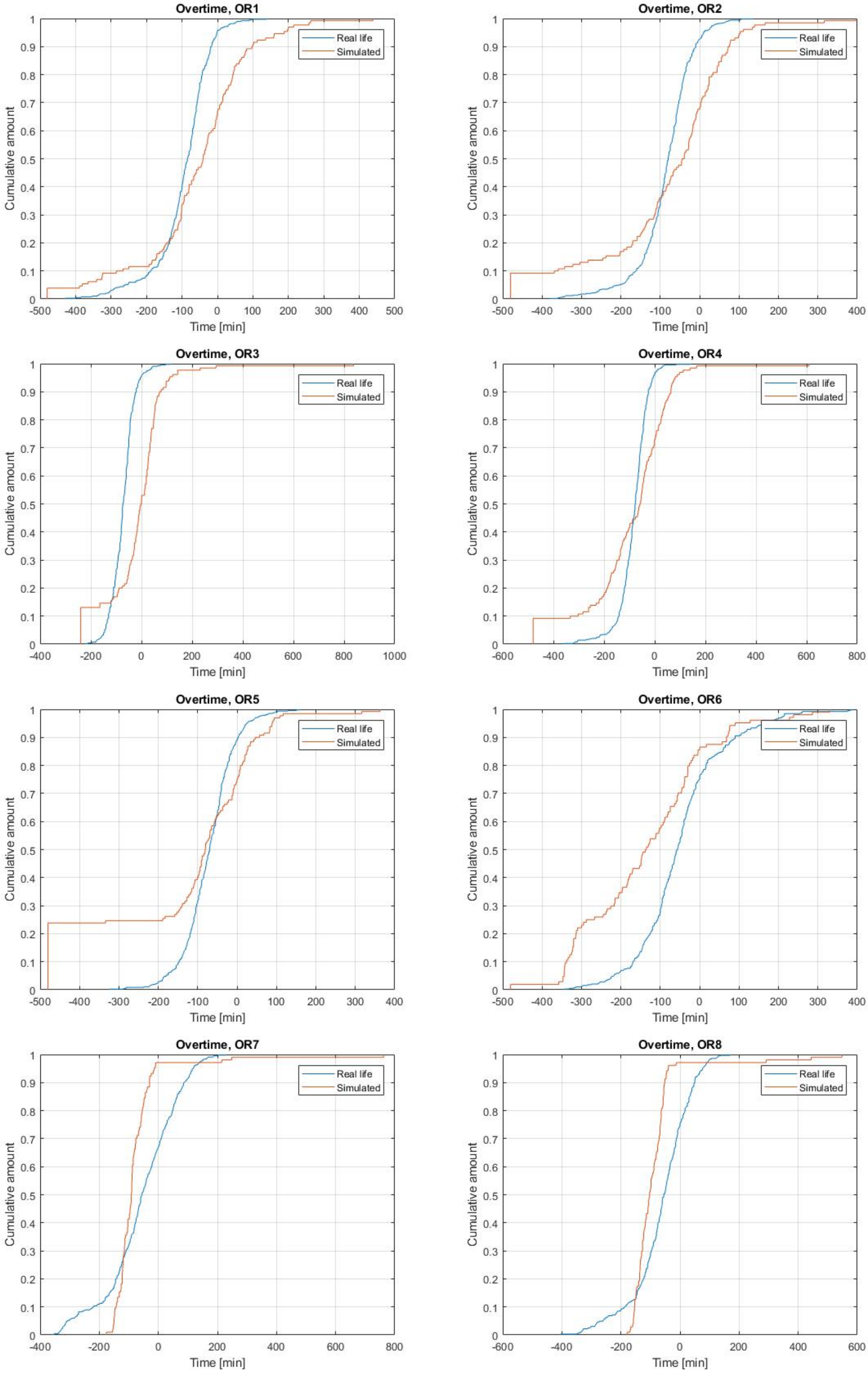


Figure 28: The cumulative distribution of the working hours at the elective ORs for MSS 1. Zero on the x-axis indicates 16.00

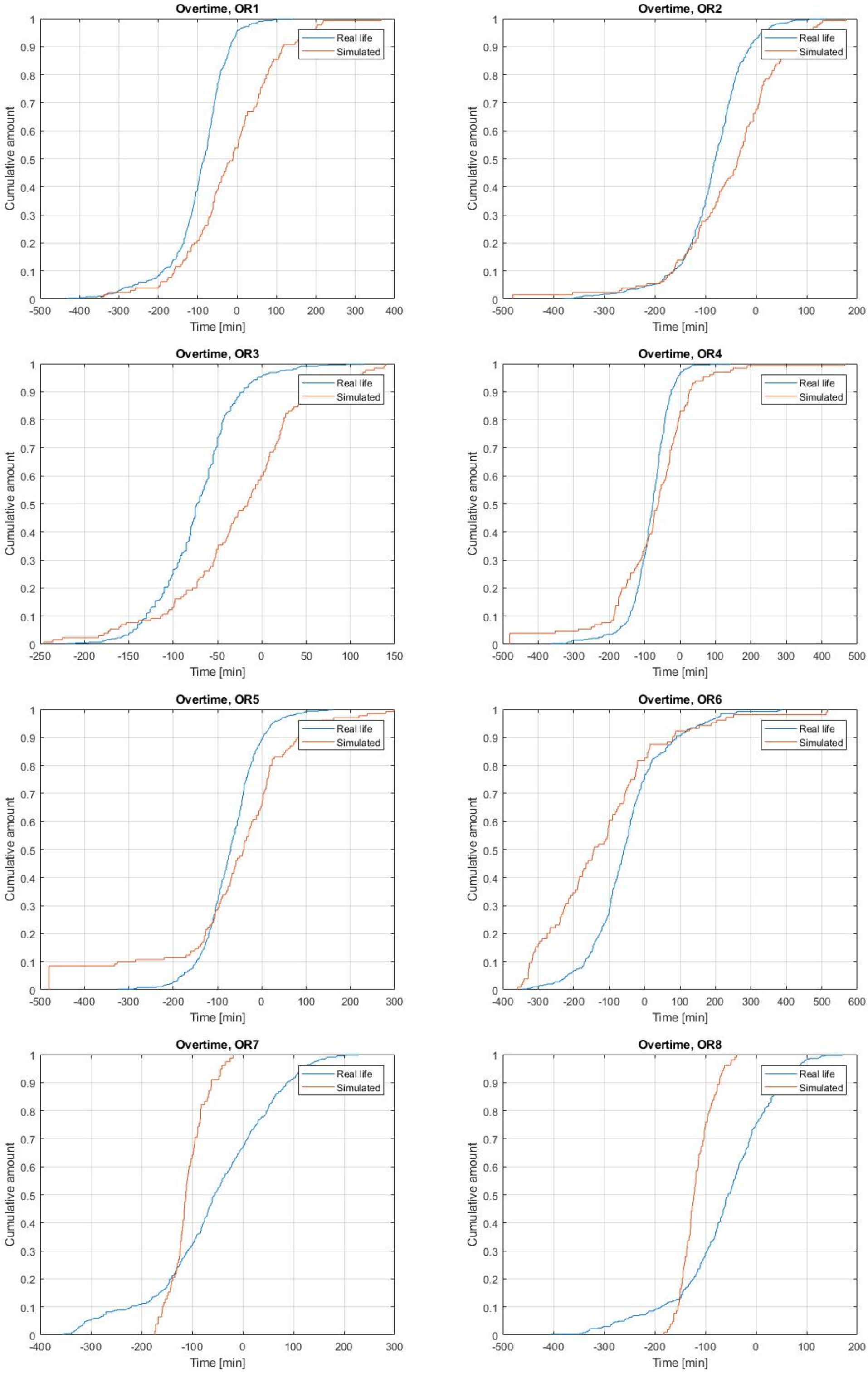


Figure 29: The cumulative distribution of the working hours at the elective ORs for MSS 2. Zero on the x-axis indicates 16.00

Table 22: The utilization of the emergency ORs at AHL for MSS 1 and MSS 2

	OR-9	OR-10
MSS 1	82 %	86 %
MSS 2	70 %	60 %

The cumulative bed loading at 12.00 for both the MSSs are illustrated in Figure 30, and the outcomes are compared to the historical data. We see that the average bed loading is higher for MSS 2 because less green inpatients are covering beds while waiting for surgery in MSS 1. Both schedules provide a bed loading that very seldom exceeds the total amount of scheduled beds available, and both outperform the historical data in terms of beds needed at the wards. Remember however that the simulation model sends more patients home during the weekend, which is not accounted for in these graphs.

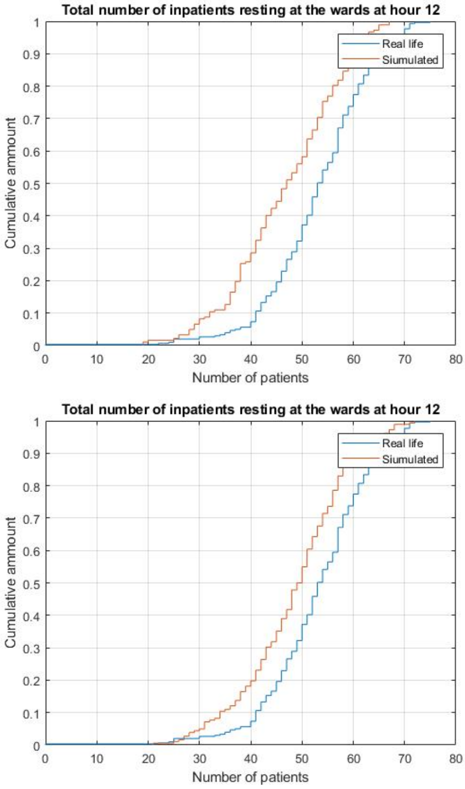


Figure 30: The cumulative distribution of the total ward loading at 12.00, MSS 1 (top) and MSS 2 (bottom)

8.5.3 A more realistic MSS

Because we schedule elective surgeries based on expected surgery duration, we seem to underestimate the OR capacity needed to handle the elective patients scheduled, resulting in much overtime work at the ORs. We now want to provide one MSS that may be more realistic in terms of the work load at the ORs. We therefore add 20 minutes of cleaning time to every surgery when scheduling the elective patients in the optimization model, and we restrict the number of flexible slots to six, providing more OR capacity to handle the electives. In addition, we require that no more than three flexible slots may be scheduled for one day, and we demand that at least two flexible slots should be scheduled for Monday and Tuesday as we know that the queue of emergencies is long following the weekend. Again, we do not allow for flexible slots at OR-7 and OR-8, and in addition we allow for the prosthesis subspecialty to access more OR slot capacity such that more prosthesis patients may possibly be scheduled for surgery.

When generating MSS 3, we use the same scenarios as we did for iteration 6 when generating MMS 1, and in the simulation model we apply the first scheduling regime. The results obtained from the optimization model may be found in Table 23. We see that when scheduling less flexible slots, we are able to schedule one more elective patient compared to the two previous MSSs. However, less green patients are scheduled to the flexible slots, resulting in more elective cancellations for this MSS. In addition, scheduling less flexible slots yields less flexible OR capacity to treat yellow emergencies. In Table 39 in Appendix B.1, we provide the number of electives scheduled through the week, and the total number of elective cancellations in the 100 scenarios.

Table 23: Outcomes form the optimization model when generating MSS 3

Objective function value of best integer solution	14.80
Upper bound	16.15
Gap	8.35%
Number of flexible slots	6
Electives scheduled each week	75/80
Green patients to flexible slots (for all 100 scenarios)	549/704
Electives cancelled (for all 100 scenarios)	220
Yellow patients to receive surgery (for all 100 scenarios)	117

The third MSS may be seen in Table 24. We see that the prosthesis subspecialty have gained access to two more surgery slots compared to the two previous MSSs, and we are now able to schedule 16/18 prosthesis patients. Furthermore, we may notice that the subspecialties elective foot, hand, plastic and arthroscopic have one more slot scheduled for elective patients, and one less as flexible. For all of these, except for the plastic subspecialty, this additional elective OR capacity is enough to schedule all the patients for these subspecialties. The back subspecialty is not responsible for any flexible slots in the third MSS, but they have access to one more elective slot compared to for the two previous MSSs, enabling us to schedule one more back patient in the third MSS.

Table 24: MSS 3

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

In Table 25, we provide the total number of elective patients that are scheduled for surgery every day. Again we may see that Monday is the day when most electives are scheduled for surgery, while there are least electives scheduled on Friday. A more detailed overview may be seen in Table 42 in Appendix B.1. In Figure 31 we may see that the expected elective bed loading is higher for MSS 3 compared to MSS 1 (and MSS 2), as more back patients and prosthesis patients are scheduled in MSS 3. However, the shape of the graph is similar to the one for MSS 1.

Table 25: The total amount of electives scheduled each day for MSS3

Mon	Tue	Wed	Thu	Fri
20	17	16	14	8

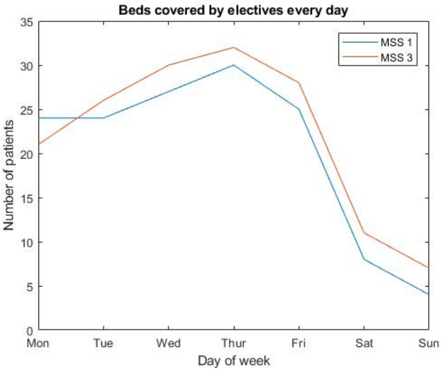


Figure 31: The expected number of beds covered by elective patients every day for MSS 1 and MSS 3

Evaluating MSS 3

The statistics form running MSS 3 for six months (following the warm-up period), 20 times in the simulation model may be found in Table 45 in Appendix B.1.

Compared to for MSS 1, the expected waiting time to receive surgery for the yellow and green patients is about three and five hours longer, respectively in MSS 3. Furthermore, six patients are cancelled on average in each run. In the following, we comment on areas of interest based on the outcomes from running MSS 3 once in the simulation model, and where it comes naturally we compare the outcomes to those obtained for MSS 1.

The cumulative waiting time to receive surgery for the green and yellow emergency patients for MSS 1 and MSS 3 may be seen in Figure 32. Not surprisingly, MSS 1 performs better than MSS 3 in terms of the waiting time for emergency patients. Still, 40 % of the green emergencies receive surgery the day after arrival in MSS 3, and very few green patients have to wait for more than five days.

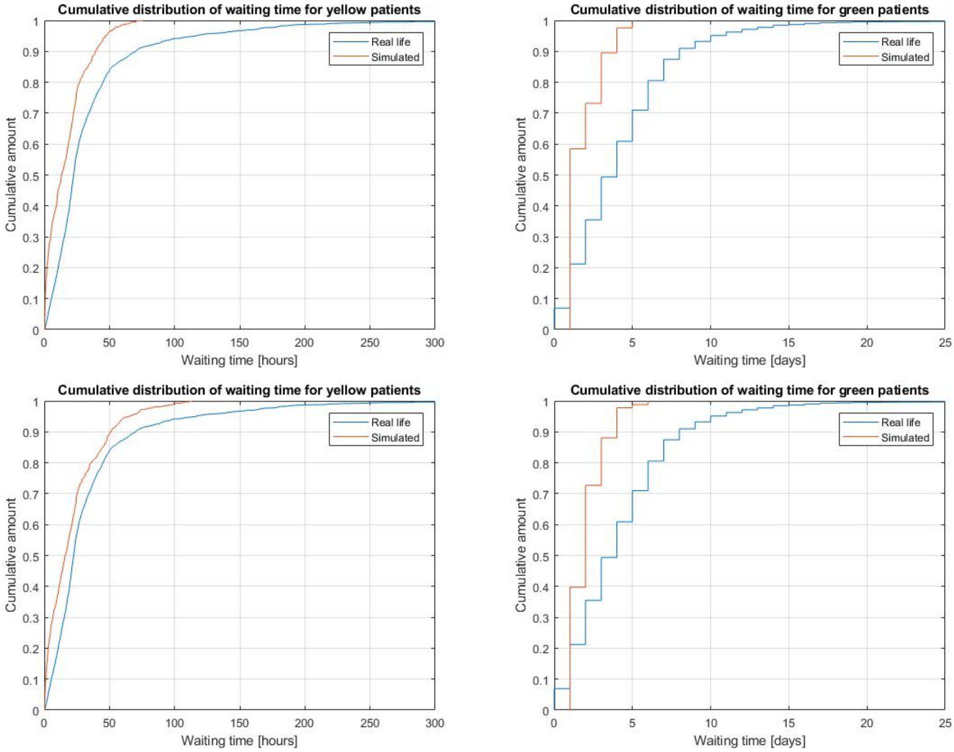


Figure 32: The cumulative waiting time to receive surgery for yellow and green emergencies, MSS 1 (top) and MSS 3 (bottom)

The queue of emergencies through the week for MSS 1 and MSS 3 may be seen in Figure 33. For MSS 3 we see that the length of the queue falls much on Thursday, indicating that the flexible slot capacity is well utilized also on the day when we have scheduled most flexible slots. Comparing to MSS 1 (and MSS 2), we see that the average number of yellow and green emergencies waiting for surgery is longer for MSS 3, and there are on average two more green emergencies waiting for surgery when we enter the weekend. In Figure 34 we plot the weekly number of green patients scheduled for BVS, and the queue of green patients at 08.00 for the same period. This plot is interesting as it illustrates how the flexible slots are used as a buffer capacity in periods of high demand for green surgeries. We see that the flexible slot capacity is often fully utilized, yielding a good OR utilization.

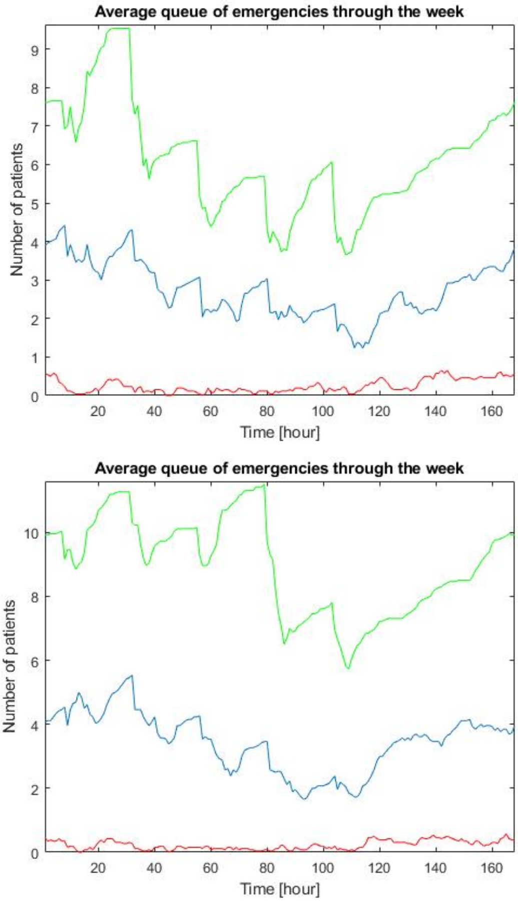


Figure 33: The mean number of emergency patients waiting for surgery through the week, MSS 1 (top) and MSS 3 (bottom)

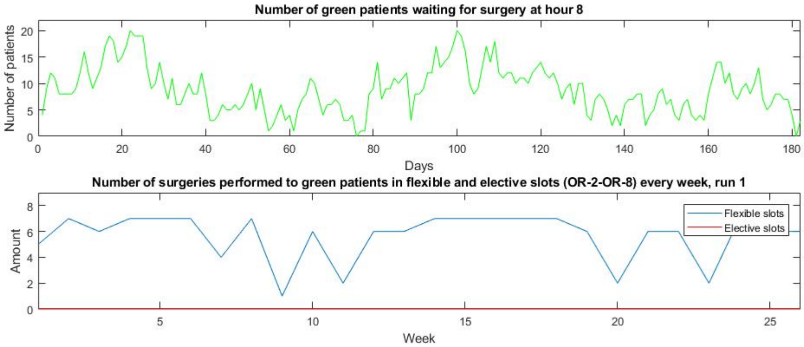


Figure 34: The number of green patients to receive surgery at flexible and elective slots at BVS for MSS 3. In the upper graph we include the queue of green emergencies for the same period.

In Figure 35 we illustrate the cumulative amount of working hours at the elective ORs and compare it to the historical data. We see that MSS 3 result in less overtime work compared to MSS 1 and MSS 2, and comparing with the historical data the simulated outcomes are quite similar for many of the ORs. In Figure 38 in Appendix B.2, we provide the cumulative amount of working hours at the elective ORs when removing only 5% of the tails for the distribution of surgery duration for the elective patients. Removing less of the tails result in some more overtime on average. For some of the ORs, like OR-3, the tail representing overtime work is longer for the case where more of the tails are removed. This indicates that some of the very long tails are caused by the scheduling of yellow and green emergencies to the ORs at BVS.

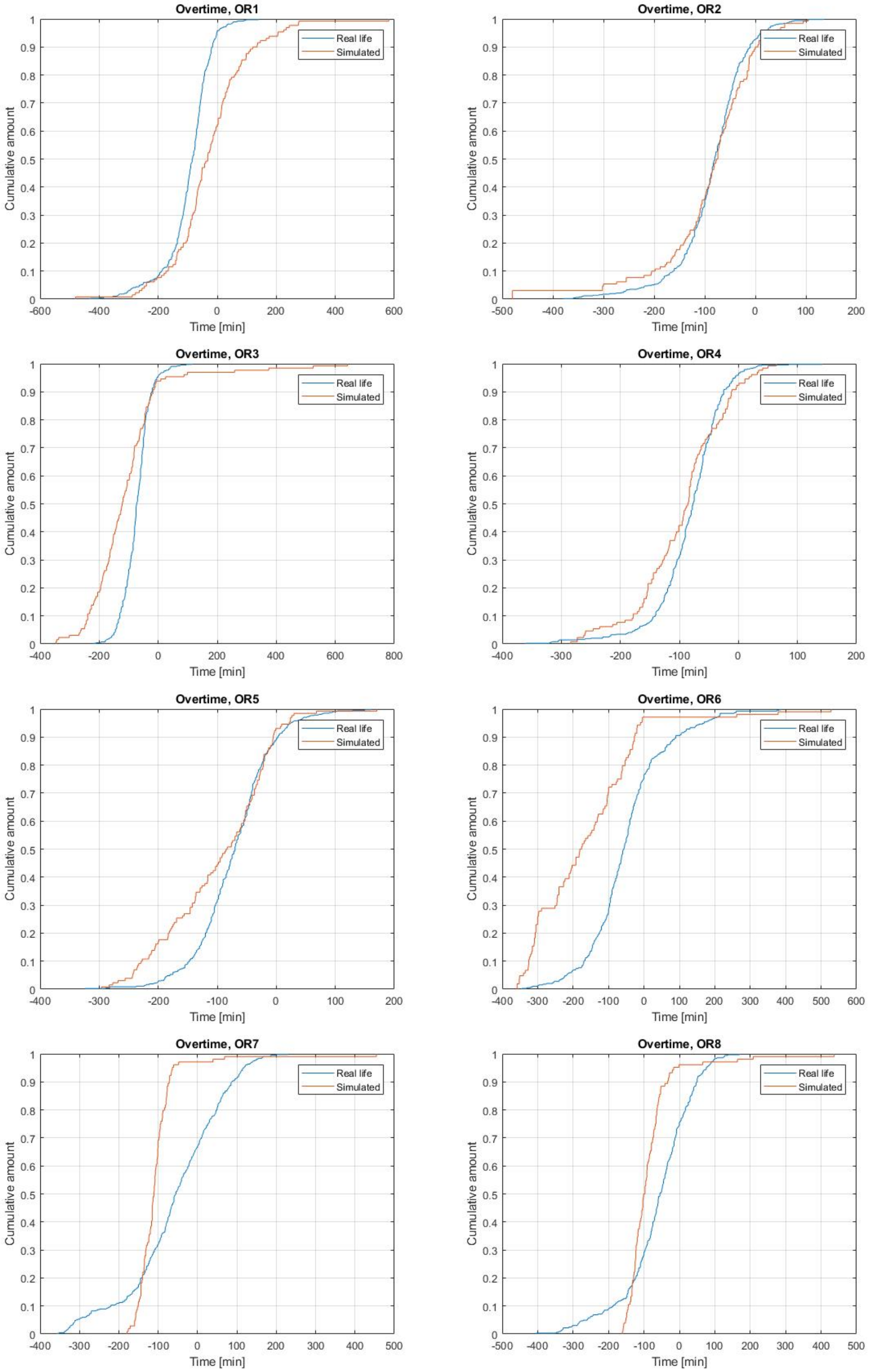


Figure 35: The cumulative distribution of the working hours at the elective ORs for MSS 3. Zero on the x-axis indicates 16.00

The cumulative ward loading at 12.00 for MSS 1 and MSS 3 may be seen in Figure 36. We see that the bed loading is higher for MSS 3 as more bed demanding electives are scheduled for surgery. We may also note that the tail is longer for MSS 3, indicating that there will be more times where rescheduling is necessary due to scarse bed capacity compared to for MSS 1.

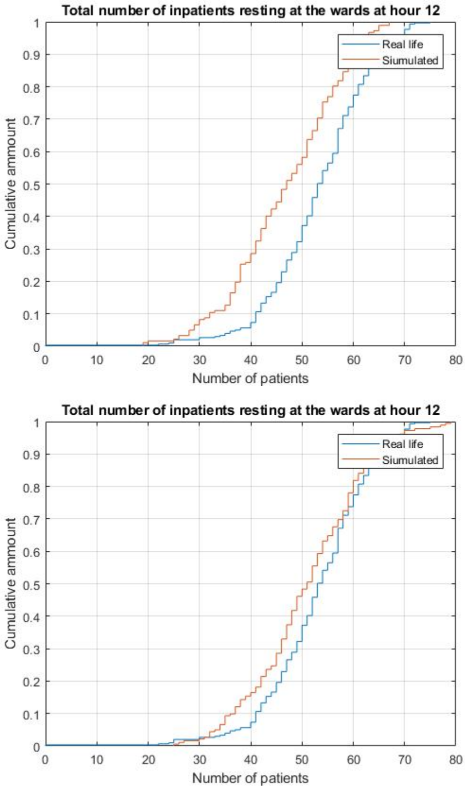


Figure 36: The cumulative distribution of the total ward loading at 12.00, MSS 1 (top) and MSS 3 (bottom)

8.5.4 The stability of the three MSSs compared to the present MSS

It is interesting to compare the outcomes from the 20 runs performed with the simulation model for each of the four MSSs (the three presented here, and the one presented when validating the simulation model). These runs are found in Tables 43, 44, 45 and 46 in Appendices B.1 and C. If the 20 runs provide outcomes that are similar to each other, we may say that the system is stable, and not dependent on the outcomes of the stochastic variables (the arrival of emergencies, the surgery duration of patients and the length of stay of inpatients). This is positive, as it indicates that the system is able to handle fluctuations. If the model is unstable, that is if the outcomes become dependent on the realizations of the stochastic variables, this may indicate that the system struggles to handle the work load and that small changes will result in very different outcomes. For each of the 20 runs we have computed the mean waiting time of emergencies to receive surgery. We now compute the mean and standard deviation obtained for theses 20 values, and plot the results for all four MSSs in Figure 37. From these plots we see that the outcomes obtained for the MSS present at the orthopaedic department varies more than for the others, especially when regarding the green emergencies. We know from the discussions above that MSS 1 is the schedule most prepared to handle the green emergencies, so it is not surprising that this model provides the most stable outcomes regarding the waiting time for green emergencies. The same goes for MSS 2 and the yellow patients.

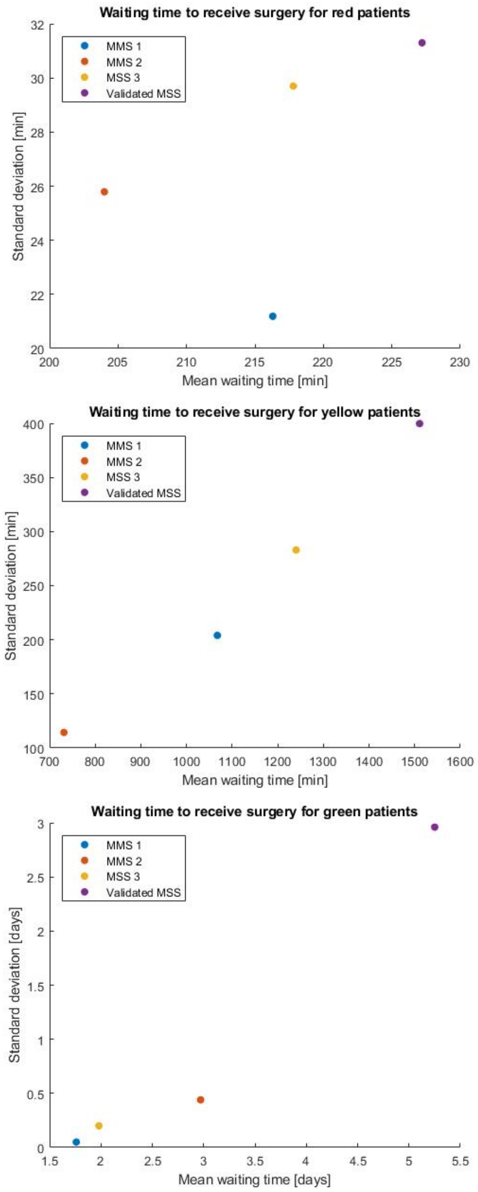


Figure 37: Comparing the waiting times for emergencies to receive surgery obtained from the 20 runs for MSS 1, MSS 2, MSS 3 and the MSS present at the orthopaedic department.

8.5.5 Summing up the case study

Scheduling flexible slots to the elective ORs yields less waiting time to receive surgery for the yellow and green emergencies. The number of green emergencies that may wait in line for the emergency ORs affects the waiting time for both the green and yellow emergencies, and there exist a trade-off concerning the waiting time for the two categories. Allowing to schedule some green emergencies to AHL results in less green emergencies waiting for surgery, and more yellow patients waiting compared to if no green emergencies are sent to AHL. Furthermore, sending some green emergencies to AHL will result in less flexible slots needed at BVS in order to handle fluctuations in the demand. Scheduling six flexible slots for BVS, while allowing for two green inpatients to wait in line at AHL results in about 40 % of the green inpatients receiving surgery the day following arrival, little elective rescheduling due scars OR capacity and the ability to provide surgery for the yellow emergency patients a bit faster than they do at the department today.

9 Concluding remarks and future research

The main problem faced by the orthopaedic department at St. Olav's Hospital today is treating the emergency patients within the dead lines proposed by the hospital. Having emergencies that wait too long for surgery is not only harming these patients, it also affects the flow of elective patients. In periods when the queue of emergency patients increases, the green emergencies are buffered to the elective ORs at BVS to leave the emergency OR capacity for the red and yellow emergency patients. As a consequence of this, elective patients may need to be rescheduled to provide OR capacity for the green emergencies. Having a large amount of emergencies waiting for surgery results in many beds being covered at the wards, and in periods of peak emergency loading, the bed capacity may become scarse. This may cause rescheduling of elective inpatients, as these inpatients need a bed following surgery.

Related to the orthopaedic department, the main contribution of this thesis is the development of a planning tool consisting of an optimization model and a simulation model. By applying these models we aim to provide a MSS where more OR capacity is devoted to green emergencies at the elective ORs, such that more emergencies receive surgery within the deadline, and the flow of elective patients is unharmed in periods of high emergency demand. The optimization model is formulated as a two-stage stochastic program with recourse. In the first stage we schedule the available OR slots as either elective or flexible, we schedule the elective patients for the elective slots, and we decide on the number of beds that should be available at the different wards on every day of the week. The uncertain parameters of the problem formulation, are the amount of emergencies resting at each ward on each day of the week, and the excess demand of green emergencies that require surgery in the same period. In the second stage we aim to perform the elective surgeries scheduled in the first stage, while scheduling all the excess demand of green emergencies without having to cancel the elective surgeries.

Applying a two-stage formulation means that we aggregate the events that provide new information about the stochastic parameters to one point in time (just before every cycle). This simplification, together with the fact that we develop a cyclic plan to handle real life fluctuations over time calls for some way to investigate how well the MSS proposed by the optimization model behaves in a real life environment. To do this, we develop a simulation model to mimic a surgery department, such as the orthopaedic department. Different scheduling regimes may be implemented in the simulation model, providing the opportunity to explore trade-offs between different scheduling strategies.

By applying the models on instances representing the orthopaedic department, we propose a MSS with six flexible slots scheduled for the elective ORs each week. This is enough to handle almost all green emergencies within the deadline, and it makes the department much better prepared to handle periods of excessive emergency demand. As a result, far less electives need to be rescheduled. Scheduling flexible slots means transferring OR capacity from the electives to the emergencies, and there exist a trade-off between the number of electives scheduled each week, and the number of electives that need to be rescheduled due to short capacity of beds or ORs.

From running the optimization model on some smaller instances, we develop general advises regarding the scheduling of flexible slots to elective ORs. As a rule of thumb we suggest that if the OR capacity is more scarse than the bed capacity, only a few flexible slots should be scheduled. Scheduling elective slots may yield weeks of low OR utilization, especially if no strategy for the use of these lots in weeks of low emergency demand are decided on. This low OR utilization is relatively more expensive if the OR capacity is scarse. If however the bed capacity is regarded as more scarse than the OR capacity, we suggest increasing the flexible slot capacity to provide more beds for the elective patients at the wards.

From an academic point of view, the main contribution of this report is the two-stage formulation applied to solve the MSSP. To our knowledge, this modeling framework has not yet been applied to the MSSP by other authors, and we are able to show practical value of applying a stochastic formulation compared to the deterministic counterpart. If we set the cost of cancelling electives low, we are able to perform more elective surgeries when applying the stochastic formulation. If however we set the cost of cancellations high, we are able to provide schedules that cancel far less electives compared to the deterministic counterpart.

Developing more efficient formulations for the MSSP is a topic for further research. The model struggles to close the LP-gap within hours, also for small instances, and the problems increase when treating the ORs as homogeneous. Developing symmetry breaking inequalities and clever ways of searching in the B & B tree should be considered to decrease the running time. If we are able to construct a more efficient formulation, we may also be able to impose more complexity to the model without increasing the running time too much. Applying more than two stages may provide additional value. One alternative formulation can take the queue of green patients on Monday morning as an input parameter (possibly stochastic), and then apply five stages, one for each weekday. In each stage we receive information about the number of green patients entering and the number of red and yellow emergencies resting at the wards. The objective function in this alternative formulation should include minimizing the number of green patients waiting for surgery on Friday afternoon.

The simulation model may be further developed to provide more realistic outcomes. Applying smaller time intervals when generating the arrival rates of emergencies will yield more accurate arrival rates through the day. In addition, adding some randomness to the prioritization rules for emergencies may provide a better fit to the real life process where the rules are often overseen. Furthermore, the scheduling of emergencies to flexible slots are made very simple, yielding an uneven loading of the flexible slots. Applying more sophisticated algorithms for the scheduling procedure may result in less overtime work and less idle time at the ORs. However, the biggest weakness of the simulation model is that we assume the wards to have infinite bed capacity. Implying rules for elective rescheduling based on the ward loading should be considered if the model is to be further developed.

In addition to developing more efficient and realistic models, we should focus more towards implementation of the models at hospital departments. As mentioned in Section 4, very few of the models developed on surgery scheduling are implemented at a hospital department. Developing guidelines for successful implementation is of great societal interest as it may play a central role when facing the challenges within health care in the years to come.

A The input data applied in the models

Here, we present the two databases used to collect data for the models, and we provide tables with the values applied for the input data both for the optimization and the simulation model.

A.1 The databases used

We have gained access to two databases in order to provide input data to our models. The first, OPPLAN, stores all relevant data related to the surgery of patients, the arrival of emergencies and the categorization of patients. The other one, Nimes, provide historical data related to the length of stay of the inpatients at the wards. As both the databases include a procedure code for the patients, it is possible to link the data on patient category level. Both databases apply identification codes for the patients registered. However, the code system applied in the two databases are not similar, so we are not able to link the databases on a patient level. We do not have access to the id-number of the patients, providing anonymous data. The historical data used for this thesis cover the period from 01.01.2015 - 27.04.2017.

There are some issues regarding the data in Nimes. We are not able to tell how many days of the total length of stay that were preoperative days, and how many were postoperative. For the elective patients this is not an issue, as most of these patients enter the hospital just before surgery. For the emergency patients this is important to know, as most of these are waiting at the wards before surgery. To estimate the preoperative days at the wards for the emergencies, we use the expected waiting time for surgery. Another issue is that the green emergency patients are not registered as an individual category in Nimes, making it hard to estimate the length of stay of these patients.

A.2 The values obtained for the input parameters

In Tables 26 to 33, we present the values applied for the input parameters in the large instances, and in Tables 34 and 35 the values applied for the small instances are provided. Note that the five patient categories included in the small instances have all the same properties as the five first patient categories in the large instances. In Table 36, the input parameters relevant for the simulation model are given.

Table 26: Values obtained for the subspecialties and the patient categories in the optimization model. T_i is the target number of elective patient category i to be scheduled each week, S_i is the expected surgery duration of patient category i , while E_{id} is the expected length of stay of patient category i that receive surgery on day d.

Subspecialty	T_i	S_i	E_{id}
Elective foot			
Aggregated group	4	143	3
Hand			
Aggregated group	8	94	0
Carpal tunnel syndrome	3	85	1
Plastic			
Aggregated group	15	95	2
Plateepitelkarsinom	2	73	1
BCC	5	142	1
Malingt melanom	4	68	0
Cancer mammae	4	97	1
Arthroscopic			
Aggregated group	6	123	2
ACL	2	186	2
Meniscus	3	173	0
Back			
Aggregated group	4	295	6
Prostheses			
Hip	7	177	4
Knee	11	174	4
Tumor			
Aggregated group	2	76	1
Emergencies			
Green inpatients	-	192	2
Green outpatients	-	131	0
Yellow patients	-	165	0

Table 27: Values of the parameters related to the ORs in the optimization model. M_{kd}^{OR} represents the number of surgery slots available at OR k on day d.

OR	M_{kd}^{OR}	Slot time [min]	Patient category
1	2	240	Elective foot, plastics
2	2	240	Plastics
3	2	240	Plastics, hand, arthroscopic
4	2	240	Arthroscopic
5	2	240	Back, tumor
6	2	240	Prosthesis
7	2	240	Prosthesis

Table 28: The OR slots available to each subspecialty each day and through the week in the optimization model

Subspecialty	N_{jd}^D	N_j
Elective foot	4	5
Hand	4	7
Plastic	4	14
Arthroscopic	4	12
Back	4	6
Prosthesis	4	14
Tumor	2	2

Table 29: Maximum number of ORs that may be covered by anesthesiologists each day in the optimization model

M_1^A	M_2^A	M_3^A	M_4^A	M_5^A	M_6^A	M_7^A
7	7	7	7	4	0	0

Table 30: The ward capacities obtained in the optimization model. A_{wd} represents the number of staffed beds available at ward w on day d, while A_w^{MAX} is the maximum number of beds available at ward w.

Ward	Name	c. A_{wd} (week)	A_{wd} (weekend)	A_w^{MAX}	Patient category hosted
1	Trauma	20	16	32	Elective foot, Hand and green inpatients
2	Reconstructive	16	16	16	Plastic, Tumor
3	Elective	10	12	12	Arthroscopic, Back
4	Fast-track	16	0	16	Prosthesis
5	Hotel-day	5	0	5	None, buffer capacity

Table 31: The objective function parameters related to the elective patients in the optimization model. P_i is the reward of scheduling more patients from the elective patient category i than the lower limit, while C_i^C is the penalty of cancelling an elective patient belonging to patient category i.

Elective patient category	P_i	C_i^C
Elective foot (aggregated)	3	3
Hand (aggregated)	2	3
Carpal tunnel syndrome	3	3
Plastic (aggregated)	3	3
Plateepitelkarsinom	3	3
BCC (aggregated)	3	3
Malingt melanom	2	3
Cancer mammae	3	3
Arthroscopic (aggregated)	3	3
ACL	3	3
Meniscus	2	3
Back (aggregated)	3	3
His prosthesis	3	6
Knee prosthesis	3	6
Tumor (aggregated)	3	3

Table 32: The values obtained the objective function parameters not relevant for the elective patients in the optimization model. $C_{ww'}^W$ is the penalty of putting a patient meant for ward w in ward w', C^{GR} is the penalty of scheduling a green patient to an elective slot, P^Y is the reward for scheduling a yellow patient to a flexible slot, and C^β is the penalty of having more patients resting at the wards than the total amount of staffed beds available.

$C_{ww'}^W$	C^{GR}	P^Y	C^β
1	2	0.5	1000

Table 33: The values obtained for the other parameters in the optimization model. $1/R_i$ is the share of patients belonging to the elective patient category i that need to be scheduled for surgery, M^{CYCLE} is the total amount of slots available through the cycle, and H_5 is the number of elective inpatients that may be scheduled for Friday.

R_i	M^{CYCLE}	H_5
2	70	0

Table 34: Values obtained for the target number of elective patients in the small instances

Patient category	Low target	High target	Subspecialty
1	8	12	1
2	32	38	2
3	18	25	2
4	21	25	3
5	12	15	3

Table 35: The four different ward capacities available for the small instances

Ward	Number of scheduled beds available							Max beds available	Config.
	Mon	Tue	Wed	Thur	Fri	Sat	Sun		
1	25	25	25	25	25	20	20	32	W1
2	20	20	20	20	20	20	20	25	W1
1	30	30	30	30	30	25	25	32	W2
2	25	25	25	25	25	25	25	25	W2
1	45	45	45	45	45	25	25	45	W3
2	40	40	40	40	40	25	25	40	W3
1	60	60	60	60	60	30	30	60	W4
2	50	50	50	50	50	25	25	50	W4

Table 36: Values obtained for the parameters used in the simulation model when trying to mimic the real life, and when developing the three MSSs in Section 8.5

Parameter	Mimic real life	MSS 1	MSS 2	MSS 3
Prob. of accessing an OR outside the opening hours (for red pat.)	0.55	0.55	0.55	0.55
Max gr. inpatients in queue for em. ORs	3	2	0	2
Max gr. outpatients in queue for em. ORs	1	0	0	0
Min time (days) ahead that we may not cancel electives	2	2	2	2
Min time (days) for a gr. pat. to wait before cancelling en el. pat.	8	8	8	8
Min time (days) to wait to schedule an el. pat. that was cancelled	10	10	10	10
Min time (min.) needed to initiate a new surgery at the em. ORs (not applicable for red pat.)	70	70	70	70
Max number of el. pat. that may be cancelled on each scheduling day to provide capacity for gr. inpatients	2	2	2	2
Max number of el. pat. that may be cancelled on each scheduling day to provide capacity for gr. outpatients	1	1	1	1
Prob. that a slot is available for a yellow pat. if excess capacity	0.25	0.25	0.25	0.25
Time (min.) used to clean the OR	20	20	20	20
Mean waiting time (min.) for the next em. pat. to be ready for surgery following previous surgery	30	30	30	30
Std. dev. of waiting time (min.) for the next em. pat. to be ready for surgery following the previous surgery	10	10	10	10

B The case study

Here we provide additional Tables and Figures relevant to the case study.

B.1 Outcomes from the optimization model

The number of electives scheduled and cancelled in each iteration when generating the two MSSs may be seen in Tables 37 and 38. Note that the cancellations are the total cancellations from all 100 scenarios in each iteration. The model only schedules 12 prosthesis patients in the first iteration compared to 14 for the other iterations, even though 14 slots are scheduled for the prosthesis subspecialty in all iterations. Scheduling more prosthesis patients in the first iteration will yield excessive cancellations of these patients, due to the limited bed capacity. Furthermore, one less plastic (aggregated) patient is scheduled for surgery in the first iteration, as two of the flexible slots are scheduled to the plastic subspecialty in this iteration, leaving restricted OR capacity for the plastic patients.

Table 37: Scheduling and cancelling of elective patients in the optimization model when generating MSS 1. T_i is the target number of electives to be scheduled from patient category i.

		Scheduled						Cancelled					
Patient category	T_i	Iterations						Iterations					
		1	2	3	4	5	6	1	2	3	4	5	6
El. foot (aggregated)	4	4	4	4	4	4	4	133	25	25	23	28	22
Hand (aggregated)	8	8	8	8	8	8	8	0	0	0	0	0	0
Carpal tunnel syndrome	3	3	3	3	3	3	3	5	0	0	0	0	0
Plastic (aggregated)	15	14	15	15	15	15	15	168	10	12	13	22	11
Plateepitelkarsinom	2	2	2	2	2	2	2	4	0	0	0	0	0
BCC	5	5	5	5	5	5	5	17	0	0	0	0	0
Malingt melanom	4	4	4	4	4	4	4	0	0	0	0	0	0
Cancer mammae	4	4	4	4	4	4	4	16	0	0	0	0	0
Arthroscopics (aggregated)	6	6	6	6	6	6	6	56	3	8	3	2	6
ACL	2	2	2	2	2	2	2	25	0	4	1	4	1
Meniscus	3	3	3	3	3	3	3	0	0	0	0	0	0
Back (aggregated)	4	2	2	2	2	2	2	102	35	39	31	37	42
Hip	7	6	4	4	7	4	4	101	1	11	13	0	12
Knee	11	6	10	10	7	10	10	46	29	15	16	20	23
Tumor (aggregated)	2	2	2	2	2	2	2	6	0	0	0	0	0

Table 38: Scheduling and cancelling of elective patients in the optimization model when generating MSS 2. T_i is the target number of electives to be scheduled from patient category i.

		Scheduled						Cancelled					
Patient category	T_i	Iterations						Iterations					
		1	2	3	4	5	6	1	2	3	4	5	6
El. foot (aggregated)	4	4	4	4	4	4	4	133	23	35	21	26	20
Hand (aggregated)	8	8	8	8	8	8	8	0	1	0	1	0	0
Carpal tunnel syndrome	3	3	3	3	3	3	3	5	0	0	0	2	0
Plastic (aggregated)	15	14	15	15	15	15	15	168	17	21	9	17	5
Plateepitelkarsinom	2	2	2	2	2	2	2	4	0	0	1	0	3
BCC	5	5	5	5	5	5	5	17	1	0	4	1	1
Malingt melanom	4	4	4	4	4	4	4	0	0	0	0	0	0
Cancer mammae	4	4	4	4	4	4	4	16	0	0	0	0	0
Arthroscopics (aggregated)	6	6	6	6	6	6	6	56	5	4	5	5	5
ACL	2	2	2	2	2	2	2	25	2	3	2	4	3
Meniscus	3	3	3	3	3	3	3	0	2	0	1	1	1
Back (aggregated)	4	2	2	2	2	2	2	102	44	51	50	36	38
Hip	7	6	4	4	4	4	4	101	8	6	5	15	2
Knee	11	6	10	10	10	10	10	46	16	18	7	5	8
Tumor (aggregated)	2	2	2	2	2	2	2	6	0	0	0	0	0

The number of patients scheduled and cancelled for MSS 3 may be seen Table 39.

Table 39: Scheduling and cancelling of elective patients in the optimization model, MSS 3. T_i is the target number of electives to be scheduled from patient category i.

Patient category	T_i	Scheduled	Cancelled
El. foot (aggregated)	4	4	26
Hand (aggregated)	8	8	0
Carpal tunnel syndrome	3	3	0
Plastic (aggregated)	15	15	15
Plateepitelkarsinom	2	2	0
BCC	5	4	0
Malingt melanom	4	3	0
Cancer mammae	4	4	0
Arthroscopics (aggregated)	6	6	9
ACL	2	2	5
Meniscus	3	3	3
Back (aggregated)	4	3	98
Hip	7	6	28
Knee	11	10	36
Tumor (aggregated)	2	2	0

In Tables 40, 41 and 42 we provide the number of patients from each patient category that is scheduled for each day of the week in MSS 1, MSS 2 and MSS 3 respectively.

Table 40: Scheduling of elective patients to the different days when generating MSS 1

Patient category	Mon	Tue	Wed	Thu	Fri
El. foot (aggregated)	0	1	3	0	0
Hand (aggregated)	2	1	0	0	5
Carpal tunnel syndrome	3	0	0	0	0
Plastic (aggregated)	5	3	0	7	0
Plateepitelkarsinom	2	0	0	0	0
BCC	1	1	3	0	0
Malingt melanom	0	0	0	4	0
Cancer mammae	2	0	2	0	0
Arthroscopics (aggregated)	2	2	0	2	0
ACL	1	0	0	1	0
Meniscus	0	2	1	0	0
Back (aggregated)	0	0	1	1	0
Hip	0	2	2	0	0
Knee	4	2	2	2	0
Tumor (aggregated)	2	0	0	0	0
Sum	24	14	14	17	5

Table 41: Scheduling of elective patients to the different days when generating MSS 2

Patient category	Mon	Tue	Wed	Thu	Fri
El. foot (aggregated)	0	1	3	0	0
Hand (aggregated)	2	2	2	2	0
Carpal tunnel syndrome	3	0	0	0	0
Plastic (aggregated)	6	3	0	6	0
Plateepitelkarsinom	2	0	0	0	0
BCC	0	2	2	1	0
Malingt melanom	0	1	0	0	3
Cancer mammae	2	0	2	0	0
Arthroscopics (aggregated)	2	2	0	2	0
ACL	1	1	0	0	0
Meniscus	0	1	0	2	0
Back (aggregated)	0	0	1	1	0
Hip	0	2	2	0	0
Knee	4	2	2	2	0
Tumor (aggregated)	2	0	0	0	0
Sum	24	17	14	16	3

Table 42: Scheduling of elective patients to the different days when generating MSS 3

Patient category	Mon	Tue	Wed	Thu	Fri
El. foot (aggregated)	0	0	4	0	0
Hand (aggregated)	0	0	2	0	6
Carpal tunnel syndrome	2	0	1	0	0
Plastic (aggregated)	3	4	0	8	0
Plateepitelkarsinom	0	0	2	0	0
BCC	1	2	1	0	0
Malingt melanom	1	2	0	0	0
Cancer mammae	3	0	1	0	0
Arthroscopics (aggregated)	3	2	0	1	0
ACL	1	1	0	0	0
Meniscus	0	1	0	0	2
Back (aggregated)	0	1	1	1	0
Hip	0	2	2	2	0
Knee	4	2	2	2	0
Tumor (aggregated)	2	0	0	0	0
Sum	20	17	16	14	8

B.2 Outcomes from the simulation model

The outcomes form running the three MSSs for six months (following the warm-up period), 20 times in the simulation model may be found in Tables 43, 44 and 45. Note that the waiting time for surgery is given in minutes for the red and yellow emergencies, and in days for the green patients.

In Figure 38 we illustrate the cumulative amount of working hours at the ORs for MSS 3 when removing only 5% of the upper and lower tail for the empirical distribution of surgery duration for each elective patient categories.

Table 43: Statistics form running MSS 1 for six months 20 times in the simulation model

Run	Cancelled	Number of emergencies				Mean waiting time			Std waiting time			Within time [%]		
		Red	Yellow	Green (in)	Green (out)	Red	Yellow	Green	Red	Yellow	Green	Red	Yellow	Green
1	0	280	831	275	425	255	1729	1.76	287	1261	1.01	74	45	100
2	0	276	731	290	393	223	1030	1.84	235	1013	1.06	79	75	100
3	0	307	702	270	382	238	1088	1.78	277	984	1.12	79	68	100
4	0	295	681	264	430	197	1185	1.76	204	1045	1.05	83	65	100
5	0	288	752	252	401	179	1158	1.71	182	1236	1.02	84	71	100
6	0	277	732	262	412	225	1185	1.66	265	1159	1.01	79	71	100
7	0	283	730	242	458	242	1085	1.79	279	966	1.06	78	70	100
8	0	282	716	273	390	212	1030	1.80	245	974	1.00	80	73	100
9	0	280	720	274	396	236	990	1.78	272	976	1.04	79	74	100
10	0	292	715	279	393	193	1009	1.76	196	879	1.05	83	73	100
11	0	287	679	274	393	221	974	1.69	225	1149	1.01	80	77	100
12	0	287	712	283	379	209	959	1.82	223	917	1.08	79	72	100
13	0	288	744	268	373	219	1128	1.69	239	1003	1.04	81	67	100
14	0	307	727	296	445	176	1005	1.79	191	943	1.09	84	74	100
15	0	311	742	232	399	233	931	1.75	242	839	1.04	77	77	100
16	0	296	763	269	383	225	1364	1.81	236	1217	1.10	80	59	100
17	0	289	731	284	414	200	807	1.81	206	755	1.04	81	83	100
18	0	308	686	286	413	207	955	1.79	216	849	0.98	81	76	100
19	0	298	721	287	419	200	944	1.82	210	992	1.02	80	76	100
20	0	285	728	251	416	236	795	1.67	288	729	0.99	79	82	100
Mean	0	291	727	271	406	216	1068	1.76	236	994	1.05	80	71	100

Table 44: Statistics form running MSS 2 for six months 20 times in the simulation model

Run	Cancelled	Number of emergencies				Mean waiting time			Std waiting time			Within time [%]		
		Red	Yellow	Green (in)	Green (out)	Red	Yellow	Green	Red	Yellow	Green	Red	Yellow	Green
1	0	279	724	282	402	188	691	2.60	217	874	1.16	85	85	100
2	0	290	722	286	364	174	718	2.36	227	887	1.20	87	85	100
3	0	271	701	278	392	203	546	3.32	234	611	1.87	82	92	86
4	0	304	707	259	406	225	851	2.13	247	985	1.00	76	77	100
5	0	293	797	275	396	214	892	2.19	217	915	0.91	80	75	100
6	0	308	729	275	398	220	630	2.44	246	728	1.11	78	87	100
7	0	299	757	274	394	201	694	2.36	210	753	1.01	82	87	100
8	0	289	727	283	396	174	859	2.66	200	926	1.16	85	79	100
9	0	263	708	282	369	194	729	2.33	228	802	1.07	83	85	100
10	0	316	758	282	393	290	930	2.69	417	1063	1.42	75	77	98
11	0	271	831	248	414	196	810	2.25	204	801	1.06	82	79	100
12	0	291	723	302	394	228	775	2.99	289	910	1.47	76	81	94
13	0	297	696	282	410	183	624	3.09	201	819	1.49	80	89	95
14	0	285	722	279	423	209	635	2.46	234	745	1.14	78	86	100
15	0	281	718	284	413	178	527	2.82	206	611	1.37	85	93	98
16	0	301	740	238	432	210	841	2.57	266	888	1.40	81	77	97
17	7	299	754	263	433	203	766	3.34	281	803	1.96	85	83	86
18	5	291	731	266	475	186	615	3.46	195	678	1.93	81	89	83
19	0	296	739	276	416	210	689	2.34	245	785	1.10	80	84	100
20	0	273	754	258	449	194	805	3.52	235	861	1.76	81	81	83
Mean	0.6	290	737	274	408	204	731	2.70	240	821	1.33	81	84	96

Table 45: Statistics form running MSS 3 for six months 20 times in the simulation model

Run	Cancelled	Number of emergencies				Mean waiting time			Std waiting time			Within time [%]		
		Red	Yellow	Green (in)	Green (out)	Red	Yellow	Green	Red	Yellow	Green	Red	Yellow	Green
1	0	311	718	244	428	210	905	1.84	245	846	1.05	80	77	100
2	0	301	724	263	408	208	1199	1.92	207	1042	1.07	81	65	100
3	5	260	745	262	396	216	1162	2.02	237	1294	1.15	78	70	100
4	0	282	776	222	418	289	1400	1.87	393	1156	1.13	75	55	100
5	49	303	792	296	423	225	1886	2.59	238	1492	1.58	79	44	94
6	41	284	753	291	403	176	1085	1.96	171	956	1.01	86	72	100
7	0	292	715	281	403	233	1176	1.86	311	1166	1.00	97	69	100
8	2	301	724	255	431	210	1136	1.97	228	1169	1.10	82	73	100
9	4	301	782	285	402	237	1686	2.20	266	1138	1.26	81	42	99
10	0	305	787	250	419	217	1201	1.90	284	1214	1.09	82	64	100
11	0	291	751	263	405	291	1322	2.01	376	1212	1.18	75	62	99
12	21	295	767	268	384	196	1328	2.41	182	1017	1.66	83	61	97
13	0	275	781	260	412	219	1293	1.80	216	1114	0.99	79	62	100
14	0	298	714	234	388	206	1015	1.74	201	893	0.96	81	71	100
15	7	286	761	276	402	193	1046	1.88	200	886	1.09	83	71	99
16	0	297	755	282	418	205	1148	1.92	205	1176	1.05	84	68	100
17	0	263	717	257	411	182	777	1.85	179	683	1.00	84	82	100
18	0	288	729	268	370	218	1176	1.91	239	1206	1.07	82	66	100
19	0	272	729	254	386	190	1036	1.91	201	907	1.06	82	71	100
20	0	313	731	273	395	234	1827	2.04	235	1277	1.20	79	40	99
Mean	6	291	743	264	405	218	1240	1.98	236	1092	1.14	81	64	99

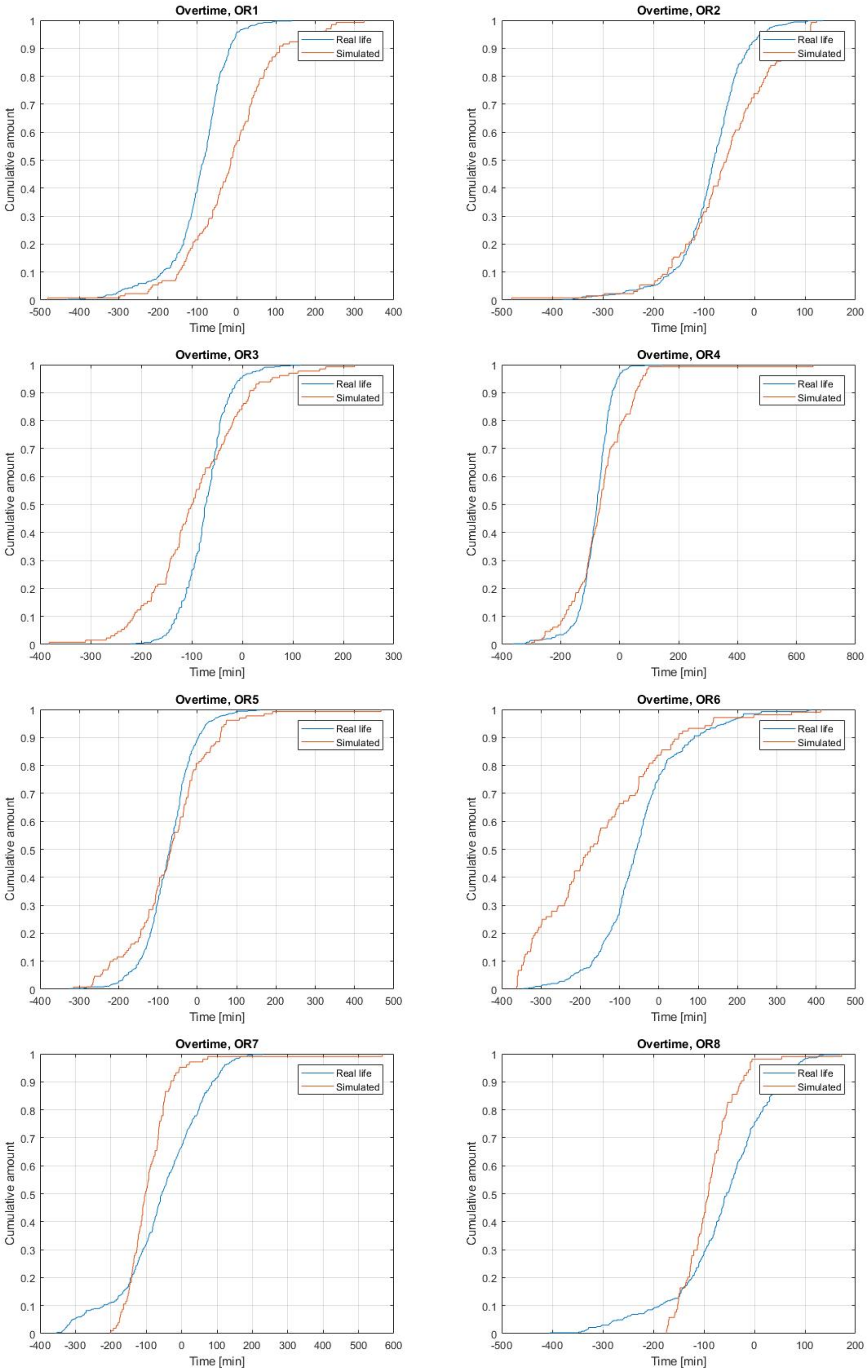


Figure 38: MSS 3: The cumulative amount of working hours at the elective ORs when removing only 5% of the upper and lower tail for the empirical distribution of surgery duration for each elective patient categories. We compare the outcomes to the historical data.

C Validating the simulation model

In order to use the simulation model as a testing tool, we need to make sure that the outcomes from the model is able to mimic the historical data in a satisfying way. To achieve this, the model is run several times while adjusting the input parameters till the outcomes are similar to those obtained from the historical data. In the following, we only report on the validated outcomes of the simulation model. In Figure 39 we provide an illustration of the flow of patients through the system. The flow is described in Section 6.2.2.

To represent a normal high production week, we simulate the MSS used at the department today, and we schedule 78 elective patients for surgery. The patients have been manually distributed, making sure that the expected surgery duration scheduled to an OR do not exceed the opening hours of the OR. Furthermore, only patients that have a high probability of being outpatients are scheduled to Friday.

The outcomes form running the MSS for six months (following the warm-up period), 20 times in the simulation model may be found in Table 46. Note that the 20 runs provide quite different outcomes for the waiting time and the number of emergencies treated within time. The big variations from one run to the next indicates that the system is unstable, and that the outcomes depend much on the emergency arrival patterns realized in each run. Note that the scheduling rules applied in the model are never changed. In the real life, these rules are fluctuating, and adapt to the present situation. However, the large variation underlines the fact that the present MSS is not well prepared to handle a fluctuating emergency demand, and that there will be many periods where the scheduling rules need to be altered in order to handle the queue of emergencies. There are, on average, 206 cancellations of elective patients due to short OR capacity, implying that more than one elective patient is cancelled each day on average.

In Table 47 we provide the average number of emergency patients arriving at the orthopaedic department for a period of six months, and the average share of emergencies to receive surgery within time. Regarding the average arrival rates of emergencies, the model is quite accurate for both the red and yellow patients, and the green inpatients. However, it overestimates the arrivals of green outpatients. When it comes to the amount of emergencies treated within time, we see that the model is able to provide surgery to more red patients within time compared to the real world. For the two other categories the outcomes are, on average, quite close to the real world.

The results presented next are representing one year (following a warm up period of 50 weeks) of running in the simulation model, and the outcomes are compared to the historical data covering the period from 01.01.2015-27.04.2017. When comparing one run from the simulation model to the historical data we may not expect to have the exact same results for such a big and complex system. However, we are able to get a feeling of how close we are to model the real world system. Note that the simulation model represents the reality as it is at the orthopaedic department today. This means having one OR at AHL running to 22.00 four days every week (instead of to 16.00). Regarding the historical data, this situation was only present

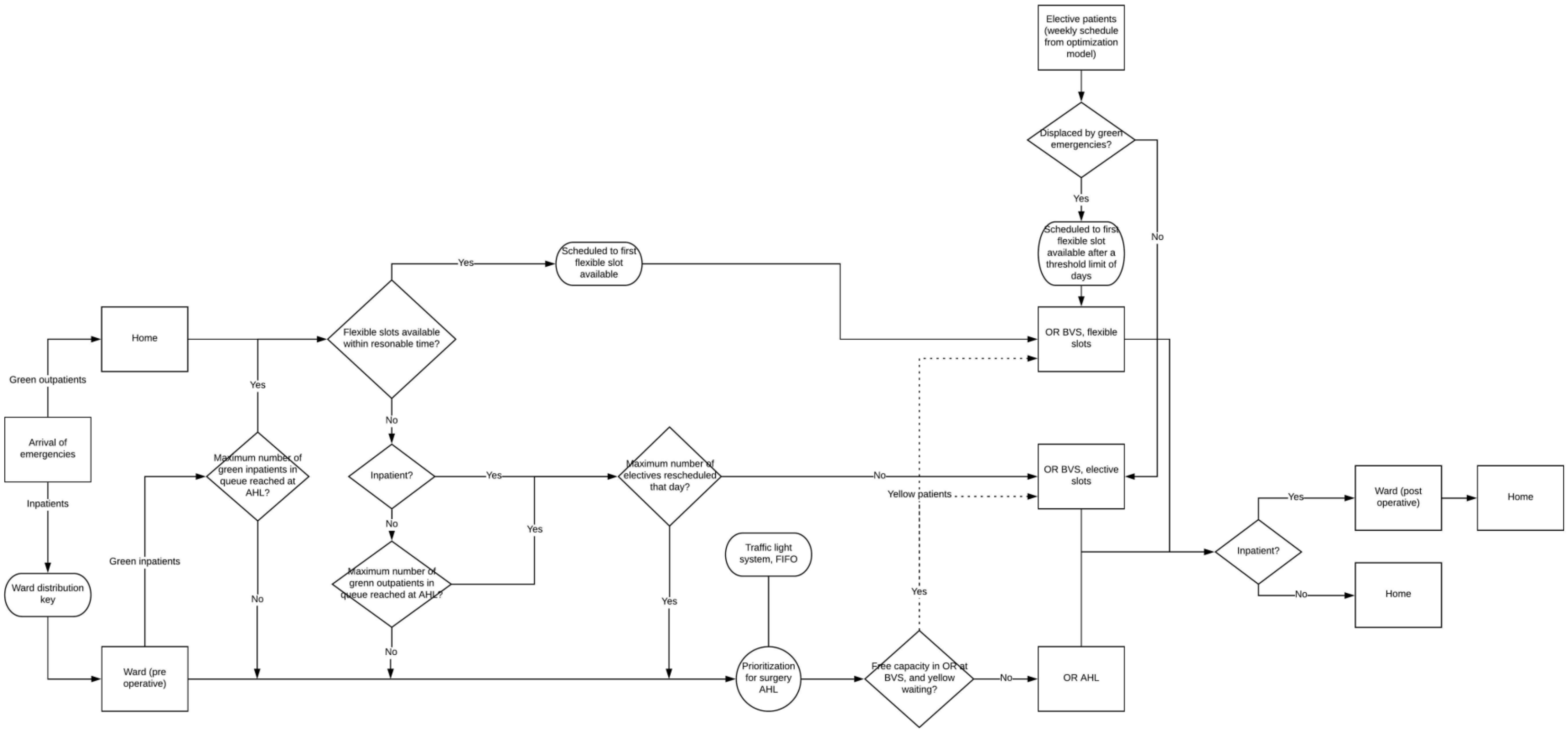

Elective patients (weekly schedule from optimization model)
Displaced by green emergencies?
Yes
No
Scheduled to first flexible slot available after a threshold limit of days
OR BVS, flexible slots
OR BVS, elective slots
OR AHL
Inpatient?
Yes
No
Ward (post operative)
Home
Home
Arrival of emergencies
Green outpatients
Inpatients
Home
Ward distribution key
Ward (pre operative)
Green inpatients
Maximum number of green inpatients in queue reached at AHL?
Yes
No
Flexible slots available within resonable time?
Yes
No
Scheduled to first flexible slot available
Inpatient?
Yes
No
Maximum number of grenn outpatients in queue reached at AHL?
Yes
No
Maximum number of electives rescheduled that day?
No
Yes
Yellow patients
Traffic light system, FIFO
Prioritization for surgery AHL
Free capacity in OR at BVS, and yellow waiting?
Yes
No

Figure 39: The flow of patients in the simulation model

Table 46: Statistics form running the MSS present at the orthopaedic department today for six months 20 times in the simulation model

Run	Cancelled	Number of emergencies				Mean waiting time			Std waiting time			Within time [%]		
		Red	Yellow	Green (in)	Green (out)	Red	Yellow	Green	Red	Yellow	Green	Red	Yellow	Green
1	202	287	752	269	411	211	1204	4.38	198	1022	3.16	78	68	75
2	144	264	752	238	388	227	1640	3.46	231	1303	2.28	78	51	89
3	269	295	762	268	427	305	1820	6.43	331	1559	4.73	69	49	59
4	204	274	732	278	394	228	1102	3.79	239	978	2.65	77	68	83
5	159	284	731	271	375	213	1111	3.39	251	937	1.73	82	71	89
6	253	301	730	286	383	210	1440	6.33	214	1125	4.41	83	55	56
7	210	293	762	238	438	203	2172	6.20	216	2114	5.02	81	50	61
8	92	311	750	253	371	202	1149	2.95	203	975	1.74	80	67	91
9	188	275	746	281	399	192	1242	4.29	193	1140	2.81	85	64	70
10	271	292	724	308	400	217	1148	4.72	224	906	3.66	82	67	72
11	161	303	703	251	424	211	1317	3.07	237	1250	1.53	80	65	93
12	227	293	667	272	430	245	1661	4.13	274	1218	2.21	79	49	75
13	165	281	704	272	395	200	1152	3.65	201	1097	2.40	83	69	86
14	170	264	729	273	376	190	1777	5.26	211	1732	5.21	84	58	77
15	220	305	692	291	422	204	1493	3.95	205	1332	2.13	82	60	80
16	271	271	708	266	438	245	1551	4.67	326	1365	2.30	82	57	66
17	178	325	707	275	404	263	1469	4.50	271	1506	3.81	72	62	77
18	163	278	706	252	399	234	1147	3.28	244	1086	2.16	78	62	87
19	248	326	795	291	378	259	2306	11.82	272	1886	8.70	75	44	36
20	316	310	777	292	405	285	2345	14.68	296	1636	8.25	72	34	28
Mean	206	292	731	271	403	227	1511	5.25	242	1308	3.54	79	59	73

Table 47: The average number of emergency patients arriving on six months based on the historical data, and the average share of emergencies treated within the dead line

	Red	Yellow	Green
Number of patients	298	748	547
Inpatients	290	704	251
Outpatients	8	44	296
Within time [%]	72	60	70

after January 2017.

C.1 The emergency patients

To compare the arrival of emergencies from the model to that of the real world, we regard the average number of patients to enter through the day. Note that for the emergencies that are registered to arrive at 00.00, we draw a random arrival time between 08.00 and 22.00, as we did when generating the arrivals in the simulation model. We see from Figure 40 that the shapes of the two arrival processes are quite similar. There are however some details that are missing in the simulated arrivals. Firstly, we see that the arrivals of yellow patients are decreasing from hour zero to hour six in the real world. This is not the case in the simulation. Furthermore, we may see that the arrivals of green patients peaks at 10.00 in the reality, and this is also missing in the simulated arrivals. Finally we see that the amount of green emergencies arriving is a little higher for the simulated outcomes compared to the real world. Dividing the day into more than three periods, and increasing the expected inter-arrival time for the green outpatients would have given a better fit to the real arrival process.

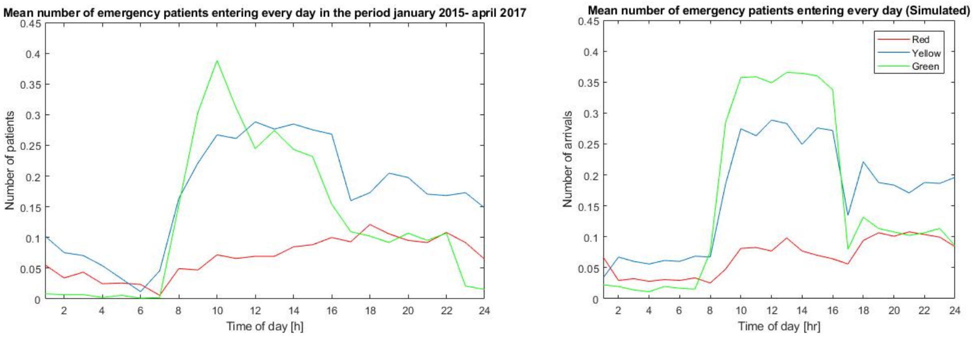


Figure 40: The hourly average number of emergencies arriving through the day in the real life and in the simulated reality. Note that the blue lines indicate the yellow patients.

Figure 41 illustrates the cumulative waiting time to receive surgery for all emergency categories in both realities. We see that the simulated waiting time is slightly less for the red and yellow patients, while it is quite similar to the historical data for the green emergencies. Note that about 8% of the green patients receive surgery on the same day as referral in the real life. This is not implemented in the simulation model, where all green patients have to wait at least one day to receive surgery. We may also note that both for the red and yellow patients the real life distributions have long tails, indicating that some patients have to wait for a long time to have their surgery. These tails represent special cases that are hard to model when applying rigid scheduling rules. One reason why the simulated results outperform the historical data, may be due to the increased emergency OR capacity at AHL implemented in the model.

The average queue of emergencies waiting for surgery every hour through the week may be seen in Figure 42. We see that the amount of green patients waiting for surgery is longer in the simulated outcomes compared to the historical data, while it is the opposite for the yellow patients. There are two reasons why the queue of green emergencies are longer in the simulated outcomes compared to for the historical data: No green emergencies receive surgery on the day of arrival, and the arrival rate is too high for the green outpatients in the simulation model. We see that the number of green patients waiting for surgery decrease towards the weekend in both cases, before it increases through the weekend when these patients are not scheduled for surgery. For the simulated outcomes the queue starts decreasing earlier in the week compared to the historical data. The small spikes seen in the historical data for the green patients is the same spike that we may see in the arrival rate at 10.00 in Figure 40. Finally, we may note that the average queue of red patients is a bit longer for the historical data compared to for the red patients. This may indicate that the red patients are not always prioritized over the other emergency categories as we have implemented in the simulation model.

In Figure 43 we illustrate the cumulative distributions of the number of patients from the different urgency categories waiting for surgery at 08.00. Also here we see that the number of green patients waiting for surgery is longer for the simulated outcomes compared to the historical data, and the opposite is observed for the yellow and red patients. Note that despite that fact that we treat the green emergencies equally fast in the model as we do in the real life, the queue of green emergencies is longer in the simulated outcomes. This means that we are able to treat more green emergencies every day in the model compared to in the real world.

In Figure 44 we provide the cumulative distribution of the weekly amount of green patients to receive surgery in elective slots. We see that the simulated outcomes yield more days with no green emergencies scheduled for the elective slots compared to the real life data.

C.2 OR utilization

The workload at the ORs is important because it provides insight in how ambitious the optimization model is when scheduling electives based on the expected values of the surgery duration. In Figure 45 we illustrate the utilization of six of the ORs at BVS obtained from the simulation model. The blue line in the upper graph illustrates the OR-utilization each day. The daily OR-utilization is calculated by

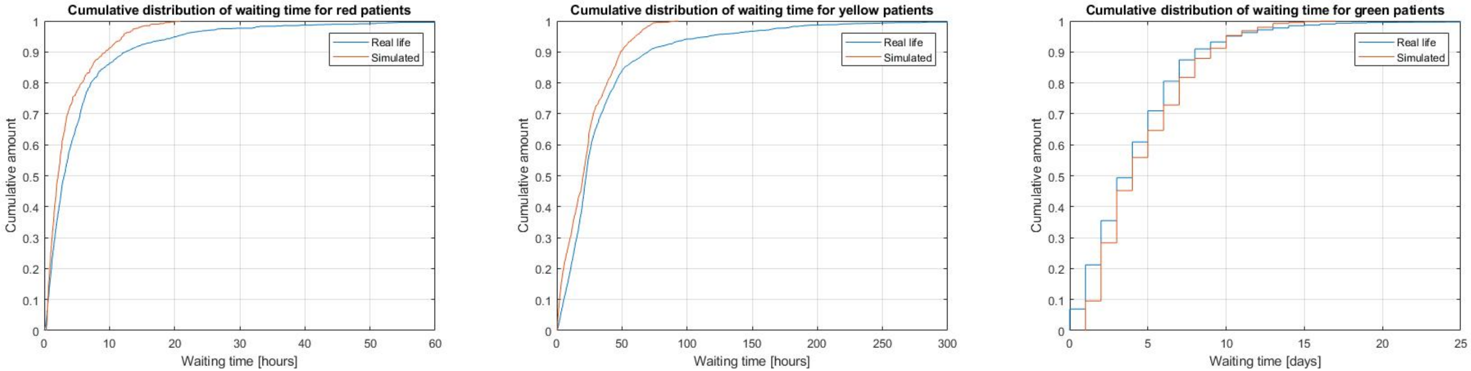


Figure 41: The cumulative distribution of waiting time for emergencies to receive surgery in the real life and in the simulated reality.

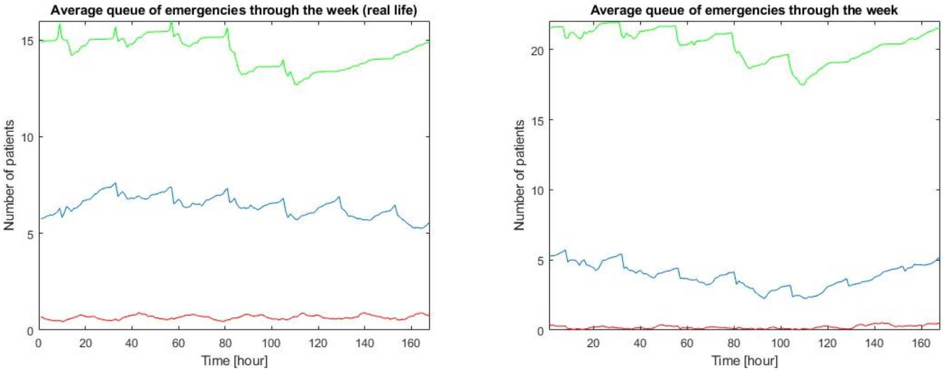


Figure 42: The average number of emergencies waiting for surgery each hour during the week in the real life and in the simulated reality

adding the surgery duration (including preparation and post time) and the cleaning of the OR following surgery, and dividing this by the opening hours available at the OR on that day. Note that surgery related time spent after the regular opening hours of the OR is not included when calculating the utilization. If an OR has not been in use on a day, the utilization is set to zero on that day. The red, discontinuous line indicates the mean utilization through the period, excluding the days when the OR is not in use. In the lower graph, the blue bars reach to the point in time where the OR closes each day. The solid blue line illustrates the scheduled opening hours each day, while the red discontinuous line provides the mean closing time through the period, excluding the days when the OR is not in use. We see that scheduling based on the mean surgery duration results in some days with much overtime at the ORs even after removing the tails of the distributions for the surgery duration for the elective patients. Note that we add 20 minutes of cleaning time between each surgery in the simulation model. This is not done in the optimization model when scheduling the patients for surgery, yielding an optimistic number of electives scheduled for each OR.

The cumulative distribution of the working hours at the eight elective ORs may be seen in Figure 46. Negative time on the x-axis indicates that the last activity at the OR on a day was done before the scheduled closing time of the OR. Compared to the historical data, the simulated outcomes have long tails towards overtime and are less steep, indicating that there is more variation in the working load from one day to the next. As we draw independent realizations of the surgery duration, and use aggregated patient categories, we are not able to provide the same stability for the OR working load in the simulation model compared to the real world.

C.3 Ward loading

Since we model a high production week, we want to compare the simulated outcomes regarding the ward loading to historical data from a high production period. To do this, we regard historical data obtained between the summer vacations in 2016 and 2017, where there were very few low production periods.

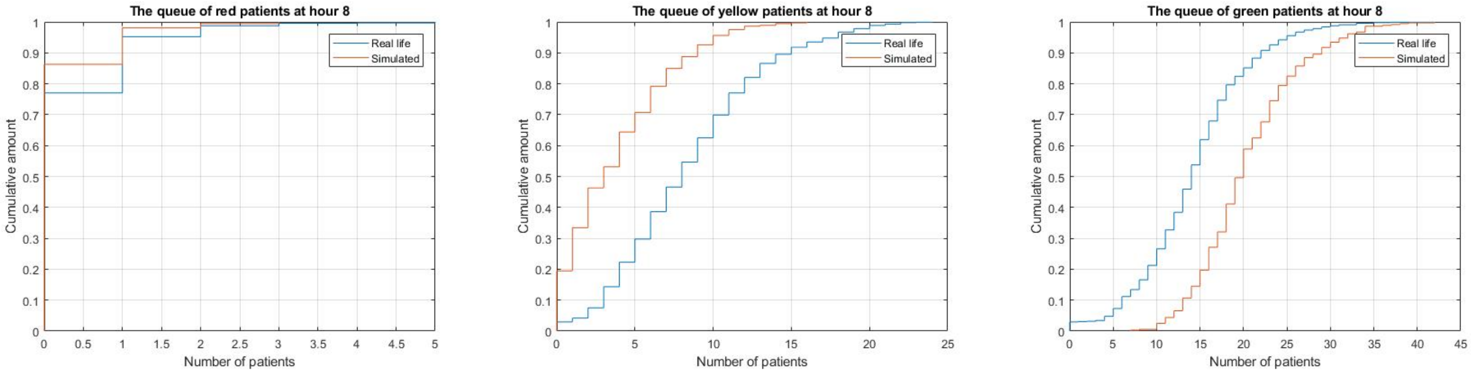


Figure 43: The cumulative distribution of the number of emergency patients waiting for surgery at 08.00 in the real life and in the simulated reality

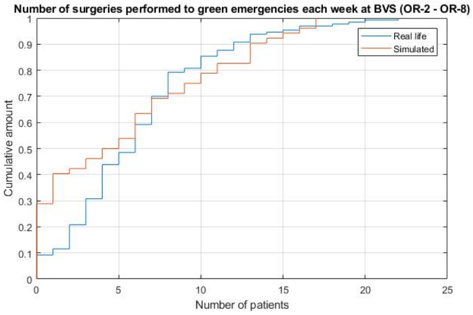


Figure 44: The cumulative distribution of the weekly number of green patients to receive surgery in elective slots at BVS in the real life and in the simulated reality.

In Figure 47 we illustrate the mean ward loading through the week. The peaks indicate that elective inpatients enter in the morning, while other inpatients leave from about 12.00. For the historical data we see that only OTS seem to have patients entering during the evening and night, even though ORS and OES also host emergency patients. In the simulation model we have distributed the emergency patients to the different wards according to a key that is not changing during the day, resulting in patients entering all three wards through the entire day. We may also see that OFT closes on Friday in the real life, while patients are still resting at the ward on Saturday morning in the simulation model. In the real life, the prosthesis patients that need to stay beyond Friday are moved to OES. The most prominent difference in the two cases are the number of patients leaving through the weekend. Only a few patients leave the wards during the weekend in the historical data, while in the simulation model we do not alter the length of stay depending on the day of the week. As a result, the mean ward loading is much less on Monday morning in the simulated outcomes compared to the historical data.

The total cumulative ward loading at 12.00 may be seen in Figure 48. We see that the simulated outcomes are similar to those from the historical data, but the tail is longer towards many beds. We see that we exceed 67 beds in about 15% of the days, indicating the need for rescheduling to handle the scars bed capacity in periods of many emergencies present.

C.4 Conclusion of the comparison

From running the model 20 times, we see that the model schedules the red patients faster compared to the historical data. This indicates that there are times in the real life when less urgent emergencies are prioritized before the red patients, or that less red patients receive surgery during the night than what we have implemented in the model. For the red and yellow emergencies, the number of patients scheduled within the deadline is similar when comparing the simulated results to the historical data. By analyzing the outcomes from one run we see that there are some differences between the model and the real world. The queue of green emergencies are longer in the simulated reality, but these patients receive surgery equally fast as in the real world. This indicates that the model performs more green surgeries every

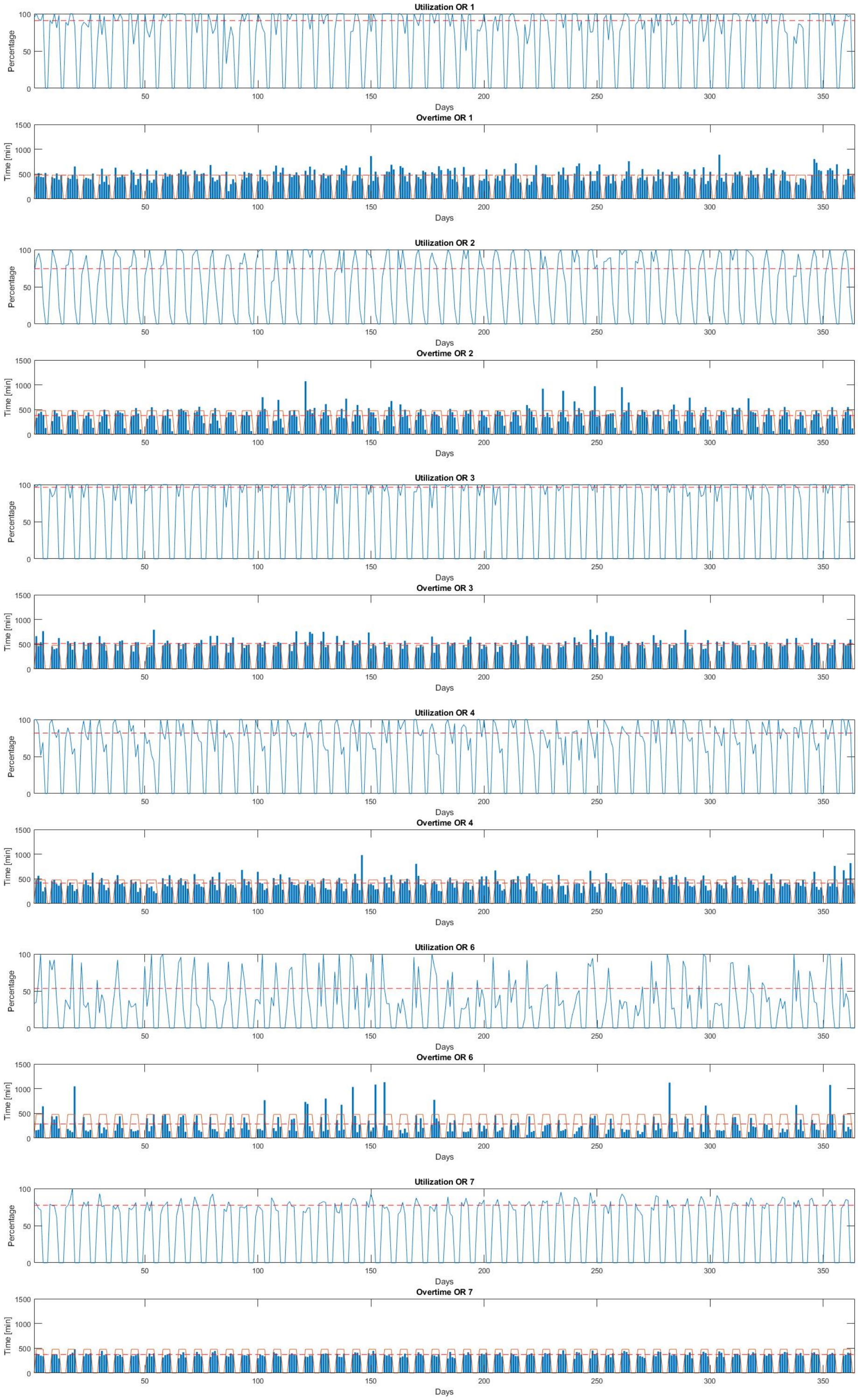


Figure 45: The OR utilization of the elective ORs in the simulation model

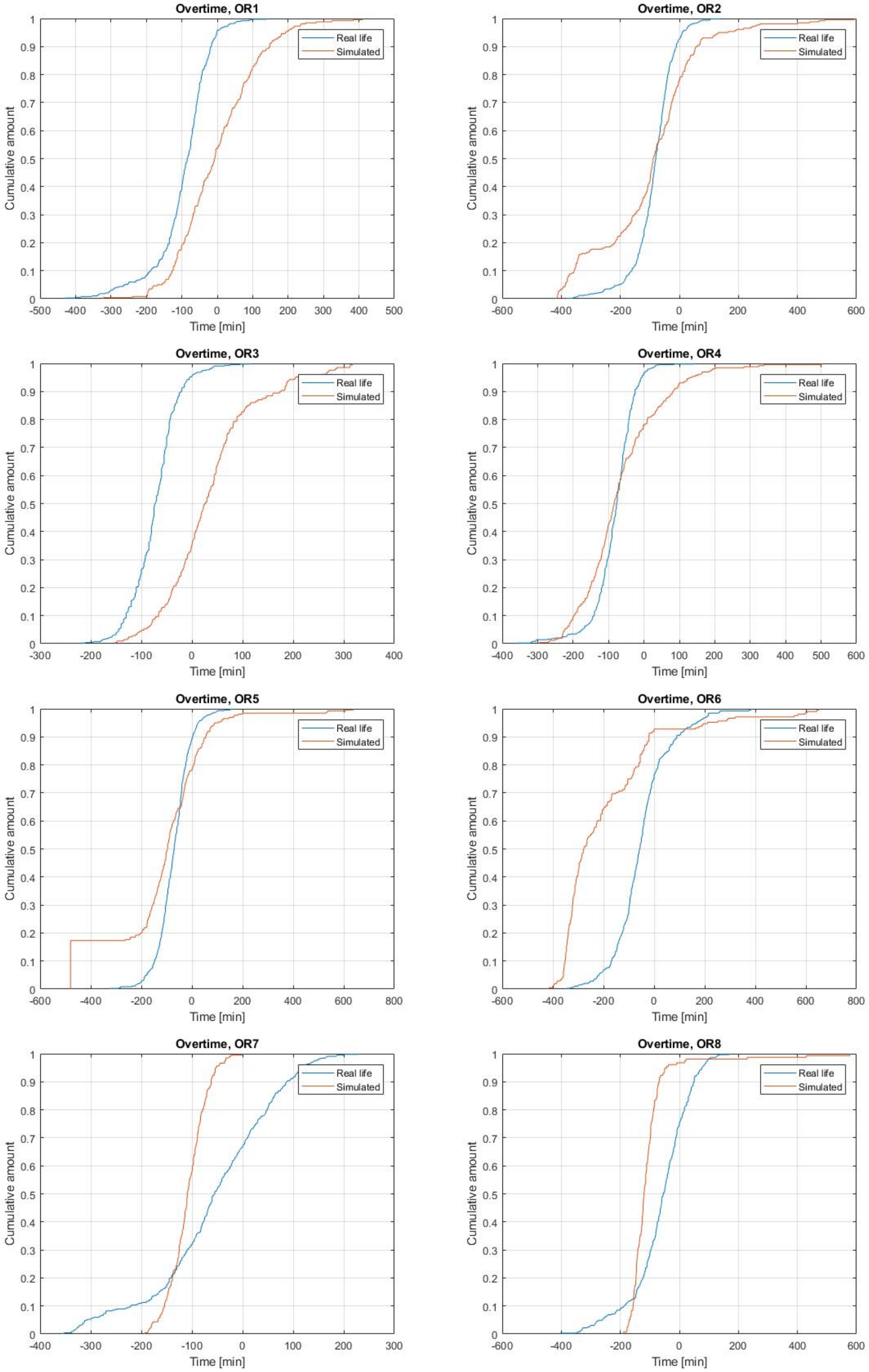
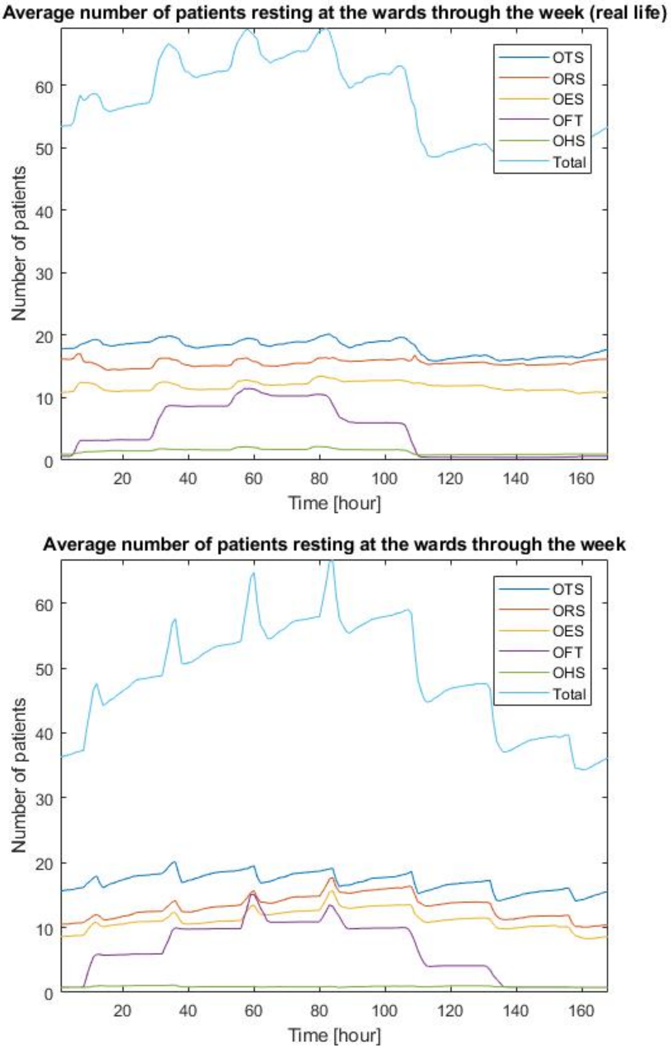


Figure 46: The cumulative distribution of the working hours in the ORs in the real life and in the simulated reality. Zero on the x-axis indicates 16.00.



Average number of patients resting at the wards through the week (real life)

Average number of patients resting at the wards through the week

Figure 47: The average ward loading every hour through the week for both the real life and for the simulated reality

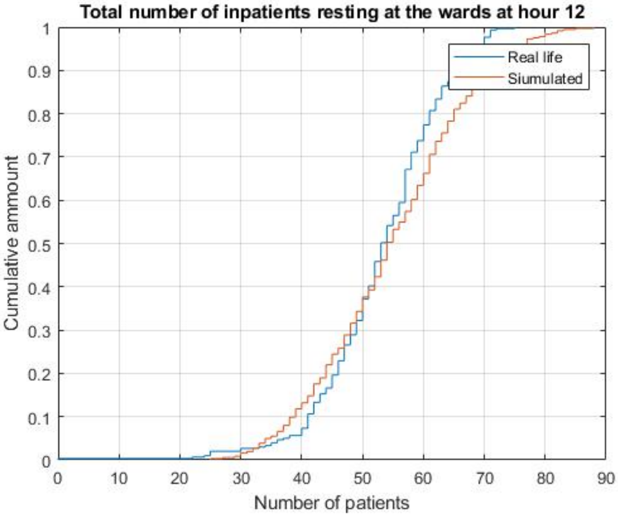


Figure 48: The cumulative distribution of the total ward loading at 12.00 for both the real life and the simulated reality

day despite the fact that less green emergencies are sent to BVS compared to for the real world. One explanation for this is the increased emergency OR capacity implemented in the model compared to the historical data gathered before 2017. Furthermore, scheduling based on expected surgery duration yield more overtime work at the ORs compared to the historical data. Regarding the ward loading, the simulation model sends more inpatients home during the weekend, resulting in less patients being present on Monday morning compared to for the real world. Furthermore, as no rescheduling is done in the model to avoid exceeding the total bed capacity, we and up with the total bed capacity being exceeded more often in the simulated results compared to the historical data.

Despite the weaknesses of the simulation model, it provides an impression of the performance of the orthopaedic department, and it should be sufficient to provide insight when testing the schedules produced by the optimization model.

References

I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129, 2008.

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.

I.J.B.F. Adan and J.M.H. Vissers. Patient mix optimisation in hospital admission planning: a case study. *International Journal of Operations & Production Management*, 22(4):445–461, 2002.

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.

M.E. Bruni, P. Beraldi, and D.Conforti. A stochastic programming approach for operating theatre scheduling under uncertainty. *IMA Journal of Management Mathematics*, 26:99 – 119, 2014.

Y. B. Ferrand, M. J. Magazine, and U. S. Rao. Partially flexible operating rooms for elective and emergency surgeries. *Decision Sciences*, 45(5):819–847, 2014a.

Yann B. Ferrand, Michael J. Magazine, and Uday 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, 2014b.

N. Freeman, M. Zhao, and S. Melouk. An iterative approach for case mix planning under uncertainty. *Omega*, 2017.

M M Günal and M. Pidd. Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4(1):42–51, 2010.

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.

Helse-midt.no. Strategi 2030 bakgrunnsnotat. 2016. URL `https://helse-midt.no/Documents/2016/Bakgrunnsnotat%20Strategi%202030.pdf`.

Helsedirektoratet.no. Drg-systemet. 2017. URL `https://helsedirektoratet.no/finansieringsordninger/innsatsstyrt-finansiering-isf-og-drg-systemet/drg-systemetom-drg-systemet`.

Julia 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.

B.L. Nelson J. Banks, J.S. Carson. *Discrete-Event System Simulation.* Prentice-Hall, 1996. ISBN 0-13-217449-9.

M. Jünger, T. Liebling, D. Naddef, G. Nemhauser, W. Pulleyblank, G. Reinelt, G. Rinaldi, and L. Wolsey. *50 Years of Integer Programming, 1958-2008.* Springer, 2010. ISBN 9783540682745.

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

A. J. King and S. W. Wallace. *Modeling with Stochastic Programming.* Springer, 2010. ISBN 9780387878171.

M. Lagergren. What is the role and contribution of models to management and research in the health services? a view from europe. *European Journal of Operational Research*, 105(2):257 – 266, 1998.

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.

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

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.

Nikolaos V. Sahinidis. Optimization under uncertainty: state-of-the-art and opportunities. *Computers Chemical Engineering*, 28(6):971 – 983, 2004. FOCAPO 2003 Special issue.

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(5):493–525, 2016.

P. Santibáñez, M. Begen, and D. Atkins. Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a british columbia health authority. *Health Care Management Science*, 10(3):269–282, 2007.

Stolav.no. Nøkkeltall for st. olav's hospital. 2017. 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, 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.

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. 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.

Maartje E. Zonderland, Richard J. Boucherie, Nelly Litvak, and Carmen L. A. M. Vleggeert-Lankamp. Planning and scheduling of semi-urgent surgeries. *Health Care Management Science*, 13(3):256–267, 2010.