Master's thesis

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

Mats Ingesen Schiøtz
Jostein Hjortland Tysse

Uncertainty and Cancellations in Advance Surgery Scheduling

An Exact Pattern-Based Solution Approach for Large-Scale Problems applied in a Rolling Horizon Simulation

Master's thesis in Industrial Economics and Technology Management
Supervisor: Anders Nordby Gullhav
Co-supervisor: Lars Hellemo
June 2023

NTNU
Norwegian University of Science and Technology

Mats Ingesen Schiøtz
Jostein Hjortland Tysse

Uncertainty and Cancellations in Advance Surgery Scheduling

An Exact Pattern-Based Solution Approach for Large-Scale Problems applied in a Rolling Horizon Simulation

Master's thesis in Industrial Economics and Technology Management
Supervisor: Anders Nordby Gullhav
Co-supervisor: Lars Hellemo
June 2023

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

NTNU
Norwegian University of
Science and Technology

Preface

This study is our master's thesis in TIØ4905 Managerial Economics and Operations Research, Master's Thesis. It culminates our academic efforts during the final spring semester in the Industrial Economics and Technology Management program under the Faculty of Economics and the Department of Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU). The thesis builds upon the specialization project we conducted in the previous semester (Schiøtz and Tysse 2022). The problem investigated in this thesis is an extension of the problem in Schiøtz and Tysse (2022); thus, parts of the thesis are based on Schiøtz and Tysse (2022).

The central focus of this research paper is to delve into the scheduling complications faced by the Orthopedics Department at St. Olavs hospital, Trondheim University Hospital. We aim to propose solution methods that consider the stochastic nature of surgery duration and how long patients need to stay in recovery beds, finding solutions that utilize the hospital's resources and give the patients a satisfactory experience.

We would like to express our sincere appreciation to our supervisors, Anders Nordby Gullhav and Lars Hellemo. The knowledge they shared and the time and effort they put into our development have significantly contributed to our learning. We also want to thank Thomas Reiten Bovim for the work he did in his master's thesis (T. R. Bovim 2018), providing valuable insights and data on the Department of Orthopedics at St. Olavs hospital, and for taking the time to meet us in person. In addition, we wish to extend a heartfelt thank you to our friends and family. Their unwavering support and belief in our abilities have been vital pillars of strength and motivation for us.

Mats Ingesen Schiøtz and Jostein Hjortland Tysse

Trondheim, June 2023

Abstract

St. Olavs hospital, central Norway's largest healthcare institution, anticipates a demand increase of 35% for surgical procedures due to demographic shifts. The shortage of healthcare professionals and resources makes efficient surgery scheduling mission critical. This study focuses on enhancing surgical scheduling at the hospital's Department of Orthopedics by improving the existing mathematical model proposed by Schiøtz and Tysse (2022) for the Advance Scheduling Problem (ASP), incorporating factors like uncertain surgery duration and recovery time.

The enhanced models aim to abide by the department's scheduling rules, prioritize the reduction of patient waiting time and schedule disruption, and account for resources like operating rooms, recovery wards, and the Master Surgery Schedule. This study will also explore the impact of integrating a cancellation rule into the hospital's scheduling system and assess how it might influence the scheduling quality while mitigating the risks associated with overtime and cancellations.

The proposed scheduling models are integrated into a simulation framework using a rolling horizon approach for quantitative comparison. Among the significant contributions of this research are incorporating cancellation rules and rescheduling of surgical cases into the models and handling large problem sizes akin to real-life scenarios at elective clinics. We propose a pattern-based Mixed-Integer Program that can solve real-life size problems to optimality, something traditional two-stage stochastic models can not. Our model considers uncertainty and manages to produce a better surgery schedule than two-stage models in multiple of our tests.

Our research results underscore the importance of computational efficiency and reveal a fundamental dilemma between schedule stability and efficiency. Interestingly, deterministic models were found to underestimate overtime, underscoring the importance of including uncertainty, especially when cancellation rules apply. Future research areas include the scalability of the pattern-based Mixed-Integer Program and advanced pattern filtering methods using machine learning techniques.

Sammendrag

St. Olavs hospital, Midt-Norges største helseinstitusjon, forventer en økning i etterspørsel på 35% for kirurgiske inngrep grunnet demografiske endringer. Mangelen på helsepersonell og ressurser gjør effektiv operasjonsplanlegging kritisk. Denne studien setter søkelys på å forbedre kirurgisk planlegging ved sykehusets ortopediavdeling ved å forbedre den eksisterende matematiske modellen foreslått av Schiøtz and Tysse (2022) for operasjonsplanlegginsproblemet ved å inkludere faktorer som usikker operasjonsvarighet og restitusjonstid.

Vi utvikler matematiske modeller som tar sikte på å overholde avdelingens planleggingsregler, prioritere reduksjon av pasientens ventetid og avbrudd i tidsplanen, og ta hensyn til ressurser som operasjonsrom, sengeposter og den overordnede operasjonsplanen ved sykehuset. Denne studien vil også undersøke effektene av å inkludere en kanselleringsregel i sykehusets planleggingssystem og vurdere hvordan regelen kan påvirke planleggingskvaliteten samtidig som risikoen forbundet med overtid og kanselleringer reduseres.

For kvantitativ sammenligning er de foreslåtte planleggingsmodellene integrert i et simuleringsrammeverk med en rullende horisont. Blant betydelige bidrag fra studien er inkluderingen av kanselleringsregler og re-planlegging av pasienter i modellene, og håndtering av store og virkelighetsnære problemstørrelser. Vi har utviklet en mønsterbasert blandet heltallsmodell som klarer å løse problemer av reell størrelse til optimalitet, i motsetning til tradisjonelle to-stegs stokastiske modeller. Modellen tar hensyn til usikkerhet, og klarer i mange tilfeller å lage bedre operasjonsplaner enn en to-stegs modell i våre tester.

Forskningsresultatene våre understreker viktigheten av beregningseffektivitet og avslører et grunnleggende dilemma mellom stabilitet i operasjonsplanen og effektivitet. Et interessant funn er at deterministiske modeller undervurderer overtid, noe som understreker viktigheten av å inkludere usikkerhet, spesielt når vi inkluderer kanselleringsregler. Skalerbarheten til den mønsterbaserte heltallsmodellen, og avanserte filtreringsmetoder for mønstre ved bruk av maskinlæringsteknikker potensielle fremtidige forskningsområder.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

St. Olavs Hospital, the largest healthcare institution in central Norway, caters to a local populace of 327,574, with its regional responsibilities stretching to cover 729,452 inhabitants (St. Olavs hospital HF 2022). However, with the predicted demographic shifts over the next decade and a half - stemming from population growth, increasing life expectancy, and urbanization - it is anticipated that demand for surgical procedures at St. Olavs could increase by 35% (St. Olavs hospital HF 2018). Additionally, there exists a shortage of healthcare professionals. G. Bovim et al. (2023) goes so far as to describe the situation as a 'ticking time bomb'. Hence, the hospital's management is faced with the crucial challenge of enhancing resource utilization and developing effective surgical schedules in a scalable fashion. G. Bovim et al. (2023) describes the situation as one we cannot pay ourselves out of and points to digitalization and automatization as one of the solutions.

In scientific literature, the task of planning surgeries is recognized as the Surgical Case Scheduling Problem (SCSP). This problem can be divided into two sub-problems: the Advance Scheduling Problem (ASP) and the Allocation Scheduling Problem. The former relates to finalizing a surgery date for the patient, whereas the latter assigns a specific operating room and sets the procedure's starting time for the surgery date (Cardoen et al. 2010).

This research study will focus on the ASP at the Department of Orthopedics at St. Olav's Hospital. The department, which handles diagnoses and treatments related to the musculoskeletal system, is divided into nine specialties, executing 15 unique and, to some extent, generalized procedures. From January 2015 to April 2017, they handled 9231 inpatients and 4885 outpatients, with 53% of these cases being elective and the remaining being emergency cases. As the numbers suggest, there are many moving pieces at the Department of Orthopedics, and efficiently scheduling surgeries is challenging. Balancing surgical schedules to align with operating room and ward capacities and ensuring adequate staffing are complex tasks. This complexity is further amplified by uncertain factors like fluctuating surgery demand, variable length of stay (LOS), unpredictable surgery

durations, and unexpected events like staff illness or patient no-shows.

The thesis has three overall goals. Firstly, it aims to enhance the mathematical model proposed by Schiøtz and Tysse (2022) by incorporating variables such as uncertain surgery duration and recovery time. The revised models will be developed to abide by the department's scheduling rules and prioritize reducing patient waiting time and schedule disruption. Regarding resources, the models will consider the operating rooms, downstream recovery wards, and the pre-established Master Surgery Schedule (MSS). The second objective is to explore the impact of incorporating a cancellation rule in the Department of Orthopedics' scheduling system. We will look into measures for mitigating various risk levels associated with overtime and cancellations and how these might influence the schedule quality. The final aim is to seamlessly integrate the models into a simulation framework using a rolling horizon approach, allowing us to compare schedule quality among various proposed models. To achieve these objectives, we will propose two Multi-Objective Mixed-Integer mathematical programs. The first program is a two-stage stochastic formulation similar to models in the existing literature. The second is a pattern-based Mixed-Integer Program (MIP) designed to precalculate the uncertain parameters, thus allowing the problem to be solved without considering scenario trees like the two-stage model does.

We provide several contributions of academic interest in this thesis. First, we include cancellation rules in our models. In practice, it is common for clinics serving elective patients to apply some kind of cancellation rule to avoid overtime. However, the consequences of those rules on the scheduling quality are unexplored territory, to the best of our knowledge. Second, we allow rescheduling of surgical cases. How this addition affects scheduling quality and the performance of stochastic models is not explored in the existing literature. Lastly, little research exists that handles problem sizes as large as the ones observed in real-life at an elective clinic, where planning horizons of up to three months are possible. No existing research within surgery scheduling has to the best of our knowledge, used our kind of pattern formulation with set-partitioning constraints and extensive precalculation of stochastic information to handle the complexity without any heuristics.

Our results emphasized the role of computational efficiency in modeling and decision-making processes. We found that the pattern-based model significantly outperformed the two-stage stochastic model, demonstrating its ability to find solutions faster and to solve the ASP to optimality for real-life size problems. We found that using stochastic models with SAA in settings where rescheduling is allowed may lead to stability complications as any stability complications are enhanced due to the inherently multistage aspects of ASP. Further, we discovered a fundamental dilemma within the healthcare sector introduced by cancellation rules and rescheduling possibilities: the balance between schedule stability and efficiency. As the demand for healthcare services increases, so does the need for efficiency. Increasing the cancellation and overtime risk increases schedule efficiency at the cost of schedule stability. This imbalance may lead to ethical issues and a worse user experience, necessitating consideration by relevant authorities. Lastly, an intriguing observation from

our study is that models not considering cancellations could maintain a low cancellation risk. This outcome is due to an implicit overestimation of overtime costs, reducing the likelihood of cancellations. However, deterministic models tend to underestimate overtime and do not exhibit this property. These findings emphasize the importance of including uncertainty in models, especially when cancellation rules apply.

Topics for further research include continued investigation of the scalability of the pattern-based MIP and more advanced pattern filtering methods. Investigating whether reinforcement learning or other machine learning methods could utilize more of the information contained in the joint probability distributions of the pattern to filter patterns and strategically handle the uncertain future demand is of particular interest.

The thesis begins with a background on St. Olavs Hospital and the Department of Orthopedics in chapter 2. chapter 3 discusses relevant literature on surgical case scheduling and existing solution methodologies. Next, chapter 4 presents the problem investigated in this thesis, the ASP. The mathematical models used to solve the ASP are presented in chapter 5 and followed by a description in chapter 6 of how the patterns used by some of the models are generated. The simulation framework used to evaluate the model is introduced in chapter 7, and a computational study is performed in chapter 8. Lastly, chapter 9 concludes the thesis and showcases our most significant findings.

Chapter 2

Background

St. Olavs hospital, the largest hospital in the region, has 36 operating rooms, 750 recovery beds, and 171 outpatient clinic rooms. However, significant demographic changes are expected in the next 15 years due to population growth, increased life expectancy, and increased urbanization (St. Olavs hospital HF 2018). One of the most critical questions on the current healthcare management agenda is how to meet the increased healthcare demand. Table 2.1 shows the projected demand and needed resources. Another way of facing the increased demand, besides increased resources, is to increase the utilization of current resources and the overall efficiency of the healthcare systems.

Table 2.1: Expected demand for somatic services at St.Olav's Hospital in 2035. The increase is due to demographic changes in the region (St. Olavs hospital HF 2018)

	2015	2035	Change (%)
Inpatient stay (days)	52.896	74.398	41
Bed days	232.681	340.842	46
Beds	750	1.099	46
Out-patient clinic (patients)	418.442	670.908	60
Out-patient clinic (rooms)	171	274	60
Surgeries (patients)	31.409	42.504	35
ORs	36	49	36

The rest of the chapter is structured as follows. First, Section 2.1 describes important terminology and aspects of surgery planning, execution, and general hospital governance. Then, Section 2.2 introduces the Orthopedic Department at St.Olav's Hospital and investigates their surgery planning procedures. Since the problem investigated by Schiøtz and Tysse (2022) is based on the same department, most parts of this chapter are identical to the background chapter in Schiøtz and Tysse (2022).

2.1 Aspects related to hospital governance

2.1.1 Hospital organization

The Norwegian Ministry of Health and Care Services (HOD) is responsible for the overall accessibility and quality of health care in Norway. Following a centralization strategy, the government has officially owned all publicly owned healthcare facilities since 2004. They govern through four Regional Healthcare Entities (RHF). Each RAF is in charge of several Local Helseforetak (HF). The hierarchical structure of health care organization is illustrated in Figure 2.1.

St. Olavs hospital, Trondheim University Hospital has three functions: (1) as a local hospital for 327 574 inhabitants, (2) a regional hospital for 729 452 inhabitants, and (3) a university hospital through collaboration and partial ownership by NTNU (St. Olavs hospital HF 2022). St. Olavs hospital occupies around 350 000 square meters of space distributed over several regional facilities. The largest facility, at 244 000 square meters, is located at Øya. The operation is divided into 21 clinics employing around 10 500 employees. An overview of the hierarchy at St. Olavs hospital is shown in Figure 2.2

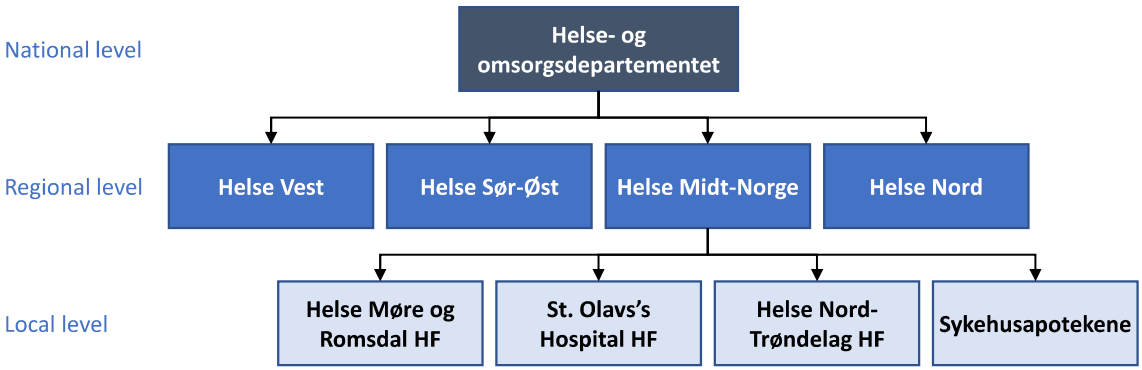


Figure 2.1: The hierarchical structure of health care organizations in Norway (T. R. Bovim 2018). St. Olavs hospital is part of the Helse Midt-Norge regional level.

A department might have several different sub-departments and clinics. We will use the Department of Orthopedic Surgery as an example, which is part of the clinic of orthopedics, rheumatology, and skin diseases. The orthopedic department consists of 14 units that can be categorized as either informational-, therapeutic-, diagnostic services, or supporting units. An overview is shown in Table 2.2

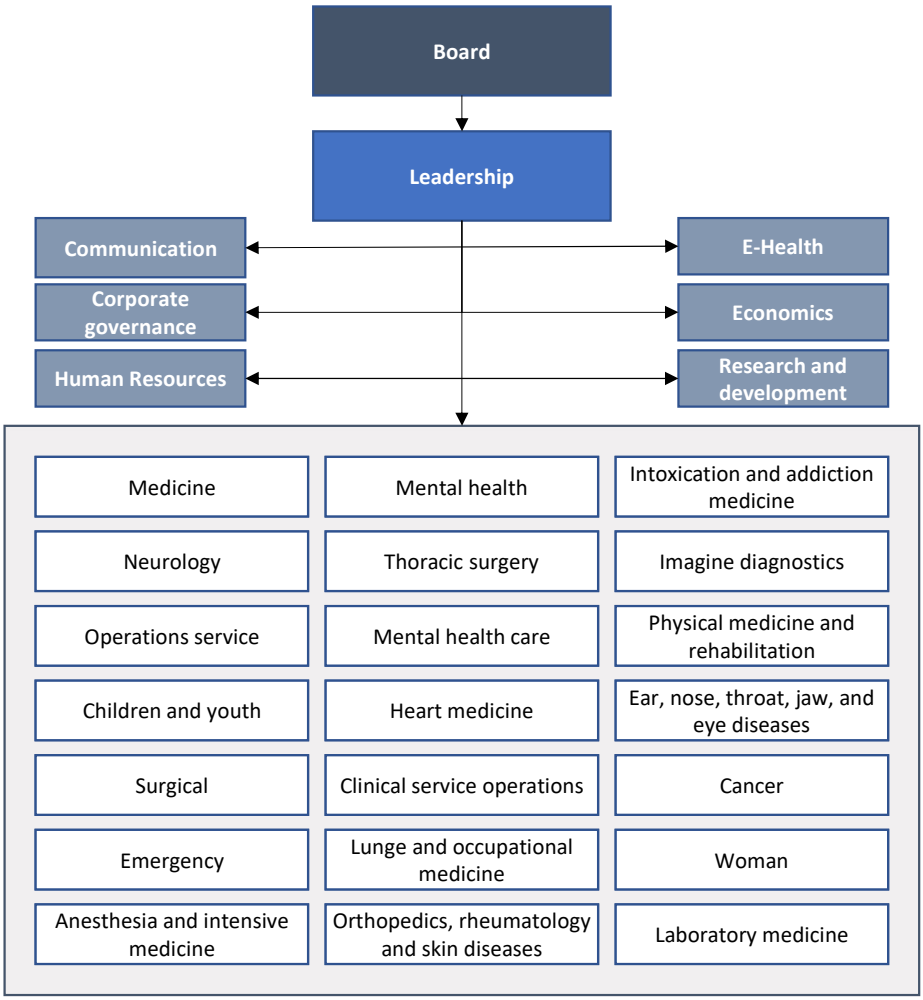


Figure 2.2: The hierarchical structure of St. Olavs hospital.

Table 2.2: The different services and units at the Department of Orthopedic Surgery (T. R. Bovim 2018)

Units
Dagkirurgisk enhet for ortopedi og plastikkirurgi
Fast-track tun ortopedi (OFS)
Felles lettpost pasienthotellet
Hotell/Dagtun ortopedi (OHS)
Inntakskontor for ortopedi og plastikkirurgi
Ortopedisk forskningssenter
Ortopedisk operasjonsavdeling
Overvåkningsenhet for ortopedi og plastikkirurgi
Poliklinikk for ortopedi og plastikkirurgi
Preoperativ poliklinikk for ortopedi og plastikkirurgi
Sengeområde for elektiv ortopedi (OES)
Sengeområde for ortopedisk traumatologi
Sengeområde for Rekonstruktiv ortopedi og Plastikkirurgi (ORS/OPLS)
Skade/Traumepoliklinikken

2.1.2 Patient and diagnoses grouping

Patients are typically classified as elective and non-elective (alternatively elective and emergency). Elective patients typically have diagnoses that develop slowly and can wait several weeks or months before surgery without significant consequences. As a result, elective surgeries are usually planned several weeks ahead.

Another standard patient classification is in- and outpatient. Inpatients are commonly characterized as patients needing to recover at the hospital several days after surgery. On the other hand, outpatients usually can receive surgery and depart on the same day. It is common for the hospital to have dedicated resources for out- and inpatients. St. Olavs hospital has dedicated outpatient clinics called 'poliklinikk' in addition to clinics that mainly serve inpatients and some outpatients. The use of outpatient clinics and hence the demand for outpatient surgery is increasing as techniques and knowledge become more and more profound. Patient types and associated research are further described in Section 3.2.3.

The Diagnosis-Related Group (DGR) patient classification system is adopted for all healthcare services in Norway. One standard system makes it easier to compare the quality of different hospitals and departments, even when treating completely different patients. The DGR consists of five central variables according to Helsedirektoratet (2022): (1) diagnosis, (2) procedure, (3) sex, (4) age, and (5) state of patient after treatment. Sometimes the DRG is combined with the associated running costs. In this way, it is possible to fund departments based on DGR targets for departments.

2.1.3 Hospital resources

Resources are needed for a hospital to run smoothly. We choose to segment the different types of resources as either staff resources or physical resources. Further, it is common to segment resources as either upstream or downstream vertically, according to when they are consumed in the patient flow. We will use the first and more simplistic segmentation for now and expand later in Section 3.2.5

Physical resources

The operating theater is where all Operating Rooms (ORs) are, and usually also preparation rooms and rooms to store surgical equipment. The preparation rooms can be used to clean patients and give anesthesia. As a result, the needed OR time for each patient can be reduced by utilizing the available preparation rooms. Some ORs might be more suitable for certain patients than others. For instance, an OR can be reserved for emergency patients; others might have installed specialized equipment for specific procedures.

Besides preparation rooms in the operating theater, other rooms might also be available to host patients outside the operating theater before surgery diagnosis, especially relevant for outpatients as they are not checked into a recovery ward.

Three types of recovery rooms are needed, and patients might use all or just some of those resources throughout their stay. The Post Anesthesia Care Unit (PACU) is dedicated to patients waking up from anesthesia. The Intensive Care Unit (ICU) is where patients in critical condition who might need help with normal body functions such as breathing are placed. Lastly, we have recovery wards where patients not in critical condition stay until they are in a state where it is safe to either go home or be transferred to other treatment facilities outside the hospital.

Staff resources

The hospital staff comprises a broad mix of people with different expertise and functions. The functions can broadly be divided into clinical, technical, supporting, and administrative. The clinical staff is especially relevant for the problem investigated in this report. A surgical team of typically 1-2 surgeons, 3-4 nurses, and one anesthesiologist is used when performing surgeries. An illustration of the flow of staff throughout surgery is showcased in Figure 2.3

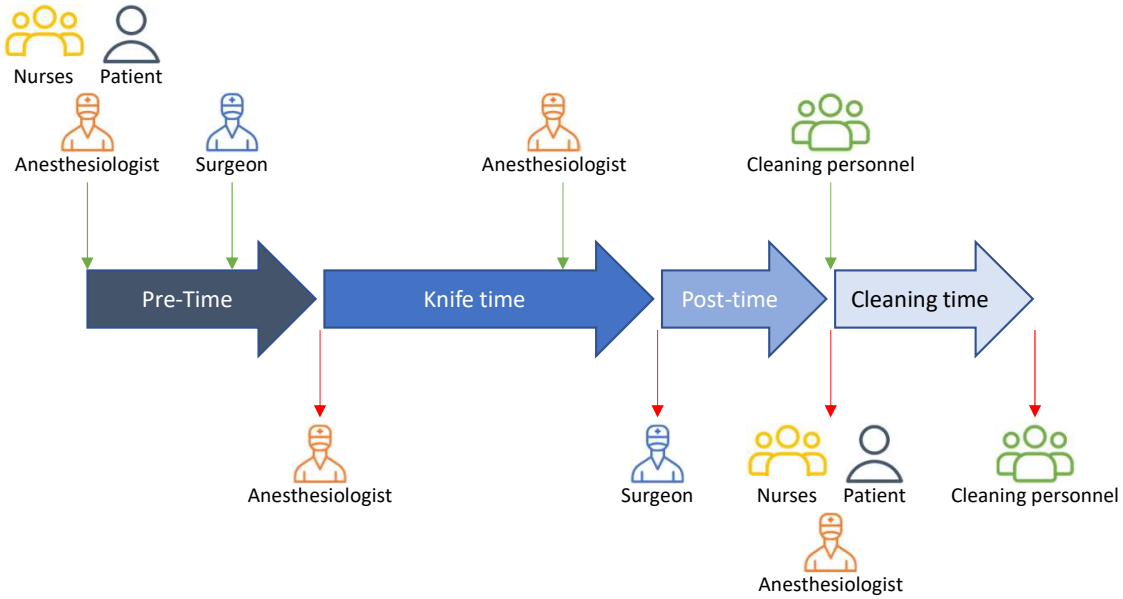


Figure 2.3: The flow of staff resources during a surgery (T. R. Bovim 2018).

Surgeons are specialized and divided into specialties. For instance, treating a knee or a back problem takes different expertise. Other than preparation, surgeons can have different levels of experience. A resident is a surgeon currently specializing, while a consultant has been specialized and is more experienced. A surgeon's experience is relevant when planning surgeries, as more experienced surgeons typically are more efficient, resulting in shorter surgery duration. Besides performing surgeries, surgeons are also needed in the wards to discharge patients and perform

other consultations.

Anesthesiologists are responsible for the anesthesia provided to patients during surgery and are needed at the start and sometimes at the end of surgeries. In addition, at St.Olav's, the anesthesiologists are one shared resource between multiple departments. As a consequence of those two factors, there are a limited number of surgeries that can start at the same time.

2.2 The orthopedic department

The case study considered in this report is about the Department of Orthopedics. Unfortunately, we were unable to perform an interview with representatives from the orthopedic department. Therefore, we have depended on numbers and insights from T. R. Bovim (2018). Additionally, we interviewed Flaata (2022), a surgery scheduler at Rikshospitalet. She further explained the surgery scheduling process and gave additional insights into the practical problems faced when scheduling surgeries.

The Department of Orthopedics is responsible for diagnosing and treating patients suffering from injuries and diseases in the musculoskeletal system, i.e., the skeleton, muscles, ligaments, joints, and other connective tissues. Most of their activity is located at Bevegelsessenteret (BVS), but they also have some available resources at Kvinne-barn-senteret (KBS) and Akutten og hjerte-lunge-senteret (AHL). Outside of Øya, they are also present at St.Olavs hospitals facilities at Røros and Orkdal, but those activities are irrelevant to this report. A map of Øya and the departments' facilities are displayed in Figure 2.4. Further, they have nine different specialties, as shown in Table 2.3. Plastics are, in theory, not part of the Department of Orthopedics, but we have included them in our report as they share the same physical resources.

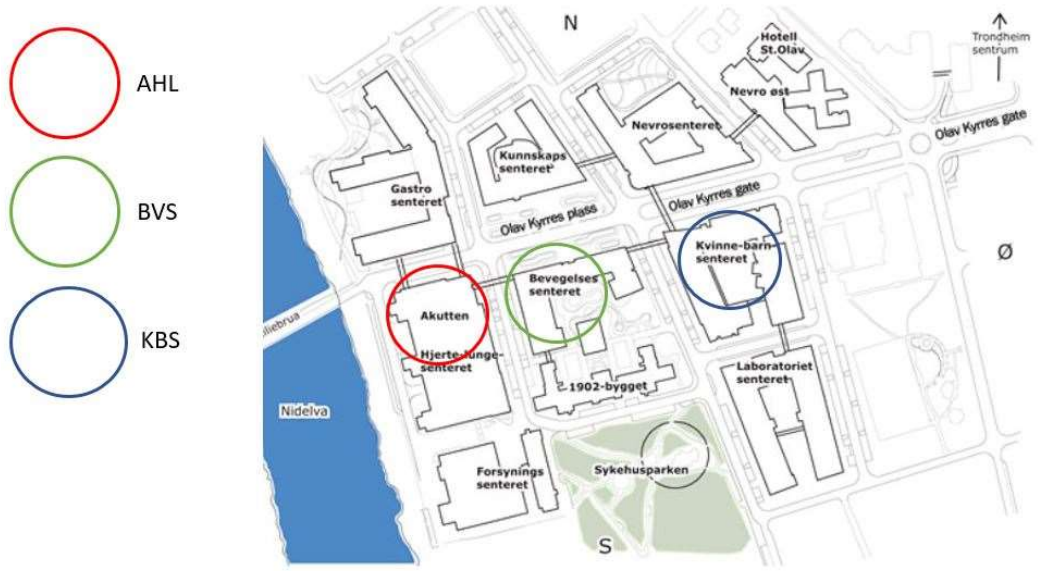


Figure 2.4: Map of St. Olav's Hospital, Øya.

Table 2.3: The specialties at the Department of Orthopedic Surgery

Specialties
Elective foot
Plastics
Reconstructive
Elective trauma
Hand
Arthroscopic
Back
Prosthesis
Children

2.2.1 Orthopedic patients

The Department of Orthopedics follows the traditional patient segmentation described in Section 2.1.2. Interestingly, they use a traffic-light system to indicate the urgency of non-elective patients. Green, yellow, and red patients should receive surgery within five days, 24 hours, and six hours, respectively. It is worth noting that the system serves as guidelines, not as strict absolute limits. The elective patients, however, do not have any guidelines on how long they can wait before receiving surgery, but three months are used as a general rule of thumb.

It is interesting to look at the share of elective vs. non-elective and inpatient vs. outpatient receiving surgery at the orthopedic department. We find almost a 50/50 split between elective and non-elective, while they treat about 50% more inpatients than outpatients, as shown in Table 2.4. For elective patients, however, we see the opposite, where outpatients outnumber inpatients by about 33%. This indicates that it is necessary to plan both types of patients in an integrated fashion, as they share the same OR resources to some extent.

Table 2.4: The distribution of patients from 01.01.15-27.04.17

Urgency	Inpatient	Outpatient	%
Elective	3 424	4 885	53
Green	1 172	1 380	16
Yellow	3 284	204	22
Red	1 351	35	9
%	59	41	100

Figure 2.5 shows a simplified illustration of the flow of elective patients at the orthopedic department. The path can ruffly be divided into five steps: (1) the patient is assessed at the outpatient clinic after receiving a referral from the patient's doctor, (2) if surgery is needed, the surgeon requests a surgery date from the scheduling administration, (3) the patient meets at the hospital before surgery, (4) surgery is performed, and (5) the patient recover before going home. Inpatients have an additional step in addition to the five steps listed above. The inpatient is invited to a preoperative assessment a few weeks before surgery. In a few cases, it is discovered that the patient no longer should receive surgery.

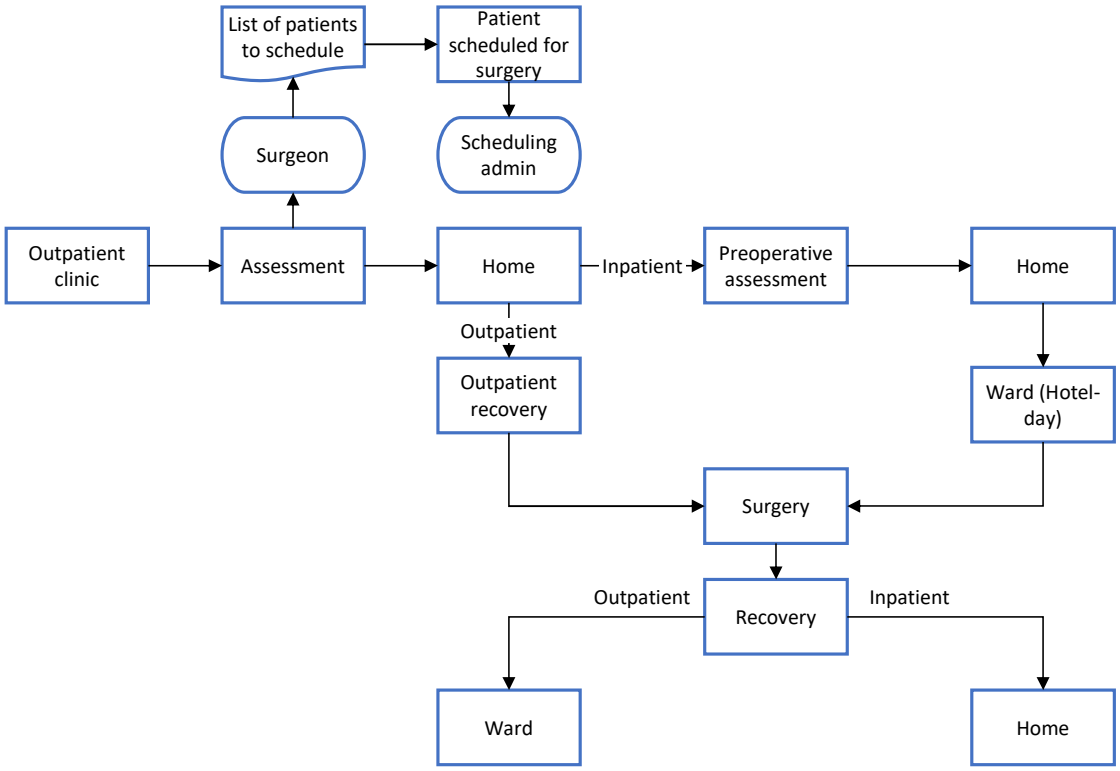


Figure 2.5: The flow of elective patients at the orthopedic department.

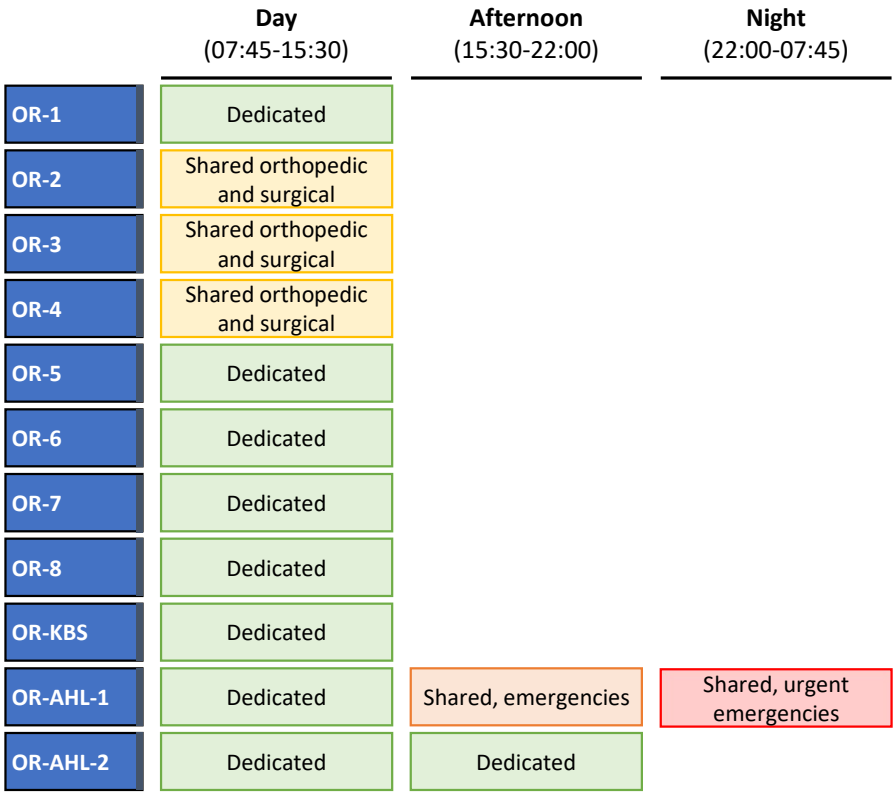


Figure 2.6: The OR opening times available to the department of orthopedics.

2.2.2 Physical resources

ORs

There are, in total, 11 ORs available to the orthopedic department, eight at BVS, two at AHL, and one at KBS. Most ORs are open between 07:45 and 15:30, as shown in Figure 2.6, while some are open in the afternoon and night to handle the demand of non-elective patients. The orthopedic department has decided to dedicate OR-2 up until OR-8 to electives, OR-KBS to children, and the rest to non-electives. This means that we, in theory, can schedule elective patients without dealing with the non-elective demand. In practice, it is not uncommon for green non-elective patients to use some of the resources dedicated to elective patients. We do not investigate these effects any further in the report. We also do not include children since they do not share any resources with the other specialties.

The wards

The orthopedic department has six wards available to them, and their capacities, relative location, and expertise are described in Table 2.5. In total, they have 70 staffed beds during weekdays. During the weekend, the capacity is lower, with a total capacity of 44, due to less staffing and fewer surgeries. Hence, the total ward capacity also depends on factors other than the number of beds theoretically available. The orthopedic department has about 90 beds, meaning around 20 beds are never used, even on weekdays. It would be possible to use those 20 beds if the staffing is increased.

Table 2.5: The wards at the orthopedic department. Each ward has a weekday capacity, weekend capacity, and some preferred patient categories.

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, arthroscopic
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

By a ward's expertise, we mean that some wards are equipped and staffed based on the expected patient types they will nurse. It is best for the patient and operational efficiency that patients within the same category share the same wards. Hence, the different expertise of the wards is something to consider while scheduling surgeries. It is, however, important to note that there are

no strict rules, and it is still possible for patients to stay in a ward typically not intended for them. This is important as it creates the flexibility to move patients as wards get full. Three possible reasons why wards might get full are: (1) the amount of inpatients receiving surgery is too high, (2) the number of emergencies arriving is higher than expected, or (3) a patient has to recover at the ward for longer than what is expected.

The first point is relatively self-explanatory. The number of inpatients recovering in the wards is strongly dependent on how many inpatients are receiving surgery in the first place. However, a key observation is that the OR and ward utilization are not directly dependent. We can have high OR utilization without high ward utilization and the other way around. The frequency of different patient types receiving surgery is a more important factor. For instance, if all patients are outpatients, no ward capacity would be used. One less extreme example of this is seen in the orthopedic department: most patients receiving surgery on Fridays are outpatients because of the lower ward capacity during the weekends. Ward capacities are more likely to become a bottleneck for inpatient throughput later in the weeks. Thus, allowing inpatients and outpatients to share OR resources can increase scheduling flexibility.

In theory, non-elective and elective patients do not share the same ward resources in the orthopedic department. However, if the non-elective demand becomes too high, the hospital might be forced to send some non-elective patients to the wards intended for elective patients. Increased non-elective demand can happen due to the inherently uncertain arrival of non-elective patients.

As previously mentioned, the frequency of different patient types receiving surgery can significantly impact ward utilization. An important factor is that patients' Length of Stay (LOS) is uncertain. The LOS depends on the procedure and other factors, such as the patient's age, overall health, and medical history. Suppose a patient has to recover in the ward for longer than expected; the risk of the wards becoming full increases. A patient's LOS can also be extended for other reasons. Before being discharged, the discharge must be approved by a surgeon. If none of the surgeons working during the weekends knows the patient, they might feel uncomfortable sending them home. As a result, patients expected to be discharged during the weekends might sometimes stay longer than expected.

Due to the uncertain demand for beds in the wards, the hospital might be forced to cancel elective surgeries. Therefore, the number of cancellations may depend on the ward utilization and capacity. One thing to note is that this does not affect outpatients since no outpatients will be canceled due to the wards' capacities. Hence, inpatients are more likely to be canceled compared to outpatients in that sense. This indicates that considering the wards while planning surgeries can decrease the cancellation rate.

Several factors make making efficient schedules challenging, including patient mix, staff resources, physical resources, economy, and overall hospital strategy and goals. How to solve this task is

described more in detail in chapter 3. However, the key takeaway is that it is simply too many factors to consider simultaneously, and the hospital must tackle the problem in an iterative and structured fashion.

Figure 2.7 shows a high-level view of the scheduling process at the orthopedic department. At first, St. Olavs hospital's board of directors set annual DRG targets for the department. The targets are often decided based on the overall hospital strategy and prioritization and budgets decided by the HOD. Then, the management at the department will use the DRG targets, their own strategy, general thoughts about the future, predictions of future demand, and their available resources to create a Master Surgery Schedule (MSS).

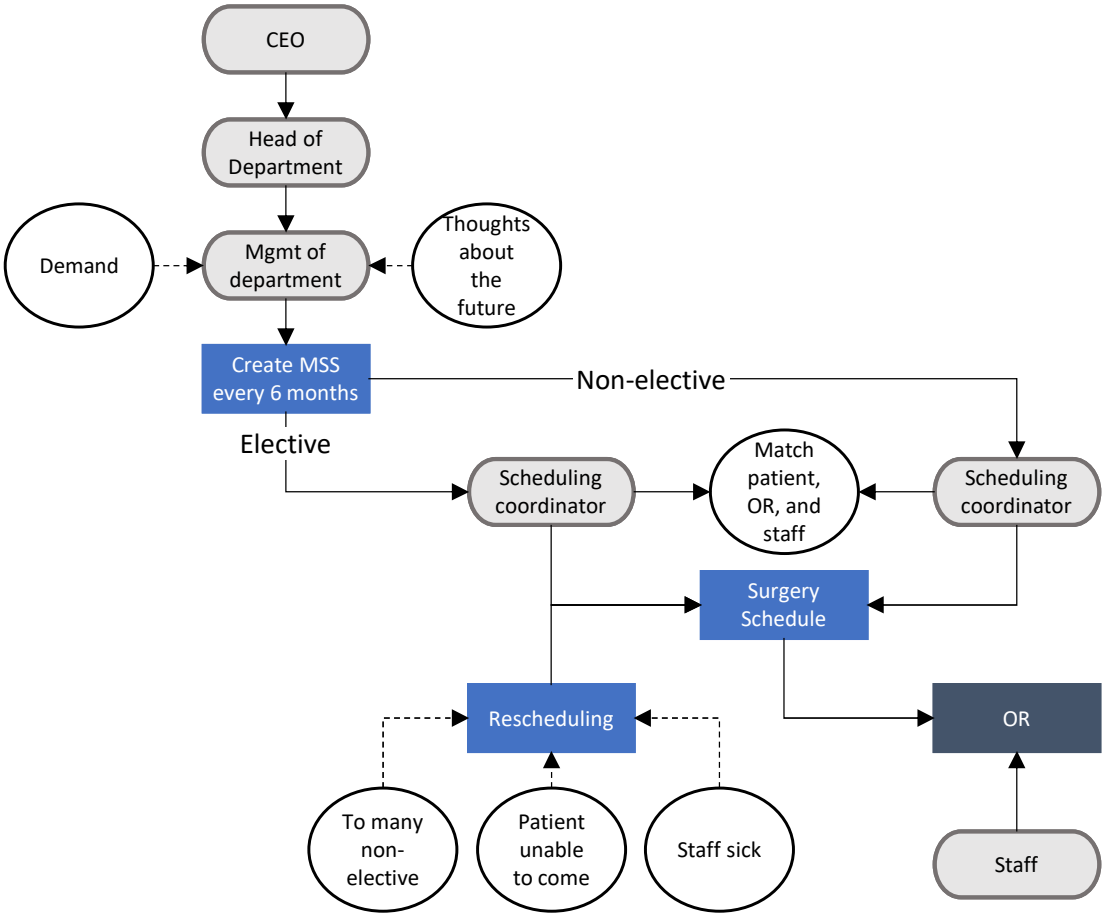


Figure 2.7: The scheduling process at the orthopedic department according to T. R. Bovim (2018).

2.2.3 Surgery scheduling

The MSS is a high-level surgery schedule where the available OR time slots are represented as blocks. The blocks connect ORs, specialties, dates, and times. This practice is regularly referred to as block scheduling in the literature. If an MSS is not used, we say they use an open scheduling policy. The MSS and alternatives are further described in Section 3.1. Figure 2.8 shows an example of an MSS at the Department of Orthopedics, and the MSS will typically be used for around six

months before a new MSS is created. Note that only one specialty is scheduled for an OR on a given day. We also observe that the resources dedicated to non-elective and elective patients are not overlapping. In this case, they are using dedicated resources to handle non-elective demand.

We will mainly focus on scheduling elective patients in this report but note that the scheduling process for elective and non-elective patients is quite similar, as shown in Figure 2.7. The major differences are that the patients, in theory, do not share the same OR resources. However, in practice, we see that non-elective, specifically green patients, often have to be scheduled in the ORs dedicated to elective patients.

The scheduling coordinators at the intake office perform the scheduling of elective surgeries. The patients are generally scheduled two to three months in advance. When a surgery request is sent, the schedulers will try to match a patient with an OR and surgeon. If they are successful, the patient is placed on the surgical schedule. However, the patient is sent to the waiting list if they cannot fit the surgery into the active surgery schedule. We call all patients sent to the waiting list deferred. Deferring a patient is the same as postponing the decision of deciding surgery date. It is our understanding that the scheduling at the Orthopedic Department today is entirely manual and is not assisted by any automated services.

	Monday	Tuesday	Wednesday	Thursday	Friday
OR-BVS-1	Green emergencies	Green emergencies	Green emergencies	Green emergencies	Green emergencies
OR-BVS-2	Elective foot	Plastics	Plastics	Plastics	Elective foot
OR-BVS-3	Plastics	Plastics	Plastics	Plastics	-
OR-BVS-4	Hand	Plastics	Hand	Arthroscopic	Hand
OR-BVS-5	Arthroscopic	Arthroscopic	Arthroscopic	Arthroscopic	Arthroscopic
OR-BVS-6	Back	Back	Back	Tumor	-
OR-BVS-7	Prosthesis	Prosthesis	Prosthesis	Prosthesis	-
OR-BVS-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 2.8: The master surgery schedule at the orthopedic department as shown in T. R. Bovim (2018).

Cancellations and rescheduling

Sometimes changes to the surgery schedule are necessary. For instance, the surgeon can be sick, there is a too high demand from non-elective patients, or patients can not make it to the appointment. In those cases, some surgeries might have to be canceled. We define a *cancellation* as

the event where a patient, for some reason, has to be removed from the surgery schedule due to unplanned events. Canceled patients sometimes need to be scheduled for a later time.

We define *rescheduling* as when a patient receives a new surgery date without being canceled. More precisely, if a patient receives a new surgery date without first being sent to the waiting list, we say the patient is rescheduled. We could not find out if the orthopedic department actively uses rescheduling to improve the efficiency of the surgery schedules. However, it is our impression that rescheduling is only done on rare occasions, if any. One reason could be that it is currently hard to identify and assess the value of potential reschedulings. It can therefore be hard to justify rescheduling patients.

One way of thinking about the difference between cancellation, deferral, and rescheduling is as follows:

- If the patient's surgery date is updated to a new one, the surgery is rescheduled.
- If a patient is moved from the surgery schedule to the waiting list, the surgery is canceled
- If the scheduling coordinator does not schedule a patient who currently is on the waiting list, then the patient is deferred.

It is important to note that these are the definitions we have decided to use. Authors have defined the terms differently, and we have yet to find one general definition adopted in the literature.

Chapter 3

Literature Review

Health Care services are constantly challenged with an increased focus on more efficient and effective solutions to a wide range of planning problems. According to Hulshof et al. (2012), the pressure on Health Care professionals is rising due to the increasing demand for and cost of healthcare. The high degree of dependency and the complexity of the different planning problems have made a taxonomy for classifying and structuring the problems beneficial. Hulshof et al. (2012) propose a bi-axial taxonomy classification of the planning problems.

In the following review, we will provide an overview of the current research on the field of Health care planning. It is necessary to shrink the size of the problems reviewed, as the set of problems is quite extensive. We will focus on research covering Surgical Case Scheduling and, more specifically, Advance Scheduling. Our review is inspired by and builds on the four literature reviews provided by Hulshof et al. (2012), Samudra et al. (2016), Harris and Claudio (2022) and Shehadeh and Padman (2022).

First, in Section 3.1, the theoretical framework of classification and structuring of planning problems proposed by Hulshof et al. (2012) is presented. Second, in Section 3.2, we present relevant literature on Surgical Case Scheduling and Advance Scheduling. Third, in Section 3.3, methods used in the literature for solving the Advance Scheduling Problem are presented. Lastly, in Section 3.4, we will position our report and describe our contribution to the literature. Section 3.1 and Section 3.2 are inspired by Schiøtz and Tysse (2022), while Section 3.3 and Section 3.4 are original to this thesis.

3.1 Taxonomy

Hulshof et al. (2012) proposes a bi-axle taxonomy. The goal of the taxonomy is to make identifying, breaking down, and classifying the different planning and control decisions easier. The taxonomy

Table 3.1: The different health care services are defined based on their characteristics.

Health Care Services		
Health care service	Description	Example
Ambulatory	Provides care without offering a room or bed in a ward	Outpatient clinics
Emergency	Treatment and evaluation of emergent medical cases	Trauma centers
Surgical	Surgeries to repair injuries, cure diseases, and fix deficiencies	Operating theaters
Inpatient	Care to patients offered in the wards during a stay	Nursing wards
Home care	Multiple and coordinated services offered at the patients' homes	Medical care at home
Residential	The assisting and monitoring of the daily life of patients	Nursing homes

has two axes; a vertical and a horizontal one. The vertical axis describes the hierarchical ordering where problems are subdivided into strategic, tactical, and operational problems based on the necessary level of detail and the planning length. The operational problems are further divided into offline and online problems. An overview of the taxonomy can be seen in Figure 3.1

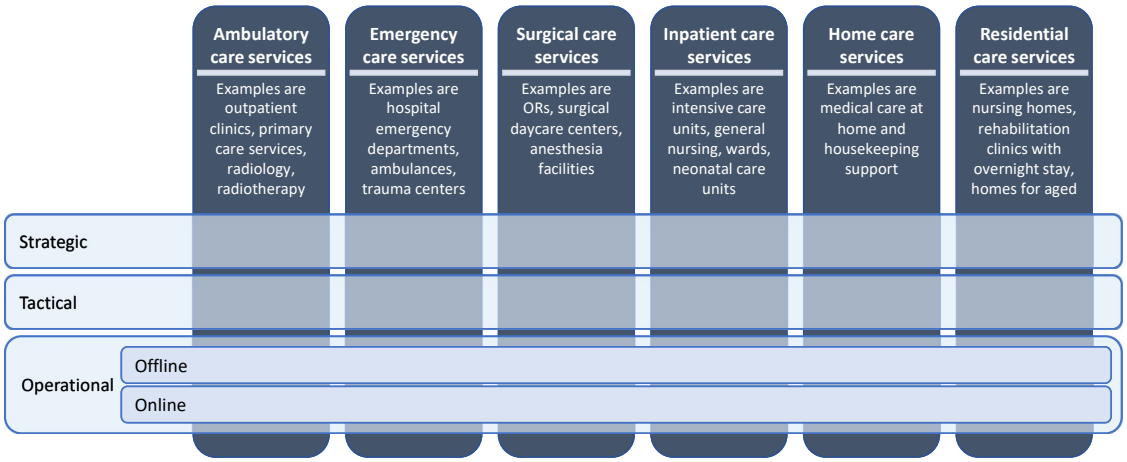


Figure 3.1: The taxonomy proposed by Hulshof et al. (2012)

Six different clusters of health care services are proposed by Hulshof et al. (2012), representing the horizontal axis. The different care services are ambulatory, emergency, surgical, inpatient, home care, or residential, described in Table 3.1. Note that the horizontal division is based on the different services, not the different service providers. As a result, some problems in different services are interrelated and share resources. For example, a diagnosis department might be used by ambulatory and emergency services (Hulshof et al. 2012). Decisions made within one healthcare service can also define some restrictions observed in other services. For instance, the number of surgical cases defines the demand for beds in the wards, while the inpatient services define the number of beds available. Some kind of alignment is hence probably beneficial.

The vertical axis defines the hierarchical structuring of different control and planning problems

for each healthcare service. The problems are clustered based on the level of detail and how information is aggregated. As the planning horizon increases, the need for simplifications increases, and information is aggregated. The problems described in order of longest to shortest planning horizon are; strategic, tactical, offline operational, or online operational. Following is a description of the hierarchy and associated planning problems for surgical care services.

Strategic problems represent the highest level of decision-making. Information is often simplified and aggregated to a high degree, and the planning horizon is long. The end goal is to define the mission and long-term strategy for the surgical units. Hulshof et al. (2012) mentions five strategic planning problems: regional coverage, service mix, case mix, capacity dimensioning, and facility locating. Capacity dimensioning is essential for our problem. In this planning problem, the goal is to match the hospitals' resources with the expected demand. According to Hulshof et al. (2012), the resources to consider are the number of operating rooms, operating time capacity, presurgical wards, ambulatory surgical wards, equipment, and staff. All of these resources affect how the orthopedic department can operate. For example, quickly changing the number of surgeons or creating a new OR is challenging. Therefore, the decisions made at the strategic level clearly define some boundaries for the departments.

The goal of tactical problems is to translate the strategic decisions into guidelines to be used at the operational level (Hulshof et al. 2012). It is typical to look at different patient groups at the tactical level and not at specific surgical cases. Different examples of tactical problems are staff-to-shift, patient routing, capacity allocation, and admission control. In staff-to-shift, the goal is to decide what kind of shift to use and how many persons of different staff types are needed. The patient routing controls how a patient moves through the system. The movement is often broken down into several steps. Some common steps are the pre-, peri-, and postoperative stages (Hulshof et al. 2012). An efficient patient routing policy can increase the bandwidth, processing speed, and overall resource utilization. An important tactical problem is the Capacity allocation problem. The goal is to subdivide the resources mentioned at the strategic level over different patient groups. It is possible to break the problem into three steps: (1) patient group identification, (2) time subdivision, and (3) block scheduling. Another term for the block scheduling problem is the Master Surgery Scheduling Problem (MSSP). The MSSP is a highly researched topic. The task is to subdivide the available OR capacity and assign dates and times to the different blocks. The resulting block schedule is relevant at the operational level. It is worth pointing out that all tactical problems are highly interrelated. For instance, block scheduling is dependent on the available staff. Some interrelationships exist between these problems and other tactical problems within the inpatient vertical. For instance, the number of planned surgeries depends on the availability of the nursing wards. Some authors have tried integrating different upstream and downstream constraints into the MSSP (Hulshof et al. 2012). Another way to handle the interrelationships is to add steps to the MSSP. Two additional steps proposed by Hulshof et al. (2012) are Temporary Capacity Change and Unused Capacity (re)Allocation. Both problems can be seen as an extension of the capacity

allocation problem where temporary changes are made to the decided capacity allocation based on short-term changes to the available resources, utilization, or demand. In that way, it is possible to increase the agility of the system and make the boundaries less rigid.

The operational planning problems are short-term and utilize the guidelines defined at the tactical level. The problems are related to the scheduling and planning based on specific entities, e.g., a specific surgeon, nurse, room, bed, or patient, unlike at the tactical level, where these entities are aggregated into more generalized groups. Hulshof et al. (2012) further divides operational planning problems into offline and online problems. The offline problems represent all the planning done on the patient and resource level in advance. It can be seen as the specification of the blueprints made at the tactical level. On the other hand, the online operational planning problems handle the unplanned events that arrive due to the inherently uncertain nature of health care systems (Hulshof et al. 2012).

The Surgical Case Scheduling problem (SCSP) is an important offline problem. The problem consists of four steps: (1) estimating the surgery duration, (2) assigning dates and operating rooms to surgical cases, (3) sequencing surgical cases, and (4) assigning starting times to surgical cases and is described by Hulshof et al. (2012). Alternatively, Magerlein and Martin (1978) propose a two-phase decomposition of the SCSP where the first problem is called the Advance Scheduling Problem (ASP) and the second problem is called the Allocation Scheduling. The description of the two problems according to Cardoen et al. (2010) is that Advance Scheduling is the process of fixing a surgery date for a patient, whereas Allocation Scheduling determines the operating room and the starting time of the procedure on the specific day of surgery.

Another relevant problem closely related to the SCSP is the Surgical Case (re)Scheduling Problem (SCrSP). The SCrSP, as an online operational planning problem, is about efficiently tackling unplanned events. Hulshof et al. (2012) says the surgical case schedule often has to be reconsidered during the day and points to emergency patient arrivals, surgery duration uncertainty, and equipment breakdown as examples of unplanned events. Hulshof et al. (2012) does not specify the planning horizon for the SCrSP, but it is our understanding that the SCrSP is aimed at solving step four in the taxonomy, assigning starting times to surgical cases. There should, however, not be any reasons why the SCrSP could not be applicable to any previous steps, including the ASP.

3.2 Aspects of Advance Scheduling

Several literature reviews have been published in the last two decades about surgery scheduling in addition to the one included in the taxonomy paper by Hulshof et al. (2012). The reviews differ in the time periods reviewed and how the existing literature is classified. Cardoen et al. (2010) classified 247 papers published between 2000 and 2009 and included six categories. Later, Samudra et al. (2016) builds on Cardoen et al. (2010) work by including papers covering 2004 and 2014.

They include all six categories and add upstream/downstream units and relationships between categories as categories. Harris and Claudio (2022) reviewed 246 papers published between 2015 and 2020. They note that increased model complexity is a major theme in the last year and illustrate this by adding a complexity score to all reviewed papers in addition to three other new categories. While the reviews by Hulshof et al. (2012), Samudra et al. (2016) and Harris and Claudio (2022) broadly cover the surgery scheduling literature, Shehadeh and Padman (2022) focus specifically on stochastic optimization approaches for elective surgery scheduling and downstream capacity planning. Shehadeh and Padman (2022) reviews the literature on the Advance Scheduling Problem and Allocation Scheduling Problem published up to 2020. A comparison of the reviews by Hulshof et al. (2012), Samudra et al. (2016), and Harris and Claudio (2022) is shown in Table 3.2.

Table 3.2: The different categories present in different existing reviews as described by Harris and Claudio (2022). Only the most relevant categories are displayed.

	Cardoen et al. (2010)	Samudra et al. (2016)	Harris and Claudio (2022)
Years covered	2000-2009	2004-2014	2015-2020
# Articles Reviewed	247	216	246
Patient type	x	x	x
Performance measure	x	x	x
Decision delineation	x	x	x
Scheduling policy			x
Planning horizon			x
Upstream/downstream units		x	x
Uncertainty	x	x	x

We have decided to base our review on Samudra et al. (2016), Harris and Claudio (2022), and Shehadeh and Padman (2022). They have all together reviewed over 500 papers published between 2004 and 2020. Their classification has been combined with Hulshof et al. (2012) taxonomy to identify papers investigating the ASP and SCrSP. We have also reviewed some papers later than 2020. A summary of all papers reviewed can be found in Figure 3.3 and Figure 3.4. The following sections present the key findings.

3.2.1 Decision delineation

The decision delineation describes what kind of decision(s) the model makes. Harris and Claudio (2022) propose a taxonomy for decision delineation of surgery scheduling using a two-axis approach. The horizontal axis represents discipline, surgeon, and patient-level decisions. The vertical axis consists of data, time, and room. The two dimensions combined describe what kind of decision and to whom the decision applies. Harris and Claudio (2022) argue that this taxonomy represents a more comprehensive way to examine the decision delineations than proposed by Hulshof et al. (2012) and found it easier to use since many authors differ in the use of the problem definitions in Hulshof et al. (2012).

We define a minimum requirement for a formulation to address the ASP: the model must assign a date on the patient level. A few of the papers identified use only the bare minimum requirement (Gul et al. (2015), Zhang et al. (2020), Wang et al. (2016)). Some papers extend the formulation by adding a case-to-block decision, which means that a surgical case is assigned to a date and an operating room (Addis et al. (2015), Al-Refaie et al. (2018), Rachuba and Werners (2017), Ozcan et al. (2017)). The possibility of deciding to open or close blocks is formulated by Zhang et al. (2020) and reused in a later formulation (Zhang et al. 2021). Al-Refaie et al. (2018) was the only one identified to include a decision of assigning cases to a downstream unit, which is surprising as several papers include downstream constraints, as further discussed in section 3.2.5. They did so by assigning each case to an ICU bed.

Most of the literature assumes a fixed tactical scheduling policy, but Moosavi and Ebrahimnejad (2020) formulates a model that solves the MSSP and the ASP in an integrated fashion. They create an MSS and schedule surgical cases to days by assigning them to blocks. Moosavi and Ebrahimnejad (2020) find that simultaneously solving the tactical and operational problems led to better assignments of surgical cases and utilization of hospital resources. This is in line with expectations based on the inter-dependency of the two problems (Harris and Claudio (2022), Samudra et al. (2016), Hulshof et al. (2012)).

Some authors also formulate a model that integrates the ASP and the Allocation Scheduling Problem to solve the SCSP effectively. For example, Haghi et al. (2017) assumes an open scheduling policy and propose a single-step formulation where cases were assigned both a surgery date and time. They also assign a surgeon to each case to ensure a surgeon is never assigned to overlapping surgeries. Ozcan et al. (2017) uses a two-step formulation where they first assign cases to blocks in an iterative manner. In the next step, they assign the admission date, OR, and surgery start time.

3.2.2 Performance measure

Harris and Claudio (2022) discuss several performance measures. The main performance measures mentioned are based on waiting time, leveling the use of resources, utilization, idle time, throughput, preferences, makespan, cancellations, and deferrals/postponement. It is common to focus only on some of the performance measures mentioned above since too many exist to be used in one formulation. The diversity of performance measures also exemplifies how the amount of different stakeholders increases the overall complexity. No single objective exists that aligns the interests of all stakeholders simultaneously. Patients want short waiting times, the staff wants to avoid overtime, and the management wants to reach financial targets, high utilization, and more.

It is common to add case-specific performance measures. Most papers minimized waiting time; however, Ozcan et al. (2017), Gul et al. (2015), and Zhu et al. (2015) do not include waiting time

as a performance measure. Ozcan et al. (2017) had a hospital-centric approach where they instead maximized the total number of patients served (throughput). They implicitly included waiting time by prioritizing patients based on the percentage of the maximum time a patient can wait without suffering detrimental consequences to their health that elapsed while waiting.

In the investigated literature, the objective value incurred for scheduling a case in the plan is generally called the waiting cost. Waiting cost can mean both the cost of getting a surgery date and the cost of getting deferred. Different implementations of waiting costs are used in the literature. Zhang et al. (2021) use a weighted approach where the combined cost of waiting and the urgency of the patients were included. They did so by having a priority score for each surgical case. Rachuba and Werners (2017) used a robust optimization approach and minimized the maximum waiting time for those cases not deferred to the next planning phase. They found that waiting time and the number of deferrals were correlated. The number of deferrals increased when using the min-max formulation compared to the more standard formulation, where the accumulated total waiting time was used. The weighted approach achieved a higher utilization than the min-max formulation. Interestingly, Moosavi and Ebrahimnejad (2020) found that the number of deferred cases decreased by using robust optimization compared to a deterministic model. Moosavi and Ebrahimnejad (2020) also uses a nonlinear waiting cost where the number of waiting days is squared, further incentivizing scheduling cases with longer waiting time first.

We find that the research differs in how the deferred patients are modeled. For example, Addis et al. (2015) and (Moosavi and Ebrahimnejad 2020) assume that all deferred cases are scheduled for surgery at the first day after the current planning horizon, denoted timestep T+1, and the cost function only minimize waiting time. Others differentiate between deferred and scheduled cases and explicitly minimize the number of deferred cases ((Kamran et al. 2018), (Wang et al. 2016), (Rachuba and Werners 2017)).

Only two of the investigated papers include staff-specific performance measures. Rachuba and Werners (2017) minimize the overtime for surgeons. They confirm the existence of a trade-off between the amount of overtime for each surgeon, the number of deferred patients, and waiting time. They also find that the min-max approach for waiting time cost decrease surgeons' overtime. They argue that this is because the min-max formulation is more conservative than the more standard formulations. Kamran et al. (2018) minimize the number of days a surgeon has to perform surgery. The reason is that surgeons want to perform surgeries continuously to stay efficient. They formulate a two-stage stochastic model where surgical cases and surgeons are assigned to blocks in the first stage, and the amount of overtime is the second stage decision. They find that the deterministic counterpart of the stochastic model often produces an infeasible amount of overtime. Adding a chance constraint improves the results considerably.

It is common to find the usage of the ORs as performance measures. Haghi et al. (2017), Gul et al. (2015), and Kamran et al. (2018) minimize the amount of OR overtime. Others minimize

OR undertime ((Ozcan et al. 2017), (Zhang et al. 2021)). Zhu et al. (2015) investigate different strategies for maximizing the OR utilization. In their formulation, operated patients have to stay in the ORs if no beds in the PACU are available. They find that operating the patients with the longest expected anesthesia recovery times reduced OR idle time. Wang et al. (2016) and Zhang et al. (2020) minimize the number of ORs that has to be opened to cover the demand. Wang et al. (2016) find that deterministic models considerably underestimate the amount of needed overtime compared to stochastic formulations. On the other hand, they do not find the impact of stochastic surgery times overly disruptive to throughput or OR utilization.

Other more uncommon performance measures are also identified. Haghi et al. (2017), Zhang et al. (2020), and Al-Refaie et al. (2018) minimize cancellation costs. Several authors include cancellations in their models but do not explicitly use them as a performance measure ((Addis et al. 2015), (Gul et al. 2015), (Wang et al. 2016), (Ozcan et al. 2017)). Moosavi and Ebrahimnejad (2018) minimize the number of additional ward beds that has to be opened. They find that increasing the number of beds in the ICU and wards can, in some cases, reduce the amount of idle OR time by 36.58 % when applied to a real case study. Adding additional beds in some wards can also potentially decrease the number of deferred patients by 58.33% (Moosavi and Ebrahimnejad 2018).

3.2.3 Patient types

Two primary patient types are identified in the literature, elective and non-elective. Elective cases are defined in different ways, but Harris and Claudio (2022) define them as cases that can be planned before the day of surgery. The elective cases are either inpatients or outpatients based on their recovery needs. Cases with a LOS over a day are defined as inpatients. Samudra et al. (2016) splits non-elective patients into two categories, urgent and emergent, where the emergent cases are assumed to need surgery right away, while the urgent cases can wait for a few hours. Both Harris and Claudio (2022) and Samudra et al. (2016) point out that a consistent definition is lacking and other patient segmentation types exist. There is not enough research on patient bulking, according to Samudra et al. (2016). Patient bulking happens when an elective patient at any point in time leaves the waiting list. The risk of patient bulking can be different for all elective cases and could indicate that other, more granular, segmentation of elective patients could be beneficial.

Several reviewed articles include non-elective cases in their formulations. Ozcan et al. (2017) includes the expected number of non-elective cases occupying recovery beds in the wards. Both Addis et al. (2015) and Kamran et al. (2018) include an occupational parameter to leave a certain fraction of the available OR time free to manage the arrival of non-elective cases. Moosavi and Ebrahimnejad (2018) include the expected non-elective demand in an offline setting. They argue that even though the demand is not known in advance, ignoring it could result in inefficient schedules. They, therefore, add the expected non-elective demand as placeholders in the surgery

schedule instead of keeping a fraction of the available capacity unutilized. This way, it could be possible to incorporate the expected emergency demand most efficiently based on the current surgery schedule. Wang et al. (2016) find that non-elective demand in a stochastic setting increases the number of surgery cancellations by 5% and results in a drop to 72.2% in OR utilization in their experiment.

3.2.4 Planning horizon

The planning horizon depends on the problem it is trying to solve, as discussed in section 3.1. Harris and Claudio (2022) identifies that it is common to find different planning horizons in papers trying to solve the same problem. They, therefore, add the planning horizon to the categorization, in contrast to Samudra et al. (2016) and Cardoen et al. (2010). Harris and Claudio (2022) distinguish between a few days or weeks, one month, 2-5 months, and more than six months. It is also sometimes hard to define the planning horizon for a given model (Harris and Claudio 2022). Most papers tackling the ASP use a planning horizon as short as one day to as many as two weeks. The exceptions are Ozcan et al. (2017) who use a one-year planning horizon, Addis et al. (2015) who use four weeks, and Wang et al. (2016) who use 26 weeks. In practice, scheduling surgical cases much longer in advance than two weeks is necessary. During our interview, a scheduler at Rikshospitalet (Flaata 2022) stated that giving an elective patient at least four weeks' notice before surgery is necessary. At the same time, it is difficult to evaluate whether the reviewed models cannot handle even longer planning horizons or whether they just have not been tested on different planning horizons.

3.2.5 Upstream and downstream units

Samudra et al. (2016) and Harris and Claudio (2022) categorize papers based on what kind of up- and downstream units are incorporated into the planning problem. The OR schedule is dependent on multiple resources outside the OR, like the recovery wards, post-anesthesia care unit (PACU), intensive care unit (ICU), and pre-surgery beds, as discussed in section 3.1. Vertical integration could therefore make the models more realistic. Harris and Claudio (2022) find that the fraction of papers considering at least one upstream or downstream unit has stagnated at around 50% in recent years, even though the overall complexity of the models has increased. They urge future research to incorporate one upstream or downstream unit and note the research regarding the upstream (preoperative) units as especially lacking. The interviewed scheduler Flaata (2022) remarks that the lack of bed capacity is a common reason surgeries must be canceled or rescheduled. This further demonstrates the potential benefits of incorporating the wards in the ASP.

Moosavi and Ebrahimnejad (2020) believes there are three pathways after surgery:

- OR → Ward → Discharge (around 95%)
- OR → ICU → Ward → Discharge (around 5%)
- OR → ICU → Discharge (above 1%)

Around 95% of all outpatients require a stay at the wards post-surgery. It indicates that the recovery wards' utilization highly depends on the surgical schedule. A visualization of the potential patient flows from arrival to discharge is displayed in Figure 3.2

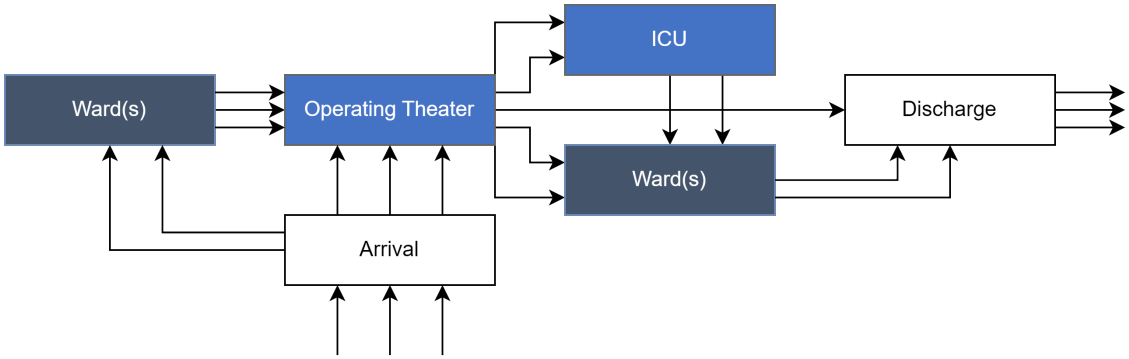


Figure 3.2: Example of patient flows from arrival to discharge.

There are multiple different formulations of downstream bed resources in the existing literature. Addis et al. (2015) only schedules with an isolated OR approach but highlights the significance of downstream bed resources in surgical case scheduling and recommends including those resources in future scheduling formulations. Al-Refaie et al. (2018) integrates the ward capacities in a simplified manner. They assume that all wards are homogeneous, substituting the typical wards constraints with a bed capacity constraint. The practical interpretation is that the hospital has a certain amount of recovery beds, and all inpatients must be assigned to one and only one bed. Zhang et al. (2020) and Zhang et al. (2021) however includes heterogeneous wards and ICU. Interestingly, they propose more lenient ward constraints than others. They include the option of admitting patients into the wrong wards and penalize it due to the expected lower quality of care.

Ozcan et al. (2017) formulate a shared bed capacity between electives and non-elective patients. They find that the ORs are the bottleneck in their case. This shows that ward capacity only becomes relevant when the OR throughput is higher than the wards can handle. Ozcan et al. (2017) also use the model to find the optimal ward capacities by identifying the different process bottlenecks. Interestingly their findings indicate that the average patient waiting time can be reduced by slightly increasing the number of beds, even when no change in OR and ward utilization is observed. The reason could be that the bed allocation becomes more flexible, enabling better combinations of patients.

We do not find the inclusion of upstream resources as regular as downstream resources. Moosavi and Ebrahimnejad (2020) include the ICU, recovery wards, and pre-surgery wards in their formulation. They allow acquiring extra beds up to a certain limit. When applied to a case study from a hospital in Iran, they find that the optimization model performs 12% better than the standard OR scheduling when it comes to deviation from the average number of beds used in the ICU and about 2.5 times better for regular recovery wards. Their findings indicate that a model including upstream and downstream units can reduce upstream and downstream bed utilization variance by 50.6% compared to traditional OR scheduling approaches. Not only does the variance become better, but the utilization also increases. Wang et al. (2016) include the same up- and downstream constraints as Moosavi and Ebrahimnejad (2018), but also include the PACU in the formulation. They find similar results as Moosavi and Ebrahimnejad (2018) where the inclusion of recovery wards has the highest impact on the original schedule. The different schedules do not significantly impact PACU utilization, which suggests that the inclusion of the PACU is not a key performance indicator in at least some instances. Further, Wang et al. (2016) also find that limited resources in downstream units result in a 23% increase in surgery cancellations in their experiment. This suggests that the number of cancellations depends on how well the surgery schedule is optimized in combination with the wards. Similarly, they also found that OR utilization dropped from 81.4% to 50.1%. This indicates that OR schedules that do not adhere to downstream resource constraints might be overly optimistic. They also find, similarly to Ozcan et al. (2017), that increasing the number of beds in wards that are not generally scarce does not improve the scheduling. Thus, the mentioned results of increased cancellations and reduced OR utilization assume limited downstream resources.

Haghi et al. (2017) propose a MIP formulation that includes both equipment and PACU resource capacity constraints. They use the model to solve the weekly planning problem and hence do not test it on longer planning horizons than one week. Nine recovery beds in the operating theater are included. While not specifying what the beds were used for, we assume they are equivalent to PACU beds since all patients were sent to the beds directly after surgery, and the expected recovery time was between 20-40 minutes.

We have searched for articles including hospital equipment, OR preparation, and cleaning. While not directly categorized as up- or downstream resources, they are still relevant aspects not usually covered in the literature. Al-Refaie et al. (2018) use homogeneous ORs, but each procedure can require some specialized equipment. Therefore, they make it necessary to ensure all required equipment is available before scheduling a case to a given OR, day, and time. One common finding in two other papers by Haghi et al. (2017) and Gul et al. (2015) is that the equipment included in the formulations is assumed to be OR specific. This means that there is no shared equipment between different ORs. In addition, certain procedures can only be performed in specific ORs. The resulting formulations thus become pretty similar to normal MSS restrictions as long as equipment requirements are assumed to be considered while constructing the MSS. Lastly, Al-Refaie et al.

(2018) interestingly explicitly divides the surgery duration into three parts: (1) OR preparation, (2) the surgery act, and (3) cleaning. Most other papers make simplifications, assuming preparation and cleaning to be included in the surgery duration.

3.2.6 Scheduling policy

Several scheduling policies are discussed in the literature and are often a key research topic in the MSSP. The scheduling policies are also relevant at the operational level as they define some boundaries on scheduling surgical cases. Block-, open-, and modified block schedules are three commonly seen policies. Open scheduling can, for instance, use a FIFO system. The open scheduling policy is the most flexible, but critics argue it increases the competition for OR time (Hulshof et al. 2012). Van Riet and Demeulemeester (2015) find that flexible scheduling policies increase utilization and decrease non-elective waiting time. However, the waiting time for elective cases and the number of cancellations also increase. Another possible scheduling policy is block scheduling. Here, OR times are subdivided over specialties or even surgeons. The benefits are that demands for OR time and downstream resources are more predictable. Modified block scheduling is a mix of open and block scheduling where some OR capacity is held open in the block schedule and divided at a later stage. Block scheduling is the most common policy found in general. Several of those papers investigating the ASP also use block scheduling ((Moosavi and Ebrahimnejad 2018), (Moosavi and Ebrahimnejad 2020), (Ozcan et al. 2017), (Zhang et al. 2020), (Zhang et al. 2021), (Rachuba and Werners 2017), (Addis et al. 2015), (Min and Yih 2010), (Zhang et al. 2019), (Zhang et al. 2020), (Neyshabouri and Berg 2017), (Shehadeh and Padman 2021)). A few uses open scheduling ((Wang et al. 2016), (Haghi et al. 2017), (Al-Refaie et al. 2018)). Modified block scheduling is the least common according to Harris and Claudio (2022). Kamran et al. (2018) is the only paper we found that uses modified block scheduling. Kamran et al. (2018) starts with an MSS similar to normal block scheduling, where each block is assigned to a subspecialty. Then, a subset of procedures that can be assigned to blocks that do not belong to their subspecialty is defined. Kamran et al. (2018) conclude that modified block scheduling policies seem more efficient than open and block scheduling policies.

According to Samudra et al. (2016), hospital scheduling policies are often static, i.e., patients are placed on a waiting list. They argue the need for increased focus on dynamic scheduling policies that adapt to changes like overbooking, rescheduling based on patient availability, and no-shows. When using a dynamic scheduling policy, the patients get their time at the moment the request is received (Harris and Claudio 2022).

3.2.7 Uncertainty, cancellations & rescheduling

There is inherently much uncertainty in the ASP, and many papers try to, in some way or another, incorporate this uncertainty into their models. Uncertain surgery duration, LOS, and non-elective arrival is the main reason for rescheduling and cancellation in the existing literature, and those effects are often researched together. Therefore, we also include some findings regarding rescheduling and surgery cancellations here. Harris and Claudio (2022) mention uncertain surgery duration, the arrival of cases, LOS, case cancellations, turnover time, and surgeon availability as causes for uncertainty. Stochastic models such as robust-, chance constraint-, and stochastic programming are more widely adopted than deterministic models in the literature, according to Zhang et al. (2021). Harris and Claudio (2022) identify, however, only a handful of papers that address the uncertain demand by elective patients. One reason for this trend could be that it is uncommon to look at specific surgical cases in strategic and tactical problems, while those papers that tackle the ASP often only have a 1-2 weeks planning horizon at max. We believe that the effect of uncertain surgery demand is lower with such short planning horizons since the waiting lists in those cases usually would be pretty full. However, the uncertain arrival of electives could affect the models more as the planning horizon increases and we move closer to a more dynamic scheduling setting seen in practice.

Wang et al. (2016) assume uncertain surgery duration, non-elective arrivals, and LOS. Comparing their stochastic model to a deterministic counterpart using only expected values, they find that the deterministic model overestimates the potential benefits and performs worse in a stochastic simulation environment. Interestingly, however, they find that if the schedule contains much overbooking, the deterministic schedule performed not only better than prior, but the schedule may accurately predict the efficiency gains even when ignoring uncertainty. Overbooking, in this case, is when a surgical case gets booked in the schedule with a surgery duration longer than expected. When ignoring uncertainty, Wang et al. (2016) recommends increasing the booked time for a random selection of 70% of the surgeries by around 30 minutes.

Gul et al. (2015) formulate a multi-stage stochastic mixed-integer program (MSSMIP) with uncertain surgery duration to solve the allocation of surgeries to ORs over a finite planning horizon under uncertainty. They analyze the trade-offs between cancellation, waiting, and overtime costs. A few interesting findings are discovered:

(1) *The ratio between the waiting and cancellation costs is the factor with the most impact on the number of cancellations.* If the waiting cost is higher than the cancellation costs, then cancellations are more likely. This is what we would expect as tighter packing of surgical cases is assumed to increase surgery throughput and, on average lower waiting times, while the risk of OR overtime, and hence cancellations, also increases.

(2) *The number of cancellations is highly sensitive to the level of uncertainty in demand and*

surgery duration. Gul et al. (2015) argue that the number of cancellations increases as the level of uncertainty also increases. This makes sense intuitively. If we expect a high level of variance, it is harder to predict the outcome before the day of surgery; hence, the risk of cancellations (and rescheduling) increases.

(3) *The number of cancellations could decrease if surgeries with similar expected duration are performed together in the same blocks.* They argue that minimizing the maximum variance of total surgery duration could benefit the number of cancellations. We believe the reason could be closely related to what Wang et al. (2016) suggested as described previously. Assuming normal distributions, we would expect procedures with similar expected duration to be as likely to go overtime as undertime. The amount of overtime and undertime could balance through a day, resulting in little expected block overtime and fewer cancellations.

In the literature, three main ways of accounting for uncertainty are observed: (1) robust optimization, (2) two-stage / multi-stage stochastic models, and (3) simulation. Addis et al. (2015) adopts a robust optimization to account for uncertain surgery duration. A rolling-horizon approach with simulation is used to evaluate the model to account for uncertain patient arrival. They note that although the surgery duration is drawn from probability distributions, their model is not stochastic. Results indicate that when starting with a deterministic model and moving towards more strict robustness requirements, the schedule achieves both lower utilization and fewer surgery cancellations. This aligns with what is found by Gul et al. (2015). Neyshabouri and Berg (2017) formulates a robust model with uncertain parameters for surgery duration and LOS. The robust formulation is inspired by Bertsimas and Sim (2004) and uses varying robustness for each patient specialty. The robust model assumes that only a subset of the uncertain parameters will deviate from their nominal value, and the objective is to minimize the worst-case costs (Neyshabouri and Berg 2017).

Min and Yih (2010) was the first to propose a two-stage stochastic model for the ASP with downstream constraints, according to Shehadeh and Padman (2022). Min and Yih (2010) include uncertain surgery duration and LOS as the second stage parameters. While exceeding the opening hours of a block is allowed by incurring an overtime cost, the ward constraint is hard in the formulation by Min and Yih (2010). Jebali and Diabat (2015) generalize the model from Min and Yih (2010) by adding the option to exceed the ward capacity by incurring a penalty to the objective, as well as penalizing OR idle time. Jebali and Diabat (2017) include a chance constraint in the formulation, considering the risk of cancellations due to lack of capacity in the ICU. Zhang et al. (2019) combines a two-stage model with a Markov Decision Process (MDP) in a two-level optimization model. By using an MDP first to select which patients to schedule each week, the complexity of the two-stage model is drastically reduced. In addition, the MDP can consider future events in a way that the standard two-stage models can not. The two-stage models found in the literature do not consider the fact that the model will be rerun later and that the possible solutions at the later stage are affected by the current actions, whereas the MDP can accomplish this to

some degree.

The relationship between uncertainty and planning horizons is investigated by Zhang et al. (2021). They assume uncertain surgery duration and LOS and compare two stochastic models: one two-stage stochastic that tackles the intra-week scheduling of surgical cases and one two-phase approach, wherein the first phase, Markov Decision processes are used to decide which cases to be scheduled for each week, and in the second phase the weekly assignment is performed. They find that models with longer planning horizons produce significantly better schedules than models with shorter ones. An essential factor discovered is hence that shorter planning horizons increase long-term waiting time and that future research should incorporate longer planning horizons compared to what is usual in existing literature (below two weeks, as described previously in 3.2.4)

The distributions used when including uncertainty are relevant for the end results, and several distributions have been proposed in the literature. Three different distributions used to model the LOS is frequently observed: normal distribution (Ozcan et al. 2017), truncated log-normal distribution (Zhang et al. 2020), and log-logistic distribution (Wang et al. 2016). Interestingly, Zhang et al. (2020) uses truncation to avoid long tails, while Wang et al. (2016) uses log-logistic distributions, which have heavier tails than log-normal distributions. Log-normal surgery duration is the most commonly used of the researched papers ((Zhang et al. 2020), (Rachuba and Werners 2017), (Addis et al. 2015)). (Zhang et al. 2020) also truncate the distributions to avoid the possible edge cases of highly unlikely long surgery duration that a typical log-normal distribution could have given. They find that the results are overly conservative when not truncating the distributions. Other distributions observed for surgery duration are the triangular distribution ((Ozcan et al. 2017), (Kamran et al. 2018)) and uniform distribution ((Moosavi and Ebrahimnejad 2018), (Moosavi and Ebrahimnejad 2020)). Min and Yih (2010) use a discrete distribution based on sampled data from real life. The distribution is discretized into 30-minute intervals. The demand, both by elective and non-electives, is exclusively modeled by a Poisson distribution in the literature ((Ozcan et al. 2017), (Zhang et al. 2021), (Rachuba and Werners 2017)). Ozcan et al. (2017) also use negative exponential distributions to predict the patient inter-arrival times.

3.3 Solution Methodologies

This section presents a selection of solution methods relevant to the ASP. While purely deterministic models like the one formulated by Schiøtz and Tysse (2022) can solve real-life sized instances of the ASP using an exact solver with methods like Simplex (Dantzig 1990) and Branch and Bound (Lawler and Wood 1966), stochastic programs of real-life size are usually too complex to solve using an exact method, even using a crude discretization of the random parameters (Zhang et al. 2019). Section 3.3.1 gives an in-depth presentation of the Sample Average Approximation (SAA) technique, commonly used for solving stochastic programs. Lastly, in Section 3.3.2, we briefly

discuss other solution techniques found in the surgery scheduling literature.

3.3.1 Sample Average Approximation

Sample Average Approximation (SAA) is a technique for solving stochastic programs. Instead of including all possible combinations of realizations of the uncertain variables in the model, a sample of N randomly generated scenarios is generated from the distributions (Shehadeh and Padman 2022). The reason for using SAA is that modern computers are not able to solve reasonably sized problems with a large number of scenarios. When $N \to \infty$, the SAA model's objective value converges to the original problem's optimal objective with probability 1 (Shehadeh and Padman 2022). We now present a variant of the SAA algorithm proposed by Min and Yih (2010).

Consider the general stochastic program with the objective function $F(x)+E[Q(x,\xi)]$, where $F(x)$ is the first stage cost, and $Q(x,\xi)$ is the second stage cost, depending on the uncertain parameter ξ. Let us assume that for any given first-stage solution $\overline{x}$, and a given realization of the uncertain variables, n, it is possible to calculate the optimal second-stage solution, and thus the objective value $Q(\overline{x},n)$. We can now perform the following steps:

Step 1: for $m = 1,\ldots,M$, perform steps 1.1 through 1.3.

Step 1.1: Generate N samples.

Step 1.2: Solve the two-stage model with these N samples, and let $\hat{z}_N^m$ and $\hat{x}_N^m$ be the objective value and first-stage solution obtained from the model, respectively.

Step 1.3: Generate N' independent random samples, where $N' \gg N$. Evaluate the true objective value $\hat{g}_{N'}^m$ and estimate of variance $\sigma^2_{\hat{g}_{N'}^m}$ by using equations (3.1) and (3.2), respectively. Note that the scenario index ξ is swapped with n to clarify that it is a sampled scenario.

$$\hat{g}_{N'}^m = F(\hat{x}_N^m) + \frac{1}{N'}\sum_{n=1}^{N'} Q(\hat{x}_N^m, n) \tag{3.1}$$

$$\sigma^2_{\hat{g}_{N'}^m} = \frac{1}{N'(N'-1)}\sum_{n=1}^{N'}\left(F(\hat{x}_N^m) + Q(\hat{x}_N^m, n) - \hat{g}_{N'}^m\right)^2 \tag{3.2}$$

Step 2: Calculate the mean objective value $\overline{z}_N^M$ and variance estimate $\sigma^2_{\overline{z}_N^M}$ by using equations (3.3) and (3.4), respectively.

$$\overline{z}_N^M = \frac{1}{M}\sum_{m=1}^{M} \hat{z}_N^m \tag{3.3}$$

$$\sigma^2_{\overline{z}_N^M} = \frac{1}{M(M-1)}\sum_{m=1}^{M} (\hat{z}_N^m - \overline{z}_N^M)^2 \tag{3.4}$$

Step 3: For each solution $\hat{x}_N^m, m = 1, \ldots, M$, Calculate the estimated optimality gap $\hat{g}_{N'}^m - \overline{z}_N^M$, and an estimated variance $\sigma^2_{\hat{g}_{N'}^m} + \sigma^2_{\overline{z}_N^M}$. Finally, select one of the M solutions based on the calculated metrics.

The values $\overline{z}_N^M$ and $\hat{g}_{N'}^m$ provide a statistically lower and upper bound of the optimal objective value, respectively (Min and Yih 2010). Since the input is the same in all the M iterations, the true optimal objective, z^* is also the same. Since $z^* \leq \hat{g}_{N'}^m \quad \forall m = 1, \ldots, M$, the upper bound for z^* can be set to the lowest value of $\hat{g}_{N'}^m$.

3.3.2 Other methods

When the size of stochastic problems increases, the SAA technique requires more sampled scenarios, significantly increasing the computational complexity (Zhang et al. 2020). Thus, several techniques have been developed to attempt to solve larger problems. Zhang et al. (2020) utilizes column generation in combination with heuristic rules to create a Column Generation Based Heuristic (CGBH). The CGBH is then combined with the SAA algorithm. The resulting CGBH-SAA algorithm significantly outperforms the original SAA algorithm's computational efficiency without reducing the solution quality (Zhang et al. 2020). As mentioned in Section 3.2.7, Zhang et al. (2019) solves a Markov Decision Process (MDP) to decide which patients to schedule each week, making it possible to solve each week individually more efficiently with SAA since many variables and constraints can be removed from the original multi-week problem. Zhang et al. (2021) combines the MDP approach from Zhang et al. (2020) with the CGBH from Zhang et al. (2019), further improving the efficiency compared to the conventional SAA algorithm.

Range et al. (2019) formulates the surgery scheduling problem as a dynamic job assignment problem with a rolling time horizon. A column generation-based method is used to solve the model, where shortest path problems with resource constraints are solved with dynamic programming to generate the relevant columns. The model has better service level and overtime metrics than a First-Come-First-Served policy. However, Range et al. (2019) does not compare the computational complexity to a normal SAA approach.

Column generation and dynamic programming are also commonly used for solving the Cutting Stock Problem (CSP) (Delorme et al. 2016), a variant of the Bin Packing Problem quite similar to the ASP. In short, the goal of the CSP is to cut *stocks* in different lengths to satisfy demand, using as few stocks as possible. The CSP can be formulated as a set covering or partitioning problem, using *patterns*. Delorme et al. (2016) describes a *pattern* as an integer array $(a_{1p}, a_{2p}, \ldots, a_{mp})$, where a_{jp} gives the number of copies of item j contained in pattern p. We will take inspiration from this use of patterns when formulating our mathematical models for the ASP. Instead of cutting a stock, we let the stock be the available time for an operating room on a given day, and instead of cutting lengths of the stock, we assign surgery types with a given duration.

Lastly, we would like to mention Benders decomposition and L-shaped decomposition. According to Kall and Wallace (1994), the two methods are in practice the same, and the term 'L-shaped' is commonly used when dealing with stochastic programs, and 'Benders' is commonly used in other areas of mathematical programming. In short, the method splits the original problem into a master problem and a sub-problem and iteratively generates feasibility and optimality cuts to get closer and closer to the optimal solution. Moving the recourse problem, i.e., second stage problem, to the sub-problem when solving stochastic programs with the L-shaped method is common. According to Rahmaniani et al. (2017), Benders decomposition makes it possible to solve larger problems and often provides a basis for designing effective heuristic solutions that would otherwise not be possible.

3.4 Our Contribution

This thesis investigates the Advance Scheduling Problem (ASP), positioned in the offline operational part within surgical care services in the taxonomy by Hulshof et al. (2012). Our problem includes downstream constraints in the form of wards, and the surgical cases have uncertain surgery duration and length of stay (LOS). We include rescheduling of cases, something that has not been thoroughly researched before. Further, we consider the cancellation rule of the hospital when creating model parameters, which we have not seen in the literature. We propose a pattern-based MIP that can handle chance constraints by filtering out patterns, which is new to the surgery scheduling literature to the best of our knowledge. We also create a two-stage stochastic model that builds on the deterministic model from Schiøtz and Tysse (2022), which we compare with the pattern-based model.

	Performance measure	Decision delinlation	Patient Type	Planning Horizon (weeks)	Scheduling policy	Staff included	Up- and downstream units	Cancellations	Rescheduling	Uncertainty
Our report	Waiting time, througput, cancellations, overtime	Patient to date, OR	General elective	1-10	Block	-	Recovery ward	Plan specific	On all patients	Surgery duration, LOS
Zhang et al. (2021)	Waiting time, OR undertime, ICU overuse, deferments	Patient to week, patient to date, Blocks to open	Elective general	1	Block	-	Recovery ward	-	-	Surgery duration, LOS
Shehadeh and Padman (2021)	OR Utilization, Ward utilization, Throughput	Patient to date, OR, time	Elective general	1	Block	-	ICU	-	-	Surgery duration, LOS
Zhang et al. (2020)	Cancellations, waiting time, OR costs	Patient to date, ORs to open	Elective general	1	Block	-	ICU	Plan specific	-	Surgery duration, LOS
Moosavi and Ebrahimnejad (2020)	Waiting time, OR over/undertime, ward demand variance	Patient to block, speciality to block	General elective, emergency non-elective	1	Block	-	Recovery ward, ICU, pre-surgery ward	-	-	Surgery duration, non-elective arrival, LOS
Zhang et al. (2019)	Throughput, Average patient waiting time, Maximum patient waiting time, Waiting list size	Patient to week, patient to date, Blocks to open	Elective general	1	Block	-	SICU	-	-	Surgery duration, LOS
Range et al. (2019)	Surgeon utilization, overtime, service level	Patient to date and surgeon	Elective general	4	Open	Surgeons	-	-	-	Surgery duration
Moosavi and Ebrahimnejad (2018)	Deferred patients, waiting time, extra ward beds	Patient to date, OR, time	General elective, emergency non-elective	1	Block	-	Recovery ward, ICU, pre-surgery ward	-	-	Surgery duration, non-elective arrival, LOS
Al-Refaie et al. (2018)	Waiting time, OR over/undertime, cancellations	Patient to date, OR, ICU	Elective inpatients	<1	Open	Surgeons	ICU, equipment	Plan specific	-	-
Kamran et al. (2018)	Waiting time, due-date violation, cancellations, OR overtime, deferments	Patient to date, block, surgeon to date, block overtime	General elective, emergency non-elective	1	Modified block	Surgeons	-	Plan specific	-	Surgery duration
Ozcan et al. (2017)	Throughput, OR undertime, ward utilization	Patient to date, time	General elective, general non-elective	52	Block	-	Recovery ward, pre-surgery ward	-	On all patients	Surgery duration, non-elective arrival, elective demand, LOS

Figure 3.3: Categorizing the investigated papers - part 1 / 2

	Performance measure	Decision deliniation	Patient Type	Planning Horizon (weeks)	Scheduling policy	Staff included	Up- and downstream units	Cancellations	Rescheduling	Uncertainty
Our report	Waiting time, througput, cancellations, overtime	Patient to date, OR	General elective	1-10	Block	-	Recovery ward	Plan specific	On all patients	Surgery duration, LOS
Jebali and Diabat (2017)	OR utilization, Overtime probability, ICU utilization, Cancellation risk	Patient to date, OR, time	Elective general	1	Block	-	ICU	Plan specific	-	Surgery duration, LOS, non-elective arrival
Neyshabouri and Berg (2017)	Ward utilization, Throughput	Patient to date, OR, time	Elective general	1	Block	-	SICU	-	-	Surgery duration, LOS
Haghi et al. (2017)	Waiting time, cancellations, OR overtime, recovery in OR	Patient to day, OR, bed, time, OR overtime	Elective general	1	Open	Surgeons	PACU, equipment	Plan specific	-	-
Rachuba and Werners (2017)	Maximal waiting time, overtime, deferments	Patient to date, OR	General elective, emergency non-elective	2	Block	-	-	Plan specific	-	Surgery duration, non-elective arrival
Wang et al. (2016)	Waiting time, deferments, OR costs,	Patient to Date	General elective, emergency non-elective	26	Open	-	Recovery ward, PACU, ICU	Plan specific	-	Surgery duration, non-elective arrival, LOS
Jebali and Diabat (2015)	OR utilization, Ward utilization, Throughput	Patient to date, OR, time	Elective general	1	Block	-	ICU, Recovery ward	-	-	Surgery duration, LOS
Gul et al. (2015)	Waiting time, cancellations, OR overtime	Patient to date	Elective general	1	Open	-	Equipment	Plan specific	Of cancelled patients	Surgery duration, elective demand
Addis et al. (2015)	Waiting Time, deferments	Patient to date	Elective general	4	Block	-	-	In advance, plan specific	Of cancelled patients	Surgery duration, elective demand
Zhu et al. (2015)	OR utilization, PACU utilization	Patient to time	Elective general	<1	Open	-	ICU	-	-	LOS
Min and Yih (2010)	OR Overtime, OR Undertime, Throughput, Average cancellations/week	Patient to date, OR, time	Elective general	1	Block	-	SICU	Plan specific	-	Surgery duration, LOS

Figure 3.4: Categorizing the investigated papers - part 2 / 2

Chapter 4

Problem Description

4.1 Problem Scope

The focus of this master thesis is the Advance Scheduling Problem with plan-specific cancellations, which involves assigning dates and operating rooms to surgical cases while considering multiple constraints. It builds on the work by Schiøtz and Tysse (2022) by including uncertain surgery duration, length of stay, and plan-specific cancellation rules. The thesis is motivated by the scheduling problem faced by the orthopedic department at St.Olav's Hospital in Trondheim, Norway, but it is also relevant to other hospitals and departments. We only consider elective surgical cases in this thesis and assume that non-elective cases have their own dedicated resources.

A surgery schedule maps surgery cases to a day and an operating room (OR). We define the *current schedule* as the surgery schedule currently in use at the hospital. There may also be patients who will receive surgery but have yet to be scheduled. We define the surgical cases for these patients as the waiting list. We define scheduling as moving cases from the waiting list into the current schedule. A point in time where scheduling is performed is defined as a *planning stage*. When scheduling in a planning stage, we use the clinic's MSS to determine which days and ORs the cases can be allocated to. Figure 4.1 shows an overview of how the schedule, waiting list, and planning stages are connected.

We also extend our problem to include advance *rescheduling* of already scheduled surgical cases. Rescheduling is defined as changing the assigned day of a surgical case in the schedule. We do not allow removing cases that already have a place in the schedule, only changing the assigned day. In this thesis, we investigate plan-specific cancellations. Plan-specific cancellations occur when the hospital decides not to start a surgery because the preceding surgeries have taken longer than expected, and there is not enough time left to perform the surgery within the department's opening hours. Cases that are canceled will be put back on the waiting list and can be scheduled again in

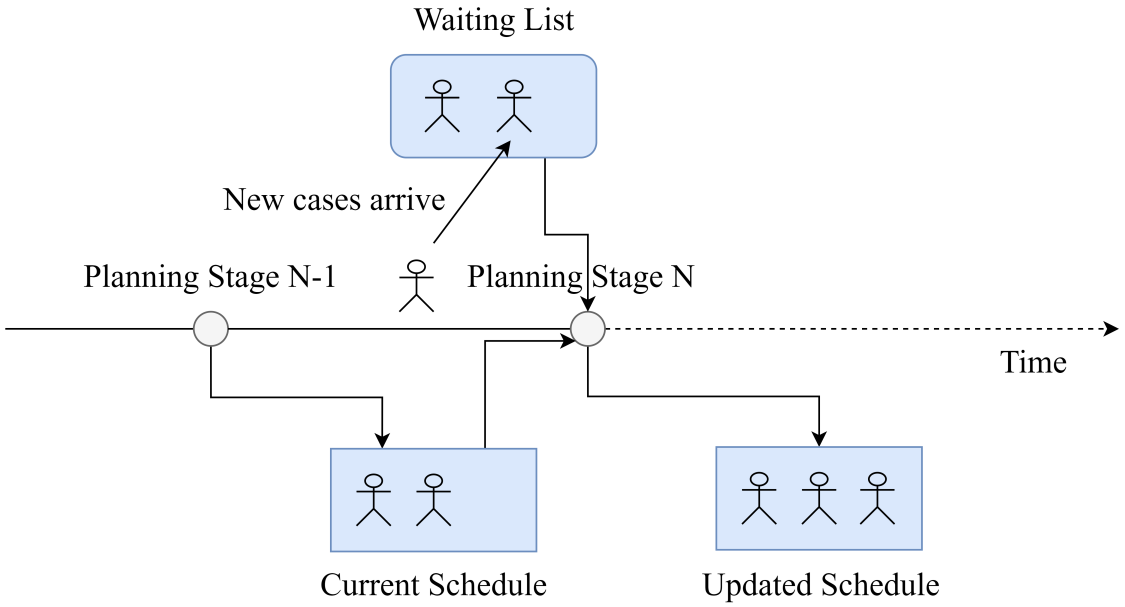


Figure 4.1: Connection between planning stages, the surgical schedule, and the waiting list.

the next planning stage.

There are no restrictions on how far in advance elective surgical cases needs to be scheduled. However, fixing the surgery dates several weeks in advance is often necessary. Fixing the schedule makes it more predictable for the clinic and the patients and can avoid unnecessary cancellations. However, scheduling too early without complete information can result in inefficient schedules or more rescheduled cases. The scheduling problem requires balancing predictability, efficiency, and agility.

New elective surgical cases are added to the waiting list throughout the week. While it is possible to enter the cases directly into the current schedule as they arrive, we wait until the next planning stage before performing the scheduling. There is a trade-off between the number of new cases in the schedule and how often the problem is solved. More cases give more flexibility when scheduling, whereas long periods between each planning stage have other downsides, such as decreased predictability.

After surgery, inpatients must stay in a ward for several days, referred to as the Length of Stay (LOS). For simplicity, we assume a shared ward capacity for all patient types. In reality, the surgery duration and length of stay are stochastic values that cannot be known until the surgery is completed and the patient is discharged. In this thesis, we assume the distributions of surgery duration and length of stay to be known.

4.2 Problem formulation

In the following section, the advanced planning problem faced by the orthopedic department at St. Olavs hospital is described in detail. Surgery cases are referred to the clinic daily, and the schedulers must frequently schedule cases for surgery by removing them from the waiting list and adding them to the active surgery schedule. However, the schedule has limited space, and finding space for everyone may be challenging. In that case, the schedulers must decide which cases should be prioritized for surgery and which to keep on the waiting list.

The schedulers must consider several factors when prioritizing cases. For instance, not all cases have waited the same amount of time for an appointment, and cases that have waited the most should be prioritized. However, simply ordering the cases based on waiting time might lead to inefficient schedules, which increases the waiting time for all cases and results in inefficient use of hospital resources.

Each surgical case is associated with a procedure, and each procedure is, associated with a specialty. There might exist several procedures associated with a given specialty. It is possible to predict the demand for surgery from surgical cases with a given procedure. We, therefore, associate each procedure with a given expected arrival rate. However, the future demand is uncertain, and the actual arrival rate might change weekly. Some weeks, only a few new cases are referred, while other weeks, we might observe many referred cases. In addition, each procedure has an expected surgery duration. However, due to uncertainty, the surgery duration will vary from case to case. Some surgeries might take longer or less than expected, and each event is as likely as the other. Similarly, each procedure is associated with an expected length of stay (LOS), but the true LOS is uncertain. However, the minimum LOS is often easier to predict since a surgical case rarely recovers much faster than expected.

The orthopedic department has several ORs at its disposal. For predictability purposes, the department has divided available surgery time slots into blocks, which are assigned to specific specialties. The block to specialty mapping is called the Master Surgery Schedule (MSS). When scheduling surgeries, we must respect the MSS and only schedule procedures associated with the respective specialty for a given block.

After surgery, the patient might need to stay at the hospital for several days to recover. There is limited space in the recovery ward. However, the clinic has extra beds available if necessary, and cancellations will, therefore, not happen due to the lack of beds. Using extra ward beds will incur an extra cost for the hospital. The capacity is reduced during the weekends, making it necessary to reduce the number of patients staying over the weekend. In some cases, the patient does not need an overnight stay. In that case, the patient is defined as an outpatient. Outpatients have dedicated recovery resources on the day of surgery and do not occupy any ward beds.

The personnel is a crucial factor for successful surgeries, and all surgeries are staffed with personnel ranging from surgeons, nurses to anesthesiologists. We call the personnel assigned to a specific surgery for the surgical team. However, employees need predictability, and the staffing schedules are therefore already created. Furthermore, the staffing is based on the MSS; hence, a surgery schedule should not have any problems with the surgical team as long as it follows the MSS.

Sometimes surgeries run overtime due to the uncertain surgery duration. In those cases, the surgical team must work outside the OR opening time and for longer than what was intended to finish the surgery. Consequently, frequently running overtime increases costs and creates a stressful work environment. Therefore, the management at the orthopedic department believes that while some overtime might be unavoidable, they should seek to limit the amount of overtime. As a result, overtime should not be used to add extra surgical cases to the schedule when the schedule is already quite full, but instead as a way of handling the unfortunate cases where a surgery runs overtime.

One way to reduce the risk of overtime is to not start surgeries but cancel them instead. The department has created a cancellation rule to help the surgical team decide whether to cancel a patient. It states that surgery is only started if they expect to finish it without running overtime. A canceled surgical case is sent home and returned to the waiting list. Canceling surgical cases is an unwanted event, and well-designed schedules should aim at keeping the risk of cancellations below a given service level. Also, while mitigating the risk reduces the probability of cancellations, the schedule should be designed to handle the consequences. This could mean keeping extra space in the schedule or reshuffling surgical cases to ensure canceled patients are not moved to the back of the queue.

Reshuffling already scheduled surgical cases is called rescheduling. While sometimes necessary, rescheduling is unwanted and should only be used when it increases the efficient use of resources. Also, rescheduling the same surgical case on multiple occasions results in an increasingly worse experience for the patient.

4.2.1 The objectives

When scheduling, several factors should be accounted for. The department believes these are the most important objectives:

1. Minimize waiting time for each patient
2. Minimize the number of reschedulings
3. Minimize the use of overtime
4. Minimize the use of extra ward beds
5. Minimize the number of cancellations

The objectives, to some degree, work against each other, and it is impossible to produce a schedule that achieves all the desired attributes. At some point, making a schedule better for one objective will only be possible by making the schedule worse for another. For instance, increasing reschedulings and overtime might be possible to reduce the overall waiting time. Finding the right balance is at the core of the problem.

Chapter 5

Mathematical Model

This chapter develops two MIP models for solving the Advance Scheduling Problem with downstream constraints. The models take in the current surgical schedule, a waiting list, and a list of mandatory cases that the models must give a slot in the surgery schedule. The models output an updated surgery schedule. Also, the models cannot remove cases from the schedule but can reschedule cases within the planning horizon.

The first model, presented in Section 5.2, is a two-stage stochastic (TSS) MIP model that incorporates uncertain surgery duration and length of stay (LOS), similar to the models developed by Min and Yih (2010) and Jebali and Diabat (2015). Two-stage models are among the most common ways to handle uncertainty within elective surgery scheduling with downstream constraints (Shehadeh and Padman 2022). However, note that the model does not handle any cancellation rules. This omission is due to two primary reasons: first, the cancellation rule at the department at St. Olav's Hospital is quite complex, which may challenge a two-stage stochastic formulation to manage the problem sizes addressed in this thesis within a reasonable time frame; second, the ASP is inherently multistage, and we have not identified any first-stage decision that could provide flexibility, a characteristic typically beneficial in a two-stage formulation. The Markov Decision Process (MDP) applied by Zhang et al. (2019) is one way to tackle the multistage properties of the ASP, but we have not attempted to formulate an MDP for our problem. We believe there exist other model formulations that can incorporate the uncertain variables in a less complex manner than a TSS model, and we will propose a pattern-based model to attempt this. The TSS model is included for benchmarking purposes, as adding uncertain surgery duration and LOS are important topics for this thesis, even in the absence of a cancellation rule.

Section 5.3 develops a pattern-based MIP model based on set partitioning using surgical patterns. We first formulate a base model with the same assumptions as the TSS model, allowing for fair benchmarking. Additionally, we propose a model extension that includes overtime and cancellation restrictions formulated as chance constraints, allowing the model to incorporate all aspects of the

problem formulation.

Another model extension using 'dummy' surgical cases to anticipate cancellations and reserve time slots for future canceled cases is also proposed. The model is included in Appendix A as the results were inconclusive due to the lack of testing and tuning time. However, it illustrates one of many possible extensions where cancellations are anticipated.

The effectiveness of the pattern-based models relies on extensive probability pre-calculations and pattern filtering rules, which are key parts of the proposed solution. For instance, filtering replaces the chance constraints and reduces time complexity. This chapter mainly focuses on describing the mathematical models, while a more in-depth look at the pattern generation method and filtering techniques is presented later in chapter 6.

5.1 Common notation

The two models share multiple indices and parameters, so we define them once in this section. A surgical case, operating room, and day are indexed i, r, and t, respectively. Additionally, we define $\underline{t}$ and $\bar{t}$ representing the first and last day of the planning horizon (see Table 5.1).

Table 5.1: Common indices used in the mathematical models.

Symbol	Description
i	Surgical Case
r	Operating Room
t	Day
$\bar{t}$	Last day in the planning horizon
$\underline{t}$	First day in the planning horizon

Table 5.2 gives an overview of the sets connected to the indices. The set $\mathcal{I}$ consists of all surgical cases in the current schedule and the cases on the waiting list. Some cases are mandatory to schedule and are denoted $\mathcal{I}^{\mathcal{M}}$. We define that all cases with a plan in the schedule, $\mathcal{I}^{\mathcal{P}}$, are also mandatory. Additional cases from the waiting list can be part of the mandatory set, for example, cases that have been canceled since the last planning stage and have been moved from the plan to the waiting list. Additionally, $\mathcal{T}$ consists of all days within the planning horizon, and $\mathcal{R}$ is the set of ORs available.

Table 5.2: Common sets used in the mathematical models.

Symbol	Description	
$\mathcal{I}$	Set of Surgical Cases	$i \in \mathcal{I}$
$\mathcal{I}^{\mathcal{M}}$	Set of Surgical Cases that must be planned in the planning horizon	$i \in \mathcal{I}^{\mathcal{M}} \subseteq \mathcal{I}$
$\mathcal{I}^{\mathcal{P}}$	Set of Surgical Cases that have an allocated slot in the current schedule	$i \in \mathcal{I}^{\mathcal{P}} \subseteq \mathcal{I}^{\mathcal{M}}$
$\mathcal{T}$	Set of days in the planning horizon	$t \in \mathcal{T}$
$\mathcal{R}$	Set of Operating Rooms	$r \in \mathcal{R}$

Further, the decision variables are shown in Table 5.3. The x-variables decide which OR and day a surgical case is assigned to, and the v-variables decide if a case is deferred, i.e., remains on the waiting list.

Table 5.3: Common Decision Variables used in the mathematical models

Symbol	Description
x_{irt}	1, if case i is scheduled in room r at day t; 0, otherwise
v_i	1, if case i is deferred to the next planning horizon; 0, otherwise

Table 5.4 show the parameters used by both models. The models' cost parameters calculation uses hyper-parameters described in Table 5.5. These hyper-parameters are the focus of Schiøtz and Tysse (2022), and for this thesis, we will use a set of hyper-parameters that showed promising behavior in (Schiøtz and Tysse 2022). Note that the cost parameters do not necessarily represent a monetary cost and do not have a specified unit.

Table 5.4: Common Parameters used in the mathematical models

Symbol	Description
C_{it}^S	The cost of scheduling case i on day t
C_{it}^R	The cost of rescheduling case i to day t
C_i^V	The cost of deferring case i to the next planning horizon
C^Φ	The unit cost of overtime
C^P	The cost of opening an extra recovery bed in the wards
K_t	Maximum ward capacity on day t
$\overline{O}$	Maximum allowed expected overtime for a block

Table 5.5: Cost Functions hyper-parameters

Symbol	Description
γ_s	Exponential growth rate parameter for scheduling cost
γ_v	Exponential growth rate parameter for deferral cost
α	Linear growth rate parameter for rescheduling
β	Base cost of rescheduling

5.1.1 Cost functions

Scheduling cost

C_{it}^S is the base cost of scheduling case i on day t. This cost reflects the waiting time for the cases and is defined in (5.1). t_i^e is the day case i entered the waiting list for the first time. It is assumed that $t > t_i^e$, i.e., all cases have entered the waiting list before the planning stage. The hyperparameter γ_s controls the exponential growth rate of scheduling costs. The use of exponential waiting cost is inspired by Moosavi and Ebrahimnejad (2020) and ensures that the marginal cost increases with the number of waiting days.

$$C_{it}^{S} := (t - t_i^e)^{\gamma_s} \tag{5.1}$$

C_i^V is the cost of deferring case i to the next planning horizon, i.e., not placing the case in the schedule (see (5.2)). The hyperparameter γ_v controls the exponential growth rate of deferral costs. We require that $\gamma_v \geq \gamma_s$, such that for all cases, the deferral cost is greater than the base cost of scheduling the patient, i.e., $C_i^V > C_{it}^S \quad \forall i \in \mathcal{I}, t \in \mathcal{T}$. When $\gamma_v = \gamma_s$, the deferral cost equals the cost of scheduling the case the first day after the current planning horizon, i.e., $C_i^V = C_{i\bar{t}+1}^S$.

$$C_i^V := (\bar{t} + 1 - t_i^e)^{\gamma_v} \tag{5.2}$$

Rescheduling cost

C_{it}^R is the cost of rescheduling case i to day t (see (5.3)). When rescheduling, we give a base cost for being moved, β. Additionally, the cost increases depending on how many days a case is rescheduled backward and how many times the case has been rescheduled in total, scaled with the hyperparameter α. t_i^f is the day case i was given an appointment the first time being scheduled, and R_i is the number of times case i has been rescheduled. If the case is not in the current schedule, the rescheduling cost is 0.

$$C_{it}^R := \begin{cases} \beta + \alpha \cdot \max\left[0, (t - t_i^f) \cdot (R_i + 1)\right] & \text{, case } i \text{ has a plan in the current schedule} \\ 0 & \text{, otherwise} \end{cases} \tag{5.3}$$

The unit costs for overtime and extra ward capacity: C^{Φ} and C^P, are set to constant values and use no hyper-parameters.

5.2 Stochastic two-stage MIP Model

In this section, we develop a two-stage stochastic MIP model. Section 5.2.1 introduces the notation, and Section 5.2.2 formulates the mathematical model. The motivation for using a two-stage model is that this type of model has shown promising results in surgery scheduling problems with uncertain surgery duration and LOS (Shehadeh and Padman 2022), outperforming deterministic models that use expected values (Min and Yih 2010), (Jebali and Diabat 2015).

5.2.1 Necessary notation

In addition to the indices and sets in Section 5.1, we have an index for the scenarios, ξ, which belongs to the set of all scenarios, Ξ (see Table 5.6 and Table 5.7). A scenario represents a single realization of the stochastic surgery duration and LOS for all the surgical cases.

Table 5.6: Indices used in the mathematical model.

Symbol	Description
ξ	Scenario

Table 5.7: Sets used in the mathematical model.

Symbol	Description	
Ξ	Scenarios	$\xi \in \Xi$

The parameters used in the stochastic model are defined in Table 5.8. $D_{i\xi}$ and $L_{i\xi}$ are the uncertain parameters for surgery duration and LOS, respectively. The availability matrix, A_{irt}, is calculated based on the MSS. If case i are allowed to be assigned to room r on day t, $A_{irt} := 1$, else $A_{irt} := 0$. By creating the A_{irt} parameters, we do not need to explicitly include procedures or specialties in the model, thus reducing the complexity of the model. The available surgery time in a given room r on day t is given by $\overline{D}_{rt}$ and is referred to as block or OR capacity. It is possible to exceed the block capacity by using overtime.

Table 5.8: Parameters used in the mathematical model.

Symbol	Description
$P(\xi)$	Probability of scenario ξ
$D_{i\xi}$	Surgery duration for case i in scenario ξ
$L_{i\xi}$	Length of stay for case i in scenario ξ
A_{irt}	Availability matrix. 1, if case i can be assigned to room r on day t; 0 otherwise
$\overline{D}_{rt}$	Available surgery time in room r on day t

Table 5.9 describes the decision variables used in the stochastic model. All variables are connected to a scenario. $y_{it\xi}$ is used to schedule patients to wards. $\phi_{rt\xi}$ and ρ_t^{ξ} are variables used for overtime and extra ward beds, respectively.

Table 5.9: Decision Variables for the stochastic model

Symbol	Description
$y_{it\xi}$	1, if case i is scheduled in a ward at day t in scenario ξ; 0, otherwise
$\phi_{rt\xi}$	Overtime in OR r on day t in scenario ξ
$\rho_{t\xi}$	Extra beds in the wards on day t in scenario ξ

5.2.2 Model formulation

We now formulate a mathematical model using the indices, sets, parameters, and variables defined in Section 5.1 and Section 5.2.1. The model is formulated by equations (5.4a)-(5.4o). To give a more concise formulation of the complete problem, we have not explicitly divided the model into a first and second stage. We will still use the terms first-stage and second-stage when discussing the model. It can be assumed that the terms with the scenario index, ξ, are considered part of the second stage. Each part of the model is discussed in more detail below the formulation.

$$\min_{z} \quad z = \sum_{i\in\mathcal{I}}\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} C^S_{it}x_{irt} + \sum_{i\in\mathcal{I}^P}\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} C^R_{it}x_{irt} + \sum_{i\in\mathcal{I}} C^V_i v_i \tag{5.4a}$$

$$+ \sum_{\xi\in\Xi} P(\xi)\Big[\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} C^{\Phi}\phi_{rt\xi}\Big] \tag{5.4b}$$

$$+ \sum_{\xi\in\Xi} P(\xi)\Big[\sum_{t\in\mathcal{T}} C^{P}\rho_{t\xi}\Big] \tag{5.4c}$$

$$\text{s.t.} \quad \sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} x_{irt} + v_i = 1 \qquad i\in\mathcal{I} \tag{5.4d}$$

$$\sum_{\xi\in\Xi} P(\xi)\phi_{rt\xi} \le \overline{O} \qquad r\in\mathcal{R}, t\in\mathcal{T} \tag{5.4e}$$

$$\sum_{i\in\mathcal{I}} D_{i\xi}x_{irt} \le \overline{D}_{rt} + \phi_{rt\xi} \qquad r\in\mathcal{R}, t\in\mathcal{T}, \xi\in\Xi \tag{5.4f}$$

$$\sum_{i\in\mathcal{I}} y_{it\xi} \le K_t + \rho_{t\xi} \qquad t\in\mathcal{T}, \xi\in\Xi \tag{5.4g}$$

$$\sum_{r\in\mathcal{R}}\sum_{t'=t^*}^{t} x_{irt'} = y_{it\xi} \qquad i\in\mathcal{I}, t\in\mathcal{T}, \xi\in\Xi, t^* = \max\{\underline{t}, t - L_{i\xi} + 1\} \tag{5.4h}$$

$$x_{irt} \in \{0,1\} \qquad i\in\mathcal{I}, r\in\mathcal{R}, t\in\mathcal{T} \quad | \quad A_{irt} = 1 \tag{5.4i}$$

$$x_{irt} = 0 \qquad i\in\mathcal{I}, r\in\mathcal{R}, t\in\mathcal{T} \quad | \quad A_{irt} = 0 \tag{5.4j}$$

$$v_i \in \{0,1\} \qquad i\in\mathcal{I}\setminus\mathcal{I}^{\mathcal{M}} \tag{5.4k}$$

$$v_i = 0 \qquad i\in\mathcal{I}^{\mathcal{M}} \tag{5.4l}$$

$$y_{it\xi} \ge 0 \qquad i\in\mathcal{I}, t\in\mathcal{T}, \xi\in\Xi \tag{5.4m}$$

$$\phi_{rt\xi} \ge 0 \qquad r\in\mathcal{R}, t\in\mathcal{T}, \xi\in\Xi \tag{5.4n}$$

$$\rho_{t\xi} \ge 0 \qquad t\in\mathcal{T}, \xi\in\Xi \tag{5.4o}$$

Equations, (5.4a)-(5.4c) define the objective function. The three terms in (5.4a) represent the first stage objective. The first term is the cost of scheduling surgical cases to a given day. This cost reflects the waiting time between the day the case entered the waiting list and the day for surgery. The second term is the cost of rescheduling cases. The third term is the cost of deferring a surgical case to the next planning horizon, i.e., not assigning a room and day to the case. The second stage

objective function is defined by (5.4b) and (5.4c). (5.4b) represents the expected overtime cost across scenarios, and (5.4c) represents the expected cost of extra ward beds across scenarios.

Constraints (5.4d) ensure that all surgical cases must either be allocated a slot in the surgery schedule or deferred to the next planning period. Constraints (5.4e) limit the use of overtime across scenarios, enforcing an average below $\overline{O}$ minutes per block. The motivation for this limit is that while a surgery schedule with a chance for a few minutes of overtime might be sensible, a schedule with high expected overtime, e.g., multiple hours, does not make sense. Constraints (5.4f) ensure that the scheduled surgery duration does not exceed the block length plus overtime. Constraints (5.4g) ensure that ward capacity is upheld, with the option of adding extra beds. Constraints (5.4h) ensure that a case is assigned to a ward at all times from the time of surgery until the ward stay is finished or until the end of the planning horizon, given by the length of stay.

Constraints (5.4i)-(5.4o) serve as variable declarations, some of which serve additional purposes. Constraints (5.4i) and (5.4j) ensure that the MSS is followed. Constraints (5.4k) and (5.4l) ensure that only cases that are not mandatory can be deferred. The second-stage variables for ward assignment, overtime, and extra beds are declared by constraints (5.4m), (5.4n) and (5.4o), respectively. Note that even though $y_{it\xi}$ and $\rho_{t\xi}$ are defined as continuous variables in the formulation, constraints (5.4h) forces $y_{it\xi}$ to be a binary variable, and constraints (5.4g) makes $\rho_{t\xi}$ take an integer value since $y_{it\xi}$ is binary, K_t is integer, and the cost C^P is assumed positive.

5.2.3 Additional comments

While creating the model, several aspects had to be considered, and some deserve further explanation. First, the reader should note that the model has Relative Complete Recourse (RCR). Models with RCR have the benefit that any feasible first-stage decision also has a feasible second-stage solution for all scenarios. The property can help solve the problem more efficiently, for instance, when using decomposition techniques.

Second, much emphasis has been made on softening some of the constraints. For instance, there are no hard restrictions on overtime and extra ward beds. Instead, overtime and extra ward beds are punished in the objective function. An alternative approach could be to allow extra ward beds and overtime up to a specific hard limit. However, such formulations would not be as soft, and the model would not have RCR (though it is still possible to achieve RCR with some modifications). Additionally, since the extra ward beds and overtime are recourse actions, hard constraints could have unwanted consequences since promising schedules could be deemed unfeasible even when the schedule is feasible for almost possible scenarios. One way to handle those issues is with chance constraints, which will be investigated in the next model. The reader should also note that tuning the cost of extra ward beds and overtime could make the restrictions softer or harder. For instance, setting the costs to zero would fully relax the constraints. A cost equal to a sufficiently

large number, M, would, on the other hand, make the restrictions 'almost' hard since it would never be beneficial to break the restrictions if other feasible solutions exist where no overtime or extra ward beds occur.

Lastly, alternative recourse problems should be mentioned. One possibility is to add cancellations as a recourse action. However, this was avoided due to a couple of reasons. First, neither the model nor the ASP generally considers ordering patients within a single block, which is necessary for the cancellation rule. As a result, allowing cancellations as a recourse decision would allow some unwanted assumptions. For instance, the surgery duration of a case is realized after surgery completion. However, a cancellation decision in the second stage would assume the known actual surgery duration of the cases. Consequently, the model would be able to know which cases it should cancel based on the realized surgery duration. In other words, the sequencing of patients might differ between scenarios. This would not have been possible in the real world, and such assumptions would have been overly optimistic and complicated to test. Formulations aimed at including ordering in the first stage problem would most likely struggle to solve the problem within a reasonable time for large enough problem sizes. Although decomposition techniques could have been tested and possibly enabled us to create a two-stage stochastic model with cancellations, we instead propose a pattern-based model with the benefit of not relying on stochastic scenarios.

5.3 Pattern-based MIP Model

This section develops a pattern-based MIP model. The use of patterns is inspired by the cutting stock problem, as discussed in Section 3.3.2. We also develop an extended model that includes chance constraints. A pattern is a template for a combination of surgical cases in a given block based on the procedure of the cases. If a block, according to the MSS, has two allowed procedures, named p_1 and p_2, an example pattern is 'three cases with procedure p_1 and one case with procedure p_2'. The motivation for using patterns is to reflect the combinatorial aspects of the problem better, as we would expect only a few of the possible combinations of procedures to be used. Also, we want to take advantage of the known distributions of the random variables for surgery duration and length of stay. For example, we can pre-calculate the true expected overtime in a block with a given pattern. The calculated parameters can then be used as input to the mathematical model, which is solved as a deterministic MIP. Two-stage models like the model in Section 5.2.2 can not usually include the complete scenario tree for the random variables and need to sample smaller scenario trees to be able to solve the model within a reasonable time, using methods like SAA. This makes the two-stage model's parameters an approximation of the true distribution, whereas the pattern model can use information from the exact distribution. Using patterns should also give a tighter formulation than the stochastic model. The reason is that the knapsack constraints related to the block capacity are no longer needed and are exchanged with set partitioning constraints.

A more detailed description of patterns, pattern generation, and other complex precalculations is given in chapter 6.

5.3.1 Necessary notation

In addition to the indices and sets in Section 5.1, we have the index p for a procedure and j for a pattern, with their respective sets $\mathcal{P}$, and $\mathcal{J}$ (see Table 5.10 and Table 5.11). We define the subsets $\mathcal{J}_{rt}$, which contains the allowed patterns for a given OR, r, and day, t. These subsets are created based on the MSS. Lastly, we define the subsets $\mathcal{I}_p$, which contains all cases from $\mathcal{I}$ with procedure p.

Table 5.10: Indices used in the pattern model.

Symbol	Description
p	Procedure
j	Pattern

Table 5.11: Sets used in the pattern model.

Symbol	Description	
$\mathcal{I}_p$	Set of Surgical Cases with procedure p	$\mathcal{I}_p \subseteq \mathcal{I}$
$\mathcal{P}$	Set of all procedures	$p \in \mathcal{P}$
$\mathcal{J}$	Set of all patterns	$j \in \mathcal{J}$
$\mathcal{J}_{rt}$	Set of possible patterns in room r and time t	$\mathcal{J}_{rt} \subseteq \mathcal{J}$

Like the two-stage model, the pattern model has variables for assigning a case to a room and day or the waiting list, x_{irt} and v_i. The model also has variables for extra ward beds, ρ_t, but with the difference that the variables are not connected to a scenario in the pattern model. The last variables in the pattern model, π_{jrt}, are used for selecting which pattern to use in each block (see Table 5.12).

Table 5.12: Decision Variables for the pattern model.

Symbol	Description
π_{jrt}	1, if pattern j is used in block (r, t); 0, otherwise
ρ_t	Extra beds in the wards on day t

Table 5.13 shows the parameters used in the pattern model. O_j is the expected overtime used when using pattern j. The parameter $Wjtt'$ gives the expected number of patients occupying wards on day t, which comes from a pattern j used on day t'. Another way to phrase the contents of the parameter is 'given that we choose pattern j on day t', how many of the patients from that exact pattern are expected to remain in the wards at day t'. Note that $Wjtt'$ might be fractional. The pattern parameter calculation methods are presented later in chapter 6.

Table 5.13: Parameters used in the pattern model.

Symbol	Description
B_{pj}	Number of patients of procedure p that can be scheduled in pattern j
$W_{jtt'}$	Expected number of patients in the ward on day t from pattern j used on day t'
O_j	Expected overtime from pattern j

5.3.2 Model formulation

We now formulate a mathematical program in equations (5.5a)-(5.5m). Indices, sets, parameters, and variables used in the formulation are defined in Section 5.1 and Section 5.2.1. The model is discussed in more detail below the formulation.

$$\min_{z} \quad z = \sum_{i\in\mathcal{I}}\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} C^S_{it}x_{irt} + \sum_{i\in\mathcal{I}^{\mathcal{P}}}\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} C^R_{it}x_{irt} + \sum_{i\in\mathcal{I}} C^V_i v_i \tag{5.5a}$$

$$+ \Phi \sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} C^{\Phi}_j O_j \pi_{jrt} \tag{5.5b}$$

$$+ \Phi \sum_{t\in\mathcal{T}} C^P \rho_t \tag{5.5c}$$

$$\text{s.t.} \quad \sum_{i\in\mathcal{I}_p} x_{irt} = \sum_{j\in\mathcal{J}_{rt}} B_{pj}\pi_{jrt} \qquad p\in\mathcal{P}, r\in\mathcal{R}, t\in\mathcal{T} \tag{5.5d}$$

$$\sum_{j\in\mathcal{J}_{rt}} \pi_{jrt} = 1 \qquad r\in\mathcal{R}, t\in\mathcal{T} \tag{5.5e}$$

$$\sum_{r\in\mathcal{R}}\sum_{t\in\mathcal{T}} x_{irt} + v_i = 1 \qquad i\in\mathcal{I} \tag{5.5f}$$

$$\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}}\sum_{t'=\underline{t}}^{t} W_{jtt'}y_{jrt'} \le K_t + \rho_t \qquad t\in\mathcal{T} \tag{5.5g}$$

$$O_j\pi_{jrt} \le \overline{O} \qquad j\in\mathcal{J}, r\in\mathcal{R}, t\in\mathcal{T} \tag{5.5h}$$

$$x_{irt} \in \{0,1\} \qquad i\in\mathcal{I}, r\in\mathcal{R}, t\in\mathcal{T} \tag{5.5i}$$

$$\pi_{jrt} \in \{0,1\} \qquad j\in\mathcal{J}, r\in\mathcal{R}, t\in\mathcal{T} \tag{5.5j}$$

$$v_i \in \{0,1\} \qquad i\in\mathcal{I}\setminus\mathcal{I}^{\mathcal{M}} \tag{5.5k}$$

$$v_i = 0 \qquad i\in\mathcal{I}^{\mathcal{M}} \tag{5.5l}$$

$$\rho_t \ge 0 \qquad t\in\mathcal{T} \tag{5.5m}$$

The objective function for the pattern model is mostly the same as for the two-stage model, with (5.5a) being the same as the first stage costs in the two-stage model, and (5.5b)-(5.5c) representing the same as the second stage costs in the two-stage model, with the minor difference that the pattern model does not use scenarios. Remember from Section 5.2 that the first stage costs are the scheduling, rescheduling, and deferral costs; the second stage costs are the overtime and extra

ward bed costs.

The operating room capacity is handled by constraints (5.5d) in combination with (5.5e), which ensure that only one pattern is chosen for each block, and (5.5f) which ensure that cases are either scheduled to only one block or put on the waiting list. Note that we no longer directly consider surgery duration in the model; constraints (5.5d) are instead limiting the number of cases scheduled to a block while the pre-calculated expected overtime from that pattern is penalized in the objective. Ward capacity is handled by constraints (5.5g) where ρ_t acts as a relaxation variable, allowing acquiring extra ward beds. Contrary to the two-stage model, this variable is not forced to be an integer. The interpretation of this variable is thus the expected number of extra beds, not an actual integer number of beds. Constraints (5.5h) limit the expected overtime per block by prohibiting the use of patterns with expected overtime above $\overline{O}$, similar to constraints (5.4e) in the two-stage model. Note that the overtime limit constraints for the pattern model are not actually implemented when using the solver in the computational study; we instead remove all patterns with expected overtime above the limit from the set of patterns, $\mathcal{J}$, before running the model.

Constraints (5.5i), (5.5j), (5.5k), (5.5l) and (5.5m) are variable declaration constraints, where (5.5l) also serves the purpose of not allowing mandatory patients to be put on the waiting list.

5.3.3 Model Extension: Chance constraints

We propose utilizing chance constraints to enable schedulers to manage the risk of overtime and cancellations. We formulate a chance constraint for overtime in (5.6), where $\mathcal{D}_i$ is the stochastic surgery duration for surgical case i, and η_o is the risk level for overtime as described in Table 5.14. If, e.g., η_o is 0.1, the interpretation of the constraint is that each block should have less than 10% probability of overtime occurring. Note that the overtime chance constraint and the overtime costs are not mutely exclusive and intend different purposes. The chance constraints limit the probability of overtime occurring. However, overtime costs penalize the amount of overtime when it occurs.

Table 5.14: Extra parameters for the extended model.

Symbol	Description
$\mathcal{D}_i$	Random variable for surgery duration for case i
η_o	Risk level for overtime
η_c	Risk level for cancellations

$$\textbf{Prob}\left(\sum_{i\in\mathcal{I}} x_{irt}\mathcal{D}_i \leq \overline{D}_{rt}\right) \geq 1-\eta_o \quad r\in\mathcal{R}, t\in\mathcal{T} \tag{5.6}$$

Instead of implementing constraint (5.6) in the model, a more elegant approach is to remove

patterns with probability for overtime larger than η_o from the set $\mathcal{J}$ before running the model. This filtering technique is discussed later in Section 6.2.2. We also implement a way to control the risk level for plan-specific cancellations, η_c. Similar to the overtime chance constraint, we do not add any constraints to the model but instead, remove patterns with a higher probability for cancellations occurring than η_c from the set $\mathcal{J}$ with the filtering technique. We do not propose a detailed equation for this constraint since it requires the implementation of the hospital's cancellation rule in the mathematical model, which would become far too complex. However, suppose we include the pre-calculated probability of cancellations from pattern j as a parameter Υ_j. In that case, constraint (5.7) achieves the same result even though it is not formulated as a chance constraint.

$$\Upsilon_j \pi_{jrt} \leq \eta_c \quad i \in \mathcal{I}, r \in \mathcal{R}, t \in \mathcal{T} \tag{5.7}$$

Chapter 6

Pattern Generation

The following section describes how we generate the patterns used by the mathematical program described in Section 5.3. The main goal is to show how the B_{pj}, $W_{jtt'}$, and O_j parameters are calculated. We will first, in Section 6.1, define what a pattern is and how a pattern is connected with surgery schedules, specialties, procedures, and surgical cases. Second, in Section 6.2, we show how to create patterns, incorporating uncertain surgery duration, LOS, and cancellations. We also show how the Markov property and Dynamic Programming can be used to speed up the algorithm. Lastly, in Section 6.3, we discuss the benefits and drawbacks of the pattern model, including a direct comparison with the stochastic two-stage model described in Section 5.2.

The reader should note that we do not discuss nor assume the shape of the surgery duration and LOS distribution of a given procedure in this chapter. The reason is that the pattern generation algorithms are general and work for all distributions. However, we assume that the distributions are discrete, making the calculations much simpler. We also assume that the distributions and their resulting probability mass functions and cumulative mass functions are known. We discuss the shape and underlying assumptions of the distributions for the procedures later in Section 7.4.2.

6.1 What is a pattern

In short, a pattern is a procedure-frequency mapping that captures the distribution of procedures among a set of surgical cases. In other words, a pattern can be seen as the shadow of a collection of surgical cases. Each case reflects their procedure, and the combined shadow is the frequency of each procedure. In statistics, another word is an absolute frequency distribution. However, the exact wording of the definition is not of importance. The key is to understand what a pattern represents. Figure 6.1 shows how patterns are created based on surgery cases and procedures.

To avoid confusion and generalize the pattern concept, we say that a pattern consists of several

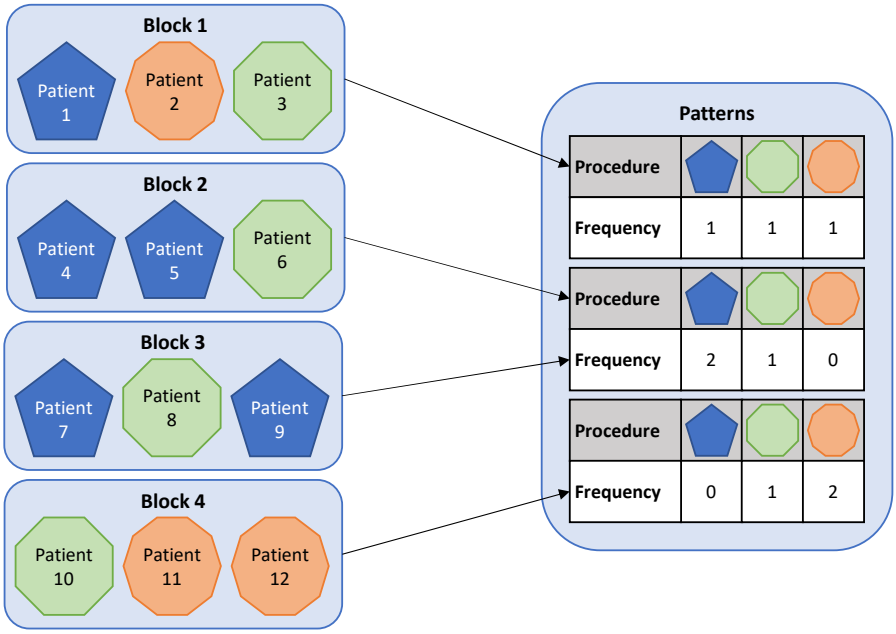


Figure 6.1: The relationship between surgery cases, procedures, and patterns

elements. An element is a generalized surgical case with only a procedure, a surgery duration distribution, D, and a ward demand distribution W. The surgery duration distribution is the same as the distribution for a surgery case with the same procedure as the element. The ward demand distribution is closely linked to the LOS probability distribution of a surgical case with the same procedure. However, instead of representing the likelihood that a surgery case has a given LOS, it is the likelihood that a surgery case will occupy a bed in the recovery ward t days after receiving surgery. In essence, we transform the LOS probability distribution into N random variables where each variable follows a Bernoulli distribution, one for each possible realization of the uncertain LOS. The success rate for the t'th random variable is defined in (6.1), where CMF_s^{LOS} is the cumulative probability of element i needing a certain number of recovery days.

$$W_{i,t} = 1 - CMF_i^{LOS}(t) \tag{6.1}$$

$W_{i,t}$ hence represents the probability distribution that element, i, needs a bed in the ward on day t. An example of the relationship between a procedure's LOS distribution and the corresponding element ward demand distribution W is shown in Table 6.1. Note how the single LOS random variable is transformed into four random variables W_t. If LOS is a discrete random variable with support R_{LOS}, then W will consist of $|R_{LOS}|$ random variables. Hence, the uncertain length of stay question is transformed from 'how many days does an element have to recover after surgery' to 'what is the probability that an element is in the ward t days after surgery'. This enables us to estimate the extra ward bed costs for a given pattern, as described shortly. Also, the reader should take note of the indexing. If an element has LOS equal to one, then it will demand a bed on day $t = 0$, i.e., the same day as the surgery.

Table 6.1: The relationship between a procedure's *LOS* and an element's ward demand distribution

(a) LOS probability distribution

LOS	L=0	L=1	L=2	L=3
pmf(LOS=i)	1/4	1/4	1/4	1/4
cmf(LOS=i)	1/4	1/2	3/4	1
1-cmf(LOS=i1)	3/4	1/2	1/4	0

(b) Ward bed demand distribution

$W_{i,t}$	$W_{i,0}$	$W_{i,1}$	$W_{i,2}$	$W_{i,3}$
$P(W_{i,t}=1)$	3/4	1/2	1/4	0
$P(W_{i,t}=0)$	1/4	1/2	3/4	1

Note how the ward bed demand distribution shown in Table 6.1b significantly differs from traditional deterministic formulations like the one used by Schiøtz and Tysse (2022). If we, for simplicity, assume that the LOS follows a uniform discrete distribution between 0 and 2, as shown in Table 6.2, then Table 6.3 shows how our transformation and addition of new random variables changes the expected demand for ward beds. One key difference is that the pattern model aims to approximate the expected demand for ward beds, not decide how many extra beds to include. The difference is important, as we only aim at estimating the expected cost of ward beds similarly to what is done in the stochastic two-stage formulation.

Table 6.2: LOS distribution assuming discrete uniform distribution between 0 and 2

l	0	1	2
P(LOS=l)	1/3	1/3	1/3

Table 6.3: Comparison between traditional ward bed demand distribution used in deterministic models and the one used in the pattern model

(a) Traditional deterministic formulation

$W_{i,t}$	$W_{i,0}$	$W_{i,1}$	$W_{i,2}$
$P(W_{i,t}=1)$	1	0	0
$P(W_{i,t}=0)$	0	1	1

(b) Pattern formulation

$W_{i,t}$	$W_{i,0}$	$W_{i,1}$	$W_{i,2}$
$P(W_{i,t}=1)$	2/3	1/3	0
$P(W_{i,t}=0)$	1/3	2/3	1

We can calculate the distributions for the sum of the elements from the surgery duration and ward bed demand distributions of a set of elements. We find the probability density function of the sum of the elements' random variables. Note how the ward distributions will become a multinomial distribution from the sum of several Bernoulli distributions. Lastly, a pattern also has an expected pattern distribution representing the expected frequency of the procedures when cancellations are considered. More specifically, it is a collection of multinomial distributions, one for each possible procedure. We can use this distribution to calculate the expected number of cancellations for each procedure in a given pattern.

6.2 Implementation

In this section, we discuss how to generate the patterns used by the pattern model. We show the approach to the setting with plan-specific cancellations, but the framework is the same for the

setting without cancellations. In short, for the setting without cancellation rules, we assume that all cases are independent of each other. As a result, the combined distributions can be found by simply combining the distributions of the cases. Algorithm 1 shows a high-level pseudo code for the generations of patterns in a setting with cancellations.

Algorithm 1 Generate Patterns

```
1: procedure GENERATE-PATTERNS
2:     patterns ← GET-ALL-COMBINATIONS()        ▷ Get all possible combinations of elements
3:     SORT(patterns)                            ▷ Sort patterns from smallest to largest
4:     for pattern in patterns do
5:         CALCULATE-JOINT-PROBABILITY-DISTRIBUTION(pattern)
6:     end for
7:     CALCULATE-PARAMETERS(patterns)            ▷ Calculate parameters for model and filtering
8:     FILTER(patterns)                          ▷ Filter based on predefined criteria
9: end procedure
```

First, we create the complete set of legal patterns at line 2. This is done by assuming all elements get the lowest possible surgery duration based on their procedure. Then, we find all combinations of elements where the combined surgery duration is less than the size of the block. It is worth noting that several other implementations could be equally sensible. For instance, the combinations could have been created based on the expected surgery duration. Also, the algorithm could have been designed without generating all patterns but simply using a bottom-up approach starting from an empty pattern. We decided to use our implementation as it was deemed sufficient from a development velocity and solution quality perspective. Nevertheless, it should be noted that our implementation restricts the decision space since all possible solutions where overtime is guaranteed, $P(overtime = 0) = 0$, are not included in the possible set of patterns. We believe this is sufficient because the department does not want to use overtime unless it is absolutely necessary. Lastly, note that the set of patterns consists of *all* possible combinations, even combinations that are not complete. We call a pattern complete if it is not a subset of any other patterns.

Sorting the patterns based on their size, at line Algorithm 3, helps increase the algorithm's efficiency. The reason is two-fold. First, note how the Markov property holds. The Markov property states that only the current state is necessary to calculate the next. In other words, the system is memoryless. A practical consequence is that after two cases have received surgery, we do not need to know the actual surgery of both cases; we only need to know the combined surgery time of the two. Thus, we need to keep track of much fewer random variables. Second, given the Markov property and that the patterns are sorted based on their size, we can use Dynamic Programming to speed up the process. Note how we only need the pattern without the last element and the procedure distributions of the last element to calculate the distributions for a given pattern.

6.2.1 Joint Probability Distribution

Algorithm 2 illustrates how to calculate the joint probability distribution for a given pattern. A joint probability distribution is a special type of random variable representing the combination of multiple other variables. In our case, it represents the probability of a certain realization of all random variables. To keep things simple, we do not show how to calculate the joint probability distribution in detail in this section.

Algorithm 2 Calculate Joint Probability Distribution for the setting with plan-specific cancellations based on expected surgery duration

```
1:  procedure CALCULATE-JOINT-PROBABILITY-DISTRIBUTION(self)
2:      if size = 1 then                          ▷ If number of elements in pattern is one
3:          HANDLE-SIZE-ONE(self)                 ▷ Handle special case where the size is one
4:      else
5:          child ← GET-LARGEST-CHILD(self)
6:          last_procedure ← GET-LAST-PROCEDURE(self)
7:          ov_threshold ← OR_SIZE - last_procedure.exp_los          ▷ Overtime threshold
8:          for i = 0 to size do                  ▷ Iterate from zero to size-1
9:              prob_cancelled ← PROBABILITY-CANCELLED(i-1, child, ov_threshold)
10:             dist_cancelled ← DISTRIBUTION-CANCELLED(i-1, child, ov_threshold)
11:             cancelled ← ⟨dist_cancelled, prob_cancelled⟩
12:             prob_not_cancelled ← PROBABILITY-NOT-CANCELLED(i, child, threshold)
13:             dist_not_cancelled ← DISTRIBUTION-NOT-CANCELLED(i, child, threshold)
14:             not_cancelled ← ⟨dist_not_cancelled, prob_cancelled⟩
15:             self.jointDistribution[i] ← COMBINE-DISTRIBUTIONS(cancelled, not_cancelled)
16:         end for
17:     end if
18: end procedure
```

First, in line 3, we handle the base case where a pattern only consists of one element. In that case, no cancellations are possible, and the total pattern distributions are the same as the ones associated with the element. The rest of the algorithm handles the cases where the pattern consists of multiple elements.

Line 5-7 defines some key variables used multiple times throughout the algorithm. At line 5, the child pattern is identified. A pattern's child is the pattern where all but the last element are the same. For instance, for a pattern with the procedure frequency (2,1) where the procedure at the second index has the shortest expected surgery duration, the child pattern would be the (2,0) pattern. However, if the procedure at the first index is the shortest, then the child pattern would have been (1,1). As previously described, we will use the joint probability distributions for the child pattern as the base of the calculations. The procedure of the last element is identified in line 6 and is used to extend the child joint probability distribution if the last element is not canceled. Last, at line 7, the overtime threshold is defined. The overtime threshold marks the total available time left in the block without running overtime. For example, if we have used 400 minutes and there were originally 480 minutes available, then the overtime threshold is 80 minutes. Note that the threshold is defined based on the cancellation rule 'no surgery is started if it is not expected

to finish without overtime'. Changing the cancellation rule to, e.g., 'no surgery is started if the probability of overtime is above X%' would be trivial by changing the threshold variable.

Line 8-16 defines the main calculations of the joint probability table for the pattern. We iterate from $i = 0$ to $i = size - 1$ to get all possible numbers of cancellations that could occur from a given pattern. So if a pattern has four elements, then a maximum of three cancellations are possible and define the support of the total number of cancellations distribution. We iterate through each possible realization of the total number of cancellations since we have conditional independence given the cancellation status. In other words, for each iteration, we calculate the joint probability distribution for the pattern, given that a total of i cancellations occur.

There are two possible events that result in the total number of cancellations equal to i: (1) the last element is canceled, and $i - 1$ cancellations had already occurred before the processing of the last element, or (2) the last element is not canceled, but i cancellations had already occurred. Line 9 handles the first case, and line 12 handles the last. If we define C_i as the random variable representing if element i is canceled, CAP as the block capacity in minutes, and $D_{1:N}$ as the random variable representing the total surgery duration of element 1 to N, then one simplified way of representing the probability of both cases is shown in (6.2) and (6.3).

$$P(C_N = 1|D_{1:N-1}) = P(D_{1:N-1} + E[D_N] > CAP) \tag{6.2}$$

$$P(C_N = 0|D_{1:N-1}) = P(D_{1:N-1} + E[D_N] \leq CAP) \tag{6.3}$$

For the case where the last element is not canceled, we get the resulting joint probability distribution of the child pattern given i cancellations. When the last element is not canceled, we sum the procedure distributions of the last element with the distributions of the child pattern given $i - 1$ cancellations and that the total surgery duration of the first N-1 elements plus the expected surgery duration of the last elements is less than or equal to the overtime threshold. We define $D_{1:N}$ as the random variable representing the total surgery duration of the first N patients. Likewise, $W_{1:N,t}$ is the total bed demand on day t for the first N patients, and $T_{1:N,p}$ represents the total number of patients with procedure p that have received surgery. Then, the updating rules representing line 10 and 13 in Algorithm 2 is defined by (6.4), (6.5), and (6.6)

$$D_{1:N} = \begin{cases} D_{1:N-1} \circledast D_N & \text{if } C_N = 1 \\ D_{1:N-1} & \text{if } C_N = 0 \end{cases} \tag{6.4}$$

$$W_{1:N,t} = \begin{cases} W_{1:N-1,t} \circledast W_{N,t} & \text{if } C_N = 1 \\ W_{1:N-1,t} & \text{if } C_N = 0 \end{cases}, \quad t \in \mathcal{T} \tag{6.5}$$

$$T_{1:N,p} = \begin{cases} T_{1:N-1,p} \circledast \delta(T_i, p) & \text{if } C_N = 1 \\ T_{1:N-1,p} & \text{if } C_N = 0 \end{cases}, \quad p \in \mathcal{P} \tag{6.6}$$

$\delta(T_i, p)$ is the Kronecker delta function which is 1 if $T_i = p$ and 0 otherwise, and $A \circledast B$ is the convolution of two random variables.

Summing distributions - Convolutions

Now, we briefly describe how we calculate the sum of distributions. Here, we look specifically at situations where the last patient's surgery is not canceled. We do this by merging the outcomes from the final surgical procedure with the outcomes of a similar scenario, but where the last patient does not exist.

In simpler terms, we are adding up the outcomes from all the other patients who either had their surgeries canceled or completed them successfully. For these calculations, we use a technique known as convolution.

If we define two discrete random variables, X and Y, with probability mass functions $p_X(x)$ and $p_Y(y)$, and support R_X and R_Y. Then the convolution equation (6.7), represents the sum of the two variables, i.e $Z = X + Y$,

$$p_Z(z) = \sum_{x \in R_X} p_Y(z - x) p_X(x) \tag{6.7}$$

We considered three different ways of calculating the convolutions: (1) brute force using (6.7), (2) Fast-Fourier-Transformation (FFT), and (3) polynomial multiplication. The brute-force method works well for smaller distributions, such as ward distributions. In theory, Fast-Fourier-Transformation works better for larger distributions since FFT's time complexity is $O(NlogN)$, while the brute force method is $O(N^2)$. However, for small values of N, the extra computational costs associated with the FFT might not be worth it compared to the more simplistic brute force method. In our tests, FFT was marginally faster than the brute-force method, but we expect the benefit of FFT to increase if the variance and range of surgery durations increase. Polynomial multiplication was tested but was not faster than FFT or the brute-force method in our case.

6.2.2 Filtering and Pattern model parameters

We now describe how the pattern model parameters are calculated. In addition to being used as input parameters to the mathematical models, these parameters can also be used to remove unwanted patterns in a filtering process before running the model, indirectly replacing the chance constraints in the extended pattern model described in Section 5.3.

Overtime

Three key statistics related to pattern overtime and surgery duration are calculated; total surgery duration distribution, expected overtime, and the probability of overtime. The total surgery duration distribution is calculated as shown in (6.8).

$$P(D_{1:N} = d) = \sum_{i=0}^{N-1} P(C_{1:N} = i)P(D_{1:N} = d|C_{1:N} = i) \tag{6.8}$$

Equation (6.9) shows how the total surgery duration distribution can be used to calculate the expected overtime. It is worth pointing out that this calculation represents the true expected overtime for a pattern. As a result, using this parameter in the pattern model should result in it representing the uncertain surgery duration as well, if not better than the stochastic model. This is discussed further in Section 6.3

$$E[O] = \sum_{d=CAP+1}^{\infty} (d - CAP) * P(D_{1:N} = d) \tag{6.9}$$

We are also interested in the probability of overtime occurring. Equation Equation 6.10 shows this calculation. Note that this parameter can be used to filter out patterns with high probabilities of overtime. In that case, the chance constraint formulated in Section 5.3 is redundant.

$$P(O) = \sum_{d=CAP+1}^{\infty} P(D_{1:N} = d) \tag{6.10}$$

Cancellations

To calculate the expected number of cancellations for a specific procedure p within a pattern, we use (6.11).

$$E[T_{1:N,p}] = \sum_{i=0}^{N-1} P(C_{1:N} = i) * E[T_{1:N,p}|C_{1:N} = i] \tag{6.11}$$

The parameter $E[T_{1:N,p}]$ represents the weighted average of the expected number of cancellations for procedure p given each possible realization of the total number of cancellations.

Note that calculating the probability of no element getting canceled is a trivial task as it is equivalent to looking up $P(C_{1:N} = N)$ in the joint probability distribution. Using this, we can filter out all patterns where the risk of cancellations is higher than the desired limit. This will, in practice, work as a chance constraint, making the chance constraint discussed in Section 5.3 redundant.

Ward bed demand

In 5.3, $W_{jtt'}$ is defined as the expected number of patients in the ward on day t from pattern j used on day t'. We want to use the joint probability distribution from a given pattern j and calculate $W_{jtt'}$. We use the ward demand distributions to calculate this. Remember, if a pattern has size N, then a maximum of $N-1$ total cancellations can occur. In that case, the joint probability distribution is broken down into N smaller joint probability distributions, one for each realization of $C_{1:N}$. The expected ward demand, $E[W_{1:N,t}$ is calculated as (6.12)

$$E[W_{1:N,t}] = \sum_{i=0}^{N-1} P(C_{1:N} = i) E[W_{1:N,t} | C_{1:N} = i] \qquad \text{for } t \in \mathcal{T} \tag{6.12}$$

We now have $|T|$ numbers, each representing the expected bed demand t days after the surgery date. It is important to note that this is the expected value of the ward demand across all possible realizations of $C_{1:N}$. As a result, adding the expected values from multiple patterns to calculate the number of extra ward beds needed will be an approximation since we do not calculate the joint probability distribution across patterns and use the distribution in the calculations. The implementation's limitations and expected consequences are discussed in the next section.

6.3 Assumptions

This section aims at connecting chapter 5 and chapter 6 by summarizing the key concepts, overall idea, and hypothesis behind the pattern model. Thus, we now discuss the underlying assumptions and expected behavior compared to the two-stage stochastic model.

We have created a pattern-based MIP that incorporates stochastic information about uncertain surgery duration and LOS while being deterministic in the sense that no stochastic parameters or scenarios are used in the mathematical model. In practice, we have three pattern models:

(P1) The base model without cancellation

(P2) Model extension with the probability of overtime as chance constraints

(P3) Extension $P2$ to also include the cancellation chance constraint.

In addition, we have two different pattern settings or *pattern files*, which can be used to test the pattern models based on different assumptions of the hospital environment:

(PC0) Uncertain surgery duration and LOS, no plan-specific cancellations

(PC1) Uncertain surgery duration and LOS, with plan-specific cancellations according to the cancellation rule

Testing different combinations of pattern files and models will enable us to test how different aspects of the problem affect schedule quality. Note also that P1-PC0 has similar assumptions as the two-stage stochastic model (S1).

6.3.1 Assumptions and limitations

Several assumptions are made when generating the patterns. First, we assume independence. For $PC0$ all surgical cases are assumed to have independent surgery duration and LOS. There is always a trade-off between accuracy and complexity. For example, one would expect that surgical cases with the same surgeon and surgery date in practice are not entirely independent, as a surgeon most likely has different efficiency levels from day to day. In the setting of $PC1$, we assume conditional independence given the cancellation status.

Second, we assume surgical cases with the same procedure to be homogeneous. The benefit is that the input variance is greatly reduced. However, it introduces some limitations. For instance, there might be better solutions for non-elective scheduling where the surgical cases are less predictable. Also, the possible combinations of elements increase substantially if the specialty has many procedures with short surgery duration. We have not had any problems generating the patterns, and the computations take a couple of minutes, even when the number of procedures is tripled from 15, as used by T. R. Bovim et al. (2020), to 45. Further, note that the number of patterns used by the model is usually considerably lower than the original set of patterns due to the filtering. The reader should note, however, that the pattern model could be expanded and included in other model formulations. For instance, the patterns could be combined with the more traditional stochastic models where, e.g., the more heterogeneous procedures are not included in the patterns. Further, decomposition techniques and column generation are also possible but are out of the scope of this thesis.

Lastly, compared to S1, the extra ward bed constraints are expected to be less accurate because the pattern model still uses the expected values of each ward demand distribution for a given day, as discussed in Section 6.2. Since the number of extra ward beds needed depends on the underlying

distributions from multiple patterns, simply summing the expected bed demand of each pattern for a given day before subtracting the original ward capacity will not yield the true expected value. The pattern model should instead underestimate the number of ward beds needed, in a similar way as deterministic models will underestimate the needed overtime. Understanding this concept is key. The issue is not that summing the expected ward bed demand from two patterns does not give the correct expected value for the joint probability distribution of the two, but that calculating the expected *extra* ward beds needed from a set of patterns cannot be accurately calculated only from the sum of the expected values. Instead, the distributions need to be considered, similar to what is seen in (6.9).

However, the reader should note that the proposed solution is expected to be considerably more accurate than those commonly seen in deterministic formulations, for example, the model developed by Schiøtz and Tysse (2022). The reason is that we transform a procedure's LOS from one random variable across days into multiple random variables, one for each day, as discussed in Section 6.2. Further, we use convolutions to calculate the sum of all the elements' distributions within a given pattern. As a result, the pattern model should be better than the deterministic models at estimating the extra ward bed demand. We believe this formulation is sufficient as the ward restrictions are already quite soft. Further, the focus of the thesis is to investigate the inclusion of a cancellation rule to avoid overtime. Therefore, getting an exact representation of the uncertain surgery duration has had a higher priority than the LOS.

6.3.2 Performance hypothesis

We now discuss our hypotheses regarding the performance and utility of the two-stage stochastic program S1 and pattern models in the context of our particular problem.

Firstly, the pattern model should effectively encapsulate the cancellation rule. The rule is highly complex, and we do not expect traditional two-stage formulations to handle the increased complexity, at least not without major complexity reduction measures.

Further, to handle uncertain surgery duration, we believe the pattern model should perform at least as well as S1, if not surpass it. The stochastic nature of surgery duration makes it a challenging variable to manage; however, there is no reason why accurately calculating the joint probability distributions should perform worse than a sample average approximation method. Therefore, we also expect the pattern model to outperform S1 regarding the stability of the schedule.

Although we anticipate S1 to excel in managing the uncertain LOS better than the pattern models, we predict that P1 will perform significantly better than traditional deterministic formulations. Given that LOS can vary greatly and unpredictably, it is vital that our model can handle this variability to some degree.

However, when it comes to larger problem instances like those encountered in the orthopedic department, we do not foresee S1 being able to handle them adequately since the computational demands and complexity may become too high.

Contrastingly, the pattern models should be considerably easier to solve than S1, thus enabling us to tackle larger problem instances. The simpler deterministic nature of these models should afford them greater efficiency. Further, the tighter formulation using set partitioning constraints instead of knapsack constraints should help reduce the optimality gap efficiently when solving the model.

Chapter 7

Simulation Framework

A simulation framework is developed to test the performance of the mathematical models. The simulation is implemented in object-oriented Python. Section 7.1 gives an outline of the simulation framework before Section 7.2 describes the implementation of the hospital simulation environment in more detail. Section 7.3 presents the evaluation scheme used to evaluate the performance of the models. Section 7.4 presents the input parameters used for the models.

7.1 Simulation Framework Outline

The purpose of the simulation framework is to simulate the surgery scheduling activity and measure the performance of the optimization model over time. We assume that the scheduling decision is made once a week and that a fixed planning horizon is used. We call this weekly scheduling activity a planning stage. The weekly planning stages are implemented in a rolling horizon, where parts of the schedule from the previous week carry over to the next planning stage. This means that at each planning stage, there may already be cases in the schedule in all weeks except the last week of the planning horizon. This is illustrated by the blue and green parts of the schedule in Figure 7.1, which shows the rolling horizon scheme with a planning horizon of four weeks. Figure 7.2 shows the cyclic nature of the simulation scheme.

After each planning stage, the schedule is updated. Two types of events can happen during the week between the planning stages. Firstly, new patients can arrive and be added to the waiting list. Arriving patients are drawn from a Poisson distribution each week as outlined in Algorithm Algorithm 3. Note that the actual surgery duration and LOS for the cases are drawn when generating the case. More details about the probability distributions for surgery duration and LOS are presented later in Section 7.4.3. Secondly, a patient in the schedule can get canceled on the day of surgery due to the preceding surgeries in the same block using enough time to make

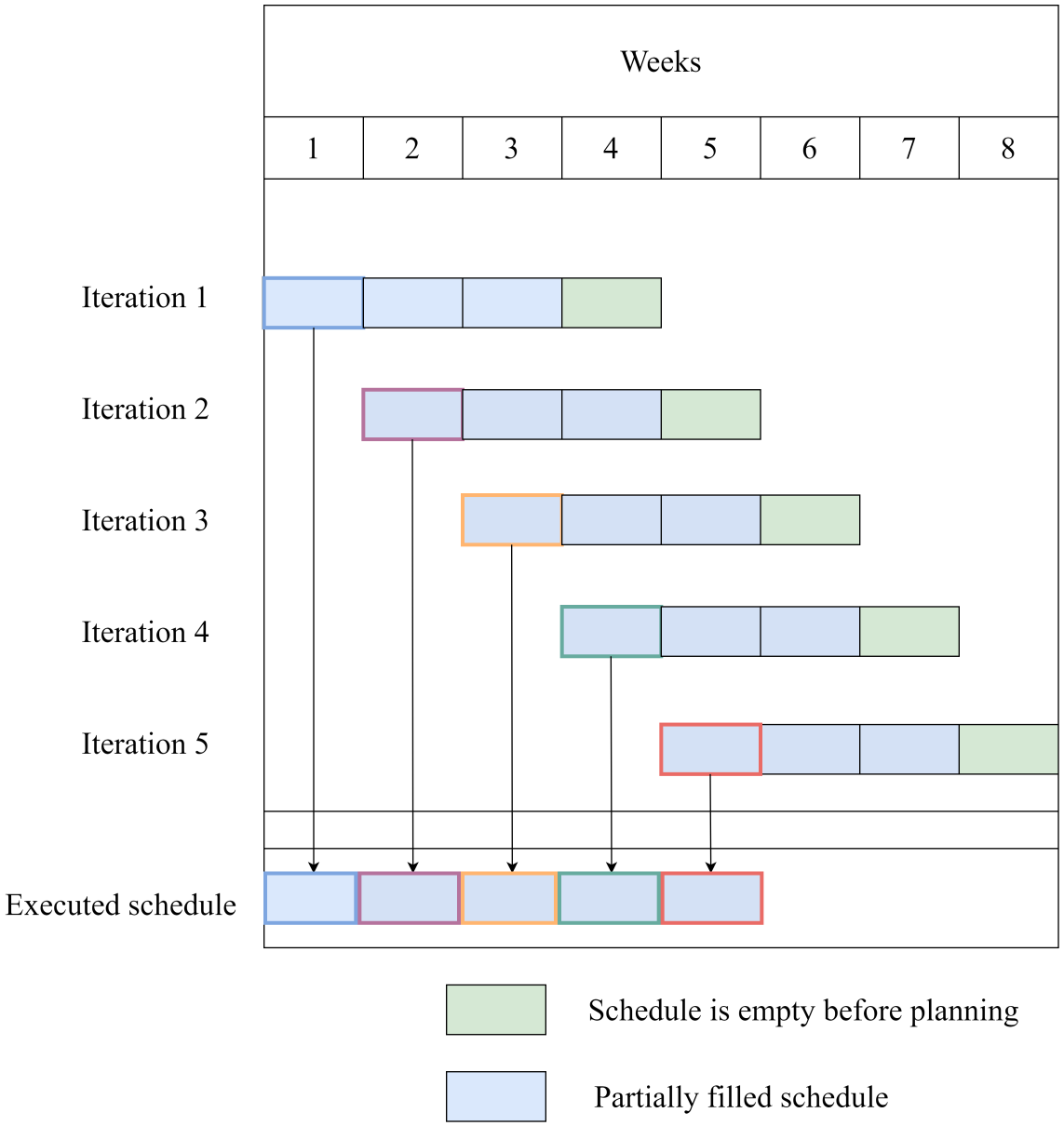


Figure 7.1: Rolling Horizon with four weeks planning horizon.

the hospital decide not to start the next surgery. We call this type of cancellation a Plan Specific Cancellation (PSC). The cancellation rule for PSC can be configured in the simulator. The default cancellation rule for cases in a block is outlined in Algorithm Algorithm 4. The standard sorting rule in the simulation is descending expected surgery duration, i.e., longest first.

At the end of each week in the simulation, all scheduled cases that are not canceled and have a surgery date in the current week are marked as completed. For simplicity, we assume that the actual LOS of a patient is revealed once the surgery is completed. We use this LOS to update the ward capacity input parameter for the next planning stage so that we do not need to include patients that have completed surgery but still need to stay in the wards in the model. The base configuration is that all cases still in the schedule are also scheduled in the following planning stage, i.e., the model cannot cancel patients by default. However, allowing the model to cancel

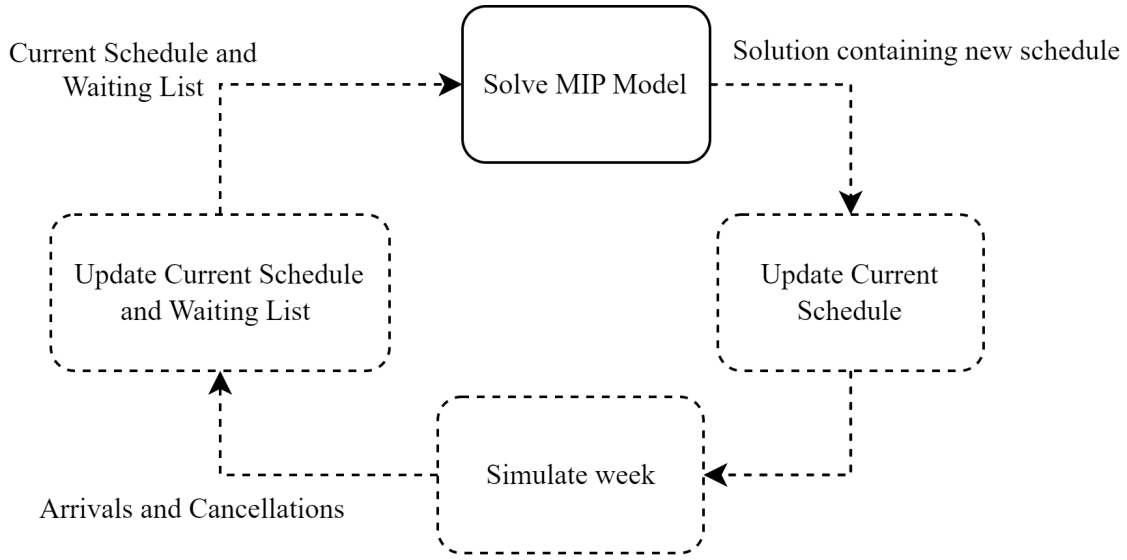


Figure 7.2: Overview of the simulation framework.

scheduled patients with a configuration setting is possible. Algorithm Algorithm 5 summarizes the rolling horizon simulation.

Algorithm 3 Generate Weekly Cases

$P \leftarrow$ Procedures
$t_0 \leftarrow$ First day of the week
$t_1 \leftarrow$ Last day of the week (excluding the weekend)
$C \leftarrow \emptyset$
for t $\leftarrow t_0, t_1$ **do**
  for $p \in P$ **do**
    $\lambda_p \leftarrow$ weekly arrival rate for procedure p
    $N \leftarrow Poission(\lambda_p/5).sample()$ ▷ Divide by 5 since we assume cases only arrive on the weekdays
    for day $\leftarrow 1, N$ **do**
      $\hat{d} \leftarrow$ actual surgery duration ▷ Drawn from the surgery duration distribution of p
      $\hat{l} \leftarrow$ actual LOS ▷ Drawn from the LOS distribution of p
      $c \leftarrow NewCase(p, t, \hat{l}, \hat{d})$
      $C \leftarrow C + c$
    end for
  end for
end for
return C

7.2 Hospital Simulation Environment

A hospital environment is created to track the simulation's events. For this case study, we make the environment similar to the Orthopedic department at St. Olavs hospital, but the environment can easily be modified by changing the configuration files.

The hospital environment the simulator interacts with consists of ORs, Procedures, Cases, and

Algorithm 4 Plan Specific Cancellations

$P \leftarrow$ Sorted list of cases scheduled in the block ▷ Can have different sorting rules
$\overline{D} \leftarrow$ Opening hours of the block
$D \leftarrow 0$ ▷ Variable for used surgery duration
for each $Case \in P$ **do**
  $E[d] \leftarrow$ Expected surgery duration for $Case$
  if $D + E[d] \leq \overline{D}$ **then**
    $Case \leftarrow$ Perform surgery on $Case$
    $\hat{d} \leftarrow$ Actual surgery duration for $Case$
    $D \leftarrow D + \hat{d}$
  else
    $Case \leftarrow$ Cancel $Case$
  end if
end for

Algorithm 5 Rolling Horizon Simulation

$S \leftarrow$ Initial schedule
$WL \leftarrow$ Initial waiting list
$N \leftarrow$ Number of planning stages
for week $\leftarrow 1, N$ **do**
  $P \leftarrow$ Generate Weekly Cases
  $WL \leftarrow$ WL + P
  $MIP \leftarrow$ Create Gurobi model with initialized variables based on S and WL
  $MIP \leftarrow$ Run MIP until configured solution gap or time threshold is reached
  $S \leftarrow$ Update S based on MIP solution
  $WL \leftarrow$ Update WL based on MIP solution
  $C \leftarrow$ Canceled patients ▷ Plan Specific Cancellations
  $S \leftarrow S - C$ ▷ Remove Canceled patients from schedule
  $WL \leftarrow WL + C$ ▷ Canceled patients added back to waiting list
  $S \leftarrow$ Remove completed surgery cases from S
end for

Schedules. Figure 7.3 shows the relation between the entities. The weekend capacity of the ward is used on Friday, Saturday, and Sunday. The reason for including Friday is that inpatient surgeries performed on a Friday need to stay in the ward until at least Saturday morning. In the simulation where LOS is measured in whole days, a case with one day LOS that is operated on Friday will occupy the ward on Friday and not Saturday since we assume the patient is sent home early Saturday.

7.2.1 Simulation Parameters

The simulator has several parameters that can be adjusted to test different behavior. The parameters are the planning horizon length, the number of planning stages, whether or not plan-specific cancellations are activated, and whether or not rescheduling is allowed. Table 7.1 summarize the parameters.

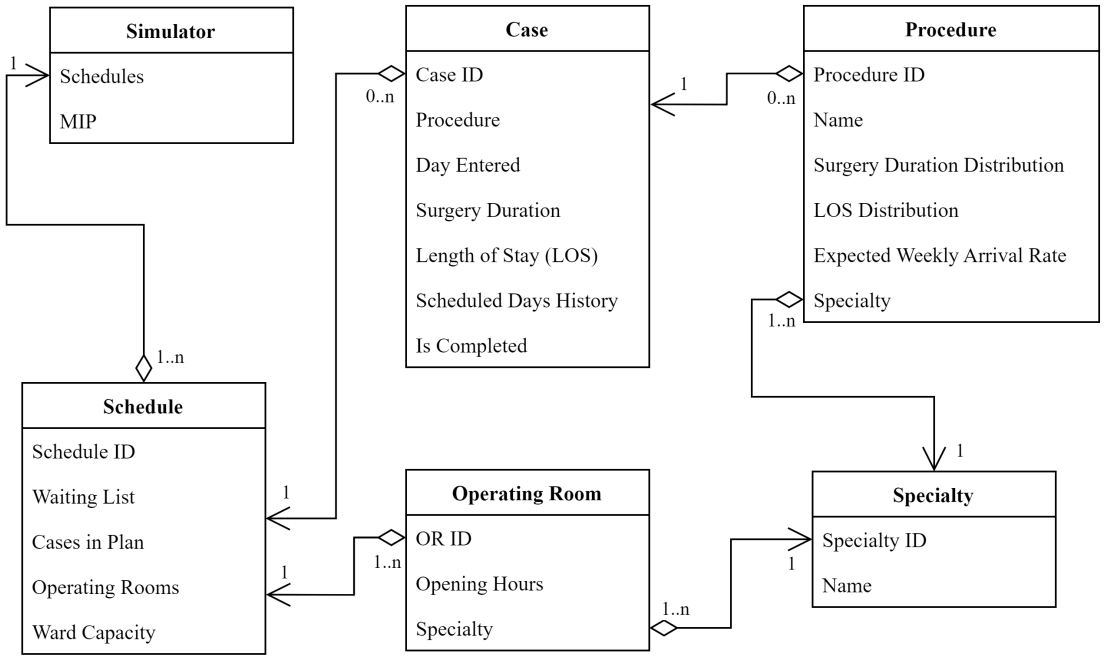


Figure 7.3: Entities and relations in the hospital simulation environment.

Table 7.1: Simulation parameters that can be modified.

Parameter	Description
Planning Horizon	Planning horizon length (weeks)
Planning Stages	Number of planning stages
Allows PCS	Whether or not plan-specific cancellations are activated
Allow Rescheduling	Whether or not rescheduling is allowed

7.3 Evaluation Scheme

We develop an evaluation scheme to compare simulation results with different parameters and models. This scheme is mostly the same as used by Schiøtz and Tysse (2022). We want to have a partially filled plan when the simulation starts. To create this plan, we simulated 60 weeks with a planning horizon of 10 weeks, starting with an empty plan on day 1. The patient arrival rates in the first week were 12 times higher than normal to initialize a realistic volume of cases in the system. We call this 60-week simulation for the *common warm-up* since this is the same plan that will be used for every evaluation. A more detailed description of parameters used in the common warm-up can be found in Appendix B.

When a specific instance is to be tested, an initial schedule is created based on the state of the common warm-up after 60 weeks. The new starting plan is filled up similarly to the common warm-up plan up to week $60 + PH - 1$, where PH is the planning horizon length in weeks for the instance to be tested. If PH is shorter than ten weeks, the planned cases for weeks after $60 + PH - 1$ are put on the waiting list. All rescheduling and cancellation history is cleared from the cases in the new schedule to avoid bias from the common warm-up behavior. After initializing the new schedule based on the common warm-up, an instance-specific warm-up is run for $PH + 1$

weeks to get most of the cases from the common warm-up out of the system to avoid bias. The model is then run until another 12 weeks are executed. These 12 weeks are the only ones to be evaluated, and the planned weeks after the 12 executed weeks are discarded. This whole process is repeated M times with different seeds, where a seed determines the weekly case arrivals. The evaluation scheme is visualized in Figure 7.4

After running the simulation M times, we can calculate an average and a confidence interval for the KPIs we want to measure. Figure 7.5 shows how the confidence interval develops when increasing M. From these results, we conclude that $M = 20$ gives a good trade-off between the confidence interval and how many times we need to run the model. If not explicitly stated otherwise, the results presented in the following chapter are created using the evaluation scheme presented in this section.

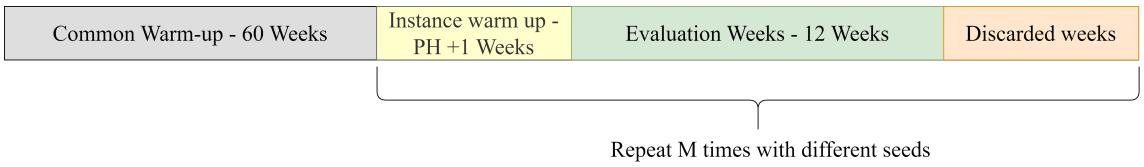


Figure 7.4: Scheme used to select weeks for evaluation.

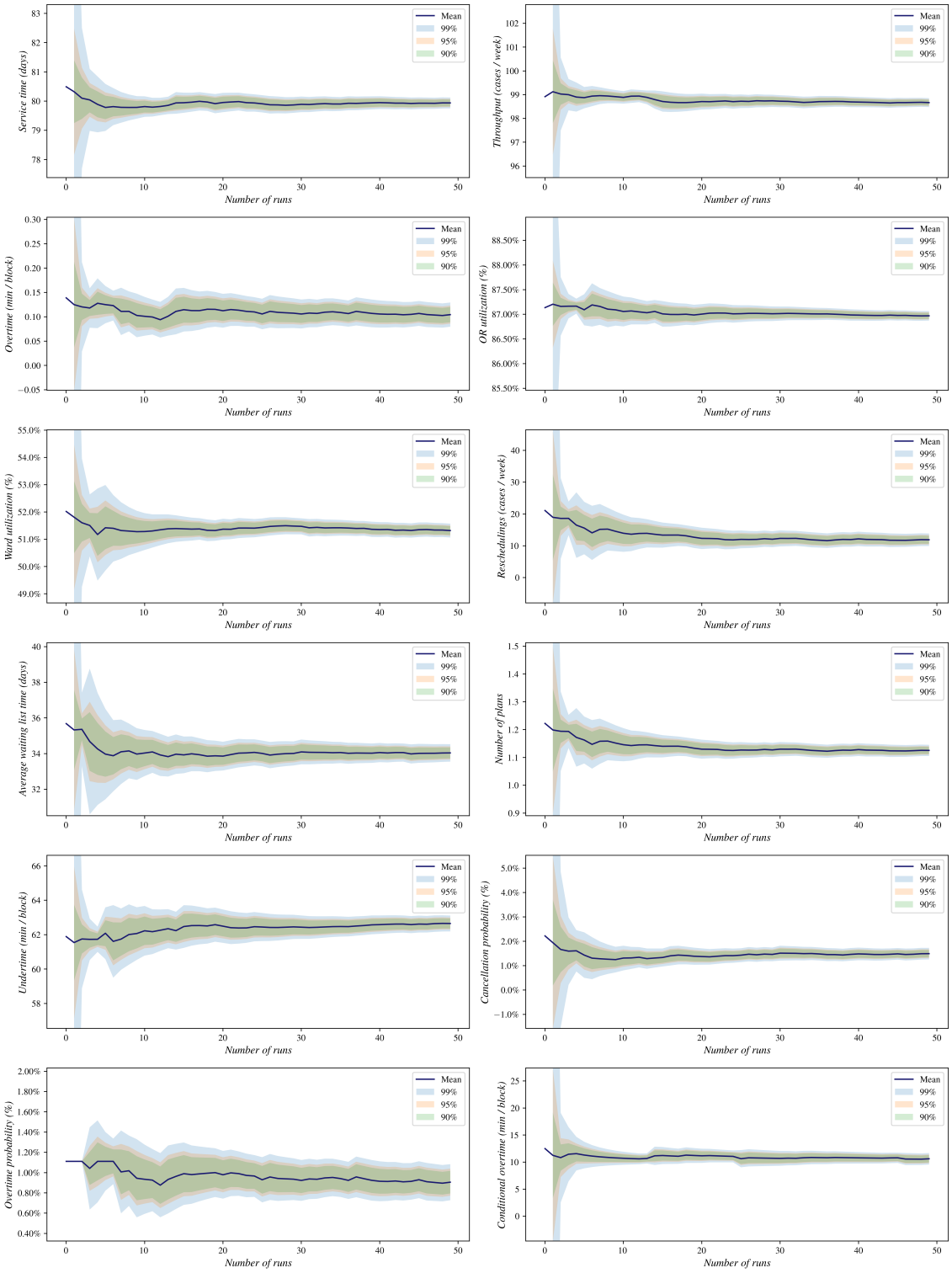


Figure 7.5: Confidence interval of KPIs when running more simulations.

7.4 Input parameters

The input parameters to the model are inspired by data from the orthopedic department at St. Olavs Hospital, gathered from T. R. Bovim (2018) and T. R. Bovim et al. (2020). We create two sets of inputs, one real-life size set and one smaller set. From now, 'large case' refers to the real-life set of parameters, and 'small case' refers to the smaller set. We will see later that the stochastic two-stage model struggles with solving the large case.

7.4.1 Hospital resources

Hospital resources included in the parameters are ORs and wards. We first introduce the ORs. In the large case, we use seven elective ORs, representing the ORs at BVs at St. Olavs Hospital. The MSS for the ORs in the large case is shown in Table 7.2. The small case uses two ORs, and we include only two specialties. Table 7.3 shows the MSS for the ORs in the small case. If an OR does not have a specialty on a day in the MSS, it is closed that day. The standard opening hours for an open OR are 8 hours (480 minutes) a day.

Table 7.2: The MSS used in the large case. This MSS represents the seven elective ORs at BVS at St. Olavs Hospital. Note that some of the ORs are closed on Thursday and Friday.

OR	Monday	Tuesday	Wednesday	Thursday	Friday
1	Foot	Plastics	Plastics	Plastics	Foot
2	Plastics	Plastics	Plastics	Plastics	-
3	Hand	Plastics	Hand	Arthroscopic	Hand
4	Arthroscopic	Arthroscopic	Arthroscopic	Arthroscopic	Arthroscopic
5	Back	Back	Back	Tumor	-
6	Prosthetics	Prosthetics	Prosthetics	Prosthetics	-
7	Prosthetics	Prosthetics	Prosthetics	-	-

Table 7.3: The MSS used in the small case. This MSS represents two elective ORs at St. Olavs Hospital. Note that these are identical to OR numbers 3 and 5 from the large case after removing all except hand and back blocks.

OR	Monday	Tuesday	Wednesday	Thursday	Friday
1	Hand	-	Hand	-	Hand
2	Back	Back	Back	-	-

We assume shared ward capacity in the hospital, with reduced weekend capacity. Table 7.4 shows the default parameter values for ward capacity used in both the large and the small case. The values for the large case are the sum of the actual ward capacities at St. Olav's Hospital in the data from T. R. Bovim et al. (2020). Values for the small case are set based on some initial testing.

Table 7.4: The ward capacities used in the large and small case.

Case	Weekday Capacity	Weekend Capacity
Large	62	44
Small	5	3

7.4.2 Procedures

Some specialties only have one procedure in the data from T. R. Bovim et al. (2020). To introduce more variability in the input data, we create two additional procedures from every original procedure, one with a higher expected surgery duration and one with lower. We let the symbols L, M, and H refer to the three resulting procedures with expected surgery duration for the lower, original, and higher expected surgery duration, respectively. Expected surgery duration for L, M and H procedures derived from original procedure p is defined as μ_p^L, μ_p^M and μ_p^H, respectively. The expected length of stay for the two new procedures is the same as the original and denoted l_p. Surgery duration and LOS distributions are described in Section 7.4.3.

The weekly arrival rate for the L, M and H procedures derived from original procedure p is defined as λ_p^L, λ_p^M, and λ_p^H, respectively. Arrival rates have been tuned to achieve a mostly stable number of surgical cases in the system, i.e., the waiting list is neither emptied nor diverging over a longer period of time. Table 7.5 shows the procedures used in the large case, and Table 7.6 shows the procedures used in the small case. Note that the large case has 45 different procedures, and the small case has nine since the original procedures' L, M, and H variants are modeled as separate procedures when generating input to the model.

Table 7.5: Procedures used in the large case. μ_p^L, μ_p^M and μ_p^H are the expected surgery duration for the L, M and H procedures derived from original procedure p, respectively. l_p is the expected LOS for all versions of procedure p. λ_p^L, λ_p^M and λ_p^H are weekly arrival rates for the L, M and H procedures derived from original procedure p, respectively.

Name	**Specialty**	μ_p^L	μ_p^M	μ_p^H	l_p	λ_p^L	λ_p^M	λ_p^H
Aggregated Foot	Foot	110	140	180	3	1.13	3.40	1.13
Aggregated Hand	Hand	70	90	120	0.1	2.02	6.06	2.02
Carpal Tunnel Syndrome	Hand	60	90	110	1	0.82	2.47	0.82
Aggregated Plastics	Plastics	70	100	120	2	3.16	9.49	3.16
Plateepitelkarsinom	Plastics	50	70	90	1	0.45	1.35	0.45
Bcc	Plastics	110	140	180	1	1.49	4.48	1.49
Malignt Melanom	Plastics	50	70	90	0.1	0.87	2.6	0.87
Cancer Mammae	Plastics	70	100	120	1	0.85	2.56	0.85
Aggregated Arthroscopic	Arthroscopic	90	120	150	2	1.62	4.87	1.62
Acl	Arthroscopic	140	190	230	2	0.70	2.09	0.70
Meniscus	Arthroscopic	130	170	220	0.1	0.96	2.89	0.96
Aggregated Back	Back	220	300	370	6	0.63	1.89	0.63
Hip	Prosthetics	130	180	220	4	1.26	3.8	1.26
Knee	Prosthetics	130	170	220	4	1.86	5.57	1.86
Aggregated Tumor	Tumor	60	80	100	1	1.16	3.47	1.16

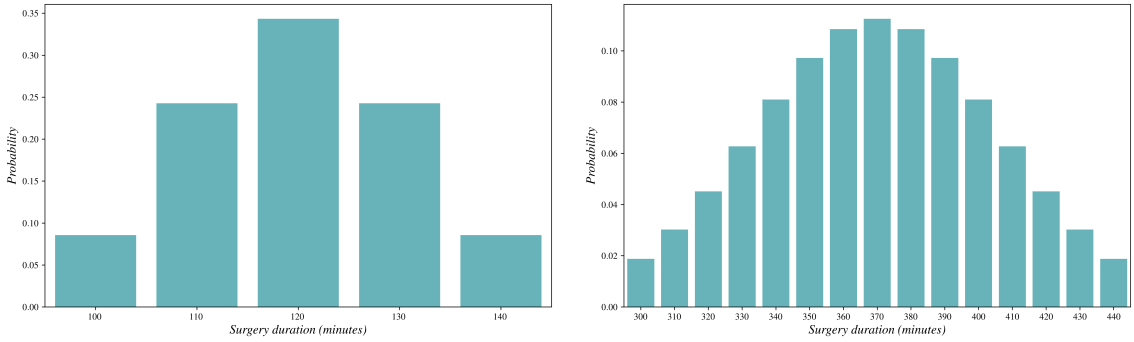
Table 7.6: Procedures used in the small case. Notation is the same as in Table 7.5

Name	**Specialty**	μ_p^L	μ_p^M	μ_p^H	l_p	λ_p^L	λ_p^M	λ_p^H
Aggregated Hand	Hand	70	90	120	0.1	2.02	6.06	2.02
Carpal Tunnel Syndrome	Hand	60	90	110	1	0.82	2.47	0.82
Aggregated Back	Back	220	300	370	6	0.63	1.89	0.63

7.4.3 Probabilty distributions for surgery duration and LOS

Similar to Min and Yih (2010) and Zhang et al. (2019), we use discrete distributions for the random variables for surgery duration and LOS of each procedure. While the surgery duration in real life is continuous, Min and Yih (2010) argues that the assumption of discrete surgery duration is reasonable because surgeries are scheduled in discrete time intervals in real life, e.g., 30 minutes instead of 29.5 or 31.5 minutes. The use of discrete distributions also makes the number of stochastic scenarios finite and reduces the computational complexity (Zhang et al. 2019). Similar to Zhang et al. (2020), we use truncated distributions to prevent the optimal solution of the stochastic problem from being over-conservative due to extremely long surgery durations or LOS that are very unlikely to happen in real life.

For surgery duration, we use a discretized truncated normal distribution with a mean equal to the expected surgery duration and a standard deviation equal to 10% of the mean. The distribution is discretized in 10-minute intervals and truncated between 80% and 120% of the mean surgery duration. The interval and boundaries are set arbitrarily and does necessarily reflect the actual distribution of surgery duration for patients at St. Olavs hospital. Regardless, we argue that the methods applied in this thesis can be used with any distribution. Thus, we chose a fairly straightforward distribution for simplicity, as finding the best distribution for surgery durations at St. Olavs is beyond the scope of this thesis. Figure 7.6 shows the surgery duration distributions for two selected procedures.



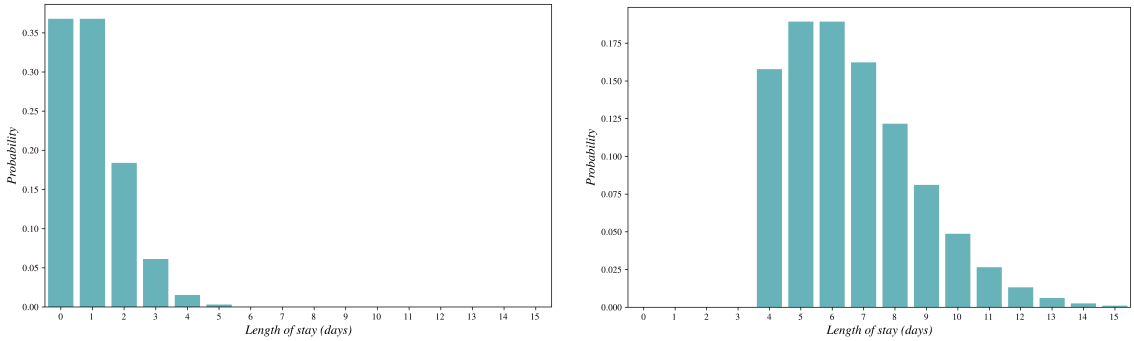
(a) Surgery duration distribution for Aggregated hand variant H with $\mu_p^H = 120$

(b) Surgery duration distribution for Aggregated Back variant M with $\mu_p^M = 300$

Figure 7.6: Surgery duration distributions of two selected procedures.

For LOS, a truncated Poisson distribution is used. Since LOS is measured in whole days, using a Poisson distribution is intuitive since it is already discrete. As with surgery duration, we truncate the LOS distributions to avoid highly unrealistic values. The lower limit of the LOS distribution is set to $\max(0, l_p - 2)$ days for each procedure p. The motivation for the lower limit is that we assume that the hospital usually requires patients to stay a minimum number of days after surgery, e.g., it is highly unlikely that a patient can get a LOS of zero days after a five-hour back surgery and walk home the same day. The upper limit for LOS is set to 15 days, an arbitrarily chosen

number. Figure 7.7 shows the LOS distribution for three selected procedures.



(a) LOS distribution for Carpal Tunnel Syndrome with $l_p = 1$

(b) LOS distribution for Aggregated Back with $l_p = 6$

Figure 7.7: LOS distributions of two selected procedures.

Chapter 8

Computational Study

This chapter extensively studies the proposed pattern-based MIP model within a rolling horizon simulation framework. Remember, three goals were defined for this thesis: (1) to enhance the mathematical model proposed by (Schiøtz and Tysse 2022) by incorporating variables such as uncertain surgery duration and recovery time, (2) explore how the cancellation rules applied at the Department of Orthopaedics's and the accepted overtime and cancellation risk levels affect scheduling quality, and (3) to incorporate the models into a rolling horizon framework. We define four research topics based on these goals, the problem description, and our proposed solutions. These topics are shown in Table 8.1.

The computational study is divided into three sub-studies: (1) technical, (2) uncertainty, and (3) cancellation rule and risk study. Table 8.1 shows a simplified view of which research topics are the main focus for the different studies. One should note that none of the studies are mutually exclusive, and the findings in one study may be used as benchmarks in another. For instance, the uncertainty study evaluates different models in simulation environments without a cancellation rule. This creates a benchmark we later use to compare the schedule quality with schedules where cancellation rules apply. However, the uncertainty study does not explicitly answer RT4, only implicitly through the cancellation study, and is therefore not marked in the table.

Including the research topics, each study has goals to guide each study and ensure the research questions are answered across the studies. The goals associated with each study are shown in Table 8.2. Note how the uncertainty study focuses on the overtime part of RT2, while the primary purpose of the cancellation rule and risk study is to investigate the cancellations.

We will compare the scheduling quality of different schedules based on a set of key performance indicators (KPIs) for the uncertainty study and the cancellation rule and risk study. Table 8.3 describes the most used KPIs throughout the study. Note that not all KPIs are relevant for both studies. For instance, the cancellation probability is irrelevant in the uncertainty study since

Table 8.1: Which research topics are most relevant for the different studies?

ID	Research Topic	Study		
		Technical	Uncertainty	Cancellation and Risk
RT1	Can a deterministic pattern-based MIP improve computational efficiency while handling uncertainty?	X	X	
RT2	How is schedule quality affected by overtime and cancellation risk?		X	X
RT3	How does the use of cancellation rules affect scheduling quality			X
RT4	What value do models that include a hospital's cancellation rule add?			X

cancellations are not part of the research topics investigated in the study.

The rest of the chapter is structured as follows: First, Section 8.1 presents an adverse selection of instances to be used in the study. We define an instance as a combination of a mathematical model, simulation setting, pattern file, and model parameters. Second, Section 8.2 performs a technical study on the pattern model, two-stage stochastic model, and the Expected Value (EV) equivalent of the two-stage stochastic model to compare the complexity and robustness of the models. Third, Section 8.3 investigates how the uncertainty affects schedule quality in an environment without cancellations to be able to test how the cancellation rule affects scheduling quality. Lastly, Section 8.4 investigates how risk acceptance for overtime and cancellations affect scheduling quality.

8.1 Experimental setup

Multiple variations and combinations of pattern files, model types, and simulation settings are used to reach each study's goals. We define one single combination of these factors as an instance. Throughout the studies, these instances will be benchmarked against each other, and we will use the differences between the instances and the resulting schedules to answer our research topics.

The hardware used for all runs is shown in Table 8.4. Using the same hardware is essential to ensure fair benchmarking. Different hardware can have different computational capabilities, such as processing speed, memory, and storage speed. With the same hardware, we minimize the risk of noise and ensure that the differences we observe are most likely due to the differences between the instances and not the hardware. Still, some possible limitation exists, like cache and system loading. We have not used the hardware for anything else when running the instances and only run one instance at a time.

Table 8.2: The overall goals of the computational study summarized

Study	Goals	Why
Technical Study	1. Benchmark the efficiency and complexity of the models	To validate the pattern model's performance against two-stage stochastic and deterministic formulations.
	2. Investigate how performance depends on problem complexity	Assess if model preference depends on problem size and if the two-stage stochastic model can manage real-world problems.
	3. Analyse the stability of the two-stage stochastic model	To ensure a fair benchmarking, by validating that the two-stage stochastic model are in- and out-of-sample stable
Uncertainty Study	1. Compare scheduling quality without cancellation rules	To validate the pattern model and identify its strengths and weaknesses compared to two-stage stochastic and deterministic formulations
	2. Investigate how uncertainty affects scheduling quality	Establishing a benchmark without cancellations allows for later comparison when they are included.
	3. Investigate how overtime risk levels affect the schedules	Analyze overtime risk acceptance without cancellations for benchmarking, which facilitates later investigation of the cancellation rule.
Cancellation and Risk Study	1. Study how including a cancellation rule affects scheduling quality	Understanding the consequences of the cancellation rules can give insights into the alternative cost
	2. Study how cancellation risk levels affect the schedules	To give insights into the cost of hedging against cancellation risk in combination with overtime risk to be used as managerial support
	3. Consider the value of including cancellations in the models	To understand what value models with cancellations provide in exchange for increased complexity.

Table 8.4: The details of the computer and Gurobi solver used during all simulations.

Component	Type
Processor	Intel Core i7-10700 @ 2.90GHz, 8 cores, 16 threads
RAM	16GB, 2933 MHz
Gurobi version	9.5.2
Python version	3.10.8
Operating system	Windows 10 Education, Version 22H2

Models

Table 8.5 show the models studied in the chapter. $S1$ is the two-stage-stochastic model as described in Section 5.2. $D1$ is the deterministic equivalent of $S1$. Deterministic equivalent means the Expected Value Problem (EVP), a setting with only one scenario where all uncertain parameters get their expected value. Lastly, three types of the pattern model, as described in Section 5.3, are used: (1) without chance constraints (the base case), (2) only with overtime chance constraints, and (3) with both overtime and cancellation chance constraints.

Table 8.3: The most used quality attributes/KPIs throughout the study

Category	Performance indicator	Measure	Unit
Hospital related efficiency	OR utilization	Number of utilized time slots out of total OR capacity	%
	Ward utilization	Number of beds used in the ward compared to the total amount of available beds	%
	Undertime	Amount of unused time slots in a block	Min
	Throughput	Number of patients treated	Cases
Patient-related efficiency	Service Time	Days between entering the waiting list and day of surgery	Days
	Maximum waiting list time	Average number of days a patient in the waiting list has waited at the end of the horizon	Days
	Waiting List Time	Number of days waiting for the patient in the waiting list who has waited the longest	Days
Stability	Overtime probability	Probability of overtime occurring in a block	%
	Conditional overtime	Expected amount of overtime given that overtime occurs	Min
	Cancellation probability	Probability that a case is canceled at the day of surgery	%
	Number of Plans	Number of different surgery dates given to a patient	Plans
	Rescheduling	Number of rescheduled patients	Cases

Table 8.5: The different models studied

Name	Model type	Constraints	Description
D1	Deterministic (EVP)	Base	The stochastic model with only one scenario equal the expected values
S1	Stochastic	Base	The two-stage stochastic model as previously presented
P1	Pattern	Base	The base pattern model without any chance constraints or cancellations
P2	Pattern	Base + Overtime	Pattern model extended with overtime chance constraints
P3	Pattern	Base + Overtime + Cancellation	Pattern model with both overtime and cancellation chance constraints

Simulation setting

The simulation settings $\{SC0, SC1\}$ represent whether cancellation rules apply and are enabled in the simulation environment. The $SC1$ settings imply that the cancellation rules defined in chapter 6 are used, while no cases will be canceled if $SC0$ is used. Then, the surgical team will instead work as much overtime as needed to finish the schedule associated with the block.

Pattern files

We also have two kinds of pattern files, $\{PC0, PC1\}$. If $PC0$ is used, the models do not assume cancellation rules. I.e., the expected overtime parameter used by the pattern models is calcu-

lated based on the assumption that all scheduled cases receive surgery. Pattern file $PC1$ assumes a cancellation rule, and the ward bed demand, expected overtime, and cancellation probability parameters are calculated as presented in chapter 6.

8.1.1 Instances

We only use one instance per model type, $D1$, $S1$, and $P1$, for the technical and uncertainty study. Cancellations are not part of the research topics relevant to these studies, and $PC0$ is thus the pattern file used during the studies. We will write $P1$ instead of $P1 - PC0$ for simplicity. In the technical study, we only run the instances for one planning stage and do not use the rolling horizon simulation. Therefore, no simulation setting exists. However, the rolling horizon setting is always used in the uncertainty study.

Several instances are included in the cancellation study, as shown in Table 8.6. P2-SC0 and P2-SC1 are included to investigate how simulated cancellations affect models which do not assume cancellations. The other instances have the same level of overtime risk, η_o, but different levels of cancellation risk, η_c. The reader should note that the overtime and cancellation risks are chosen such that it is possible to identify the differences between the instances. Hence, some instances are included even if we do not believe the resulting schedules to be advisable. For instance, P3-C50, which allows patterns with up to 50% cancellation, is not expected to create high-quality schedules but enables us to investigate what happens when cancellation risk is unimportant. Lastly, in the risk study, we use both P2 and P3 with even larger variations of risk parameters to identify the general trends.

Table 8.6: Instances for cancellation study

Name	Compact form	Model	Sim enviroment	Pattern file	η_o	η_c
D1-SC1-PC0-O10	D1-SC1	Deterministic 1	With cancellations	No cancellations	-	-
P2-SC0-PC0-O10	P2-SC0	Pattern 2	No cancellations	No cancellations	10%	-
P2-SC1-PC0-O10	P2-SC1	Pattern 2	With cancellations	No cancellations	10%	-
P3-SC1-PC1-O10-C1	P3-C1	Pattern 3	With cancellations	With cancellations	10%	1%
P3-SC1-PC1-O10-C10	P3-C10	Pattern 3	With cancellations	With cancellations	10%	10%
P3-SC1-PC1-O10-C25	P3-C25	Pattern 3	With cancellations	With cancellations	10%	25%
P3-SC1-PC1-O10-C50	P3-C50	Pattern 3	With cancellations	With cancellations	10%	50%

8.2 Technical Study

In this section, we investigate the complexity and stability of the mathematical models. First, the stability of the S1 model with different numbers of scenarios is analyzed in Section 8.2.1. Section 8.2.2 then analyzes the computational complexity of the D1, P1 and S1 models for different problem sizes.

In the complexity analysis, the models are tested on both the large and small cases introduced in

Section 7.4, with the planning horizon ranging from 1 to 10 weeks. The initial state of the hospital environment contains a waiting list and a schedule where the last week of the planning horizon is empty, but the other weeks are partially filled with surgical cases. The total number of cases in the initial environment is 104 in the small case and 1075 in the large case. The split of surgical cases between the current schedule and the waiting list is given by the planning horizon. The models are run 20 times each using the same input to get an average solution since we experience some variation in the solving time between runs. Only a single planning stage is executed, not a complete rolling horizon simulation. The Gurobi time limit is set to 10 800 seconds, and the gap limit is 0.01%. The other parameters are set to the default values, shown in Appendix B.

8.2.1 Stability of the two-stage model

We will now investigate how many samples we need to get a good approximate solution using SAA. The test is performed on the small case with a four-week planning horizon. An advantage of the problem formulation of the two-stage model is that for a given first-stage solution $\overline{x}$, the second-stage cost can be calculated without running the optimization model. For a given feasible solution $\overline{x}$, the optimal amount of overtime and extra ward beds in a given scenario can be found using (8.1) and (8.2), respectively, where $W_{t\xi}$ is the number of patients in the wards at day t in scenario ξ, calculated using Algorithm 6. This makes it possible to evaluate a solution on a much larger number of scenarios than when running the two-stage model, getting a better estimate of the true objective value.

$$\phi_{rt\xi} = \max\left(0, \sum_{i \in \mathcal{I}} D_{i\xi} x_{irt} - \overline{D}_{rt}\right) \qquad r \in \mathcal{R}, t \in \mathcal{T}, \xi \in \Xi \tag{8.1}$$

$$\rho_{t\xi} = \max(0, W_{t\xi} - K_t) \qquad t \in \mathcal{T}, \xi \in \Xi \tag{8.2}$$

Algorithm 6 Calculate number of patients in wards in a given scenario

```
ξ ← current scenario
W_tξ ← Number of patients in the wards at day t in current scenario, initialized as 0 for all t
Cases ← All scheduled cases from the first stage solution, i.e., all i where v_i = 0
for each i ∈ Cases do
    L_iξ ← LOS for case i in current scenario
    t̂ ← Day of surgery for case i
    t* ← t̂ + L_iξ − 1                    ▷ Last day where case i occupies a ward bed
    for t = t̂, ..., t* do
        W_tξ ← W_tξ + 1
    end for
end for
return W_tξ
```

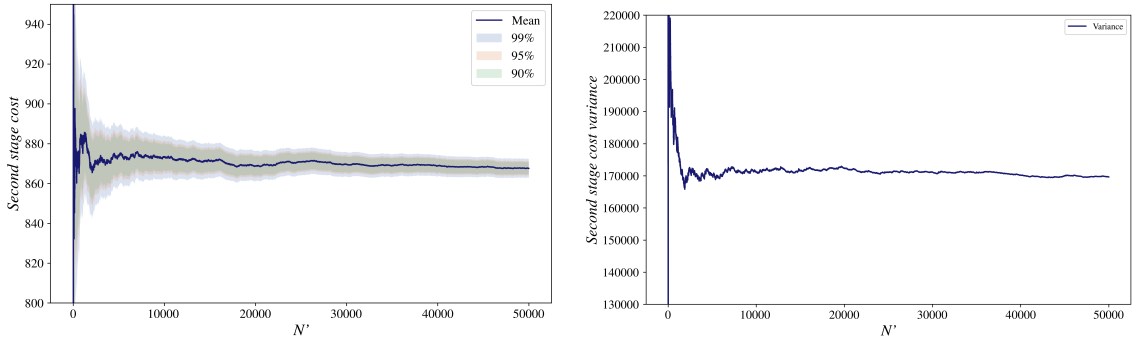
To evaluate the stability and convergence of the SAA solutions with different numbers of sampled scenarios, we use the procedure described in Section 3.3.1. For a given first-stage solution $\overline{x}$, let the

first-stage costs, i.e., objectives (5.4a) in Section 5.2.2 be defined as $F(\overline{x})$. For the solution $\overline{x}$ and a given scenario ξ, let the second stage costs of this scenario be denoted $Q(\overline{x}, \xi)$, and calculated as shown in (8.3), where $\phi_{rt\xi}$ and $\rho_{t\xi}$ is calculated as described above.

$$Q(\overline{x}, \xi) = \omega_\Phi \sum_{r \in \mathcal{R}} \sum_{t \in \mathcal{T}} C^\Phi \phi_{rt\xi} + \omega_P \sum_{t \in \mathcal{T}} C^P \rho_{t\xi} \tag{8.3}$$

To investigate the convergence of the objective value for different numbers of scenarios, we want to calculate the lower and upper bounds described in Section 3.3.1.

First, we decide which value of N' to use for evaluating the true objective value. We choose a first-stage solution $\overline{x}$ and calculate the mean, variance, and confidence interval of the second-stage costs, $\sum_{n=1}^{N} Q(\overline{x}, n)$, for multiple values of N'. Figure 8.1 shows how the mean, confidence intervals, and variance change when N' increases. We conclude that $N' = 40000$ is adequate to evaluate the true objective value and will use this value of N' when doing the rest of the calculations in this section.



(a) Cumulative mean of second stage cost for N' samples with confidence intervals.

(b) Variance of the second stage cost for N' samples.

Figure 8.1: Stability of the estimated true objective of the second stage

Similar to Min and Yih (2010), we use $M = 10$ in the SAA procedure. Figure 8.2 shows the upper and lower bounds obtained for different values of N. Additionally, the expected value problem (EVP), i.e., the D1 model, is solved on the same input, and a true objective value for the solution obtained by the EVP is also shown in the figure. The two-stage model clearly outperforms the EVP when looking at the true objective value estimate. For $N \geq 250$, we can see that the value of the stochastic solution (VSS), i.e., the difference between the true objective value of the EVP and the stochastic model, is around 500. An overtime cost of 8 was used in this particular model, so using a two-stage model, in this case, potentially saved around 60 minutes of overtime during the planning horizon compared to the EVP. Alternatively, since the extra ward bed cost is 200, we potentially saved renting 2.5 extra beds. Figure 8.3 shows two metrics we can use to evaluate how many scenarios we need to sample, the optimality gap $\hat{g}^m_{N'} - \overline{z}^M_N$, and an estimated variance $\sigma^2_{\hat{g}^m_{N'}} + \sigma^2_{\overline{z}^M_N}$. For each number of scenarios, N, we chose the solution m based on the lowest $\hat{g}^m_{N'}$ in step 3 of the procedure described in Section 3.3.1. Based on the results in Figure 8.2 and

Figure 8.3, we could argue that $N = 250$ is sufficient to get a good approximation of the complete set of scenarios. However, based on the variance estimators in Figure 8.3a, we will use $N = 500$ when running the model for the rest of this thesis if nothing else is explicitly mentioned.

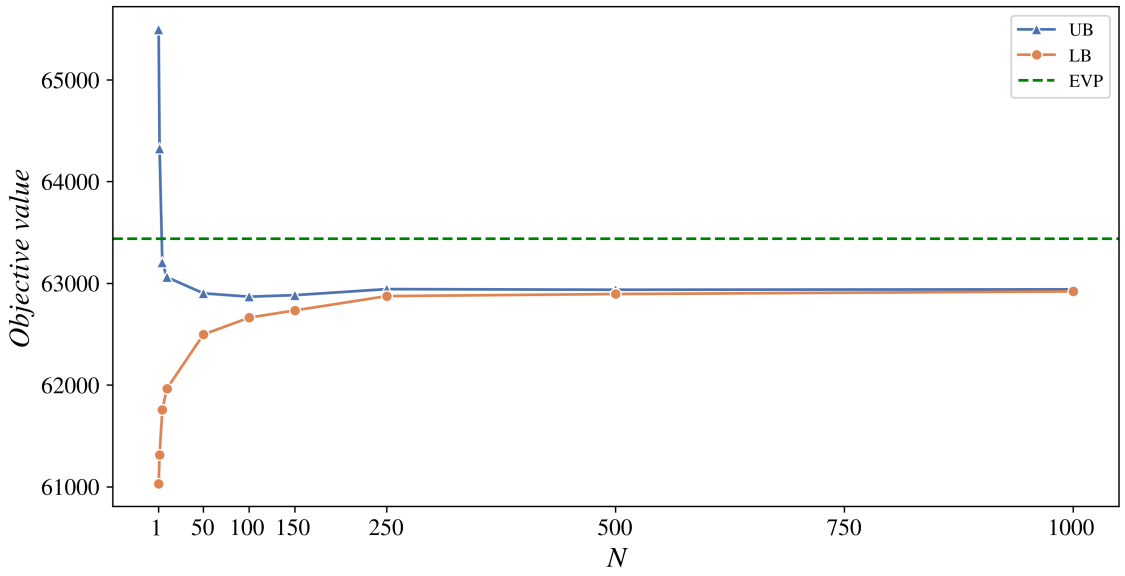
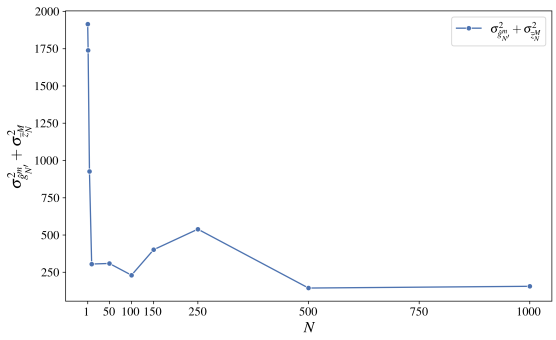
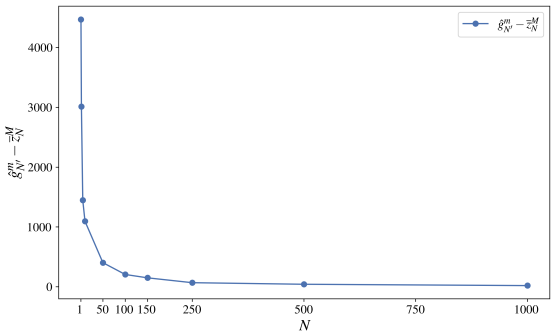


Figure 8.2: Upper and lower bounds for the true objective value and the true objective value of the EVP solution



(a) $\sigma^2_{\hat{g}^m_{N'}} + \sigma^2_{\overline{z}^M_N}$ for a selected solution m for different N.



(b) Estimated optimality gap $\hat{g}^m_{N'} - \overline{z}^M_N$ for a selected solution m for different N

Figure 8.3: Gap and variance estimators for different values of N

8.2.2 Complexity

To evaluate the complexity of a model, we can measure the time to find a feasible solution and the gap over time. The gap is defined in (8.4), where Z_P is the primal bound, and Z_D is the dual bound of the solution. The Gurobi model requires a gap limit to be defined as the optimal solution. We set this limit to 0.01% and thus count a gap below this threshold as an optimal solution. For reference, the models usually have an objective value of around 60000 as we saw in Figure 8.2, so a gap of 0.01% corresponds to a difference between the primal and dual bound by about the

cost of one minute overtime for the whole planning horizon, which we consider insignificant. We start by comparing the D1, P1, and S1 models with 500 scenarios on the small case, using different planning horizons. The models are then tested on the large case, representing the real-life problem faced by St. Olav's Hospital. Finally, we show how the complexity of the S1 model relates to the number of scenarios used in the SAA algorithm. All models are run until the optimal solution is found or a time limit of 10800 seconds is reached.

$$gap = \frac{|Z_P - Z_D|}{|Z_P|} \tag{8.4}$$

Small case

Figure 8.4 shows the gap development for different planning horizons for the S1 model. We do not show the corresponding plots for the D1 and P1 models since they both reach very small gaps in a couple of seconds for all planning horizons in the small case. However, we would like to note that the time to reach the optimal solution for the D1 and P1 models increases with the planning horizon, similar to the trend we see in Figure 8.4 that S1 takes a longer time to close the gap with longer planning horizons. A detailed overview of the time used to reach a feasible solution, optimal solution, and selected gaps can be found in Appendix C.1 for all planning horizons 1-10 for all models. We will now look closer at the four-week planning horizon case since most of the results presented in the upcoming sections are made using a four-week planning horizon.

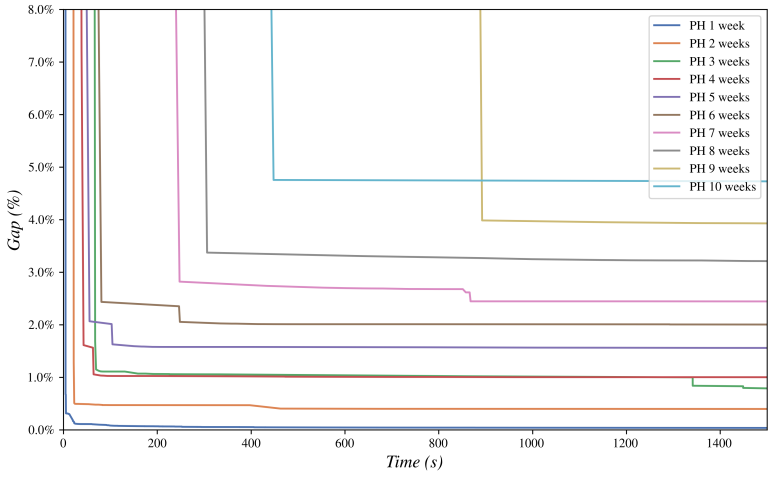


Figure 8.4: Gap over time for S1 for planning horizons 1-10 weeks in the small case.

Table 8.7 compares the three models on the number of seconds used to reach certain gaps with a planning horizon of four weeks. It also shows the time to find the first feasible and optimal solutions. We see that both the deterministic and pattern-based model is able to find the optimal solution within a few seconds, while the two-stage model never reaches the optimal solution and uses almost an hour to reach a 1% gap. However, we see in Figure 8.4 that S1 is actually quite

close to 1% in less than 100 seconds but then uses a long time to finally get under 1%. We believe a gap close to 1% provides a good surgery schedule not too far from the optimal, so in the upcoming sections, we will compare the schedules created by the D1, S1, and P1 models with a four-week planning horizon for the small case.

Table 8.7: Time, in seconds, to reach solutions and gaps for the D1, P1, and S1 models with a four-week planning horizon in the small case.

Model	First Feasible	20%	5%	1%	Optimal
D1	0.60	0.60	0.60	0.68	2.18
P1	0.11	0.27	0.27	0.27	2.37
S1	14.50	38.46	42.76	3472.21	-

Large case

Figure 8.5 shows the development of the optimality gap for the D1 model with different planning horizons for the large case in the first 600 seconds. The values represent the smoothed average of 20 runs of the same model with identical input, which explains why some graphs are not strictly declining. We see that the model is able to find a gap of less than 1% in less than 300 seconds for all planning horizons. When increasing the planning horizon, we would assume the problem complexity to increase. However, interestingly a nine-week planning horizon seems more difficult for the model to solve than ten weeks. This may be because, with ten weeks, it is possible to fit all the cases in the schedule without leaving any on the waiting list. Thus, the model does not need to find out which cases to leave on the waiting list and can focus on shuffling the cases in the schedule to a favorable solution. For long planning horizons, the model seems to struggle with closing the gap completely. We believe a reason for this may be the difference in magnitude of the objective function terms. For instance, in a typical solution, the overtime term is in the range of 400-500, while the scheduling and deferral costs are around 30 000 each. This may lead the solver to get relatively good gaps in the start by focusing on the large cost terms but then stagnating when needing to do more fine-tuned optimization to close the last percentage gap.

We perform a similar analysis on the P1 model as the D1 model. Figure 8.6a shows the smoothed average gap over time development, and Figure 8.6b shows the gap development of a single run. Similar to the D1 model, we see that the P1 model can find small gaps in a reasonable time. In Figure 8.6b, we see that the model has an interesting step-wise improvement of the gap with quite large steps. Figure 8.6b only shows a single run, but we consistently observe this behavior from the P1 model. These results can be explained by the tighter formulation of the P1 model, as it seems to cut away infeasible solutions efficiently. We also see that the problem with the ten-week planning horizon is easier than the nine-week, similar to the D1 model.

We do not show the gap over time development for the two-stage model for the large case. The reason is that when attempting to use 500 scenarios in the large case with four weeks planning horizon, the computer runs out of memory and thus can not solve the model. In Section 8.2.1,

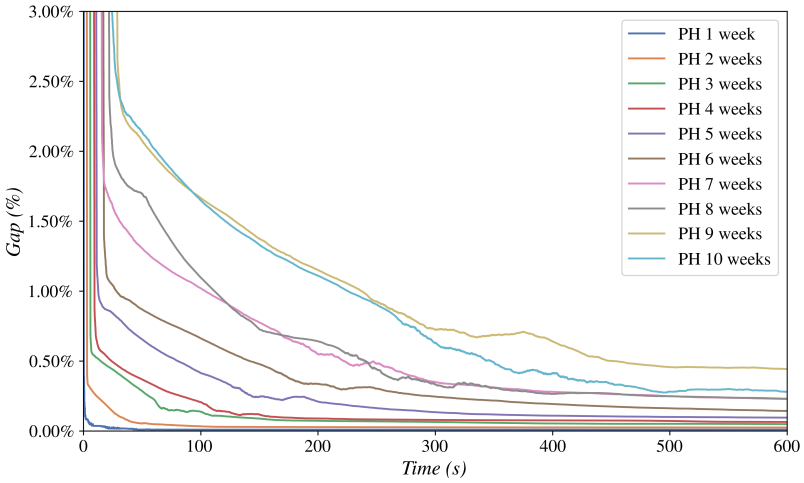
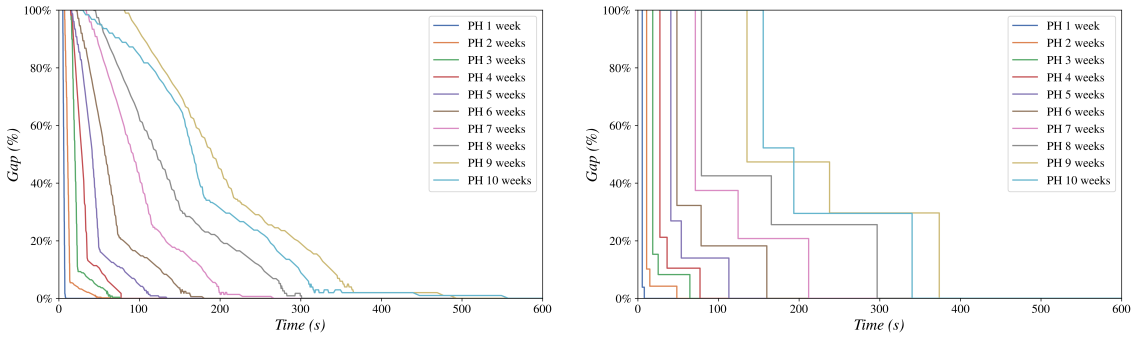


Figure 8.5: Smoothed average gaps for D1 for planning horizons 1-10 weeks.



(a) Smoothed average gaps for P1 over 20 runs.

(b) Gaps for a single run of P1.

Figure 8.6: Gaps for P1 for different planning horizons.

we concluded that we would prefer 500 scenarios in the small case to get a satisfactory stable solution. We probably need even more in the large case due to the increased number of variables and parameters. Thus, we conclude that the S1 model is not suited for solving the real-sized problem with four weeks planning horizon. However, we are able to run the model with two weeks planning horizon, which we will show soon. Even though we could show the gap development for the two-week case here, we argue that four weeks is a minimum to get a somewhat realistic problem size. Thus the two-week plot is not of interest or comparable to the previously shown D1 and P1 plots.

Table 8.8 shows the time to reach gaps for the different models in the large case with a four-week planning horizon. The results for all planning horizons 1-10 can be found in Appendix C.2. We see that the D1 model is much faster than the P1 model at the start, but P1 is able to find the optimal solution in the end, which D1 is not. Interestingly, D1 was only able to find the optimal solution within the time limit for one and two weeks planning horizon, while P1 can find the optimal solution for all 1-10 weeks planning horizon. Again, we refer to Appendix C.2 for the complete overview of these results.

Table 8.8: Time to reach solutions and gaps for the D1, P1, and S1 models with four weeks planning horizon in the large case.

Model	First Feasible	20%	5%	1%	Optimal
D1	0.37	0.79	0.79	8.94	-
P1	8.92	35.09	77.17	77.17	149.97
S1	-	-	-	-	-

Scenario complexity

Since the two-stage models' complexity usually is heavily dependent on the number of scenarios, we perform some tests on our S1 model. We run gap analyses for a two- and four-week planning horizon for different numbers of scenarios on both the small and the large case.

Table 8.9 and Table 8.10 show the time to reach gaps for the different number of scenarios for two and four weeks planning horizon in the small case, respectively. Note that for certain numbers of scenarios, the model struggles with reaching a 1% gap. However, we have included the 1.5% gap in these tables to highlight that the model is able to find relatively small gaps in a reasonable time. The challenge with closing the gap could be related to the same difficulties with the objective function discussed earlier when analyzing the gap for the D1 model in the large case.

Table 8.9: Time, in seconds, for S1 to reach gaps for the small case with a two-week planning horizon for multiple numbers of scenarios.

Scenarios	First Feasible	20%	5%	1.5%	1%	Optimal
2	0.03	0.05	0.05	0.09	0.09	4.42
50	0.61	0.86	0.86	1.13	1.13	-
100	1.28	1.90	1.90	2.29	2.67	-
150	2.03	3.26	3.26	3.26	3.90	-
250	3.61	6.54	6.54	6.54	7.23	-
500	6.80	21.55	21.55	22.72	22.72	-
1000	13.79	27.23	27.23	27.23	1020.61	-
2000	30.52	70.17	70.17	70.17	76.14	-

Table 8.10: Time, in seconds, for S1 to reach gaps for the small case with a four-week planning horizon for multiple numbers of scenarios.

Scenarios	First Feasible	20%	5%	1.5	1%	Optimal
2	0.05	0.13	0.35	0.35	0.88	-
50	1.31	3.47	4.90	11.69	457.28	-
100	2.79	8.32	10.69	197.57	1184.66	-
150	4.76	17.76	20.43	20.43	861.65	-
250	7.69	51.00	55.56	55.56	1904.54	-
500	14.91	35.62	40.04	61.13	22.72	-
1000	32.82	68.87	77.46	77.46	1020.61	-
2000	89.17	194.64	220.57	70.17	76.14	-

Table 8.11 and Table 8.12 show the time to reach gaps for the different number of scenarios for two and four weeks planning horizon in the large case, respectively. Note that the computer runs out of memory when attempting over 1200 scenarios for a two-week planning horizon and 400 scenarios for a four-week planning horizon. Even if we accept using 400 scenarios, the two-stage model uses

548 seconds to find the first feasible solution. Compared to the pattern-based model, which uses only 8.92 seconds, this is a more than 60 times difference. On the other hand, the results indicate that if a 1% gap is acceptable and the time budget of the operational planner is multiple hours, it could be possible to use two weeks planning horizon and enough scenarios to get a stable solution, applied in, i.e., a rolling horizon metaheuristic to generate longer plans by solving two weeks at the time. However, due to the limited time aspect of this thesis, we conclude that the two-stage model is too slow to be extensively tested for the large case, and we will only perform tests on the small case with this model in the upcoming sections

Table 8.11: Time, in seconds, for S1 to reach gaps for the large case with a two-week planning horizon for multiple numbers of scenarios.

Scenarios	First Feasible	20%	5%	1%	Optimal
100	37.11	99.91	99.91	106.87	-
200	103.76	227.66	227.66	239.27	-
300	178.25	750.27	750.27	769.12	-
400	256.63	1556.60	1556.60	1583.19	-
500	330.93	2219.00	2219.00	2254.44	-
600	405.24	2156.85	2156.85	2193.16	-
700	479.49	1581.25	1581.25	1626.83	-
800	558.72	2168.15	2168.15	2219.61	-
900	636.26	2107.36	2107.36	2166.43	-
1000	729.78	2206.93	2206.93	2275.84	-
1100	815.54	2570.91	2570.91	2622.57	-
1200	897.96	2818.99	2818.99	2880.08	-
1300	-	-	-	-	-

Table 8.12: Time, in seconds, for S1 to reach gaps for the large case a four-week planning horizon for multiple numbers of scenarios.

Scenarios	First Feasible	20%	5%	1%	Optimal
100	79.19	317.94	357.21	1545.34	-
200	219.03	919.38	973.16	2222.27	-
300	380.13	2834.12	2905.38	-	-
400	548.20	5757.22	5880.53	-	-
500	-	-	-	-	-

Comments on pattern generation complexity

As mentioned in Section 3.3, it is common to use techniques like column generation to iteratively generate patterns while solving pattern-based models in problems like the Cutting Stock Problem. The reason is that it is often not possible to enumerate all patterns with modern computers, and if one is able to do that, the models often get too complex. However, we have not experienced any computational problems when generating patterns, and our model is able to find optimal solutions even when all possible patterns are used. The number of patterns depends on how many different procedures can be scheduled in a given block and the duration of the procedures. When generating the patterns, we assume every procedure gets the shortest possible surgery duration and fully enumerate all combinations that keep the sum of these shortest durations below the block

duration limit. For the large case, this gives a total of 99324 patterns. However, almost all of these patterns are for the Plastics specialty, which can be seen in Table 8.13. If it was the case that we could not solve the model, an alternative strategy could be to isolate the Plastics blocks to be solved with a different type of model and solve the rest with the pattern-based model. Another option is to group together procedures within the Plastics specialty, as many of them have very similar distributions for surgery duration and LOS.

Table 8.13: Number of patterns per specialty in the complete set of legal patterns.

Specialty	Number of patterns
Arthroscopic	415
Back	8
Foot	31
Hand	1217
Plastics	97421
Prosthetics	87
Tumor	145

Even though there are 99324 patterns in total, the model is rarely run with all these. When filtering out patterns with expected overtime above 30 minutes for the P2 model, we are left with 30990 patterns. Filtering out patterns with a probability for cancellations above 10% for the P3 model leaves 17216 patterns. Thus, the filtering drastically reduces the number of variables needed and thus reduces the complexity of the model. Another technique we use to reduce the number of variables is to only create indices and variables for patterns that are possible to choose each day based on the MSS.

8.3 How does uncertain surgery duration and LOS affect scheduling quality

This section investigates how uncertain surgery duration and LOS affect scheduling quality in an environment without cancellations. Later sections will look at the case with cancellations, but it is also interesting to see how uncertainty affects the different models before cancellation rules are included. In that way, it is easier to compare the cost of the cancellation rule in the case study. The simulation framework presented in chapter 7 is used to evaluate the quality of the schedules over an extended time period.

The section is structured as follows. First, we investigate schedules with varying overtime costs in Section 8.3.1. Second, Section 8.3.2 looks at varying extra ward bed cost levels. Third, in Section 8.3.3, we use the overtime chance constraint formulation, P2, to test how risk levels affect the schedules. Lastly, the key findings are summarized in Section 8.3.4, and the hypothesis from Section 6.3 is updated.

Note that we use the small problem case for the first two sections. The reason is that the time

complexity of S1 does not permit us to test for the large case. However, using a smaller test instance should not be a disadvantage based on the goal of investigating how the uncertainty and models affect the schedules. However, we use the large case in Section 8.3.3 and do not include the S1 model in the section.

8.3.1 Uncertain surgery duration

By varying the overtime cost, C^{Φ}, and the expected overtime limit, $\overline{O}$, we can investigate how the schedules resulting from the different models behave. Figure 8.7 shows how the KPIs are affected by the overtime cost, C^{Φ}. For $C^{\Phi} = 0$, overtime is not penalized, and the block capacity constraints are relaxed. In that case, overtime is only restricted by the expected overtime constraint. On the other hand, the constraint is 'hard' for $C^{\Phi} = M$. No overtime is allowed, and we, in practice, get a robust formulation. In that case, the expected overtime constraint becomes redundant since no expected overtime will ever be possible. The constraint is soft for all other values, i.e., $0 < C^{\Phi} < M$, but the overtime cost regulates the softness. The value of M is set to 1000000.

Some general trends are observed. Service time and OR undertime increases, while throughput and overtime decreases when overtime is penalized more. Except for overtime, one could argue that the schedule gets worse for all KPIs. However, Ward utilization and extra beds do not look to be affected by the different levels of overtime costs. A possibility could be that the ward costs must be higher to affect the models and are dominated by the other objectives or that varying the overtime cost does not affect the patient mix related to LOS and extra ward beds. We will investigate the ward costs further in Section 8.3.2.

Further, it is interesting to see how the schedule characteristics differ between the soft, robust ($C^{\Phi} = M$) and relaxed formulations($C^{\Phi} = 0$). We observe significant changes in the robust formulation. The average service time increases by approximately six days for S1 and P1. Overtime reaches zero, and the undertime increases. This indicates that a robust formulation where overtime is never allowed might have unwanted consequences for the Department of Orthopedics. We will investigate the possibility of formulating the block capacity constraint using a chance constraint approach in Section 8.3.3.

Service time and throughput

D1 looks to achieve better service times and throughput than the other models. The OR utilization can explain this. Generally, the model with the best service time for a given overtime cost also has the highest OR utilization. In Section 6.3, we argue that D1 will underestimate the overtime cost. That could explain the high throughput observed for D1. On average, D1 packs tighter, increasing the probability of overtime. Therefore, we should be careful to conclude that D1 is *better* for service time and throughput; instead, the higher throughput and lower service time result from

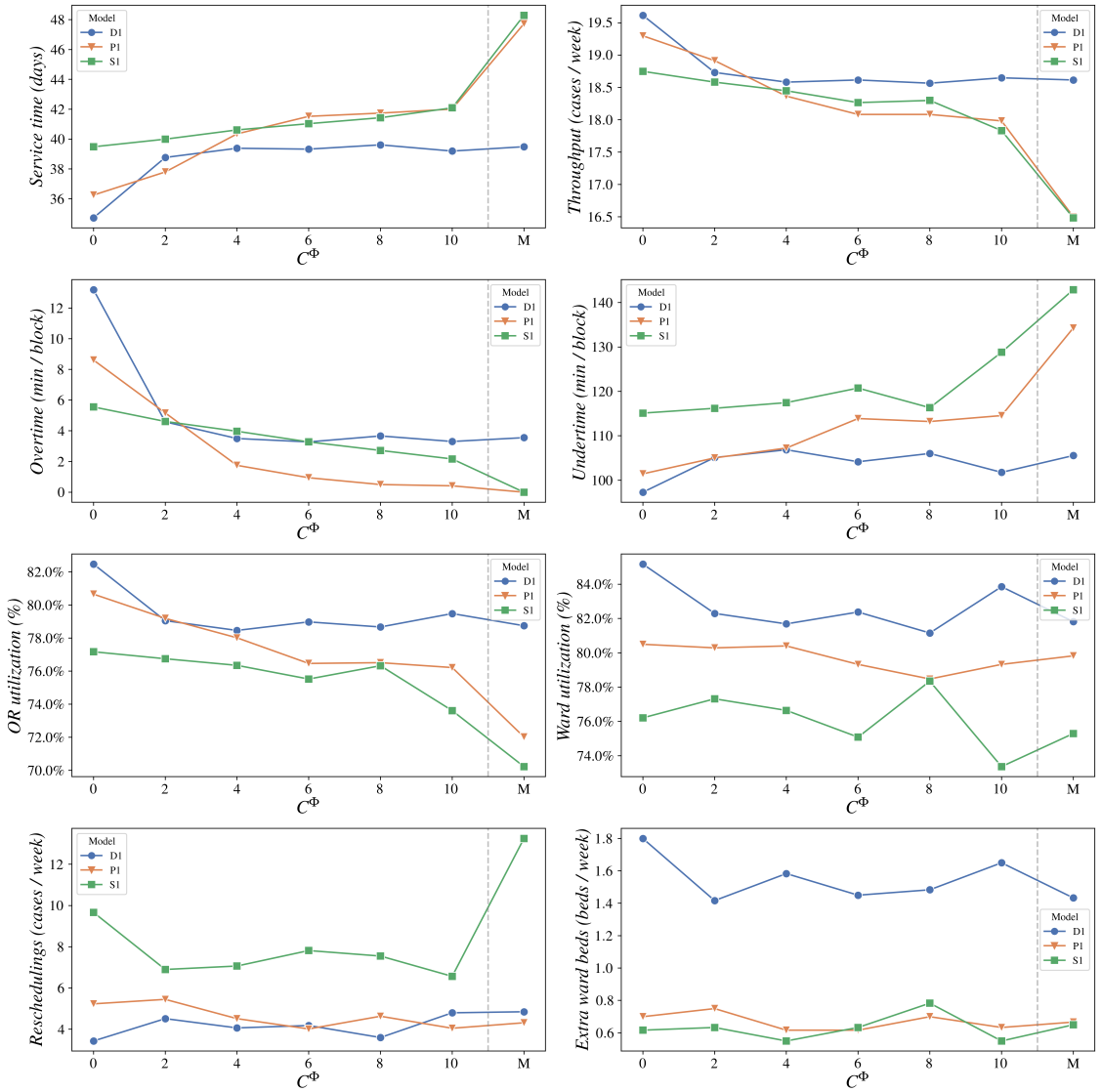


Figure 8.7: How does overtime costs affect scheduling quality

underestimating overtime and packing too tightly. Similarly, we observe that we also increase the throughput for S1 and P1 by lowering the overtime costs.

Rescheduling

It is harder to explain why S1 has significantly more reschedulings than P1 with the robust formulation. However, the stability of S1 could be one possibility. Remember, new scenarios are drawn between planning stages. When overtime costs become so high that no overtime is preferred, slight differences between scenario trees can create significant changes. For example, think about a case where the combination of a set of cases has a 0.001% probability of overtime, but there is no overtime according to the scenario tree. Then the model might use that solution since no overtime is expected. However, if in the next planning stage, the new scenario tree gives the solution just a few minutes of overtime in one of the scenarios, then the existing schedule becomes much worse

according to the model, and the plan would most likely change. This could explain why we see much higher levels of reschedulings in the S1 model when the overtime cost is equal to M.

The results are quite remarkable, indicating that the scheduling quality is affected by the parameters, the number of planning stages, and the stability of the underlying mathematical model. The consequences might become more apparent if we imagine the OR capacity constraint as hard without flexibility. Then, we could risk that a solution is deemed feasible in one planning stage but infeasible in another. This could indicate that the schedules become more sensitive to the stability of the model as the block capacity restriction becomes harder and harder. Thus, we recommend avoiding modeling block capacity as a hard constraint if stability is a potential issue, considering both feasibility issues and the scheduling quality.

Overtime

As expected, the amount of overtime decreases when overtime costs increase. However, the sensitivity of the models with respect to overtime costs differ. For D1, the effect flattens out as long as $C^{\Phi} \geq 4$. This is interesting as overtime decreases for larger values in S1 and P1. The finding is consistent for other KPIs as well. One possible reason is that the overtime cost never becomes a dominant objective for D1 when the overtime cost is less than four or that the overtime cost is treated similarly for all the cases. This could make sense since D1 underestimates the overtime considerably. Figure 8.8 strengthen this hypothesis. We see that the amount of overtime does not change, even when no expected overtime is permitted, $\overline{O} = 0$. This is an important finding as it illustrates the consequence of D1 underestimating overtime. D1 is unable to achieve zero overtime. Consequently, adding stochastic information to the model should be seen as necessary if the OR capacity restrictions are strict.

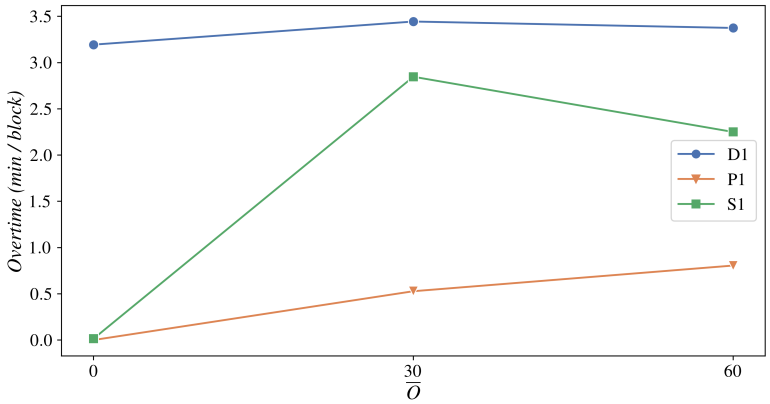


Figure 8.8: Expected overtime limit

Further, as C^{Φ} increases, both P1 and S1 converge towards zero. It appears that both models are able to avoid overtime. More importantly, it indicates that both models can adequately describe

the uncertain surgery duration. However, further inspection shows that the S1 cannot perfectly describe the uncertainty. In fact, for $\overline{O} = 0$, S1 still uses 0.014 minutes of overtime per block. This has two implications. First, for all practical purposes, S1 describes the uncertainty well enough. Second, P1 describes the uncertainty correctly and slightly better than S1. This strengthens our hypothesis from Section 6.3.

However, it is interesting that except for the robust and relaxed cases, S1 has overtime much closer to D1 than P1. S1 and P1 should describe the uncertain surgery duration equally. The higher time complexity of S1 is a possible explanation. In Section 8.2.2, we saw that P1 is much faster at closing the gap than S1. When overtime is not a dominating objective, other costs might be more important to close the gaps and prioritized by the solver, for instance, patient waiting time. Therefore, S1 might have more overtime than P1 simply because it needs more time than P1 to find an equally good schedule and is not able to close the gap within the time limit.

The higher gaps of S1 could also explain why overtime decreases when we do not allow reschedulings, as seen in Figure 8.9. For the case without reschedulings, the models can only influence the last week of the planning horizon. The rest is already decided in previous planning stages. With only a one-week planning horizon in practice, the time complexity of S1 falls dramatically, as seen in Appendix C.1. This could explain why S1 is much closer to P1 without rescheduling since both have been able to close the gaps sufficiently within the time threshold we gave both models.

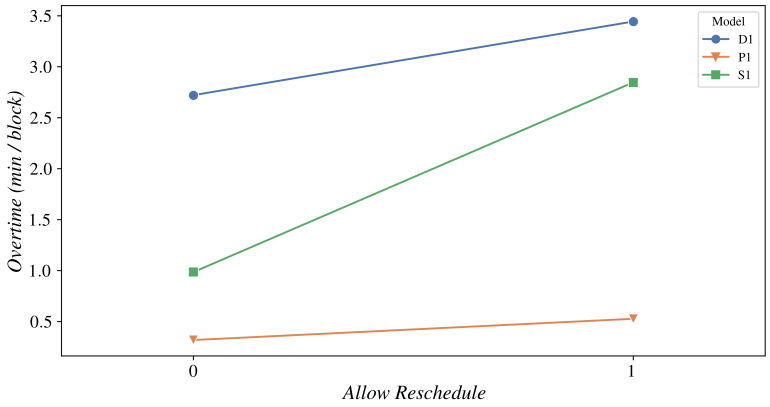


Figure 8.9: Rescheduling allowed

We now see a recurring theme. S1 and P1 have the same amount of overtime in those cases where minimizing overtime becomes a dominating factor for the overall solution. However, when overtime is not a dominating factor, S1 has more overtime than P1, but a shorter planning horizon (and lower problem complexity) reduces the difference. This indicates that P1 might be better than S1 at minimizing overtime due to its stronger and less time-complex formulation, not because the underlying assumptions regarding surgery duration differ.

8.3.2 Uncertain LOS

As we did with overtime, we can formulate the ward capacity constraint as relaxed, robust, and soft when $C^P = 0$, $C^P = M$, and $0 < C^P < M$, respectively. Figure 8.10 summarizes the results. M is set to 1000000.

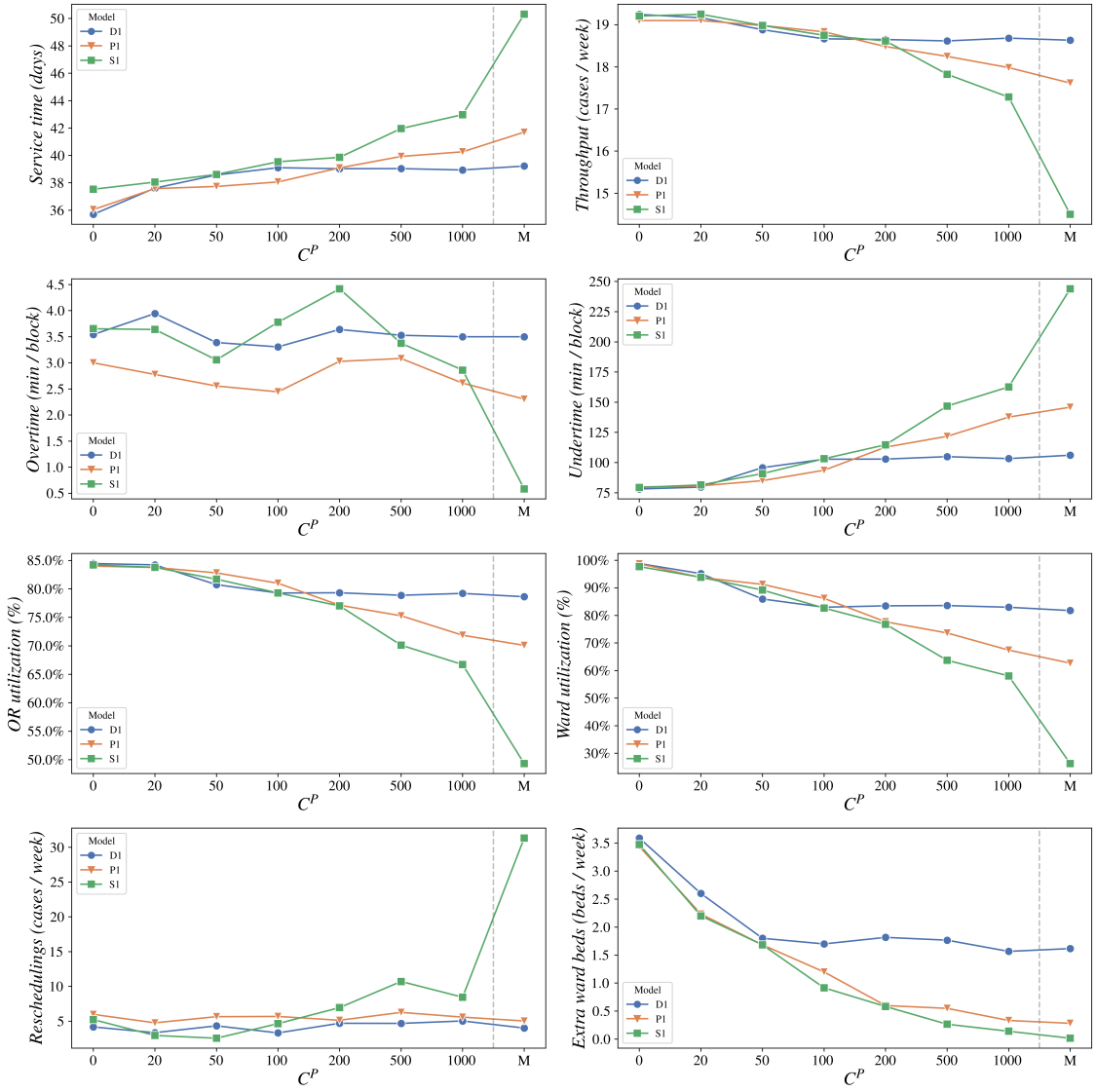


Figure 8.10: How does extra bed cost affect scheduling quality

We observe the same trends as in Section 8.3.1. As C^P increases, the schedule quality gets worse for all KPIs except extra ward beds and, to some degree, overtime. Further, D1 again flattens out at around $C^P = 50$, and S1 becomes significantly worse than the other for the robust formulation. However, for the case where the restriction is relaxed, we observe that the three models are considerably more similar than when overtime costs are equal to zero.

When $C^P = 0$, we hypothesized that P1 and S1 should represent the uncertainty equally. Therefore, one would expect the two models to have similar schedules when relaxing the extra ward bed

restriction. However, Table 8.14 shows that the two models differ considerably for some KPIs. Note that the two models are still considerably more similar than when the block capacity restriction was relaxed. P1 has better service time and overtime, while S1 has fewer reschedulings. Interestingly, S1 has fewer reschedulings, as we saw the opposite for the case where overtime cost was zero. As previously discussed, the gaps can explain some of these differences. For example, P1 could pack the patients more effectively and thus get better service times.

Table 8.14: Relaxing the ward bed restriction

Model	Extra ward bed cost = 0			
	S1	P1	Change	%
Throughput	19.21	19.10	-0.11	-0.6%
Service time	37.53	36.05	-1.49	-4.0%
Undertime	79.42	79.79	0.38	0.5%
Overtime	3.65	3.00	-0.65	-17.9%
OR utilization	84.22%	84.00%	-0.21%	-0.3%
Ward utilization	97.72%	98.76%	1.04%	1.1%
Reschedulings	5.23	5.98	0.75	14.4%
Extra ward beds	3.48	3.44	-0.04	-1.1%

Extra ward beds

In Section 6.3, we argued that P1 could not perfectly describe the bed demand, but it would be considerably better than D1. Figure 8.10 strengthens our hypothesis. When C^P increases, P1 follows S1 much closer than D1. Since the ward demand restriction is robust for $C^P = M$, we should expect the number of extra beds to be zero, given a perfect representation of uncertain LOS. Table 8.15 shows the values for $C^P = M$.

Table 8.15: Extra ward bed usage when ward bed capacity is robust

Model	Extra ward beds
D1	1.62
P1	0.28
S1	0.02

As expected, we still see 0.28 extra ward beds in the robust formulation for P1. This further strengthens our hypothesis as it indicates that we get 0.28 extra beds in the case where P1 plans without the use of extra beds. Also, note how we still do not reach zero for S1. Similarly, as with the overtime costs, using SAA is one likely reason. However, we argue that 0.02 is, for all practical purposes, the same as zero for the problems investigated in this thesis. Interestingly, the extra ward bed usage for D1 is 5.8 times higher than P1. This indicates that P1 is much better at representing the uncertain LOS than D1.

These findings have some important implications. First, remember that P1 has a different way of representing the uncertain LOS variable. D1 uses the expected value from a Poisson distribution. However, P1 transforms the Poisson distribution into N Bernoulli distributions and uses the

expected value of those distributions, as described in Section 6.1 and Section 6.2. This transformation and bed demand formulation is still fully deterministic and could be used in all deterministic formulations, including D1. Our findings indicate that the P1 formulation is considerably better than the more traditional formulation, as seen in Schiøtz and Tysse (2022).

8.3.3 Chance Constraint

An alternative way of modeling overtime is by using a chance constraint. Figure 8.11 shows how scheduling quality changes for different risk levels of overtime for P2 on the large test instance. Note how this is different from punishing overtime. Overtime costs are still included since we want to minimize overtime for those scenarios where it occurs. Restricting the risk levels, however, guarantees that overtime does not occur with a given probability.

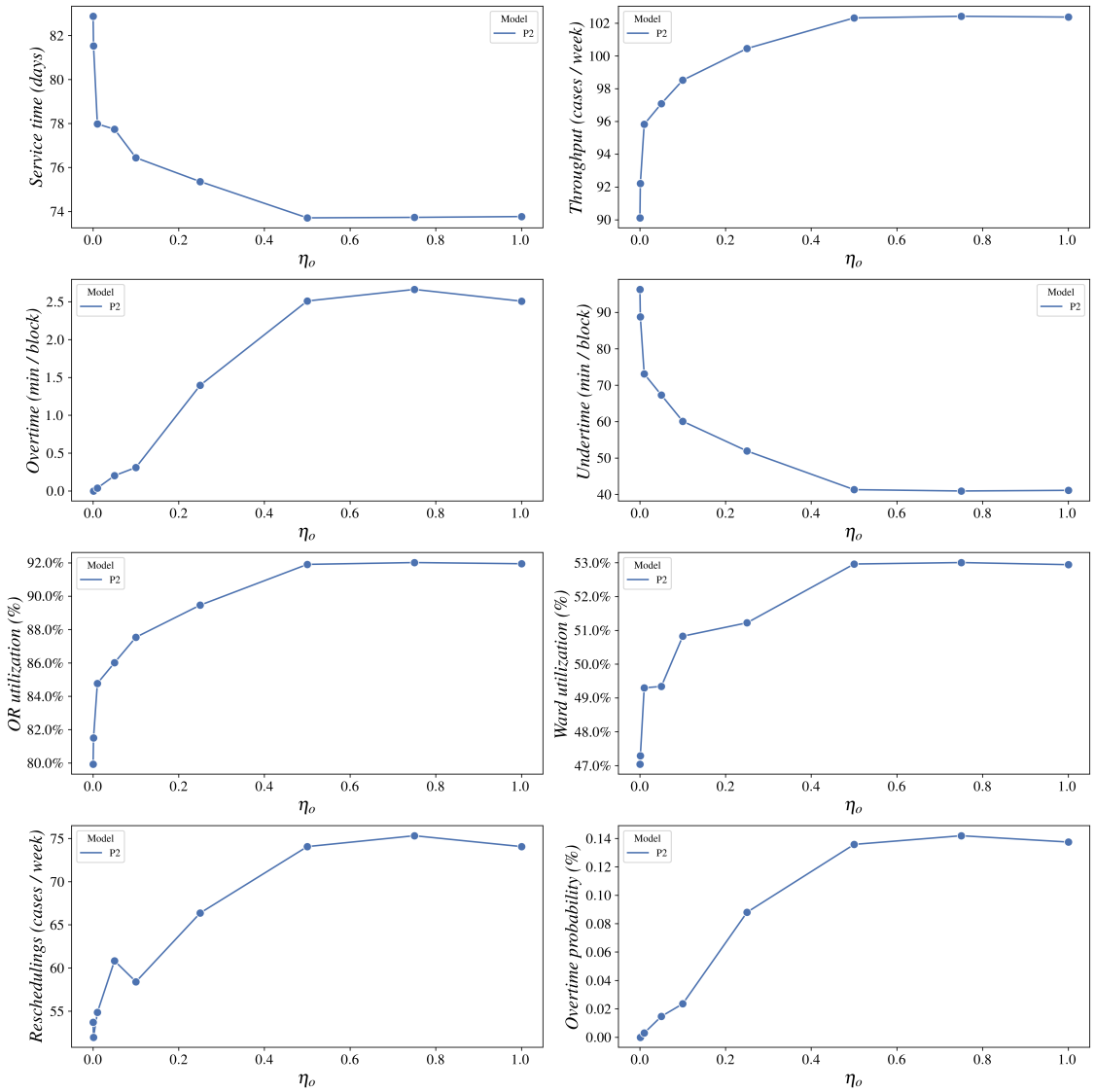


Figure 8.11: Probability of overtime as a chance constraint. For P2 on the large test instance

Increasing the risk of overtime above zero can have tremendous results on scheduling quality. First, we observe decreasing marginal change in the KPIs when we increase the risk levels. For instance, note how we observe the same levels of overtime and probability of overtime for $\eta_o = 0.5$ and $\eta_o = 1$. Multiple reasons could explain this. However, remember that we still have overtime costs and are restricting the amount of expected overtime. As a consequence, the feasible region could be entirely or close to identical for both cases, and the expected overtime constraint is the active constraint for all or most of the blocks in both cases. In more generalized terms, the probability of the expected overtime constraint becoming active increases as η_o increases, in turn decreasing the marginal contribution of η_o.

Second, note how even a slight increase in risk from tolerance from zero can have tremendous results for schedule quality. Table 8.16 shows the percentage-wise change in the KPIs when we increase the risk compared to $\eta_o = 0$ for selected values of η_o. We see that overtime is almost zero, even when increasing η_o by 1 percent. Interestingly, while we see little change in overtime and the probability of overtime, the other aspects of schedule quality are massively improved. For instance, the throughput is increased with over 5.5 cases each week, and the average service time is decreased by almost five days. The quality of service from the surgical cases side is thus massively improved, while it could be argued that the work environment for the staff has not changed at all.

Table 8.16: Percent wise change in scheduling quality for increased risk of overtime relative to the case with zero risk of overtime

	Overtime probability risk, η_o												
	0%	0.1%			1%			5%			10%		
	Value	Value	Change	%	Value	Change	%	Value	Change	%	Value	Change	%
Throughput	90.13	92.22	2.09	2.3%	95.83	5.71	6.3%	97.09	6.97	7.7%	98.52	8.40	9.3%
Service_time	82.88	81.53	-1.35	-1.6%	77.98	-4.89	-5.9%	77.74	-5.14	-6.2%	76.45	-6.43	-7.8%
Undertime	96.32	88.77	-7.56	-7.8%	73.12	-23.21	-24.1%	67.31	-29.01	-30.1%	60.09	-36.23	-37.6%
Overtime	0.00	0.00	0.00		0.04	0.04		0.20	0.20		0.31	0.31	
Overtime_prob	0.0%	0.0%	0.00		0.3%	0.00		1.5%	0.01		2.4%	0.02	
OR Utilization	80%	82%	0.02	2.0%	85%	0.05	6.1%	86%	0.06	7.6%	88%	0.08	9.5%
Ward Utilization	47%	47%	0.00	0.5%	49%	0.02	4.8%	49%	0.02	4.9%	51%	0.04	8.0%
Reschedulings	53.71	51.98	-1.73	-3.2%	54.86	1.15	2.1%	60.84	7.13	13.3%	58.41	4.70	8.8%

Finally, it is worth mentioning that the actual probability of overtime is significantly lower than the risk threshold denoted by η_o. This is consistent with expectations, as scheduling is influenced by more than just the amount of overtime. For example, there may be situations where it is more efficient to adopt a pattern with a lower risk of overtime than specified by η_o, as the frequency of the procedure allows the model to decrease the waiting times further. Even though such an outcome is foreseeable and unsurprising, it should be noted how large the differences are. For example, if the schedulers aim to maintain the average probability of overtime under 2.5%, then a η_o up to 10% would suffice. However, it is essential to consider that alternative approaches beyond chance constraints could optimize the expected probability more effectively. This is because chance constraints do not directly penalize the overtime probability. Instead, they ensure a specific level of service quality. Determining the optimal value for η_o falls outside the scope of this thesis, as it necessitates a deeper understanding of the specific domain. However, we recommend performing

simulations with varying model parameters, particularly η_o, to identify the combinations that yield the desired scheduling attributes.

8.3.4 Summary of key findings

In this section, we have examined three formulation types: soft, robust, and relaxed, observing significant changes to schedule quality with the robust formulation for the models considering uncertain surgery duration and LOS. The robust formulations resulted in increased service time and undertime and reduced overtime.

D1 appeared to perform better than the other models for service time and throughput due to higher OR utilization. However, this is likely due to D1 underestimating overtime costs. Also, the model with the best service time often had the highest OR utilization. The number of reschedulings for S1 with the robust formulation varied considerably compared to P1, possibly due to S1 being less stable.

Overtime decreases as overtime costs increase, with D1 flattening out when overtime costs reach a certain point. This could indicate that D1 does not plan for any overtime as long as the overtime costs are over a certain threshold or that D1 is not affected by an increased overtime cost over a certain threshold. Additionally, P1 and S1 can avoid overtime, suggesting they sufficiently describe the uncertain surgery duration. D1, however, was not able to avoid overtime entirely.

Regarding uncertain LOS, similar trends were observed. As the ward capacity constraint increases, the schedule quality worsens except for extra ward beds and, to some extent, overtime. D1 again flattens out at a certain point, while S1 gets considerably worse under the robust formulation. Further, we found P1 to better represent uncertain LOS than D1.

Modeling the risk of overtime with a chance constraint indicates that even a slight increase in the risk of overtime can significantly improve scheduling quality, with diminishing returns as the risk levels increase. Further, the actual probability of overtime is often lower than the defined risk threshold, possibly due to other factors influencing scheduling. Therefore, simulations are recommended to find the optimal parameters, including the risk threshold, to achieve desired scheduling qualities.

In conclusion, the findings demonstrate that including uncertainty in the models improves the schedules by better handling the overtime and ward capacity. In addition, the pattern model can describe the uncertain surgery duration better than the two-stage stochastic model with SAA. Further, the results indicate that the pattern model models the uncertainty LOS sufficiently. As a result, we believe the pattern model creates sound schedules, and the added benefits from S1 regarding the more exact LOS representation do not look to outweigh the added computational complexity compared to P1 for the problem investigated in this thesis.

8.4 Cancellation study

This section investigates how hedging against overtime and cancellation affects schedule quality. We provide a detailed comparison between the instances described in Section 8.1 for several KPIs. Further, an analysis of how varying levels of overtime risk acceptance affect scheduling quality is also included.

8.4.1 Cancellations

The average cancellation probabilities for the different instances are shown in Figure 8.12. P2-SC1 has a 1.81% cancellation probability, while P2-SC0 obviously has no cancellations since cancellations are not included in the simulation. This illustrates that models which do not account for cancellation rules underestimate the cancellation risks, as expected.

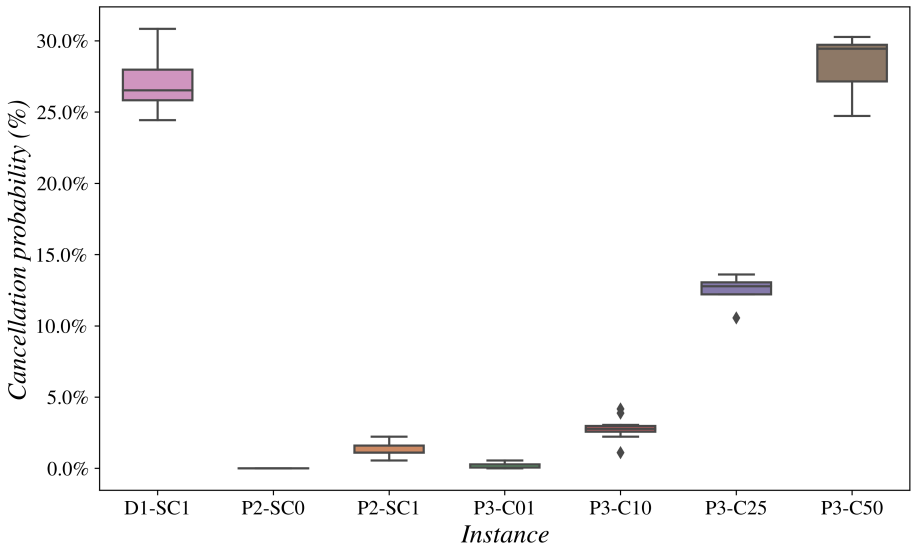


Figure 8.12: Cancellation probability per block (%)

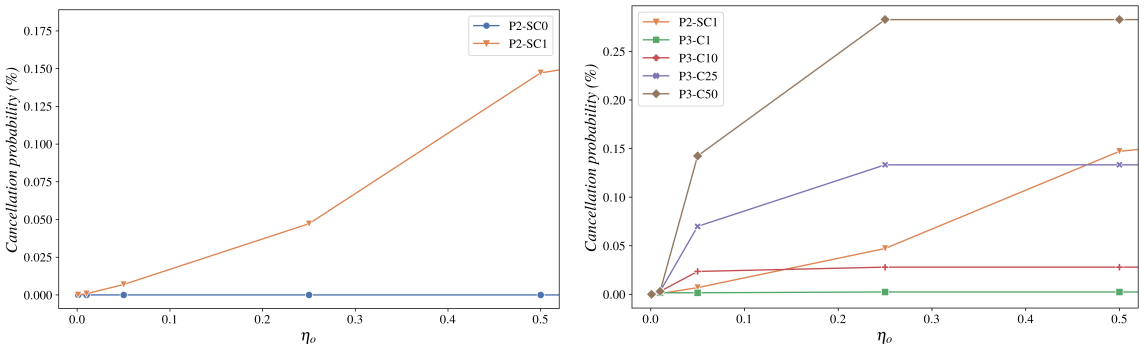
However, one should note that P2-SC1 has a lower cancellation probability than all P3 models except P3-C1. This is interesting as it shows that, on average, we get more cancellations with the chance constraint formulations than with formulations that do not assume any cancellations. Now, this is not sprightly a good thing. The reader should note that with the chance constraint, we do not punish cancellations; we hedge against the risk. For P2-SC1, we do not hedge against any cancellation risk. However, we still get low cancellation rates on average, likely due to the overtime costs preventing the models from overbooking too much.

However, say the orthopedic department is okay with a cancellation probability of 10%. Then, both P2-SC1, P3-C1, and P3-C10 are, on average feasible schedules, and the performance for other KPIs needs to be evaluated before any ranking of the three models is possible. That being said, we only know that the cancellation probability *for each block* is under 10% for P3-C1 and P3-C10.

P2-SC1 has no such guarantee.

Interestingly, D1-SC1 has a significantly higher cancellation probability than P2-SC1, while none assumes cancellations when scheduling. This again shows how much the deterministic models underestimate overtime. Since it underestimates overtime, then cancellations are not implicitly punished either. The results indicate that while not including cancellations in the models might produce sound schedules in some cases, the models still need to handle the uncertainty and correctly represent the overtime.

Figure 8.13a shows how the cancellation probability for P2-SC1 increases linearly as η_o increases. However, from $\eta_o = 0.5$ and up, no more increase is observed. In those cases, we believe the expected overtime constraints become binding, not the overtime chance constraints. This is a good property, as limiting the expected overtime for models without cancellation rules in practice also limits the cancellation probability.



(a) PC=0 (b) SC=1

Figure 8.13: Cancellations for varying levels of overtime risk

From Figure 8.12, one could argue that the performance of P2-SC1 is better than expected. However, Figure 8.13b illustrates how the instance cannot guarantee any risk limit. In contrast, the P3 instances limit the risk by increasing η_o. We observe that the cancellation probability becomes independent of η_o when η_o increases. Note how the cancellation probability becomes independent of η_o for different values for the P3 instances. P3-C1 and P3-C10 reach the limit earlier than P3-C25 and P3-C50.

Table 8.17 shows how a surgical case's cancellation probability depends on its expected length relative to other similar procedures. A lower expected surgery duration significantly increases the cancellation probability, while longer procedure variants have a lower risk of cancellation.

Table 8.17: Cancellation probability given procedure variant.

Procedure length	P2-SC0	P2-SC1	P3-C01	P3-C10	P3-C25	P3-C50
M	0.00%	0.32%	0.03%	0.67%	3.02%	5.37%
H	0.00%	0.63%	0.00%	0.35%	0.48%	2.94%
L	0.00%	1.31%	0.00%	2.24%	11.57%	22.08%

Similarly, procedures with a short expected surgery duration are again the most likely to be canceled, as seen in Table 8.18. Interestingly, we observe significant differences even for procedures with identical surgery duration. There could be multiple reasons for this. For instance, the sequencing in the simulation might not be completely random given equal surgery duration.

Also, the marginal increases are different within a procedure. For instance, compare aggregated hand with Carpal Tunnel Syndrome (CTS). When the cancellation risk increases from 10% to 25%, Aggregated Hand only slightly increases 0.04% while CTS increases with 4.44%. It is not clear exactly why this occurs. One possible reason is that CTS has a longer expected LOS than Aggregated Hand. The differences in LOS make the days the cases are scheduled not random. As a result, it could mean that CTS becomes more unlucky with the patterns it is placed in on average. For instance, we might want cases with longer LOS in patterns with lower throughput, as it puts less strain on the wards. However, patterns with lower throughput but similar total surgery duration have a higher probability of cancellations. In turn, cases with longer LOS might also have higher cancellation probability for some pattern filtering combinations. One should be careful to conclude without further investigation, but it is essential to remember that changing the cancellation risks could discriminate between procedures.

Table 8.18: Cancellation probability per procedure (%)

	E[DUR]	P2-SC0	P2-SC1	P3-C1	P3-C10	P3-C25	P3-C50
Arthroscopic	-	**0.00**	**0.21**	**0.00**	**0.21**	**3.13**	**8.21**
ACL	190	0.00	0.37	0.00	0.00	0.00	0.00
Aggregated arthroscopic	120	0.00	0.00	0.00	0.35	6.64	23.86
Meniscus	170	0.00	0.26	0.00	0.28	2.75	0.77
Back	300	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Foot	140	**0.00**	**0.00**	**0.00**	**0.00**	**3.15**	**4.76**
Hand	-	**0.00**	**1.11**	**0.00**	**1.39**	**3.63**	**8.22**
Aggregated hand	90	0.00	1.39	0.00	1.39	1.43	8.96
Carpal tunnel syndrome	90	0.00	0.83	0.00	1.39	5.83	7.47
Plastics	-	**0.00**	**1.17**	**0.03**	**2.01**	**6.21**	**9.45**
Aggregated plastics	100	0.00	0.54	0.07	1.54	4.02	8.23
BCC	140	0.00	0.00	0.00	0.00	0.63	0.00
Cancer mammae	100	0.00	0.15	0.00	1.60	2.45	7.74
Malignt melanom	70	0.00	1.63	0.10	3.87	10.54	23.89
Plateepitelkarsinom	70	0.00	3.54	0.00	3.05	13.39	7.37
Prosthetics	-	**0.00**	**0.07**	**0.00**	**0.00**	**10.58**	**25.00**
Hip	180	0.00	0.00	0.00	0.00	11.84	26.02
Knee	170	0.00	0.14	0.00	0.00	9.32	23.99
Tumor	80	**0.00**	**2.46**	**0.00**	**2.82**	**3.32**	**8.88**

8.4.2 Overtime

Figure 8.14 shows the probability of a block running overtime. Interestingly, significant differences are observed between the instances. Remember that all instances have restricted the risk of overtime at 10%, and the unit overtime costs are identical. Still, the realized overtime probability ranges from 0.19% and 5.17%.

The overtime probability is reduced from 2.23% to 1.11% when cancellation rules are added to the simulations, as we see by comparing the two P2 instances. This effect happens since P2-SC1 overestimates the overtime probability. Remember, the instance uses the pattern file calculated without cancellations, where all cases are expected to receive surgery. However, when cancellation occurs in the simulation, the actual overtime decreases.

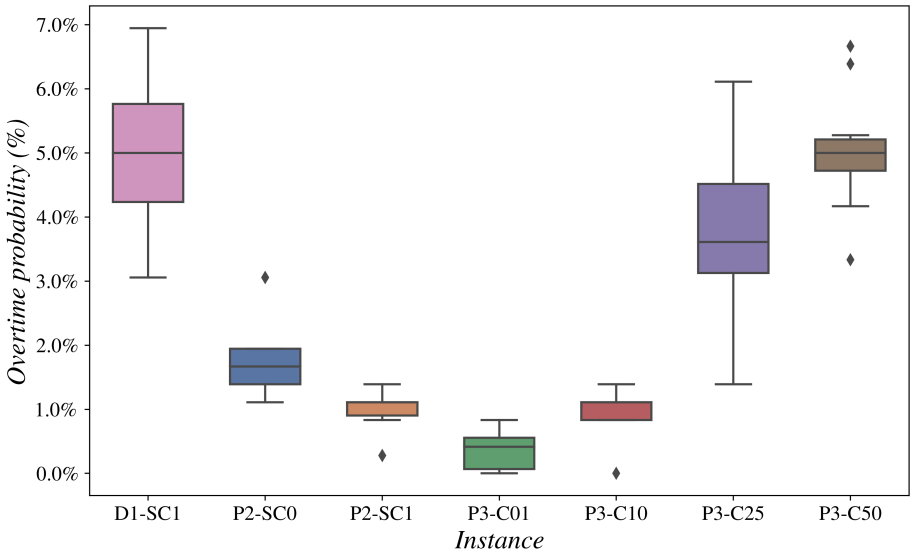


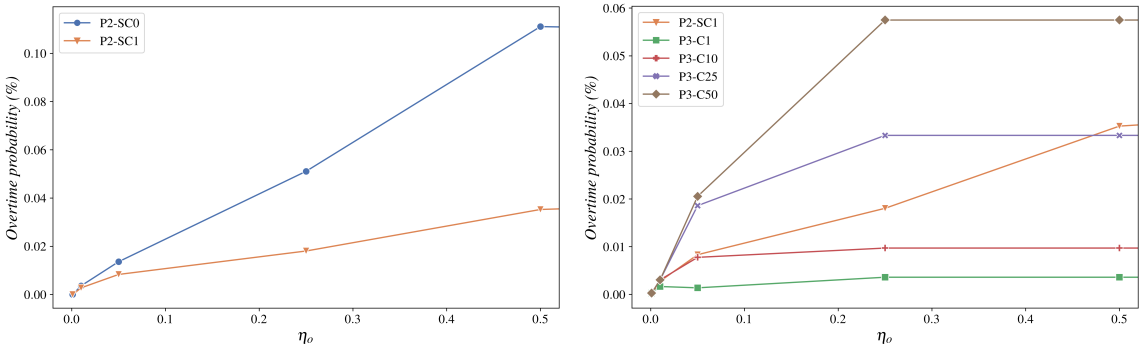
Figure 8.14: Overtime probability (%)

This observation could explain why the cancellation probability for P2-SC1 was lower than expected, as shown in Section 8.4.1. Patterns with high expected overtime in PC0 have a higher probability of cancellations. Hence, when expected overtime is penalized, we implicitly penalize cancellations instead. For the PC1 file, however, the expected overtime is correctly calculated, and the instances no longer penalize cancellations. There are both good and bad aspects of this property. On one side, the result implies that models that do not include cancellations still handle cancellations if used in an environment with cancellation rules. On the other hand, penalizing cancellation implicit through the overestimation of overtime entangles the two KPIs, making the interpretation of the costs much harder. Further, penalizing cancellations should not be seen as a benefit compared to using chance constraints, only as another way of handling the problem.

An unintended consequence of not including the cancellation rule in the formulations is that we may restrict the problem more than intended. Since overtime is overestimated, models not accounting for cancellations might find a pattern infeasible, even if the actual risk of overtime is within the accepted risk level. As a result, we recommend including the cancellation rules in models that use chance constraints and overtime to handle uncertain surgery duration.

Further, we observe that including a cancellation rule decreases the overtime probability, as expected. However, Figure 8.14 shows how adding cancellation rules adds another layer of complexity for the hospital. Finding an acceptable overtime risk level can be a difficult task. However, the

cancellation risk level also affects overtime. As a result, the average overtime is dependent on both risk parameters. Increasing the risk of cancellations can also increase the overtime risk, as seen in Figure 8.15b.



(a) PC=0 (b) SC=1

Figure 8.15: Overtime probability for varying levels of overtime risk (%)

As allowed overtime risk increases, the overtime also increases. However, the overtime probability only increases up to a certain threshold. For P3, the threshold is met when the cancellation probability becomes restricting, at around 0.05 or 0.1, depending on the cancellation risk level. However, the P2 instances do not meet the threshold before $\eta_o = 0.75$ when the expected overtime constraint becomes restricting.

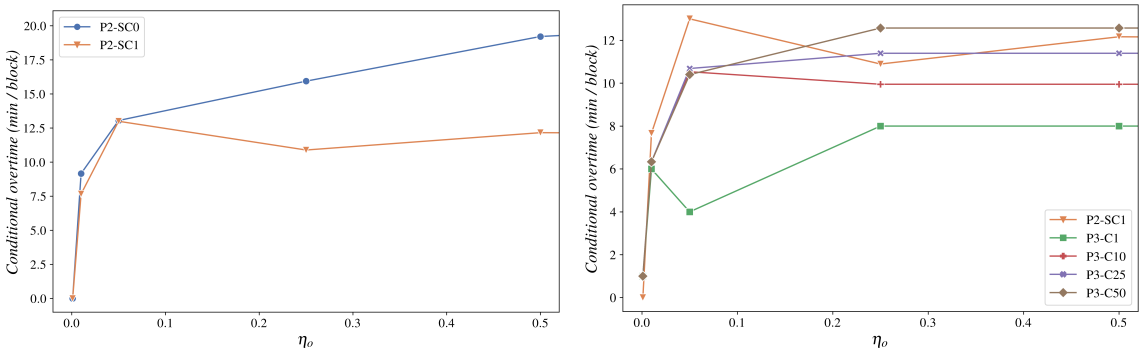
The reader should note how models not accounting for cancellations become increasingly worse at handling overtime correctly as overtime risk increases. Intuitively, this makes sense. If no overtime is allowed, no cancellations can occur either; however, as observed in Section 8.3.3, the utilization of resources becomes considerably worse when not allowing any overtime risk. As a result, given that the hospital accepts the risks, models that do not handle the cancellation rules become worse at correctly handling overtime and increasingly worse as acceptable risk levels are increased.

Conditional overtime

Another appropriate method of evaluating overtime is by looking at the amount of overtime given that overtime occurs. In other words, given that overtime occurs, how bad does it get? One could compare it to conditional Value at Risk (CVaR), often used in financial models. Conditional overtime is especially relevant to investigate when using chance constraints since the constraints do not necessarily restrict the worst-case scenarios.

From Figure 8.16a and Figure 8.16b, we see that the worst case expected overtime is about 12% when the hospital uses the cancellation rules, while 17.5% expected conditional overtime occurs without the cancellation rule. Thus, a cancellation rule decreases the overtime probability and the amount needed when overtime occurs. One possible reason is that the surgery duration is assumed to follow a truncated normal distribution. Although no strict limit for overtime occurs, the chance

constraint still implicitly handles the overtime since the distributions are assumed truncated.



(a) PC=0 (b) SC=1

Figure 8.16: Conditional overtime for varying levels of overtime risk

While overtime probability increases linearly as risk is increased, the amount of conditional overtime looks to rapidly increase to a certain level before flattening out as long as some risks are allowed. This is interesting as there appears to exist a hard limit on the possible amount of overtime, even when a 50% risk of overtime and 50% risk of cancellations is allowed. Also, the limit is approximately the same for the instances as long as any risk is permitted, except for P3-C1, where very low levels of cancellation risk are allowed. There could be two reasons for this. First, remember that overtime is still penalized, so the extra overtime cost in one pattern might never become worth it compared to an alternative pattern with lower overtime costs. Another, and potentially more dominating reason, is that as long that overtime is limited in any way, either with expected overtime limits or cancellation rules, then the maximum possible overtime is based on the range between expected surgery duration and maximum surgery duration for the case with the shortest expected surgery duration. For most specialties, this limits overtime considerably. However, if no maximum surgery duration exists, which might be a more realistic assumption, then the conditional overtime between the instances might become more varied.

Evaluating what risk levels the orthopedic department should use is outside the scope of this thesis. However, the overtime and cancellation risk levels should be considered together, as they both affect the amount of overtime used in the department.

Overtime per speciality

Table 8.19 shows the overtime probability for each specialty at the Department of Orthopedics. Observe how the probability is not strictly increasing when increasing cancellation risk. This illustrates that accepting a higher risk of cancellations does not necessarily reduce the overtime probability, although it is the general trend across all specialties combined. The reason could be that new patterns with a higher probability of cancellations but significantly lower overtime probabilities are included. Increasing the cancellation risk can reduce the overtime probability in that case.

Further, the number of available blocks for each specialty is different. As a result, while the overtime probability risk is guaranteed with the chance constraint, the total overtime for surgical teams might differ considerably between specialties. For instance, if there is a 1% probability of overtime, and a specialty has only one block each week, then the likelihood of no overtime occurring is 99%. However, if the specialty has ten blocks, the likelihood of no overtime across the week is reduced to $0.99^{10} \approx 90\%$. As a result, we recommend individually tuning the overtime and cancellation risk parameters for each specialty, not only at the department level.

Table 8.19: Overtime probability per block for each specialty

	P2-SC0	P2-SC1	P3-C01	P3-C10	P3-C25	P3-C50
Arthroscopic	1.94%	1.39%	0.56%	1.25%	3.33%	6.39%
Back	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Foot	0.00%	0.42%	0.42%	0.00%	4.58%	3.75%
Hand	3.61%	2.78%	0.83%	0.83%	4.72%	6.94%
Plastics	2.92%	1.25%	0.52%	1.98%	3.54%	3.75%
Prosthetics	0.48%	0.00%	0.00%	0.36%	5.60%	7.62%
Tumor	5.83%	0.83%	0.00%	0.83%	0.83%	0.83%

8.4.3 Patient-related quality attributes

The instances create schedules with significantly different user experiences from the patient's point of view. Several KPIs affect a patient's user experience, for instance, the number of times it is rescheduled, how long it has to wait on the waiting list, and the service time in general.

Reschedulings

Figure 8.17 shows the average weekly number of reschedulings. We see that one instance is clearly different from the others, P3-C50. Due to the large number of cancellations in the schedules made by this instance that allows 50% risk of cancellations, the number of plans naturally increases since the canceled cases must get a new plan.

Additionally, since the previously canceled cases likely have waited longer than several cases in the current schedule, the model will attempt to perform reschedulings to get the canceled cases back into the schedule, preferably in one of the first weeks. Thus, we get a double effect on the number of plans due to both cancellations and additional reschedulings.

Interestingly, P2-SC0 is quite similar to P3-C10 and P3-C25 as it illustrates how there is not a 1:1 relationship between the number of reschedulings and cancellations. For instance, P3-C1 has fewer reschedulings on average than P2-SC0. However, the confidence intervals are overlapping, and we should be careful to draw conclusions other than that the introduction of cancellation rules does not automatically drastically increase the number of reschedulings.

A case might have multiple plans from first entering the system to after receiving surgery. A case

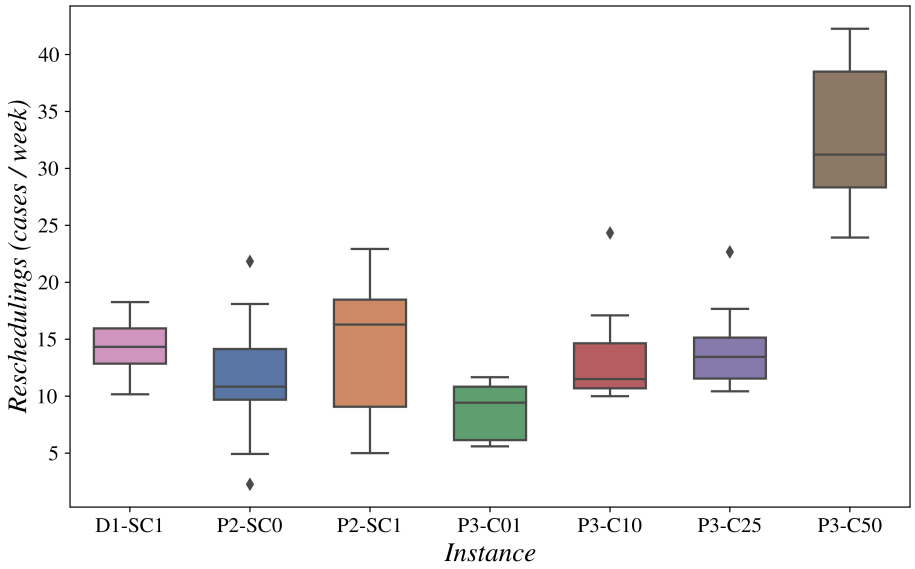


Figure 8.17: Average number of reschedulings

gets a new plan both when rescheduled and when planned again after getting canceled. Table 8.20 shows the likelihood of a case having a certain amount of plans for each instance. It also shows the likelihood that a plan is not changed, given the current number of plans the case has. $P(X \mid X)$ means the probability of having X plans before receiving surgery, given that you have X plans today. $P(X \cup Y)$ represents the likelihood that a case will end up with X or Y plans before receiving surgery.

Table 8.20: Likelihood of having a certain number of plans before receiving surgery

	Avg	Max	$P(1 \mid 1)$	$P(2 \mid 2)$	$P(3 \mid 3)$	$P(4 \mid 4)$	$P(1)$	$P(1 \cup 2)$	$P(1 \cup 2 \cup 3)$	$P(1 \cup 2 \cup 3 \cup 4)$
D1-SC1	1.2	10	82.5	35.2	63.3	41.2	82.5	88.7	95.8	97.6
P2-SC0	1.1	4	86.5	42.0	87.0	100.0	86.5	92.2	99.0	100.0
P2-SC1	1.2	5	83.7	40.9	78.0	89.0	83.7	90.3	97.9	99.8
P3-C01	1.1	4	89.4	43.8	84.3	100.0	89.4	94.0	99.1	100.0
P3-C10	1.1	5	85.3	39.2	75.9	78.1	85.3	91.1	97.9	99.5
P3-C25	1.2	7	84.6	36.6	73.2	63.2	84.6	90.2	97.4	99.0
P3-C50	1.4	9	72.8	32.0	67.0	42.6	72.8	81.5	93.9	96.5

P3-C50 has significantly fewer cases with only one plan, most likely due to the high levels of cancellations. Still, it is interesting that D1-SC1 has a 10% higher probability of one plan since they have roughly equal cancellation probabilities. This is due to D1-SC1 having a much lower number of reschedulings.

A general pattern emerges when comparing P2-SC1 with P3-C1, P3-C10, and P3-C25. The P3 models have a higher probability of only having one plan, while P2-SC1 are better than P3-C10 and P3-C25 at handling cases with multiple plans. This could indicate that the procedures that get canceled are more random for P2-SC1, while the P3 models discriminate more, meaning most cases have few plans at the expense of a few procedure types. This is similar to what we found regarding the cancellation probability.

Lastly, note how all instances have an interesting property: the probability of getting a new plan in

the future increases if a case has more than one plan. One reason for this could be the sequencing rule. If a case is canceled once, it most likely was because the case was scheduled to receive surgery last that day. The results indicate that this is likely to happen at least twice if it first happens once. This is an important finding as it indicates that the schedules should be able to handle the sequencing itself, or the sequencing rules must handle the cancellation probability to avoid discrimination. However, this would either increase the complexity of the models substantially or make for a less effective sequencing rule.

Service time

Average service time decreases as cancellation risk increases, as seen in Figure 8.18. Increased cancellation risk usually means tighter packing of cases. Tighter patterns are thus filtered out using the chance constraints for low levels of η_c. However, increasing η_c will, on average, increase the number of cases that receive surgery and reduce the average service time. We know from Section 8.3.1 that D1-SC1 has low service time due to underestimating overtime packing more tightly. Comparing D1-SC1 with P3-C25 and P3-C50 further strengthens the possibility that tighter packing could explain the service times.

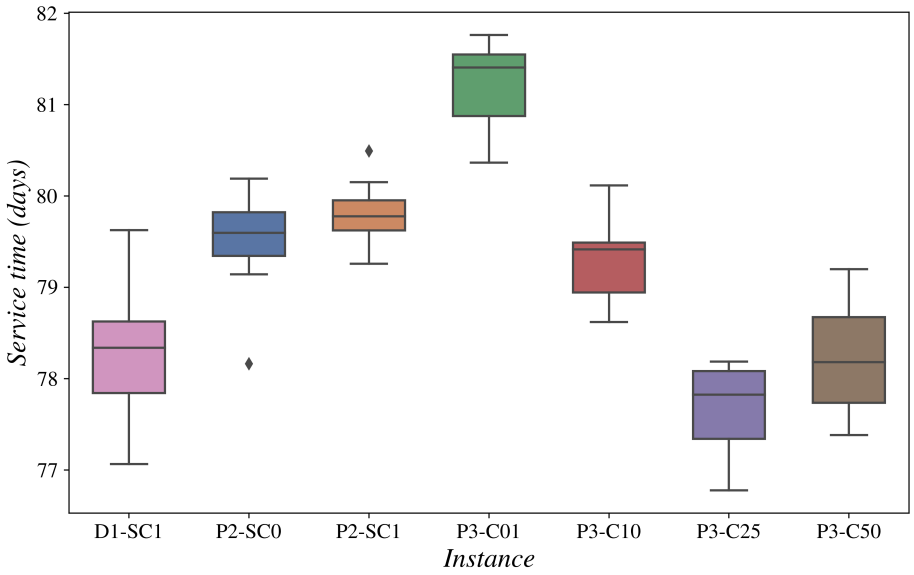
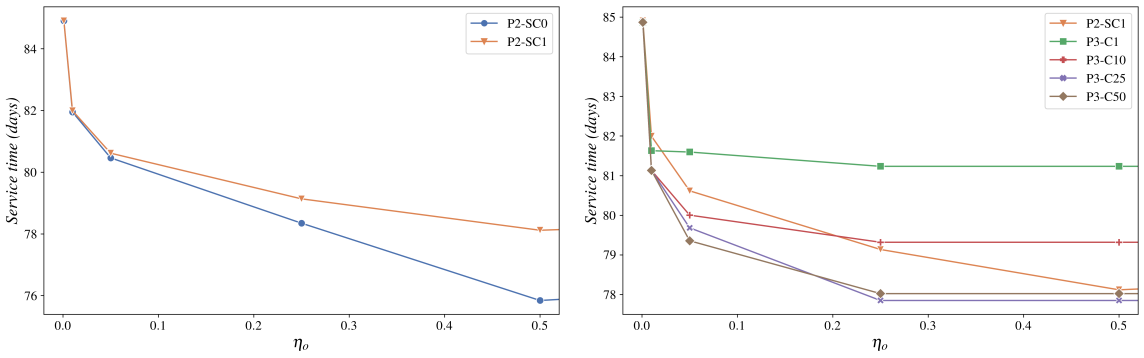


Figure 8.18: Service time (days)

Figure 8.19 shows that increasing the risk level for overtime generally improves the service time. Again, a slight increase in risk level from 0.1% to 1% drastically improves the KPI.

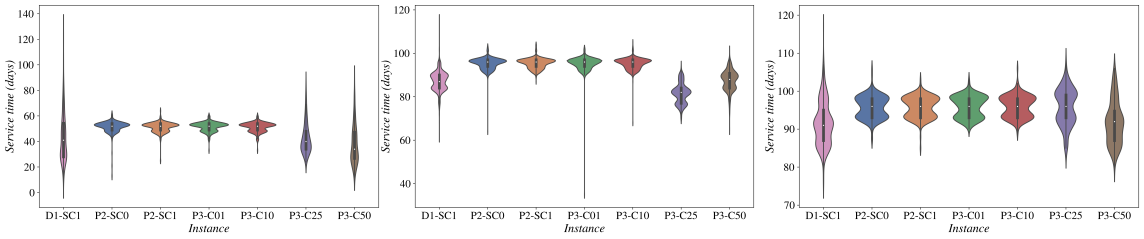
We make some interesting observations by investigating the service time on the procedure level. While the service time distribution for a given procedure might be nearly identical between some of the instances, the same instances might give significantly different results for a different procedure. Figure 8.20 shows the service time distributions for the three Aggregated Knee variants.

P2-SC0, P2-SC1, P3-C1, and P3-C10 all have similar service time distributions. However, the



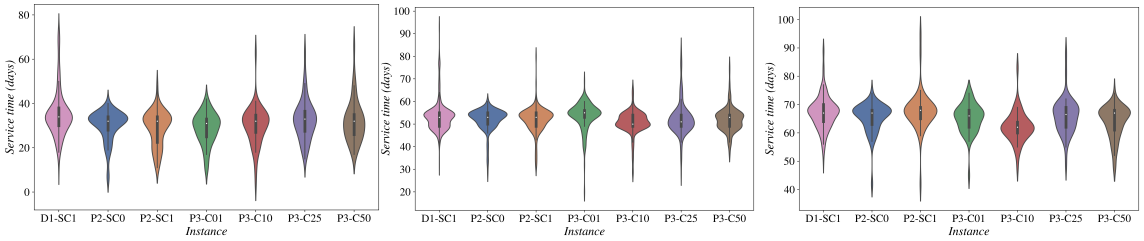
(a) PC=0 (b) SC=1

Figure 8.19: Service time for varying levels of overtime risk



(a) Variant L (b) Variant M (c) Variant H

Figure 8.20: Service time distributions for variants of the Aggregated Knee procedure



(a) Variant L (b) Variant M (c) Variant H

Figure 8.21: Service time distributions for variants of the Plateepitelkarsinom procedure

distributions for the Plateepitelkarsinom procedures shown in Figure 8.21 show more significant differences between the four instances for that specific procedure. We also note that P3-C25 and P3-C50 are quite different from the other four instances for knee procedures.

When analyzing the distributions for all procedures, we do not find any strict correlation between expected surgery duration and the models having worse or better service times. However, we believe the change in service time comes from the different patient mix, which happens when patterns are added or filtered away. This illustrates that one should diligently investigate how the risk levels affect the service times before implementing the schedules to ensure that the quality of care for cases with specific procedures is not unintentionally reduced.

Waiting list

Figure 8.22 and Figure 8.23 show the average and max number of days a case has been on the waiting list for the different instances, respectively. It is a good sign that the average number of days is less than half the service time for every instance, as this indicates that the model is prioritizing scheduling cases that have waited longer.

If any original cases from the warm-up schedule still have not been scheduled, the theoretical max waiting list time is 300 days. We see that all instances have around 100 days, meaning all cases eventually get through the system. When observing the distribution of the waiting list, we see even clearer that the models prioritize the cases with longer waiting times.

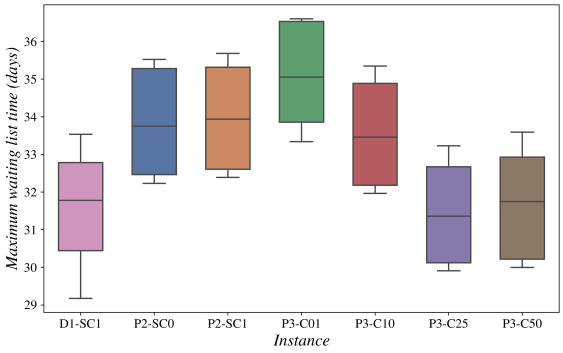


Figure 8.22: Average waiting days for cases on the waiting list (days)

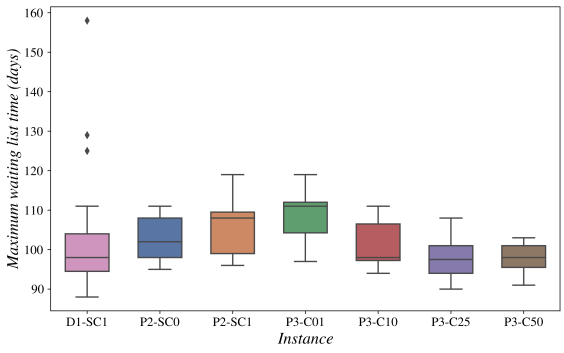


Figure 8.23: Maximum waiting days for cases on the waiting list (days)

Figure 8.24 shows the waiting list distribution for P3-C10. The distributions for the other instances have similar shapes and can be found in Appendix D. Interestingly, all instances perform similarly concerning the waiting list KPIs. Schiøtz and Tysse (2022) found that minor changes to cost function hyper-parameters could significantly change the waiting list aspects; however, this does not seem to be the case for changes to cancellation and overtime risks.

8.4.4 Operational Efficiency and Resource Utilization

Operational efficiency and good resource utilization are critical aspects of scheduling from a governance point of view. Excluding direct patient-centric concerns and making sure the sparse resources are used to their full potential are key success factors.

Throughput

Figure 8.25 shows the average weekly throughput for the six instances. As expected, the introduction of cancellations in the simulation environment slightly reduces throughput for P2-SC1 compared to P2-SC0. P3-C01 has the lowest throughput, likely due to the low cancellation risk of 1%. P3-C10 achieves approximately the same throughput as P2-SC0 and P2-SC1, while P3-C25

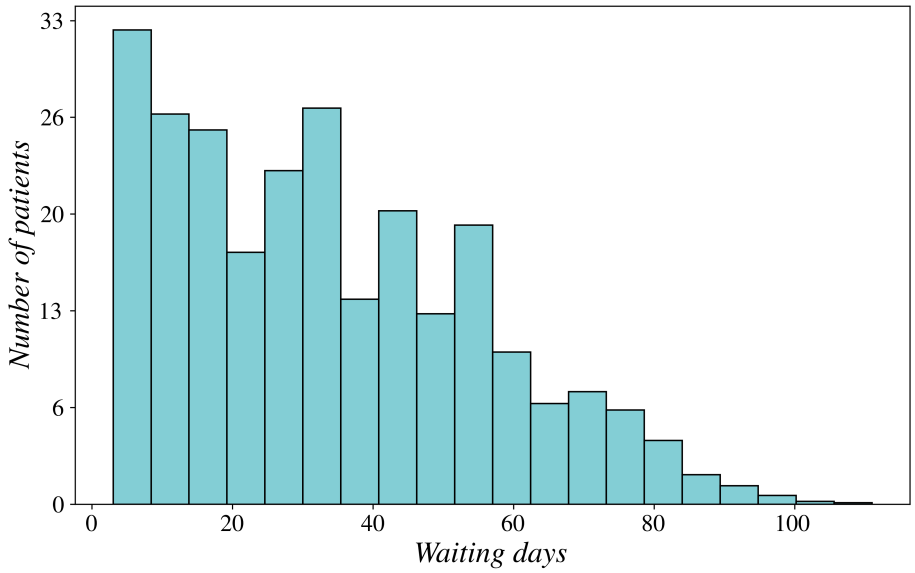


Figure 8.24: Waiting days for cases at the waiting list for P3-C10 (days)

and P3-C50 have significantly higher throughput than the other instances. We see that we do not gain any higher throughput by increasing the cancellation risk from 25% to 50%.

One might expect P2-SC0 to have higher throughput than the other models since the environment is without cancellations. However, since the expected overtime is calculated differently between P2 and P3, and the overtime cost is relatively high, we believe that P2-SC0 overbooks less than, e.g., P3-C25 and P3-C50, since P2-SC0 will get a higher total cost of overtime for the same schedule.

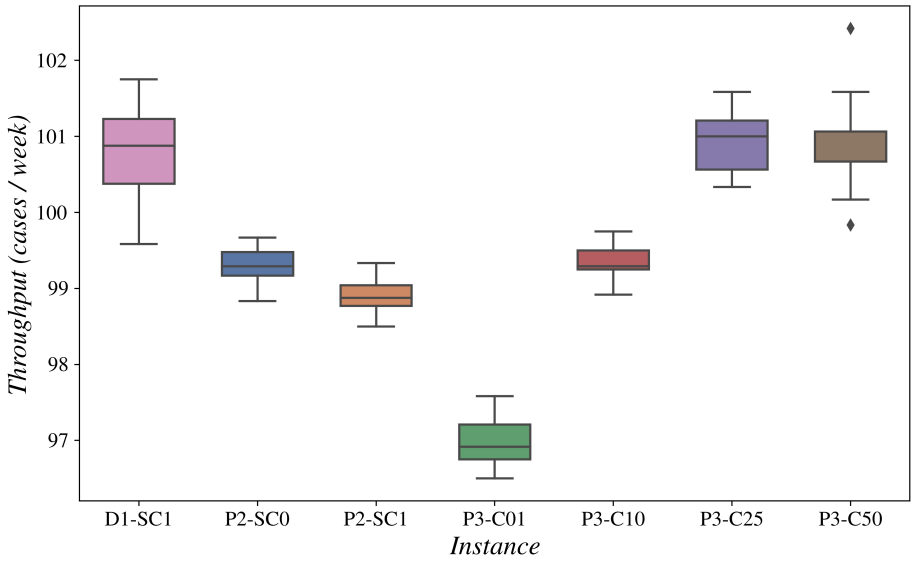
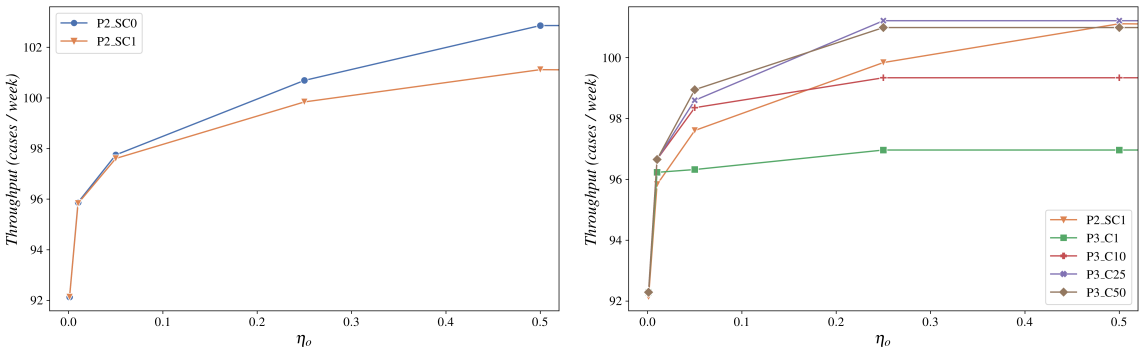


Figure 8.25: Throughput (Cases)

Figure 8.26 shows how the throughput increases when the overtime risk level is increased. We see a significant increase in throughput from a slight overtime risk increase at the start, but this effect quickly diminishes, especially after 10%.



(a) PC=0 (b) SC=1

Figure 8.26: Throughput for varying levels of overtime risk

The operating rooms

The OR utilization for the different instances is shown in Figure 8.27. We note that the shape of the plot is very similar to the throughput, which is expected. The instances achieve utilization of 85-90% utilization, which we would argue is quite good in an uncertain environment. Also, remember that the Back specialty only has cases with long surgery duration and is likely responsible for a considerable amount of the undertime since often only one case is scheduled to a Back block.

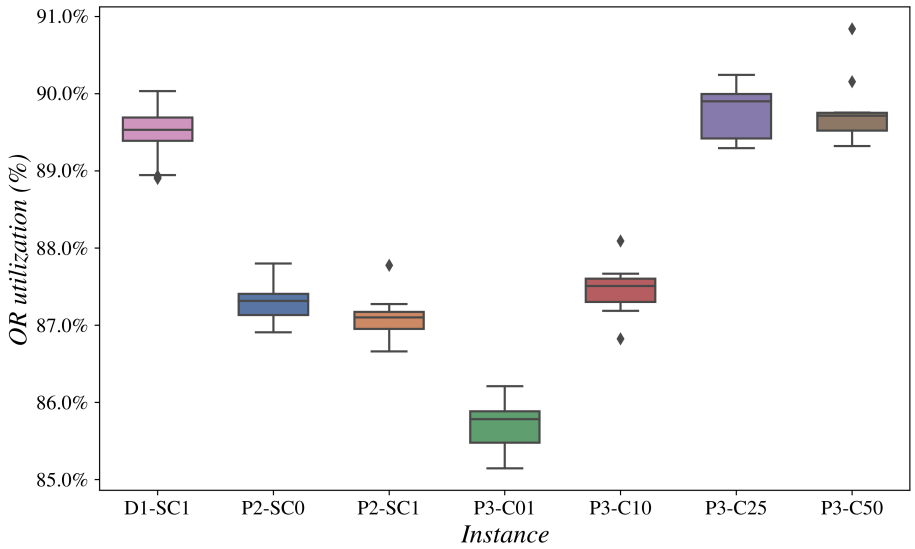
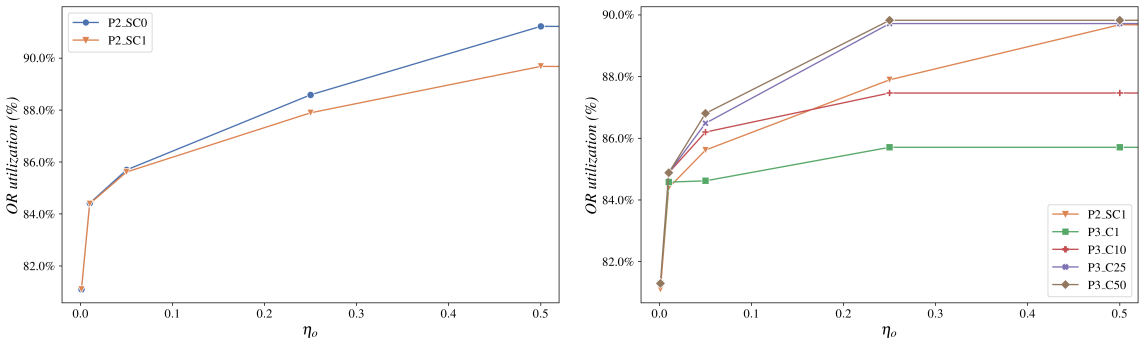


Figure 8.27: OR utilization (%)

When increasing the overtime risk, we see in Figure 8.28a that we can achieve over 90% in the simulation environment without cancellations. However, around 89% appears to be the maximum in the environment with cancellations. These findings indicate that introducing a cancellation rule to avoid overtime directly reduces the utilization of available OR resources.

Interestingly, we see in Figure 8.28b that P2-C1 is not able to achieve higher than 85% utilization even when no risk restriction on overtime is set. This happens due to the low cancellation risk level of 1%, which has removed patterns where more utilization is possible.



(a) PC=0 (b) SC=1

Figure 8.28: OR utilization for varying levels of overtime risk

The wards

Looking at the average ward utilization, we see in Figure 8.29 that all instances have around 50% utilization. P3-C25 and P3-C50 have slightly higher ward utilization than the other instances, which the higher throughput can explain. Interestingly, although P3-C1 has significantly lower throughput than the other instances, the ward utilization is around the same. This can be explained due to the different patient mixes selected by the instances. For example, since P3-C1 has lower throughput and higher average ward utilization than P3-C10, we infer that P3-C1 schedules fewer cases than P3-C10, but the cases have longer LOS compared to cases scheduled by P3-C10.

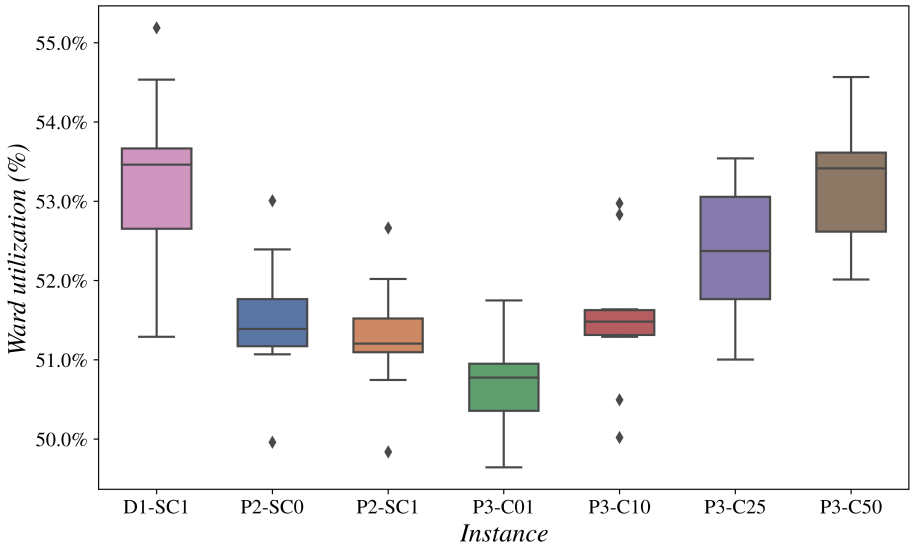


Figure 8.29: Ward utilization (%)

Figure 8.30 shows the ward utilization for the different instances during the week. We see that the utilization increases from Monday to Friday and decreases from Friday to Monday again. The decrease during the weekend is natural because no surgeries are performed on Saturday or Sunday.

Since the ward capacity on Friday is lower than the other weekdays, there might be fewer cases in the wards on Friday versus Thursday, even though the utilization is higher on Friday. This is

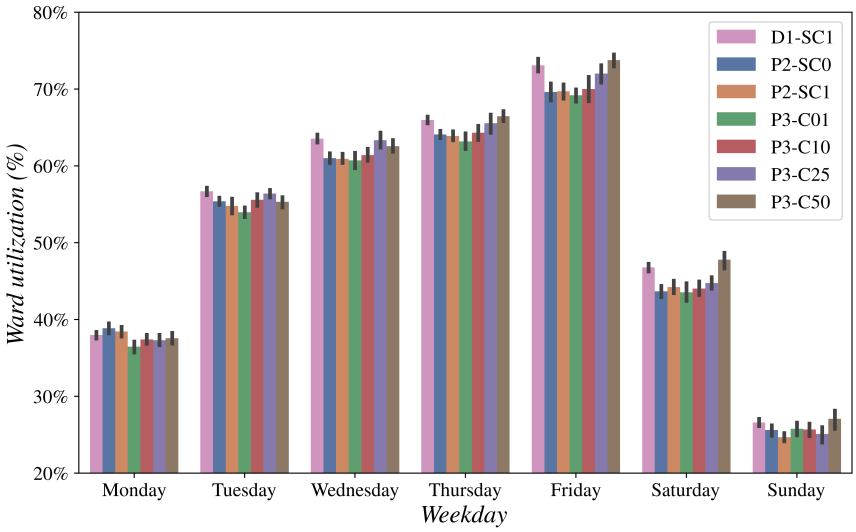


Figure 8.30: Ward utilization per day (%)

shown in Figure 8.31, which shows the number of cases occupying the wards over a period of time from a selected simulation. Since the ward capacity is lower on the weekends, it is intuitive to believe that all instances want to schedule cases with high LOS early in the week.

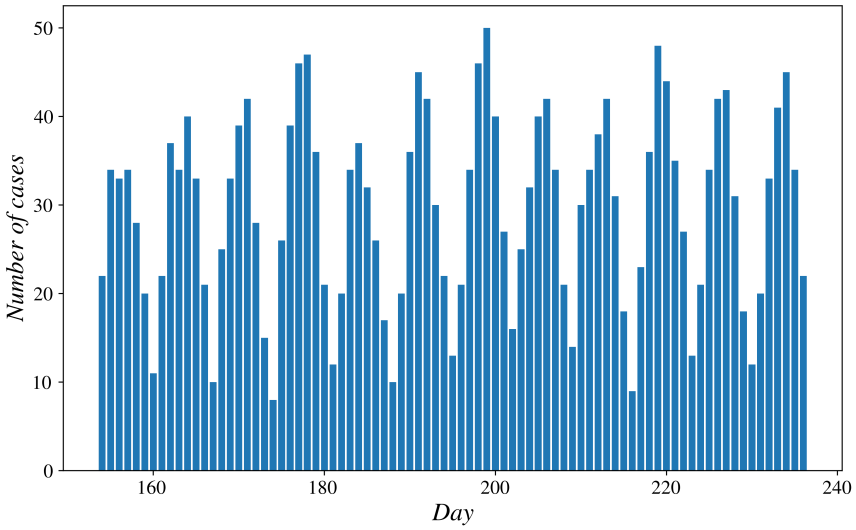


Figure 8.31: Number of cases in wards each day (Cases)

We see in Figure 8.32 that this is indeed the case. However, this result is also affected by the MSS since the LOS of patients is often an important factor when creating the MSS (T. R. Bovim et al. 2020). Back patients, for example, have the highest LOS and only have available blocks on Monday, Tuesday, and Wednesday. We do not see any notable differences between the instances regarding the per-day distributions.

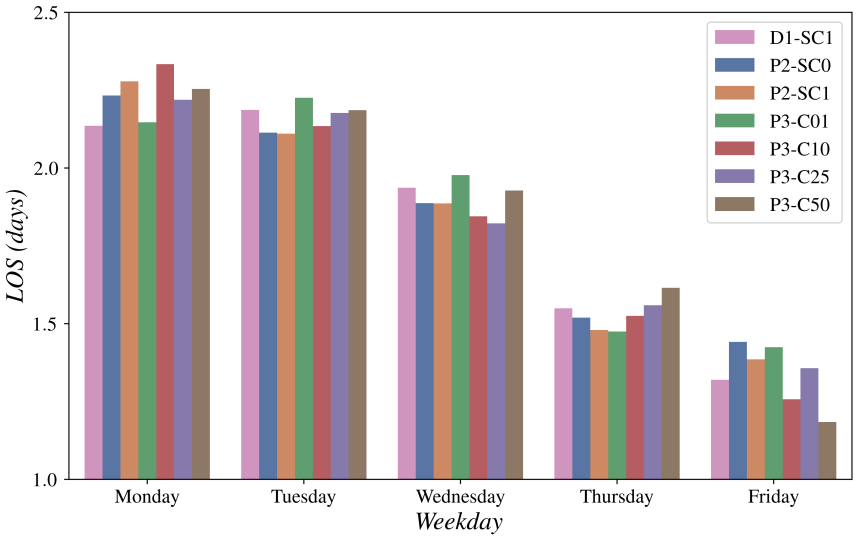


Figure 8.32: Average LOS per day (days)

8.4.5 A combined view

Table 8.21 shows how the instances perform against each other for a set of KPIs. The colors represent the instance performance relative to the maximum and minimum values across all the instances. The colors should be interpreted cautiously, as the values for some KPIs are quite similar. For instance, there are most likely little practical differences between 50.72% and 53.38% ward utilization. A few key observations emerge from the table based on the color indicators.

Table 8.21: Summary of findings for the instances based on key performance indicators

	D1_SC1	P2_SC0	P2_SC1	P3_C01	P3_C10	P3_C25	P3_C50
Throughput	100.79	99.28	98.92	96.97	99.34	100.94	100.86
OR utilization	89.51%	87.30%	87.09%	85.71%	87.47%	89.77%	89.78%
Ward utilization	53.28%	51.51%	51.28%	50.72%	51.52%	52.40%	53.27%
Overtime prob	4.95%	1.83%	0.94%	0.36%	0.97%	3.72%	5.03%
Conditional overtime	12.78	12.79	10.78	8.00	9.95	11.53	12.78
Cancellations	26.86%	0.00%	1.25%	0.25%	2.81%	12.58%	28.39%
Reschedulings	14.26	11.48	14.56	8.74	13.38	14.24	32.55
Number of plans	1.22	1.12	1.15	1.09	1.14	1.18	1.41
Service time	78.30	79.50	79.78	81.24	79.32	77.66	78.20
Average waiting list	31.59	33.84	33.98	35.08	33.56	31.45	31.65
Max waiting list	102.13	102.90	105.60	108.60	101.10	97.60	97.80

The KPIs can be categorized based on efficiency and stability, where efficiency is further segmented into hospital and patient-oriented efficiency. A clear trade-off between the efficiency and stability of the schedules exists. When stability KPIs - i.e., cancellation probability, overtime, reschedulings, and the number of plans - are prioritized, we observe a decrease in efficiency KPIs such as utilization, throughput, service times, and waiting lists, and vice versa. This finding is interesting as it illustrates an important issue for the healthcare sector. The demand for health care is expected to increase drastically, so the sector will need to use its resources more efficiently. However,

our findings indicate that increased efficiency could worsen the user experience for patients and employees.

In many ways, this trade-off represents an ethical problem. While hedging against cancellation risk might benefit individual cases, it could, on average, lead to worse outcomes for all. We recommend diligently tuning the models if they are to be used in real life. Continual adaptation, ethical guidelines, and transparent communication with stakeholders are advisable. Note that this might be a general issue, also with manual scheduling. However, the potential consequences could be more severe if the situation is not handled carefully.

A promising finding is that we can create schedules representing a middle ground between extremes by independently adapting the cancellation and overtime risk parameters. For example, both P3-C10 and P3-C25 have a nice mixture of effectiveness and stability. Deciding the optimal combination of risk parameters is outside the scope of this thesis. For instance, if patient-related stability is not the most important, an increase in the cancellation risk is likely recommendable. On the other hand, cancellations often increase the service times for the specific case, and one should hedge more against cancellation risk if the surgeries have strict timelines.

Introducing a cancellation rule decreases the system's efficiency, as seen when comparing P2-SC0 and P2-SC1. However, it accomplishes its primary goal of reducing the overtime probability at the clinic. At the same time, the differences are minor for many KPIs, and if the ability to proactively cancel cases to avoid overtime is essential, then the rule might have some merit. In one way, it is surprising that the rule's negative effects are not more significant.

It is possible to create sound schedules for an environment with cancellations, even when the model does not assume any cancellations. P2-SC1 is an example of this, which is quite similar to P3-C10. The model achieves this by indirectly penalizing cancellations through overestimating overtime. Still, depending on the specific use case, the added versatility and increased interpretability of decoupling overtime and cancellation risk might be preferable. For example, understanding how the risk is handled based on the parameters might be especially important based on the ethical side of the problem.

Our findings indicate that fully deterministic models that do not assume any uncertain surgery duration have significant disadvantages when applied in an environment with cancellations. This is because the underestimation of overtime transfers into an increased probability of cancellations, and in our case, D1-SC1 performance matches P3-C50 the most. In other words, the deterministic model performs similarly to one that only guarantees a $< 50\%$ probability of cancellations.

8.5 Limitations of the study and future research

To wrap up the computational study, we highlight some limitations and offer suggestions for future research to build on the work in this thesis.

We have not thoroughly tested the scalability of pattern generation when the number of procedures increases. Other hospitals and departments may have a much larger number of procedures than we have in our case study of the Department of Orthopedics at St. Olavs hospital. How an increased number of procedures affects the complexity of the pattern generation and the model complexity should thus be further investigated. An approach for handling this increase is to investigate which patterns are actually used over a longer time period and only continue to use these. Another approach can be to use machine learning techniques like reinforcement learning to identify and generate good patterns. We believe that only a fraction of the available patterns are used by the model, and removing patterns that are never or rarely used can thus drastically reduce the complexity and allow for solving larger problems.

The planning activity is performed with a fixed interval in our case study. While adding cases to the waiting list during the week and postponing the scheduling activity until we have as much information as possible may provide the best plans from the hospital's point of view, the patients may prefer to get to know the appointment time much earlier. We suggest further investigating an approach to dynamic scheduling where the scheduling activity is done as new cases arrive.

The problem investigated in this thesis does not consider non-plan-specific cancellations. A non-plan-specific cancellation can occur for many reasons, like a patient calling in to cancel an appointment or a surgeon getting sick. This type of cancellation was investigated by Schiøtz and Tysse (2022), but a very simplified approach was used. Non-plan-specific cancellations can significantly impact hospital resource utilization, and methods to predict these cancellations should be further investigated.

The models developed in this thesis assume the distributions of the random variables for surgery duration and LOS are known. Finding a good distribution to represent the real world can be quite challenging. We used pretty simplified distributions in our case study, and we advise finding better distributions if the models are to be applied in a real-life setting. With the increased accessibility of data in the digitalized world, methods like case-based reasoning or other machine learning techniques may be used to provide realistic distributions.

None of our models consider that the scheduling activity is repeated and that decisions in one planning stage will have consequences for the next. To be able to take future events like uncertain patient arrivals into account when making the decision, we believe reinforcement learning may provide good solutions. With modern graphic processors, deep reinforcement learning is commonly available and can solve large problems where table-based reinforcement learning techniques are

unsuitable.

Chapter 9

Concluding Remarks

In the conclusion of our thesis, we reflect on our accomplishments in achieving the defined goals. Firstly, we aimed to build on the proposed mathematical model by Schiøtz and Tysse (2022) to account for uncertain surgery duration and recovery time. We developed revised models that align with the department's scheduling rules, emphasizing the reduction of patient waiting time and schedule disruptions. Resource considerations in these models included operating rooms, recovery wards, and the pre-existing Master Surgery Schedule (MSS). Secondly, we investigated the implications of integrating a cancellation rule into the Department of Orthopedics' scheduling system and explored how the risk levels of cancellations and overtime affected scheduling quality. Lastly, our objective was to smoothly incorporate the proposed models within a simulation environment using a rolling horizon approach, enabling a quantitative comparison of scheduling quality across various proposed models.

To achieve these objectives, we proposed two Multi-Objective Mixed-Integer mathematical programs. The first is a two-stage stochastic formulation similar to the existing literature. The second is a pattern-based Mixed-Integer Program (MIP) designed to precalculate uncertain variables, enabling deterministic problem-solving.

Based on the goals, and the proposed solutions, we defined four research topics to be the focus of our computational study: (1) Can a deterministic pattern-based MIP improve computational efficiency while handling uncertainty? (2) How do overtime- and cancellation risk influence schedule quality? (3) How does the use of cancellation rules affect scheduling quality?, and (4) What value do models that include the cancellation rule add?

Regarding computational complexity and performance, the pattern-based model significantly outperforms the two-stage stochastic model when a reasonable number of scenarios is used in the SAA algorithm. For real-life size problems, the two-stage model was not able to run with the required number of scenarios. When reducing the number of scenarios so that we can run the model, the

pattern-based model is still more than 60 times faster than the two-stage stochastic model at finding a feasible solution and was able to find the optimal solution within the time limit in all our tests, contrary to the two-stage model. Even the deterministic counterpart of the two-stage model with only one scenario could not reach the optimal solution for more than a planning horizon of two weeks in the large test case, highlighting the computational advantage of the developed pattern-based model. When applied to a toy-sized problem, the two-stage stochastic model can find reasonably good solutions in an acceptable amount of time but is still outperformed by the pattern-based model. Our results indicate that the complexity of the two-stage stochastic model is too high for solving real-life-sized problems. However, column generation and Benders decomposition could significantly improve the two-stage stochastic model's computational performance. Our pattern-based model is an alternative approach to these methods, and we have shown that our method is more than efficient enough to solve real-life size problems.

The pattern and two-stage stochastic models showed promising results, managing to avoid overtime, which suggests their effectiveness in capturing the uncertainties in surgery duration. However, the deterministic (EVP) model struggled to avoid overtime. The pattern-based model did not accommodate uncertain Length of Stay (LOS) as effectively as its two-stage stochastic counterpart, but it significantly outperformed the deterministic model. This suggests that future deterministic models might benefit from relaxing the integrality restriction for extra ward beds and approximating the extra ward bed costs.

The introduction of a chance constraint to model the risk of overtime suggested a clear trend: even a minimal increase in the risk of overtime can lead to a substantial improvement in scheduling quality. However, this improvement displayed diminishing returns as risk levels increased. As a result, hospitals should be careful when deciding their accepted risk levels and avoid fully hedging against overtime.

Further, we observed that stochastic models using SAA became significantly worse when overtime and ward restrictions became hard. This is likely due to the model needing to be fully stable. When applied in a multistage rolling horizon setting where rescheduling was allowed, potential stability complications look to be enhanced. As a result, our research indicates that one should be careful combining SAA with models that allow rescheduling cases.

Intriguingly, models that do not account for cancellations can still maintain a low cancellation risk when analyzed within a framework with cancellation rules. This occurs due to the models' tendency to overestimate the overtime costs, implicitly penalizing cancellations in the objective function. Additionally, restricting the expected overtime also implicitly reduces the likelihood of cancellations.

On the other hand, deterministic models that do not consider uncertainty tend to underestimate overtime and hence do not exhibit this property. As a result, our research strongly advocates

including uncertainty when cancellation rules apply. Even though cancellations can be implicitly handled by restricting overtime, models that account for cancellations might be preferable due to their enhanced interpretability and versatility, especially in light of the ethical considerations associated with the problem.

Finally, we uncovered a critical balance between schedule stability and efficiency. Given the mounting pressure for enhanced efficiency in healthcare due to rising demand, this observation raises an essential dilemma. Pursuing greater efficiency may compromise stability, potentially leading to ethical concerns and a deteriorating user experience. Relevant authorities must address these issues while optimizing healthcare operations.

In conclusion, our thesis provides valuable insights into the value of uncertainty and the consequences of cancellation rules when optimizing the advance scheduling of surgeries. We illustrate the potential issues with models not accounting for uncertain surgery duration and recovery times. Further, we showcase the value of a pattern-based MIP using set partitioning constraints and extensive precalculation of stochastic information compared to two-stage stochastic counterparts using SAA. Our thesis underscores the need for robust, adaptable, and ethically-conscious schedules. Our work lays the foundation for future research in the ASP with uncertain surgery duration and recovery times that account for real-world complexities like a cancellation rule to avoid overtime. We hope this exploration contributes to more resilient healthcare systems, ultimately bringing us closer to meeting the challenges faced in the industry.

Bibliography

Addis, Bernardetta, Giuliana Carello, Andrea Grosso and Elena Tànfani (2015). 'Operating room scheduling and rescheduling: a rolling horizon approach'. eng. In: *Flexible services and manufacturing journal* 28.1-2, pp. 206–232.

Bertsimas, Dimitris and Melvyn Sim (2004). 'The Price of Robustness'. In: *Operations Research* 52.1, pp. 35–53.

Bovim, G. et al. (2023). *Tid for handling - Personellet i en bærekreftig helse- og omsorgstjeneste.* Vol. 4. Norges Offentlige Utredninger.

Bovim, Thomas Reiten (2018). *Stochastic Master Surgery Scheduling for the orthopaedic department at St. Olav's Hospital.*

Bovim, Thomas Reiten, Marielle Christiansen, Anders N. Gullhav, Troels Martin Range and Lars Hellemo (2020). 'Stochastic master surgery scheduling'. In: *European Journal of Operational Research* 285.2, pp. 695–711.

Cardoen, Brecht, Erik Demeulemeester and Jeroen Beliën (2010). 'Operating room planning and scheduling: A literature review'. In: *European Journal of Operational Research* 201.3, pp. 921–932.

Dantzig, George B (1990). 'Origins of the simplex method'. In: *A history of scientific computing*, pp. 141–151.

Delorme, Maxence, Manuel Iori and Silvano Martello (2016). 'Bin packing and cutting stock problems: Mathematical models and exact algorithms'. In: *European journal of operational research* 255.1, pp. 1–20.

Flaata, Linda (Sept. 2022). *Personal Interview.* Information provided directly by the author.

Gul, Serhat, Brian Denton and J.W. Fowler (Nov. 2015). 'A Progressive Hedging Approach for Surgery Planning Under Uncertainty'. In: *INFORMS Journal on Computing* 27, pp. 755–772.

Haghi, M., S.M.T. Fatemi Ghomi and P. Hooshangi-Tabrizi (2017). 'A novel deterministic model for simultaneous weekly assignment and scheduling decision-making in operating theaters'. In: *Scientia Iranica* 24.4, pp. 2035–2049.

Harris, Sean and David Claudio (Mar. 2022). 'Current Trends in Operating Room Scheduling 2015 to 2020: a Literature Review'. In: *Operations Research Forum* 3.

Helsedirektoratet (2022). *DRG-systemet*.

Hulshof, Peter J. H., Nikky Kortbeek, Richard J. Boucherie, Erwin W. Hans and Piet J. M. Bakker (Dec. 2012). 'Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS'. In: *Health Systems* 1.2, pp. 129–175.

Jebali, Aida and Ali Diabat (2015). 'A stochastic model for operating room planning under capacity constraints'. In: *International Journal of Production Research* 53.24, pp. 7252–7270.

— (2017). 'A Chance-constrained operating room planning with elective and emergency cases under downstream capacity constraints'. In: *Computers & Industrial Engineering* 114, pp. 329–344.

Kall, Peter and Stein Wallace (Jan. 1994). *Stochastic Programming*. Vol. 46.

Kamran, Mehdi A., Behrooz Karimi and Nico Dellaert (2018). 'Uncertainty in advance scheduling problem in operating room planning'. In: *Computers & Industrial Engineering* 126, pp. 252–268.

Lawler, Eugene L and David E Wood (1966). 'Branch-and-bound methods: A survey'. In: *Operations research* 14.4, pp. 699–719.

Magerlein, J M and J B Martin (1978). 'Surgical demand scheduling: a review'. en. In: *Health Services Research* 13.4, pp. 418–433.

Min, Daiki and Yuehwern Yih (2010). 'Scheduling elective surgery under uncertainty and downstream capacity constraints'. In: *European Journal of Operational Research* 206.3, pp. 642–652.

Moosavi, Amirhossein and Sadoullah Ebrahimnejad (2018). 'Scheduling of elective patients considering upstream and downstream units and emergency demand using robust optimization'. eng. In: *Computers & industrial engineering* 120, pp. 216–233.

— (2020). 'Robust operating room planning considering upstream and downstream units: A new two-stage heuristic algorithm'. eng. In: *Computers & industrial engineering* 143, p. 106387.

Neyshabouri, Saba and Bjorn P. Berg (2017). 'Two-stage robust optimization approach to elective surgery and downstream capacity planning'. In: *European Journal of Operational Research* 260.1, pp. 21–40.

Ozcan, Yasar A., Elena Tànfani and Angela Testi (2017). 'Improving the performance of surgery-based clinical pathways: a simulation-optimization approach'. eng. In: *Health care management science* 20.1, pp. 1–15.

Rachuba, Sebastian and Brigitte Werners (Apr. 2017). 'A fuzzy multi-criteria approach for robust operating room schedules'. In: *Annals of Operations Research* 251.1, pp. 325–350.

Rahmaniani, Ragheb, Teodor Gabriel Crainic, Michel Gendreau and Walter Rei (2017). 'The Benders decomposition algorithm: A literature review'. In: *European journal of operational research* 259.3, pp. 801–817.

Range, Troels Martin, Dawid Kozlowski and Niels Chr Petersen (2019). 'Dynamic job assignment: A column generation approach with an application to surgery allocation'. In: *European Journal of Operational Research* 272.1, pp. 78–93.

Al-Refaie, Abbas, Mays Judeh and Toly Chen (Oct. 2018). 'Optimal multiple-period scheduling and sequencing of operating room and intensive care unit'. In: *Operational Research* 18.3, pp. 645–670.

Samudra, Michael et al. (Oct. 2016). 'Scheduling operating rooms: achievements, challenges and pitfalls'. In: *Journal of Scheduling* 19.

Schiøtz, Mats Ingesen and Jostein Hjortland Tysse (2022). *A rolling horizon approach to the Advance Scheduling Problem with downstream constraints: A case study on a Norwegian Hospital.*

Shehadeh, Karmel S. and Rema Padman (2021). 'A distributionally robust optimization approach for stochastic elective surgery scheduling with limited intensive care unit capacity'. In: *European Journal of Operational Research* 290.3, pp. 901–913.

— (2022). 'Stochastic optimization approaches for elective surgery scheduling with downstream capacity constraints: Models, challenges, and opportunities'. eng. In: *Computers & operations research* 137, p. 105523.

St. Olavs hospital HF (2018). *Utviklingsplan 2019–2035.* St'Olav's hospital.

— (2022). *Utviklingsplan 2023–2026.* St'Olav's hospital.

Van Riet, Carla and Erik Demeulemeester (2015). 'Trade-offs in operating room planning for electives and emergencies: A review'. In: *Operations Research for Health Care* 7. ORAHS

2014 - The 40th international conference of the EURO working group on Operational Research Applied to Health Services, pp. 52–69.

Wang, Shuo, Vahid Roshanaei, Dionne Aleman and David Urbach (2016). 'A discrete event simulation evaluation of distributed operating room scheduling'. In: *IIE Transactions on Healthcare Systems Engineering* 6.4, pp. 236–245.

Zhang, Jian, Mahjoub Dridi and Abdellah El Moudni (2019). 'A two-level optimization model for elective surgery scheduling with downstream capacity constraints'. In: *European Journal of Operational Research* 276.2, pp. 602–613.

— (2020). 'Column-generation-based heuristic approaches to stochastic surgery scheduling with downstream capacity constraints'. eng. In: *International journal of production economics* 229, p. 107764.

— (2021). 'A two-phase optimization model combining Markov decision process and stochastic programming for advance surgery scheduling'. eng. In: *Computers & industrial engineering* 160, p. 107548.

Zhu, Yue, Yulin Zhang, Zhuhan Jiao and Dong Li (2015). 'Surgical scheduling under patients' uncertain anesthesia recovery time'. In: *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 1–4.

Appendix

A Alternative Pattern Model Extension: Reserve capacity

We formulate an extended version of the pattern model that features reserving capacity using dummy surgical cases. If a pattern is expected to give n cancellations in a given week, the model should reserve capacity for these n cases later in the planning horizon. The model mostly uses the same notation as the standard pattern model, so we only show the new notation and model changes here. We introduce the set W, a subset of the days in the planning horizon consisting of all Sundays in the horizon (see Table A.1). The variable d_{prt} is used for reserving dummy cases (see Table A.2), and the parameter G_j is the expected number of cancellations from pattern j (see Table A.3). Details on how G_j is calculated is presented in chapter 6

Table A.1: Sets used in the extended pattern model.

Symbol	Description
$\mathcal{W}$	Set of Mondays (first day in week) in planning horizon $t \in \mathcal{W} \subseteq \mathcal{T}$

Table A.2: Extra Decision Variables for the extended pattern model

Symbol	Description
d_{prt}	Number of dummy patients with procedure p scheduled in block (r, t)

Table A.3: Parameters for the extended pattern model

Symbol	Description
G_{pjt}	Expected cancellations of patient type p from pattern j on day t

The objective function is identical to the pattern model's, (5.5a)-(5.5c). Constraints (5.5d) are replaced with (1), which adds the opportunity to add dummy patients d_{prt} to match the pattern capacity.

$$\sum_{i \in \mathcal{I}_p} x_{irt} + d_{prt} = \sum_{j \in \mathcal{J}_{rt}} B_{pj} \pi_{jrt} \quad p \in \mathcal{P}, r \in \mathcal{R}, t \in \mathcal{T} \tag{1}$$

We add the constraints shown in (2), which states that the expected number of canceled patients with a given procedure and week must be exceeded by the number of scheduled dummy patients, d_{prt}, scheduled in the following week.

$$\sum_{j\in\mathcal{J}}\sum_{r\in\mathcal{R}}\sum_{t'=t}^{t+6} G_{pjt}\pi_{jrt'} \leq \sum_{r\in\mathcal{R}}\sum_{t^*=t+7}^{t+13} d_{prt^*} \quad p\in\mathcal{P}, t\in\mathcal{W} \tag{2}$$

Lastly, we add the variable declaration constraints for the dummy patient variables in (3). The variables can take positive integer values, including 0.

$$d_{prt} \in \mathbb{Z}^+ p\in\mathcal{P}, r\in\mathcal{R}, t\in\mathcal{T} \tag{3}$$

All other sets, indices, parameters, constraints, and variables from the original pattern model are included and unchanged.

We test the model against another pattern model, both using chance constraints ensuring less than 10% cancellation risk. We could not draw any specific conclusions from the results shown in Figure A.1 on whether the extended model is better than the P3-C10 model.

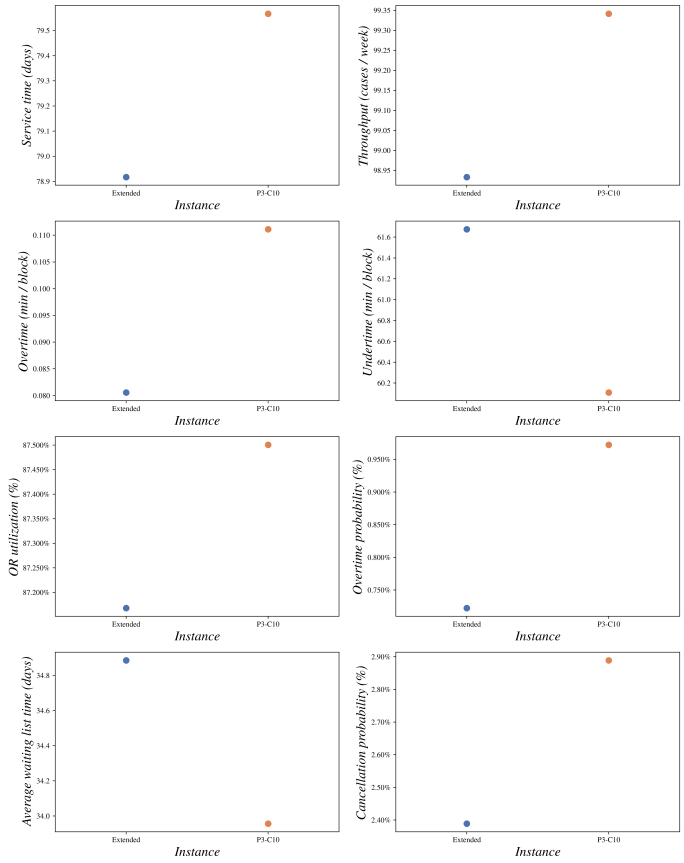


Figure A.1: KPIs for the extended model and P3-C10 model in a setting withouth rescheduling.

B Model and simulation parameters

Table B.1 shows the parameters used when generating the warmup schedule. Table B.2 shows the default parameters used for the rest of the technical study. In most performed tests, one or more parameters differ from the default. The other parameters can be assumed to take the default values.

Table B.1: Parameters used when creating the warmup schedule.

Parameter	Value
Model	P1
Gurobi Time Limit	500 seconds
Gurobi Gap Limit	1%
Planning Horizon	10 weeks
Simulation Length	60 weeks
Allow PCS	False
Allow Rescheduling	True
C^P	1
C^Φ	3
γ_s	1.333
γ_v	1.383
α	0.25
β	12
η_o	1
η_c	1
$\overline{O}$	30 minutes

Table B.2: Default parameters in the computational study.

Parameter	Value
Gurobi Time Limit	300 seconds
Gurobi Gap Limit	1%
Planning Horizon	4 weeks
Simulation Length	17 weeks
Allow PCS	True
Allow Rescheduling	True
C^P	200
C^Φ	8
γ_s	1.333
γ_v	1.383
α	10
β	12
η_o	1
η_c	1
$\overline{O}$	30 minutes

C Technical study - Results

C.1 Small case

Table C.1, Table C.2 and Table C.3 shows the time to reach gaps for the D1, P1 and S1 model, respectively.

Table C.1: Time, in seconds, to reach gaps for D1 in the small case.

Planning Horizon	First Feasible	20%	5%	1%	Optimal
1 week	0.016	0.016	0.016	0.016	0.053
2 weeks	0.031	0.031	0.031	0.031	0.122
3 weeks	0.433	0.449	0.449	0.457	1.527
4 weeks	0.604	0.604	0.604	0.682	2.175
5 weeks	0.705	0.708	0.708	0.848	4.528
6 weeks	0.751	0.772	0.772	1.003	29.587
7 weeks	1.036	1.067	1.067	1.577	24.455
8 weeks	1.051	1.082	1.526	2.512	158.5
9 weeks	3.091	3.138	3.851	5.169	102.496
10 weeks	1.402	1.443	2.015	3.969	222.954

Table C.2: Time, in seconds, to reach gaps for D1 in the small case.

Planning Horizon	First Feasible	20%	5%	1%	Optimal
1	0.046	0.059	0.059	0.087	0.108
2	0.055	0.087	0.158	0.158	0.846
3	0.087	0.139	0.254	0.254	1.148
4	0.106	0.267	0.267	0.267	2.374
5	0.142	0.352	0.352	0.352	5.964
6	0.156	0.465	0.465	0.465	13.547
7	0.186	0.593	0.593	1.274	18.769
8	0.215	0.809	0.809	1.557	234.03
9	0.239	0.872	0.872	1.809	415.622
10	0.269	1.071	1.071	2.341	405.67

Table C.3: Time, in seconds, to reach gaps for S1 in the small case.

Planning Horizon	First Feasible	20%	5%	1%	Optimal
1	2.98	4.39	4.39	4.39	-
2	7.0	21.77	21.77	22.87	-
3	10.99	67.44	67.44	1341.23	-
4	14.5	38.46	42.76	3472.21	-
5	18.33	48.23	55.48	-	-
6	22.7	70.55	80.8	-	-
7	26.45	231.77	247.38	-	-
8	31.74	289.85	306.07	-	-
9	38.06	159.41	892.34	-	-
10	41.49	447.78	447.78	-	-

C.2 Large case

Table C.4 and Table C.5 shows the time to reach gaps for the D1 and P1 model, respectively. A table for the S1 model is not included here, as we have not performed additional tests to those presented in the main report. Remember that the S1 model has problems solving the large case due to hardware limitations.

Table C.4: Time, in seconds, to reach gaps for D1 in the large case.

Planning Horizon	First Feasible	20%	5%	1%	Optimal
1 week	0.09	0.16	0.16	0.36	146.97
2 weeks	0.19	0.31	0.31	1.06	6427.99
3 weeks	0.28	0.53	0.53	3.05	-
4 weeks	0.37	0.79	0.79	8.94	-
5 weeks	0.47	1.05	4.59	12.09	-
6 weeks	0.57	1.41	7.00	30.27	-
7 weeks	0.66	1.86	8.20	85.26	-
8 weeks	0.75	2.11	8.70	73.62	-
9 weeks	0.76	2.38	9.42	265.75	-
10 weeks	0.83	2.70	10.42	255.48	-

Table C.5: Time, in seconds, to reach gaps for P1 in the large case.

Planning Horizon	First Feasible	20%	5%	1%	Optimal
1 week	3.03	5.19	5.19	7.49	13.53
2 weeks	5.33	10.0	13.77	45.56	261.2
3 weeks	8.32	17.03	60.62	60.62	93.48
4 weeks	8.92	35.09	77.10	77.10	77.19
5 weeks	11.04	49.85	106.69	106.69	149.75
6 weeks	13.55	74.38	152.07	152.07	740.20
7 weeks	16.01	199.43	199.43	199.43	1855.37
8 weeks	20.19	278.10	278.10	278.10	2544.53
9 weeks	22.87	350.11	350.11	350.11	4344.98
10 weeks	25.88	323.13	323.13	323.13	2167.91

D Additional figures for the computational study

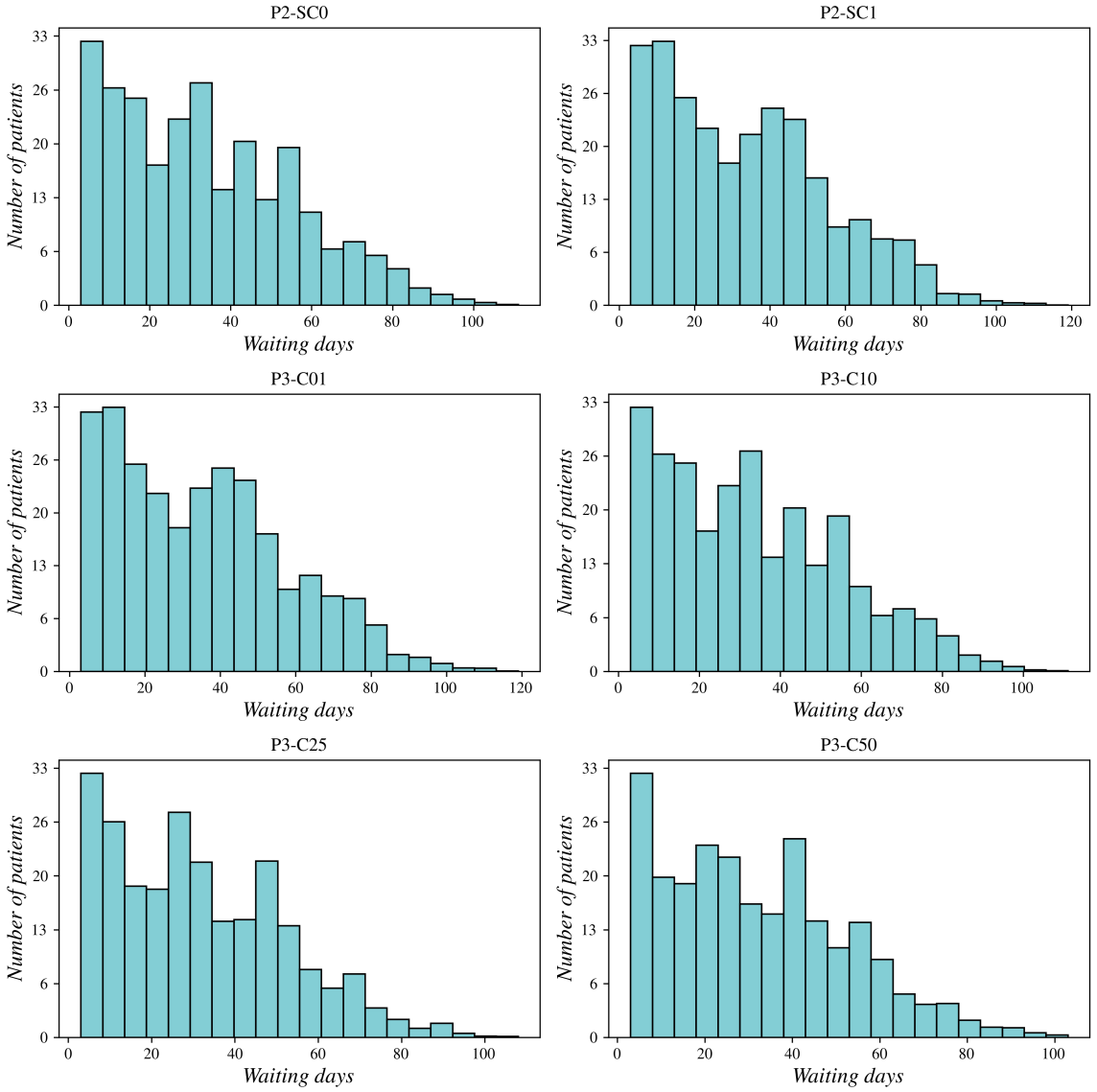
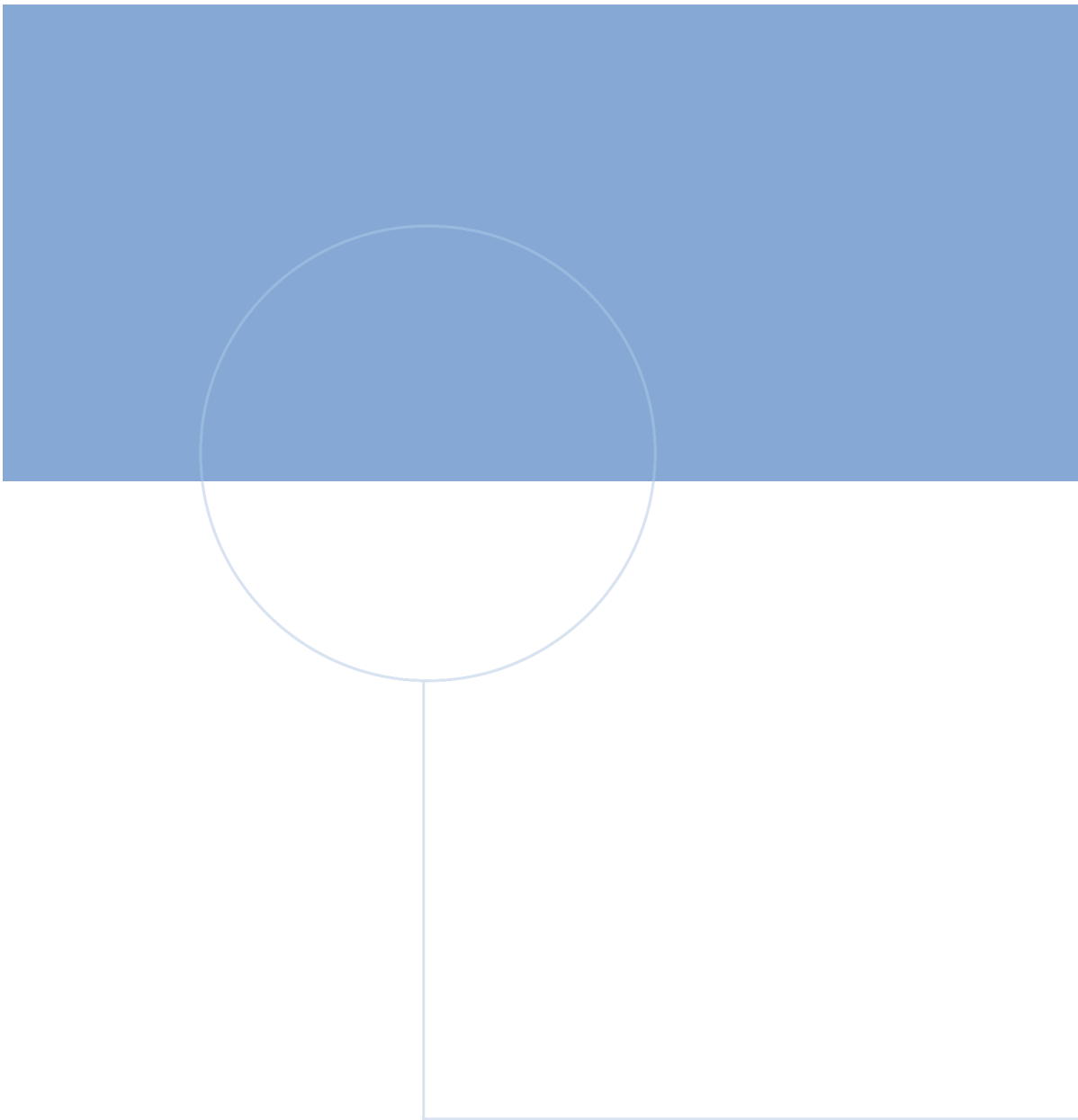


Figure D.1: Waiting list distribution for the instances P2-SC0, P2-SC1, P3-C01, P3-C10 P3-C25, and P3-C50.



NTNU

Norwegian University of
Science and Technology